Benford's Law: An Empirical Investigation and a Novel Explanation

CSM Technical Report 349

P. D. Scott & M. Fasli
Department of Computer Science, University of Essex

Abstract

This report describes an investigation into Benford's Law for the distribution of leading digits in real data sets. A large number of such data sets have been examined and it was found that only a small fraction of them conform to the law. Three classes of mathematical model of processes that might account for such a leading digit distribution have also been investigated. We found that based on the notion of taking the product of many random factors the most credible. This led to the identification of a class of lognormal distributions, those whose shape parameter exceeds 1, which satisfy Benford's Law. This in turn led us to a novel explanation for the law: that it is fundamentally a consequence of the fact that many physical quantities cannot meaningfully take negative values. This enabled us to develop a simple set of rules for determining whether a given data set is likely to conform to Benford's Law. Our explanation has an important advantage over previous attempts to account for the law: it also explains which data sets will not have logarithmically distributed leading digits. Some techniques for generating data that satisfy Benford's law are described and the report concludes with a summary and a discussion of the practical implications.

1. Introduction

The motivation to carry out the study described in this report arose as part of an investigation to discover what characteristics of data sets make them difficult for classification learning procedures. This study required numerous sets of data of known characteristics. The most efficient way to obtain such data is to generate it. However, results obtained using such artificial data sets are open to the objection that real data are in some way different from artificial data. Consequently it is important that artificial data generators produce data sharing as many of the properties of real data as possible.

The notion that about 30% of all numbers obtained from real numerical data start with the digit 1 has been in circulation for over a century. Following an empirical study, Benford (1938) proposed a rule for the distribution of all nine possible leading digits in real data:

$$P(\text{First significant digit} = d) = \log_{10}(1 + 1/d)$$

This is now known as Benford's Law. For many years its status was little more than a numerical curiosity but practical implications began to emerge in the 1960s when it was recognised that the suggestion that almost 1/3 of the numbers processed began with a 1 could have implications for the design of efficient computers (Hamming 1970, Knuth 1981). Varian (1972) suggested it could be used to assess the authenticity of economic models and in recent years it has been used to detect fraudulent financial data (Nigrini 1993, 1996). If Benford's Law is true, that is, if it describes a property that is common in real data sets, there are strong implications for the generation of artificial data sets. Techniques must be developed that produce data complying with the law.

However, it is a somewhat surprising law. The most obvious prior hypothesis is surely that all digits would be equally probable since the leading digit ought to depend on the unit of

measurement chosen. Benford himself recognised the unexpected nature of his rule and entitled his paper "The Law of Anomalous Numbers". Consequently we decided to examine the evidence before embarking on the development of "Benford data generators".

Although the literature contains several theoretical papers that attempt to explain *why* Benford's Law is true, there is very little empirical investigation of *whether* it is in fact true. We therefore decided to carry out such an empirical study ourselves. The primary objective of the study was to determine whether Benford's Law is true of some, or even all, real numerical data. Our strong expectation was that only some sets of data would conform to the law. If this proved to be the case, our secondary objective was to identify what types of data were most likely fit the law.

This report provides an account of this empirical study and the explanation that we developed which led us to a convenient method for identifying data that is likely to conform to the law. In Section 2 we briefly review the origins of Benford's Law and subsequent empirical and theoretical research. In the following section we re-examine the data presented in Benford's original paper and find that only about half the data sets provide reasonably close matches to his law. Section 4 provides an account of our own much larger empirical study of 230 data sets. Despite deliberately seeking out data that looked as if it might have logarithmically distributed leading digits we found that only a small percentage of our data sets were reasonably close matches. We also identified a category of data that, while clearly not having the leading digit distribution prescribed by Benford's Law, was similar in that leading digit frequency was a monotonically decreasing function of digit value. In the next section we investigate three classes of mathematical models of processes that could produce data conforming to the logarithmic law for digit distributions. In Section 6 we take the most promising of these as a starting point to investigate the leading digit distributions of lognormal distributions. We demonstrate that lognormal distributions with shape parameters greater than 1 have leading digit distributions that conform closely to Benford's Law. This leads directly to the most important conclusion of this report: that Benford's Law is a consequence of the fact that many physical quantities can only have positive values. This provides the basis for a simple set of rules for identifying data sets that are likely to conform to the law. Section 7 discusses artificial data generators for producing scalars and vectors whose leading digits fit closely to the logarithmic distribution. The concluding section summarises our findings and discusses their practical implications

2. Benford's Law

Newcomb (1881) observed that the earlier pages of logarithmic tables were more worn than later pages. He concluded that numbers with low valued leading digits arose more frequently in calculations[1]. This led him to examine how numbers in natural data were distributed and ultimately to propose that the "probability of the occurrence of numbers is such that the mantissae of their logarithms are equally probable". From this he inferred that the distribution of leading digits is such that

$$P(\text{First significant digit} = d) \;=\; \log_{10}(1 + 1/d)$$

Newcomb presented no empirical evidence or theoretical proof of this rule so, at this stage in its history, it might reasonably be called "Newcomb's Conjecture".

The rule was rediscovered six decades later by Benford (1938) who was apparently unaware of Newcomb's work. Benford provided empirical support by publishing the distributions of leading digits in 20 data sets taken from a wide variety of sources. For

[1] There is another obvious explanation of Newcomb's observation. If users paged through the tables from the beginning then the earlier pages would inevitably receive more wear, however the numbers were distributed. The story has something of the flavour Newton's apple about it.

convenience, these results are reproduced in Table 1. Ever since, these have formed the core of the experimental support for what is now known as Benford's Law. Most of the subsequent empirical evidence takes the form of isolated observations that particular data sets are in close accord with the law. Hill (1995a) provides a useful summary of these findings. Raimi (1976) is worthy of special mention since he includes two examples of data sets drawn from natural data that manifestly do not conform to Benford's Law.

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Weight	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	n^{-1},$\sqrt{n}$,…	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	n^1,n^2,….n!	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Predicted		30.1	17.6	12.5	9.69	7.92	6.70	5.80	5.11	4.58	

Table 1: The distribution of leading digits in Benford's (1938) data sets expressed as percentages. The final row of the table indicates the percentages predicted by Benford's Law.

Despite this rather slender empirical support, many very distinguished mathematicians and computer scientists have accepted the truth of Benford's Law (e.g. Knuth 1981, Hamming 1970, Weaver 1963). There is disagreement about whether it is a necessary mathematical truth or a contingent property of nature. Weaver (1963) held that it was a "built-in characteristic of our number system", while Knuth (1981) conjectured that the law was a close approximation to reality and that the true distribution might be changing as the universe expands. As a result of this widespread acceptance of the logarithmic law, research has been largely concerned with answering the theoretical question of why the law is true.

The mathematical explanations proffered fall into three groups (see Raimi 1976 and Hill 1995a for reviews). Benford's own explanation is an example of those based on the notion of counting upward through the natural numbers until you reach values whose representations have many digits. Suppose that in such a counting process you have reached the value of 9999. At this point the distribution of leading digits among the numbers generated will be uniform; each will have occurred about 11% of the time. Now suppose you continue counting until you reach 19999. All the additional numbers will have 1 as their leading digit so its frequency will rise to about 55%. The frequency of the other digits will decline proportionately. As the counting proceeds up to 99999, there will be no more 1s as leading digit. Consequently its frequency will decline once more to about 11%. Benford (1938) showed that the average value for the frequency of 1 as leading digit is 30.1%, as

predicted by his law. It can also be shown that such a counting process generates the other eight leading digits with the predicted frequencies. Benford's paper is less clear on the type of process that would emulate this procedure and hence produce numbers conforming to his law. This leads to both theoretical and practical difficulties in applying this explanation. This point will be discussed further in Section 5.3.

The second group of mathematical explanations is based on the notion producing a number by multiplying a lot of numbers together. This approach is appealing because of its similarity to the Central Limit Theorem. This states that the distribution of the sum of a number of independent random variables tends to the normal distribution as the number of variables is increased. This theorem is the reason for the ubiquity of the normal distribution in science and statistics. Consequently the notion that Benford's Law might embody a similar general rule for the product of a number of random variables is very appealing.

