# Detecting Evidence of Non-Compliance In Self-Reported Pollution Emissions Data: An Application of Benford's Law

Selected Paper
American Agricultural Economics Association Annual Meeting
Tampa, FL, July 30-August 2, 2000

Christopher F. Dumas*
Assistant Professor
University of North Carolina, Wilmington

John H. Devine
Student Research Assistant
University of North Carolina, Wilmington

University of North Carolina, Wilmington
Department of Economics and Finance
601 South College Rd.
Wilmington, NC 28403
TEL: 910-962-4026
FAX: 910-962-7464
EMAIL: dumasc@uncwil.edu

July 30, 2000

* Author to whom correspondence should be addressed: Dr. Christopher F. Dumas, University of North Carolina, Wilmington, Department of Economics and Finance, 601 South College Rd., Wilmington, NC 28403. UNCW Cameron School of Business Working Paper Series No. 2000-02-009.

# ABSTRACT

The paper introduces Digital Frequency Analysis (DFA) based on Benford's Law as a new technique for detecting non-compliance in self-reported pollution emissions data. Public accounting firms are currently adopting DFA to detect fraud in financial data. We argue that DFA can be employed by environmental regulators to detect fraud in self-reported pollution emissions data. The theory of Benford's Law is reviewed, and statistical justifications for its potentially widespread applicability are presented. Several common DFA tests are described and applied to North Carolina air pollution emissions data in an empirical example.

Key Words: Benford's Law, Digital Frequency Analysis, Pollution Monitoring, Pollution Regulation, Enforcement

JEL Codes: Q25, Q28

# 1. INTRODUCTION

Federal and state environmental agencies collect pollution emissions data to verify permit compliance and to assess emissions fees. Typically, these data are reported to the agency by emissions sources via self-administered reporting forms. Agencies usually do not have the resources to conduct frequent, on-site audits of firms' emissions reports. For example, the U.S. Environmental Protection Agency conducts "a limited number" of data quality inspections in support of its Toxic Release Inventory program, but the data are not independently verified [25]. Similarly, state environmental agencies do not have the resources to conduct frequent on-site inspections to verify reported emissions numbers [20].

Given infrequent inspections, incentives exist for sources to underreport pollution emissions. The probability of getting caught is relatively low, and the benefits of lower pollution emissions include reductions in permit application and annual permit renewal fees, reductions in emissions fees, avoidance of costly command-and-control plant modification requirements and better public relations. A method of determining the relative likelihood of fraudulent underreporting across pollution sources would improve the efficiency of compliance monitoring and enforcement by allowing regulators to better identify and target potentially fraudulent sources for earlier or more frequent inspections. This paper applies new techniques recently developed in the field of accounting to the problem of detecting potential underreporting in pollution emissions data. The techniques are based on a statistical property, known as "Benford's Law," that is exhibited by many types of data sets.[1]

The paper is divided into five sections. The following section of the paper describes Benford's Law and explains why it likely applies to pollution emissions data. The third section presents several statistical tests based on Benford's Law that can be used to detect evidence of non-compliance in self-reported data. Section four applies the tests in an empirical case study of criteria air pollution emissions data from North Carolina. The final section concludes with a summary of findings, several caveats, and a discussion of future research possibilities.

## 2. BENFORD’S LAW

### 2.1 Background

Benford's Law [4] describes a property of the numbers found in many empirical data sets.[2] For many large data sets, the relative frequency with which the first digit in each of the numbers in the data set takes on each of the possible (base 10) values 1 through 9 is not the naive estimate of 1/9, but rather follows "Benford’s distribution," as shown in the first column of Table 1. Under Benford’s distribution, it is much more likely that the first digit in each number in the data set will be a "1" than a "9." Benford's Law is the equation that gives the relative frequencies, f(p), of the first digits of the numbers in a data set as a function of the first digit value, p, i.e.:

$$f(p) = \log_{10}[(p + 1)/p], \; p = 1, 2, \ldots, 9. \qquad (1)$$

Similar relative frequency distributions hold for the second digit in each number in the data set, the third digit, etc., and, indeed, even for joint distributions of the digits [10].

Empirically, many data sets have been found to be consistent with Benford's Law [4, 22, 23, 26]. Recently, tax accountants have begun to use consistency with Benford's Law as a test for evidence of income tax evasion. These tests are based on tax reporting models that give rise to income tax data distributed according to Benford's Law under truthful reporting [6, 24, 7, 17].

If an underlying data-generating mechanism is assumed to be consistent with Benford's Law, then deviation of an observed data set from Benford's Law in these models constitutes evidence of ex post data manipulation. If observed tax return data deviate significantly from Benford's Law, then tax authorities take this as evidence of potential tax fraud and reallocate regulatory auditing effort accordingly.

An environmental agency's problem of detecting fraud in self-reported pollution emissions data is analogous to the problem of detecting fraud in self-reported income tax return data. In both instances, reporters have incentives to underreport, and regulatory agencies face the problem of allocating limited audit resources.

## 2.2 Characterizing the Distribution of First Digits

Following Goudsmit and Furry [9], consider a large set, X, of self-reported, non-negative data, where x is an element of X. Let f(x)dx be the fraction of observations in the interval between x and x + dx; then,

$$\int_{0}^{\infty} f(x)dx = 1. \tag{2}$$

Write each observation as:

$$x = p \cdot 10^{m}, \tag{3}$$

where p, $1 \leq p \leq 10$, indicates the significant figures of x and m, an integer, is the order of magnitude of x. Benford's law concerns the distribution of the proportions of observations with p lying between two consecutive integer values. For a fixed value of m, via a change of variables the fraction of observations with p lying between p and p + dp may be expressed as:

$$f(p \cdot 10^{m}) \cdot 10^{m}\, dp\,. \tag{4}$$

Summing over all values of m, the density function of first digits p, b(p), is:

$$b(p) = \sum_{m=-\infty}^{\infty} f(p \cdot 10^{m}) \cdot 10^{m}\,. \tag{5}$$

## 2.3 Theoretical Sources of