An informal argument that multiplying random numbers will produce Benford's distribution is provided by considering a circular slide rule (Raimi 1969, Boyle 1994). This is a disc with a base 10 logarithmic scale around its perimeter. Multiplying numbers is equivalent to adding their logarithms. So multiplying a set of numbers is equivalent to moving around the perimeter of the circular slide rule. In its crudest form the argument says that if you make enough such random moves, you are equally likely to end up anywhere on the perimeter. Hence the leading digits will be distributed logarithmically as Benford's Law predicts. A more rigorous version of this argument, which does not appeal to the circular slide rule, was given by Boyle (1994) who showed that the logarithmic distribution is the limiting distribution of leading digits when random variables are repeatedly multiplied, divided or raised to integer powers.

The third approach to deriving Benford's Law is quite different in character. It could be termed the "ontological approach" because it asks what form would a first digit law take if such a law existed. The arguments begin by assuming that any such law should be independent either of the units of measurement used (scale invariance) or the base of the number system (base invariance) since these are contingent aspects of our culture rather than fundamental properties of nature. Both scale invariance and base invariance have been used to derive Benford's Law. The derivations from scale invariance (e.g. Pinkham 1961) have been criticised (Knuth 1981, Hill 1995b) for lack of rigour because they make unwarranted assumptions about the distributions of numbers. Hill (1995b) has shown that base invariance uniquely implies Benford's Law.

The first two groups of derivations postulate processes that may produce the numbers found in natural data sets. For the first group this process is counting and there are clearly some numbers that are derived in this way. An obvious example from Benford's original data is provided by atomic weights. Successive elements include greater number of protons and neutrons (which have approximately equal masses) and these are by far the largest factors in determining the mass of the resultant atom. Similarly, many quantities may be the result of the multiplicative effect of a number of factors which is the underlying process for the second group of derivations. Neither derivation provides any insight into how many naturally occurring data sets are produced by the type of process that they postulate. Thus unless one has a prior model of the processes that led to the numbers in a data set, there is no reason to expect that they would conform to Benford's Law. The third group of explanations has even less to say about whether real data should conform to the logarithmic law. In essence, it states that if there is a first digit law then it must be Benford's Law but makes no contribution to assessing the truth of its premise.

The situation is thus such that, if it should turn out to be the case that Benford's Law is valid, then there are several alternative mathematical explanations of why this should be so. On the other hand, none of them imply that the logarithmic law is necessarily true. It remains a contingent truth that can only be settled empirically. In fact, it is obvious that Benford's

Law is not a universal property of naturally occurring numbers. Consider, for example, the heights of an adult population measured in imperial units (feet). The most frequent leading digit will be 5, and the next two will be 4 and 6. It is extremely unlikely that there will be any values with a leading digit of 1. In contrast, if the same data were expressed in metric units, far too many (well over 90%) of the values would have 1 as their first digit.

Thus the real empirical questions are "Do a substantial proportion of natural data sets conform to Benford's Law and, if so, which ones?". These are the questions we set out answer in our experimental study.

3. Benford's Original Data – A Reassessment

When Benford (1938) carried out his original empirical study, considerable labour was necessary to obtain the data. He used a wide variety of sources ranging from tabulations of mathematical functions, through physical constants and street addresses to numbers appearing in newspapers. The result was about 20,000 numbers distributed across 20 data sets. This body of data is still the most substantial piece of published evidence for Benford's Law. It represents a formidable amount of tedious work counting the frequencies of leading digits. It should nevertheless be recognised that 20 sets of data with an average of 1000 numbers each is in fact a rather small sample from which to derive a law.

Having counted his data, Benford simply expressed them as percentages and presented them in a table alongside the frequencies predicted by the logarithmic law. His claim that the data did indeed conform to his law rests entirely on the apparent similarity of the numbers. He made no attempt assess how good the fit was. In fact, closer inspection shows that, for some data sets, the digit frequency is not even a monotonically decreasing function of digit magnitude for higher valued digits. This could of course be due to chance variation in the sample. However, the evidence would have been much more convincing if some measure of the statistical significance of the differences had been made.

Although the raw data is not available, and in many cases the original source is extremely obscure, Benford's paper tabulates the percentage frequencies and the total counts for each data set. These are sufficient to allow an assessment of how closely the data conforms to the logarithmic law. There are two standard tests for measuring the goodness of fit between a set of data and a hypothetical distribution from which it may be derived: the chi-square test based on differences between observed and expected frequencies, and the Kolmogorov-Smirnov test, based on deviations between cumulative distributions. The former is applicable to binned data, while the latter is more suited to continuous data. Digit frequencies fall into nine bins so we used the chi-square measure as our criterion for assessing goodness of fit. (The Kolmogorov-Smirnov test was also applied to some of the data we studied but did not prove particularly useful).

Table 2 shows the results of applying chi-square tests to Benford's original data. The fourth column shows the probability that a value of χ^2 at least this big would be produced in a sample drawn from a population conforming to Benford's Law. It can be seen that three of the data sets (D, F, and R) are remarkably close fits. A further eight (A, G, I, M, O, P, Q, and T) satisfy the standard 5% significance criterion (i.e. there is a 1 in 20 chance that a χ^2 this large would occur in a sample from a population conforming to the logarithmic law). The remaining nine sets of data cannot be regarded as conforming to Benford's Law.

Benford's original data is not available and most of his data sets cannot be readily reconstructed. The exception is set J, Atomic Weights. The atomic weights of twelve transuranic elements have been measured since Benford's paper was published. All of these have 2 as their leading digit. We assessed the goodness-of-fit for this enlarged data set in order to see whether advances in physics had led to closer conformation to Benford's Law. The figures are given in the penultimate row of Table 2. The additional elements almost

double the value of χ^2 and the corresponding probability dropped to 0.00102, thus removing atomic weights from the set of data sets that satisfy the 5% significance criterion. In fact, the digit distribution is too heavily weighted towards low value leading digits; the skewness is 1.37 compared with the 0.79 to be expected from the logarithmic law. It will take the discovery of substantially more elements to redress this bias and bring the periodic table into line with Benford's Law.

Group	Title	χ^2	Probability
A	Rivers, Area	4.962	0.762
B	Population	118.6	< 0.000001
C	Constants	24.44	0.00193
D	Newspapers	0.1602	0.999998
E	Spec. Heat	111.2	< 0.000001
F	Pressure	1.27	0.996
G	H.P. Lost	3.46	0.902
H	Mol. Wgt.	125.8	< 0.000001
I	Drainage	11.14	0.194
J	Atomic Weight	17.25	0.0276
K	n^{-1},$\sqrt{n}$,…	440.8	< 0.000001
L	Design	19.21	0.0138
M	*Digest*	3.227	0.919
N	Cost Data	15.6	0.0485
O	X-Ray Volts	5.426	0.711
P	Am. League	14.59	0.0675
Q	Black Body	9.523	0.300
R	Addresses	1.297	0.996
S	n^1,n^2,….n!	24.99	0.00156
T	Death Rate	7.555	0.478
J*	Extended Atomic Wts	26.07	0.00102
Average		84.10	< 0.000001

Table 2: Results of chi-square tests for the data sets presented in Benson (1938) (see text).

Benford also computed the average values for each of the digit frequencies (see penultimate line of Table 1). Several authors have argued that the apparent close fit provided by these averages is evidence for the view that Benford's Law reflects the leading digit distribution to be expected when data from many sources are pooled. However, Benford appears to have simply computed the average percentages for each digit and taken no account of the differing sizes of the 20 data sets. The final line of Table 2 shows the results of a chi-square test applied to the pooled data, thus avoiding over-representation of smaller data sets. It is clear that the pooled data does not conform closely to Benford's Law.

One possible objection to this reassessment of Benford's evidence is that using χ^2 as a measure of goodness-of-fit is too strict a criterion. It is well known that a χ^2 test will always detect a significant difference if a large enough sample is used. However, these are not particularly large data sets: the mean size is 1011. Furthermore, the criterion is only being used to identify data sets that do not deviate sufficiently to justify rejecting Benford's Law: that is, there is at least a modest chance they were drawn form a population whose leading digits has a logarithmic distribution.

So what conclusions can be drawn from Benson's own empirical evidence? It is certainly not the convincing demonstration of the truth of his law that many have claimed it to be. Half the data sets proved to be very poor matches to the prescribed distribution. On the other hand, some of those that did match matched extraordinarily well. Furthermore, even

those data sets that do not match the law have distributions that are very heavily skewed to the low valued digits, rather than the uniform distribution that one might have expected. Thus Benson's data certainly provides evidence for the idea that leading digit frequency is often a monotonically decreasing function of digit value. It provides some evidence that this function has the logarithmic form asserted in Benson's Law but gives little grounds for believing that this is always the case.

4. Natural Data Sets

Modern computing techniques have dramatically reduced the effort required to carry out the type of empirical investigation that forms the core of Benford's work. We decided to carry out a similar study using a substantially larger collection of data sets.

This presented us with a methodological problem that we have still not satisfactorily resolved. How should one set about selecting such a collection of data sets? One of the questions we would have liked to answer is what proportion of natural data sets conform to Benford's Law. This suggests that we should take a random sample of all such data sets. Unfortunately, it is not clear either whether the concept of the set of all data sets has any meaning or, if it does, how one may set about sampling it in an unbiased way.

Indeed it is far from clear how Benford set about choosing his sample. The fact that all his examples appear, at least superficially, to conform to the law is a little suspicious. It is very easy to find data that departs drastically from the frequencies predicted by the logarithmic law. The absence of any such data sets from Benson's study suggests that he may have had a, perhaps unconscious, bias towards data in which numbers starting with the digit 1 were particularly common. In one sense our problem was very different from Benford's. Whereas he had to painstakingly assemble the majority of his data sets from whatever potential sources he could find, we had access to the overwhelming quantity of data freely available on the web.

It is generally recognised that artificially constructed numbers, such as telephone numbers or serial numbers, do not conform to Benford's law, so this type of data was excluded from our study. Preliminary investigations suggested that data sets conforming to the law were not common. Since we also wished to gain some insight into what types of data did fit the logarithmic law, we decided to deliberately choose data sets that included variables that appeared plausible candidates: that is, their most frequent leading digit was 1. Such data sets often included other variables with very different leading digit distributions and these too were normally examined. Other data sets were chosen because previous authors have asserted that data of that type provided a good fit to the logarithmic law. Clearly, this approach precludes an estimate of the proportion of all data sets that conform to the law, although it could be regarded as evidence for an upper bound. Furthermore, it provides empirical evidence on how well those data sets that appear to conform actually fit the prescribed frequencies, and should also provide information about the type of data that conforms to Benford's Law.

We computed the chi-square statistic and the associated probability for each set of data examined. We also computed quantities related to the first four moments of the digit distribution (mean, variance, skewness and kurtosis) to provide a further basis for comparison between the actual frequencies and those predicted by Benford's Law.

We investigated the distribution of leading digits in 230 data sets, all of which can be accessed on the web. Details of these data sets, including the relevant URLs, are provided in Appendix 1. They ranged in size from 132 to 23484 items. In total well over half a million numbers were examined.

Of the 230 data sets examined 29 (i.e. 12.6%) satisfied the 5% significance criterion for conformity to Benford's Law. The results for these data sets are presented in Table 3,

which is ordered by declining goodness-of-fit. All but one of the remaining 201 data sets failed to fit at the 1% significance level. The vast majority of them were extremely poor fits with probabilities of less than 0.00001 that they were samples from distributions conforming to the logarithmic law. Given that the selection of data was biased towards including variables that appeared to fit the law, it seems reasonable to conclude that the majority of data sets do not have leading digit distributions characterised by Benford's Law.

Data Set	Variable	No of Items	Modal Digits	Mean	Var	Skew	χ^2	Prob
Flag Data (MLDB)	Population	194	1,2,3	3.30	5.88	0.826	3.376	0.9086
USDA Crop Data	tb8286 variable 2	4167	1,2,3	3.39	5.98	0.824	4.973	0.7605
USDA Crop Data	ry8286 variable 6	2607	1,2,3	3.45	6.05	0.795	5.654	0.6859
USDA Crop Data	ry87 variable 5	453	1,2,3	3.42	5.53	0.816	6.127	0.6330
USDA Crop Data	tb8286 variable 4	4167	1,2,3	3.45	6.11	0.807	6.535	0.5876
USDA Crop Data	ar8286 variable 5	784	1,2,3	3.19	5.91	0.948	7.205	0.5147
USDA Crop Data	ry8286 variable 2	2612	1,2,3	2.05	5.84	0.779	7.627	0.4708
Boston Housing (MLDB)	Crime per capita	506	1,2,3	3.61	6.63	0.67	7.923	0.4410
USDA Crop Data	br87 variable 2	2713	1,2,3	3.35	6.01	0.858	8.123	0.4216
USDA Crop Data	ry88 variable 5	388	1,2,4	3.62	6.65	0.692	8.428	0.3929
USDA Crop Data	ry7276 variable 6	1966	1,2,3	3.49	6.06	0.750	8.785	0.3608
USDA Crop Data	ry88 variable 3	388	1,2,3	3.45	5.52	0.770	8.871	0.3533
USDA Crop Data	ar8788 variable 3	311	1,2,(3,4)	3.55	5.58	0.687	9.147	0.3300
USDA Crop Data	ry87 variable 3	453	1,2,3	3.34	5.74	0.826	10.84	0.2112
USDA Crop Data	ar8788 variable 4	311	1,2,3	3.32	5.59	0.852	10.91	0.2068
USDA Crop Data	br87 variable 5	2676	1,2,3	3.52	6.14	0.770	11.18	0.1918
USDA Crop Data	br88 variable 2	2592	1,2,3	3.39	5.98	0.827	11.51	0.1746
Wisconsin Breast Cancer	wdbc variable 9	556	1,2,3	3.18	5.93	1.01	11.70	0.1649
USDA Crop Data	ry88 variable 2	388	1,2,3	3.56	6.94	0.707	12.06	0.1483
USDA Crop Data	br87 variable 3	2676	1,2,3	3.34	5.91	0.849	12.48	0.1310
Flag Data (MLDB)	Area	194	1,2,3	3.28	6.80	1.02	12.58	0.1271
USDA Crop Data	ar7781 variable 4	764	1,2,3	3.30	5.65	0.938	12.59	0.1268
USDA Crop Data	br88 variable 5	2536	1,2,3	3.40	5.84	0.817	12.67	0.1239
Wisconsin Breast Cancer	wpbc variable 17	198	1,2,4	3.85	6.12	0.527	13.50	0.0949
USDA Crop Data	tb7781 variable 4	4100	1,2,3	3.49	6.01	0.775	13.86	0.0854
USDA Crop Data	ar8788 variable 2	311	1,2,3	3.60	5.58	0.636	14.43	0.0713
USDA Crop Data	cp7276 variable 2	230	1,3,2	3.35	5.49	0.891	14.87	0.0618
USDA Crop Data	ry7276 variable 3	1967	1,2,3	3.36	5.76	0.879	15.06	0.0579
USDA Crop Data	cp7276 variable 6	226	1,3,2	3.50	6.68	0.741	15.53	0.0497

Table 3: Variables providing the best fits to Benford's Law from a sample of 230. The fourth column lists the most frequent leading digits in declining order of frequency. The next three columns list the mean, variance and skewness of the observed distributions. The eighth column is the value of χ^2. The right hand column is the probability that a value of χ^2 at least this large would be observed in a random sample drawn from a population conforming to the predicted values.

Variables from the USDA Crop Data sets account for 24 of the 29 variables that conformed closely to the logarithmic law. This is only partially explained by the fact that 73 of the 230 sets of data were taken from this source. One third (32.8%) of the USDA variables conformed to the law; the remaining two thirds included a large number of very poor fits. The remaining five sets of data that conformed to the predicted distribution were taken from the UCI Irvine Machine Learning Repository (Blake & Mertz 1998). Both the numeric variables from the Flags data set, which list the populations and areas of the world's countries, provided good fits. Three of the 63 variables taken from the Wisconsin Breast Cancer data sets and the one variable we examined from the Boston Housing data set (crime per capita) also had leading digit distributions that conformed, by the 5% significance criterion, to Benford's Law.

The results from the 87% of data sets that did not appear to conform to Benford's Law are also worthy of consideration. These included all the variables in the climatological and financial time series data sets. The absence of examples of the latter, from the list of those that provided reasonably close fits, is surprising, since it has often been suggested that Benford's Law provides a good model of such data. Table 4 lists the results we obtained for the Dow Jones data sets. The three sets that span the twentieth century are all very similar, the differing values for χ^2 simply reflecting the different sample sizes. It is clear why they are such a poor fit to Benford's Law: the second and third most frequent leading digits are 8 and 9 respectively. The fourth data set, covering the decade 1983-93 is clearly different. Although 1, 2 and 3 are the most frequent leading digits the fit is still very poor. The problem here is that there are far too few leading digits greater than 3. This is reflected in the low moment values. The corresponding values for a distribution following Benford's Law are: mean 3.44, variance 6.07 and skewness 0.79.

Data Set	No of Items	Modal Digits	Mean	Var	Skew	χ^2	Prob
Dow Jones 1900-99 Daily	27011	1,8,9	4.44	8.79	0.190	7999	<0.00000
Dow Jones 1900-99 Weekly	5118	1,8,9	4.51	8.80	0.157	1650	<0.00000
Dow Jones 1900-99 Monthly	1177	1,8,9	4.51	8.76	0.152	371	<0.00000
Dow Jones 1983-1993 Daily	2782	1,2,3	1.81	0.57	0.330	2076	<0.00000

Table 4. Results obtained for the four Dow Jones index data sets. (See caption to Table 3 for explanation of columns).

A more interesting failure to fit the prescribed distribution arises in the Global Ocean Wind Stress data sets. The results for the first four of these (which are typical) are shown in Table 5. It is clear that these four digit distributions are very similar to each other. Furthermore their χ^2 values are all well under 100. Thus, although none of them come close to satisfying the criterion for matching Benford's Law (for which a maximum χ^2 of about 15 is required), they do provide a much better fit than the financial data just discussed.

Data Set	No of Items	Modal Digits	Mean	Var	Skew	χ^2	Prob
Global Ocean Wind Stress m1v1	2161	1,2,3	3.29	6.12	0.837	51.27	<0.00000
Global Ocean Wind Stress m1v2	2161	1,2,3	3.26	6.32	0.874	65.03	<0.00000
Global Ocean Wind Stress m1v3	2161	1,2,3	3.29	6.31	0.840	61.76	<0.00000
Global Ocean Wind Stress m1v4	2161	1,2,3	3.24	6.17	0.902	53.48	<0.00000

Table 5. Results obtained for four of the Global Ocean Wind Stress data sets. (See caption to Table 3 for explanation of columns).

Figure 1 shows the observed frequencies for these four data sets. This provides a further demonstration that these data sets have very similar leading digit distributions which resemble that prescribed by Benford's Law but differ from it in a consistent fashion. The leading digit is 1 too often and the frequencies of the next few possible values are reduced roughly proportionately.

Leading digit distributions such as these raise some interesting questions. They are clearly orderly: that is, their frequency declines smoothly as the digit value increases. Furthermore, a superficial inspection suggests that they conform to Benford's Law but a closer scrutiny shows that this is not the case. The problem is not that the chi-square criterion is too strict. It is evident from Figure 1 that there is a systematic deviation from the logarithmic law.

Such approximations to Benford's Law were common in our data sets. They suggest a number of possibilities. One is that leading digit distributions follow some other law that

produces frequencies that are similar to those prescribed by the logarithmic law. Both the theoretical and empirical evidence are against this. The fact Benford's Law is the only function that is base invariant (Hill 1995b) implies that any alternative law would fail if we use any base but 10 for our number system. Although such approximations to Benford's Law are common in our data, they do not all show the same patterns of deviation as those illustrated in Figure 1. An alternative possibility is that these distributions that approximate Benford's Law are produced by processes which would, if continued further, produce distributions which were much closer matches. This possibility will be explored further in the next section.

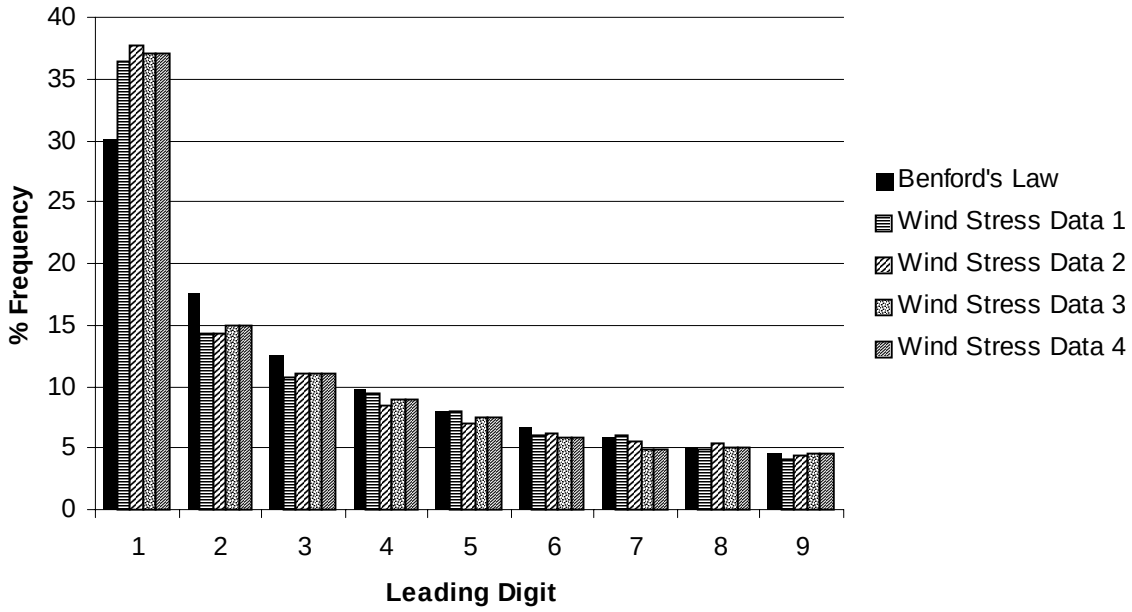

Figure 1: First digit frequency distributions as prescribed by Benford's Law and as observed in a typical examples from the Wind Stress data sets.

The main conclusions to be drawn from our empirical study are:

(1) That only a small proportion of real data sets conform closely to Benford's Law.

(2) That many real data sets have leading digit distributions that are radically different from those prescribed by the logarithmic law.

(3) That there are a significant number of real data sets that definitely do not conform to the law but have leading digit distributions that are broadly similar. In particular, leading digit frequency is a monotonically decreasing function of digit value.

In the next section we will investigate mathematical models of processes that could explain both close and approximate matches to Benford's Law.

5. Mathematically Generated Data Sets

Benford (1938) believed that there was "a distinct tendency for [items] of a random nature to agree better with the logarithmic law than those of a formal or mathematical nature". We decided to investigate how several sets of data generated by mathematical functions conformed to the law. Our motivation was not to discover mathematical relationships but to identify processes that might give rise to logarithmic leading digit distributions in real data.

5.1 Recurrent products

The first type of process considered was that in which successive numbers are generated by repeatedly multiplying the current number by either a constant or a random number. The results obtained are shown in Tables 6 and 7.

Data Set	Series	No of Terms	χ^2	Prob
1	$t_{n+1} = t_n + t_n$-1 (Fibonacci)	1476	0.0470	1.00000
2	$t_{n+1} = t_n \times (1 + \sqrt{5})/2$	1475	0.0423	1.00000
3	$t_{n+1} = t_n \times (1 + \sqrt{3})/2$	2276	0.436	0.99992
4	$t_{n+1} = t_n \times e/2$	2314	1.56	0.99153
5	$t_{n+1} = t_n \times (1 + \sqrt{3})/3$	7934	0.0519	1.00000
6	$t_{n+1} = t_n \times e/3$	7530	0.0318	1.00000
7	$t_{n+1} = t_n \times \sqrt{2}$	2048	0.0725	1.00000
8	$t_{n+1} = t_n \times \sqrt{3}$	1293	0.0466	1.00000
9	$t_{n+1} = t_n \times \sqrt{5}$	883	0.100	1.00000
10	$t_{n+1} = t_n \times \sqrt{6}$	793	0.246	0.99999
11	$t_{n+1} = t_n \times \sqrt{7}$	730	0.403	0.99994
12	$t_{n+1} = t_n \times \sqrt{8}$	683	0.732	0.9994
13	$t_{n+1} = t_n \times 2$	1001	0.161	1.0000
14	$t_{n+1} = t_n \times 5$	5000	0.0930	1.0000
15	$t_{n+1} = t_n \times 9$	5000	0.026	1.000
16	$t_{n+1} = t_n \times 9.9$	5000	0.766	0.999
17	$t_{n+1} = t_n \times 9.99$	5000	137	0.0000
18	$t_{n+1} = t_n \times 9.999$	5000	17561	0.0000
19	$t_{n+1} = t_n \times 10$	5000	11609	0.0000
20	$t_{n+1} = t_n \times \sqrt{10}$	5000	18315	0.0000

Table 6: Conformity to Benford's Law of sequences generated by repeated multiplication by a constant.

The first row of Table 6 shows the results for the first 1476 terms of the Fibonnaci sequence. It confirms the well known result that its leading digit frequencies conform very closely to Benford's Law. The remaining rows show the effect of using a range of constants as multipliers. As can be seen, in the majority of cases the resulting distribution is a very close fit to the logarithmic law. The exceptions arise when the multiplier is an exact integral power or root of 10. This is to be expected since multiplying by 10 does not change the leading digit. Values close to such roots or powers of 10 will only converge very slowly on the logarithmic distribution.

Table 7 show the effect when the multiplier is not a constant but a uniformly distributed random variate. This type of generation provides a simple Markov model for many naturally occurring processes. For example, successive stock market closing prices could be considered as generated by multiplying the previous closing price by a random factor.

It is clear that such sequences converge on the logarithmic distribution except in those cases where the mean is an integral power or root of 10 and the standard deviation is small. This raises the question of why we found that the Dow Jones index data (see Table 4) did not conform to Benford's Law. There are two obvious reasons. First the simple Markov model is not strictly appropriate since the change between two successive closing values will not be independent of preceding changes. More significantly, the average ratio between successive

values will be close to 1 and the change will be a small fraction of the value. This is exactly the situation in which we would not expect to see convergence on the logarithmic distribution even in a sample of many thousands. Thus it would be surprising if such financial indices did conform to Benford's Law except in spectacularly volatile markets.

Series	Mean	Stnd Dev	No Terms	χ^2	Prob
$t_{n+1} = t_n \times \text{urand}(0.0,1.0)$	0.5	0.289	10000	6.66	0.574
$t_{n+1} = t_n \times \text{urand}(0.5,1.5)$	1.0	0.289	10000	3.67	0.886
$t_{n+1} = t_n \times \text{urand}(0.9,1.1)$	1.0	0.0577	10000	502	0.000
$t_{n+1} = t_n \times \text{urand}(0.95,1.05)$	1.0	0.0289	10000	1002	0.000
$t_{n+1} = t_n \times \text{urand}(0.99,1.01)$	1.0	0.00577	10000	28576	0.529
$t_{n+1} = t_n \times \text{urand}(0.95,1.1)$	1.025	0.0433	10000	21.9	0.005
$t_{n+1} = t_n \times \text{urand}(0.5,2.0)$	1.25	0.639	10000	7.86	0.448
$t_{n+1} = t_n \times \text{urand}(1.0,3.0)$	2.0	0.577	10000	9.70	0.287

Table 7: Conformity to Benford's Law of sequences generated by repeated multiplication by a random variate. urand(a,b) is uniformly distributed in the range a..b. Columns 2 and 3 show the mean and standard deviation of the random variate.

5.2 Products of random variates

The next type of process we considered was that in which each number in the data set is the product of several random variables. Such processes are of interest because they are the basis for one of the theoretical explanations of Benford's Law. Furthermore, as noted above, Boyle (1994) has shown that the logarithmic distribution is the limiting distribution of leading digits when random variables are repeatedly multiplied, divided or raised to integer powers. Thus we would expect to find close conformity to Benford's Law in such data.

	Number of terms in product					
Data Set	10 terms		20 terms		50 terms	
	χ^2	Prob	χ^2	Prob	χ^2	Prob
Π(urand(0.0,1.0)	8.21	0.414	-	-	-	-
Π(urand(0.0,0.5)	7.86	0.447	-	-	-	-
Π(urand(0.5,1.5)	15.0	0.0589	6.82	0.55	5.26	0.729
Π(urand(0.9,1.1)	16096	0.0000	10694	0.0000	4930	0.0000
Π(urand(2.0,8.0)	6.99	0.537	6.74	0.565	1.64	0.990
Π(urand(4.0,6.0)	6086	0.0000	2026	0.0000	142	0.0000
Π(urand(4.5,5.5)	17056	0.0000	10947	0.0000	5037	0.0000
Π(urand(4.9,5.1)	108344	0.0000	80260	0.0000	63075	0.0000
Π(norm(0.0,1.0)	2.90	0.935	3.25	0.918	11.0	0.201
Π(norm(5.0,1.0)	707	0.0000	20.4	0.00884	8.87	0.354

Table 8: Conformity to Benford's Law of sets of numbers, each of which is the product of several identically distributed random variates. urand(a,b) is uniformly distributed in the range a..b. norm(m,s) is a normally distributed random variate with mean m and standard deviation s. Columns 2 and 3 show the χ^2 and associated probability when 10 random variates are multiplied. The remaining columns show the same information for products of 20 and 50 random variates respectively. Each data set contained 10000 items.

Our results shown in Table 8 show that there is indeed convergence towards the logarithmic distribution in all cases, and that for some distributions this convergence is rapid. However, it is also clear that in other cases the deviation may be very slow. Two factors

clearly influence the rate of convergence: the variance of the mantissa of the random variate and the deviation of the random variate's distribution from Benford's Law. This illustrated in Figure 2 in which $\log_{10}(\chi^2)$ is plotted against the number of terms in the product of random variates. It is clear that the reduction is approximately linear until values are reached where the fit is good. (Note that the variation in χ^2 once values of 1 have been reached is to be expected. In interpreting this and subsequent data, a helpful rule is that about 25% of samples drawn from a population conforming to Benford's Law would have a χ^2 in excess of 10, and that about 50% would have values more than 7)

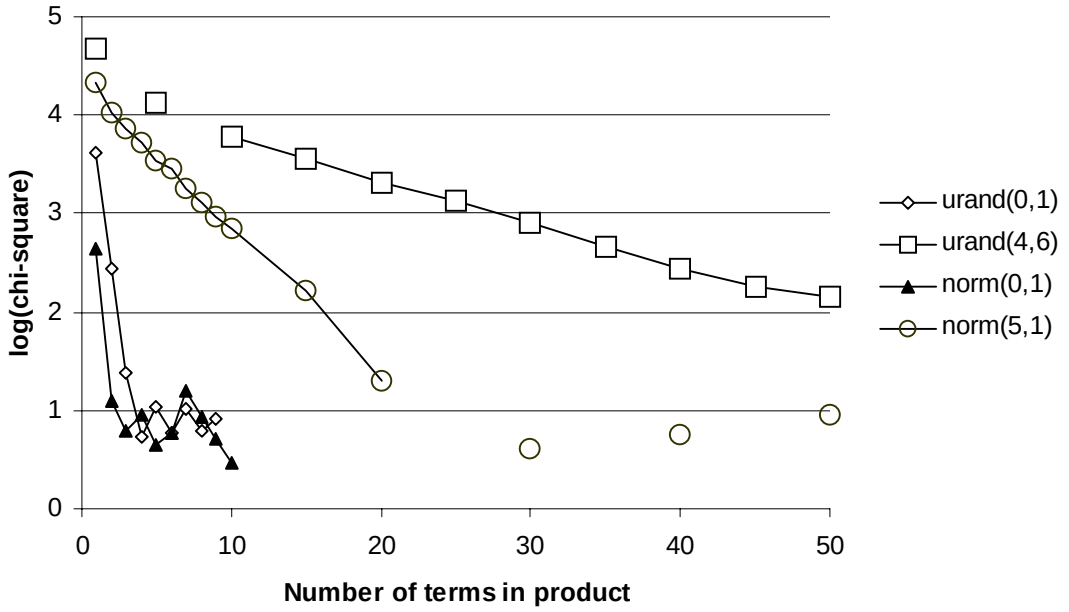

Figure 2: Convergence of the product of random variates on logarithmic distribution prescribed by Benford's Law.

These results suggest a possible explanation for our finding that many data were approximate matches to Benford's Law: that is, their leading digit distributions did not conform to the logarithmic law but nevertheless declined monotonically as the digit value increased. A process that can be modelled by the multiplication of random variates might produce an approximation to Benford's Law if the number of such variates was too few to produce complete convergence.

Figure 3 shows how the frequency of 1 as the leading digit changes as the number of terms in the product of normal random variates (mean 5.0, s.d. 1.0) is increased. Other digits exhibit a similar pattern. It is clear that distributions with either too few or too many 1s as leading digit could arise as the product of small number of random variates. This would explain the type of approximation to Benford's Law illustrated by the Wind Stress data sets (see Figure 1).

Thus, the explanation of Benford's Law as a consequence of the multiplication of random factors could provide an explanation not only for the minority of data sets that actually do conform to the prescribed distribution but also of those which, while clearly not conforming, do have similar distributions.

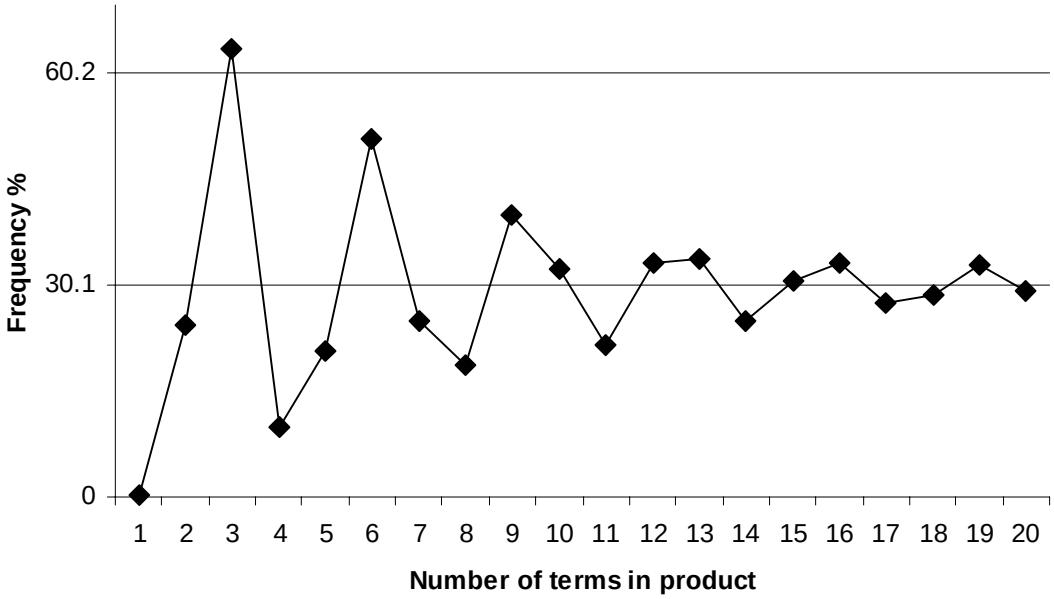

Figure 3: Frequency of 1 among leading digits for the product of random variates having a mean of 5 and a standard deviation of 1.

5.3 Modelling Benford's explanation

Section 2 included a brief account of Benford's own explanation of his law (Benford 1938). This was based on the observation that if you begin counting at 1 and continue indefinitely the average leading digit frequencies will fluctuate in such a manner that their average values are those prescribed by the logarithmic law. It is not clear what forms of physical or social process could be modelled by such a procedure. In an attempt to emulate it, we produced data sets using the following simple method:

```
repeat
      upper_bound = urandint(1,Grand_Upper_Bound)
      output(urandint(1,upper_bound))
until enough data generated
```

where `urandint` is a procedure returning a random integer uniformly distributed in the range specified by its parameters. This algorithm seems a reasonable replication of the account Benford gives for the distribution in his Addresses data. Each person lives at a random house number in a different street, and the streets are of different random lengths. The results we obtained are shown in Table 9 for different values of the Grand Upper Bound.

Grand Upper Bound	χ^2	Prob
10000	150	0.000
20000	111	0.000
30000	145	0.000
40000	140	0.000
50000	133	0.000
60000	137	0.000
70000	147	0.000
80000	153	0.000
90000	151	0.000

Table 9. Conformity to Benford's Law of data sets produced using a procedure based on Benford's explanation (see text). Each data set contained 5000 items.

These results provide little evidence that Benford's explanation applies to the data sets he considered. The fit to the logarithmic distribution is poor. On the other hand, all the distributions obtained by this process decreased monotonically as the value of the digit increased. There is also something deeply unsatisfactory about the need to introduce the `Grand_Upper_Bound` parameter. Postulating that the upper bounds of data sets are uniformly distributed between 0 and some finite maximum is really a very crude way of approximating the distribution of the maxima of all data sets over the range $-\infty \ldots +\infty$. There is in fact no meaningful way in which this distribution can be defined and this fact is at the root of the theoretical objections to Benford's explanation.

5.4 Conclusions

We have examined three classes of model for processes that might give rise to data conforming to Benford's Law. Of these, the third, based on Benford's own explanation, is the least satisfactory for both empirical and theoretical reasons. The success of the first method, based on recurrent multiplication is somewhat marred by its failure to provide a useful model of financial indices. Nevertheless this model is worth serious consideration when the data items are successive members of a sequence. However, in the majority of data sets the individual items have a much greater degree of independence. For these, only the second model, based on each item being the product of several random variates, appears plausible. This model has the additional advantage of offering an explanation for the occurrence of data sets that roughly approximate Benford's Law. The implications of this model are explored further in the next section.

6. An Explanation for Benford's Law

The evidence in the preceding two sections suggests that the multiplication of random factors is the most plausible explanation for a data sets conformity to Benford's Law. Multiplying numbers together is equivalent to adding their logarithms. The central limit theorem states that the sum of independent random variates tends to a normal distribution as the number of variates is increased. Thus the logarithm of the product of random variates should also tend to a normal distribution.

6.1 The Lognormal Distribution and Benford's Law

A probability in which the logarithm of the variable is distributed normally is called a lognormal distribution (Evans, Hastings & Peacock 2000). Lognormal distributions, are unimodal, positively skewed, have a range from 0 to ∞, and are defined by two parameters. The first is a scale parameter, m, whose value is the median of the distribution. The other is a shape parameter, σ, which is the standard deviation of the logarithm of the variate – in other words, it is the standard deviation of the associated normal distribution.

It therefore seems pertinent to investigate under what circumstances the lognormal distribution gives rise to leading digit distributions that conform to Benford's Law. Table 10 shows the results of estimating the goodness of fit to Benford's Law of a range of lognormal distributions with various median and shape parameters.

These results clearly show that conformity to Benford's law is independent of the median of the lognormal distribution. This result is to be expected since it is a scale parameter and the law is scale invariant. They also show that the goodness of fit is a function of the shape parameter σ. In particular, distributions whose shape parameters exceed 1.2 appear to be very good fits. In contrast, as the value of the shape parameter is reduced below 1.2, the fit deteriorates rapidly.

These results also demonstrate that not every lognormal distribution has a leading digit distribution that accords with Benford's Law. Conversely, it is not the case that every

distribution whose leading digits are distributed according to the logarithmic law is lognormal. Counterexamples are readily generated. Consider a sample from a lognormal distribution that does fit Benford's Law. Now multiply 50% of the items by -1. The result will be a distribution that satisfies Benford's Law (because the leading digits are not changed) but is not lognormal (because it is symmetric).

	Shape Parameter σ									
Median	**0.2**	**0.5**	**0.8**	**1.0**	**1.2**	**2.0**	**3.0**	**4.0**	**5.0**	**10.0**
0.2	11507	2773	194	33.0	7.04	7.64	10.1	4.39	10.4	7.72
0.5	21507	3152	225	45.3	13.1	10.1	5.49	7.89	5.47	7.89
0.8	21360	2738	201	27.3	15.1	8.96	9.05	7.23	10.7	14.0
1.0	14525	2801	199	38.0	14.1	11.6	9.29	5.98	9.86	6.71
2.0	11507	2773	194	33.0	7.04	7.64	9.99	4.38	10.8	9.02
3.0	18023	2940	213	36.3	11.4	10.1	6.27	11.4	19.6	5.41
4.0	20673	3105	211	35.6	8.44	11.1	6.33	7.30	10.9	8.40
5.0	21507	3152	225	45.3	13.1	8.79	5.54	7.68	5.62	6.96
Mean	17576	2929	208	36.7	11.2	9.49	7.76	7.03	10.4	8.27

Table 10: Conformity (χ^2) of lognormal distributions with various median and shape parameters to Benford's Law. Each data set contained 10000 items. The final row shows the average χ^2value for each shape parameter.

Lognormal variates have been widely used to model physical quantities where the normal distribution is inappropriate because their values are necessarily positive. For example negative weights, heights and time durations are usually meaningless. In some cases the normal distribution is still acceptable because the mean is many standard deviations above zero; adult human height is such an example. In such a situation, the best fitting lognormal distribution would have a very small shape parameter. In contrast, when the mean is closer to zero, the lognormal distribution provides a much better model than the normal distribution. In such circumstances the shape parameter will be much larger.

6.2 Explaining Benford's Law

Hence we are now in a position to characterise data sets that we would expect to give rise to digit distributions satisfying Benford's Law. Data whose distributions conform to a lognormal distribution whose shape parameter exceeds 1.2 should give rise to leading digit distributions satisfying the logarithmic law. Data that are likely to satisfy this criterion will:

(1) Have only positive values.

(2) Have a unimodal distribution whose modal value is not zero.

(3) Have a positively skewed distribution in which the median is no more than half of the mean. (This constraint ensures that the shape parameter of the lognormal distribution will exceed 1.2)

In the many cases these criteria could be assessed, without actually examining the data, simply by considering the properties of the quantities that the numbers represent. For example, it is obvious that adult height is a positive quantity with a unimodal distribution. However, it is certainly not sufficiently skewed to have a median only half of its mean (in fact they are essentially the same). Hence we should not expect height to conform to Benford's Law. In contrast, consider the distribution of annual salary among all the employees of a large company. Once again this is necessarily positive and almost certainly unimodal. Since there will be far more people on low pay than receiving large salaries the distribution will be positively skewed. Whether it will be sufficiently skewed for the median to be less than half the mean is less obvious; at this stage inspection of the data would be necessary. However,

even without such an inspection, it is clear that this quantity is much more likely to conform to Benford's Law.

It is interesting to consider how this explanation applies to the Address data that Benford used to illustrate his own explanation (see Section 5.3). Street addresses are by convention positive integers. All streets start at number 1 but some are longer than others, so the highest number address will vary. Hence street numbers are positive and their distribution will be positively skewed. Since 1 will be the modal value but some streets are quite long the median will be much less than the mean. Hence, by the rules given above we would expect a random collection of addresses to approximate to a lognormal distribution with shape parameter greater than 1.2. We would therefore expect their leading digits to conform to Benford's Law. This explanation avoids any need to make dubious postulates about the distribution of street lengths.

It is also worth noting that many other probability distributions that have a positive range and a large positive skew can be closely approximated by a lognormal distribution. Such distributions could therefore be expected to be reasonable approximations to Benford's Law. This is readily confirmed by experiment. For example, the negative exponential distribution yields χ^2 values in the range 53…77; a fit similar to that of the wind stress data discussed in Section 4.

The fundamental conclusion to be drawn from this is that Benford's Law is not a necessary mathematical truth or a deep mystical property of our universe. It is a straightforward consequence of the way in which we quantify our observations of that universe. Measurements that cannot meaningfully take values less than zero give rise to Benford's Law. Not all of them do. If the range of measurement is such that zero falls well outside the range of practical consideration, then the leading digits will not conform to the law. But many of the quantities that we measure are necessarily positive and have ranges that include significant numbers of items close to zero. According to our explanation, it is these that give rise to Benford's Law.

7. Generating Data that Conforms to Benford's Law

This investigation was originally motivated by our wish to generate artificial data sets whose characteristics closely resembled those of real data. Having established that many real data sets either conform to Benford's Law or have leading digit distributions that are roughly similar, we must now consider how to generate artificial data with these characteristics.

There are two ways in which this can be done. The first relies on our finding that the leading digits of a lognormal distribution with a shape parameter greater than 1 appear to distributed according to the logarithmic law. Lognormally distributed numbers are readily generated by transforming the output of a normal random number generator (Evans, Hastings & Peacock 2000). Specifically,

$$m\mathrm{e}^{\sigma N(1,0)}$$

where $N(0,1)$ is a standard normal variate, is lognormally distributed with median m and shape parameter σ.

This relationship to normal variates offers a further advantage that is of particular relevance in our own application of evaluating the performance of machine learning procedures. We have previously described a random vector generator that produces vectors of normally distributed numbers in which the interdependency of the elements is specified with a covariance matrix (Scott & Wilkins 1999). By transforming the output of such a generator it is thus possible to produce vectors of numbers satisfying Benford's Law with any desired degree of mutual interdependence.

The transformation of a normal variate offers a convenient way of generating artificial data whose leading digits conform closely to the logarithmic law. However, many real data sets only conform approximately. Generating such data presents more of a problem because there are many ways in which data could deviate while still remaining an approximate match. Numbers distributed lognormally with a shape parameter less than 1.2 appear to produce leading digit distributions that have too many 1s and too few high value digits. An alternative approach would be needed to produce other types of deviation.

One such alternative approach to generating data conforming to the logarithmic law would be to directly exploit the finding, discussed above, that the product of random variables converges on this distribution. Such a generator would simply produce the product of a pre-defined number of random variates. (A similar approach was once used to generate approximations to normally distributed variables before the widespread adoption of the Box-Muller method). This method offers no advantages over the transformational method if the goal is data that conforms closely to Benford's Law. However, by varying the choice of random distribution and the number of terms in the product, data that only approximately conforms to Benford's Law could be produced (see Section 5.2).

8. Conclusions

This investigation has achieved more than we intended when we began. Our initial intention was to discover whether Benford's Law applied to a significant number of real data sets and, if so, to identify common features of those data sets. In fact our investigation has lead us to a novel explanation of the law and a simple set of rules for identifying data sets that are likely to conform or approximate to the logarithmic distribution of leading digits.

Re-examination of Benford's own data revealed that only about half of his data sets provided evidence for his law. Examination of a much larger collection of data sets showed that only a small though significant minority conformed to the law. A larger group conformed roughly in that the leading digit distribution was a monotonically decreasing function of digit value. The remainder had digit distributions that were radically different from those prescribed by Benford's Law. This last group includes types of data, such as financial indices, that previous authors have cited as exemplars of the law.

Three mathematical models of data generating processes were then explored. We concluded that one based on Benford's own explanation was flawed both in practice and theory. A second, based on recurrent multiplication by random factors produced good results but was only applicable to data in which successive items are strongly interdependent. The third, based on generating each item by multiplying several random factors, was applicable when the data items were independent. It gave good results, and had a sound theoretical basis in Boyle's (1994) proof that such processes converge on the logarithmic distribution. It also provided an explanation of those distributions that provided an approximate match to Benford's Law.

This process led us to consider the distribution of leading digits in lognormal distributions. We demonstrated that lognormal distributions with shape parameters greater than 1.2 had leading digit distributions that conformed closely with the logarithmic law. Hence it followed that any set of data that could be modelled by such a distribution would conform to Benford's Law. Hence the task of characterising data that would fit the law was transformed into the simpler one of stating the properties of such a lognormal distribution.

From this it was immediately apparent that the distribution must be confined to positive values. Hence Benford's Law arises because many naturally occurring quantities are measured in such a way that negative values have no meaning. Not all such necessarily positive quantities will be distributed according to the law. Only those where the median of

their distribution is much closer to zero than the mean. A large number of naturally occurring quantities have these characteristics.

This explanation offers more than a solution to a longstanding numerical curiosity; it also has considerable practical value. In recent years Benford's Law has been increasingly employed to identify fraudulent financial data (Nigrini 1993, 1996). Both the validity and the utility of this technique depend upon knowing when data ought to conform to Benford's Law. The criteria we propose can be applied simply and rapidly by anyone with an understanding of what the numbers in a data set represent about the world. It is also immediately clear that data conforming to Benford's Law can be produced readily by generating lognormally distributed random numbers.

Further research would strengthen the argument propounded in this report. The claim that the leading digits of a variable distributed lognormally with a shape parameter greater than 1.2 conform to Benford's Law is crucial. It rests at present on simulation evidence. Clearly a proof would be preferable. The argument demonstrates that a large class of naturally occurring quantities can be expected to conform to the law. This does not preclude the possibility that there are other types of natural quantity that are not distributed lognormally but also conform to the law. It is possible to construct such distributions but it is an open question whether they provide models of quantities occurring in real data. We conjecture that, if any such quantities exist, they are uncommon and that the empirical finding that a significant proportion of real data sets conform to Benford's Law is essentially a consequence of their being necessarily restricted to positive values.

Acknowledgements

The work reported in this paper was supported by the EPSRC under grant GR/M44705. Thanks are also due to Dave Hales for some of the implementation.

References

Benford, F. (1938). The Law of Anomalous Numbers. *Proc American Philosophical Society,* 78(4) pp 551-772.

Blake, C.L. and Merz, C.J. (1998). *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Boyle, J. (1994). An Application of Fourier Series to the Most Significant Digit Problem. *American Mathematical Monthly,* 101 pp 879-976.

Evans, M., Hastings, N. & Peacock, B. (2000). *Statistical Distributions*. 3rd ed. Wiley. New York.

Flehinger, B. J. (1966). On the Probability that a Random Digit has Initial Digit A.. *American Mathematical Monthly*, 73 pp 1056-1061.

Hamming, R. (1970). On the distribution of numbers. *Bell System Technical Journal* 49 pp 1609-1625.

Hill, T. P. (1995a). A Statistical Derivation of the Significant-Digit Law. *Statistical Science*. 10(4) pp 354-363.

Hill, T. P. (1995b). Base-invariance implies Benford's Law. *Proc. American Mathematical Society*. 123(2) pp 887-895.

Knuth, D. (1981). *The Art of Computer Programming 2: Seminumerical Programming*. 2nd ed. pp 239-249. Addison-Wesley, Reading, MA.

Newcomb, S. (1881). Note on the frequency of the use of the digits in natural numbers. *Amer. Jour. Math*. 4 pp 39-40

Nigrini, M. J.(1996). A Taxpayer Compliance Application of Benford's Law. *Journal of the American Tax Association*, Spring 1996 pp 72-91.

Nigrini, M. J.(1993). *The Detection of Income Tax Evasion Through an Analysis of Digital Distributions*. Ph.D. Dissertation, Univ. of Cincinnatti.

Pinkham, R. (1961). On the distribution of first significant digits. *Ann. Math Statists*. 32 pp 1223-1230.

Raimi, R. A. (1976). The First Digit Problem. *American Mathematical Monthly*. 49 pp 521-538.

Raimi, R. A. (1969). The Peculiar Distribution of First Digits. *Scientific American*. 221 pp 109-120.

Scott, P. D. & E. Wilkins, E. (1999). Evaluating Data Mining Procedures: Techniques for Generating Artificial Data Sets. *Information and Software Technology*. Special Issue on Data Mining. June 1999 vol 41, no 9, pp 579-589

Varian, H. (1972). Benford's Law. *American Statistician* 26 pp 65-66.

Weaver, W. (1963). Lady Luck: *The Theory of Probability*. pp 270-277 Doubleday, New York.

Appendix: Sets of Natural Data

The natural data sets used in the study were obtained from the web. This appendix gives a brief description of each and the url of the site from which it was obtained.

UCI Repository of Machine Learning Databases. (Blake & Mertz 1998)

http://www.ics.uci.edu/~/mlearn/MLRepository.html

Flags: (2 variables, 194 cases)

The two numeric variables are the populations and areas of the worlds countries.

Boston Housing: (1 variable, 506 cases)

The single variable examined is the crime per capita by township.

Wisconsin Prognostic Breast Cancer: (32 variables, 198 cases)

Wisconsin Diagnostic Breast Cancer: (30 variables, 569 cases)

Abalone (8 variables, 4176 cases)

Echocardiogram (1 variable, 132 cases)

Fraser River Flow Data (6 variables, 156 cases)

http://libstat.cmu.edu/datasets/fraser-river

Mean monthly flow of the Fraser at Hope B.C., March 1913 – December 1990.

Climate Datasets

http://dss.ucar.edu/datasets

This site contains a heterogeneous collection of climatological data sets. Three were used in this study:

Global Ocean Wind Stress 1870-1976 (ds232.0) (45 variables, 2161 cases)

Monthly Antarctic Ice Analyses 1973-1990 (ds234.0) (9 variables, approx 800 cases)

Shea's Climatology Atlas 1950-1979 (ds290.0) (8 variables, 23484 cases)

USDA-NASS Crop County Data (73 variables, cases range from 226 to 17091)

http://usda.mannlib.cornell.edu/data-sets/crops/9X100/

Agricultural productivity data. Data items are planted and harvested acreage, yield per harvested acre, and total production by county.

Financial Indices

Dow Jones

http://www.quoteline.com

Dow Jones Index 1900-99 (Daily) (1 variable, 27011 cases)

Dow Jones Index 1900-99 (Weekly) (1 variable, 5118 cases)

Dow Jones Index 1900-99 (Monthly) (1 variable, 1177 cases)

Dow Jones Index 1983-93 (Daily) (1 variable, 2782 cases)

S&P (3 variables, 648-6789 cases)

http://sciapp.com/fhdata.html

Consumer Credit Data (6 variables, 681 cases)

http://www.big.frb.fed.us/releases/G19/hist/cc_hist_cb.html

Historical data for US consumer credit outstanding for commercial banks and finance companies.

Consumer Price Index (1 variable, 645 cases)

http://www.stls.frb.org/fred/data/cpi/cpiaucns

US consumer price index 1946-1999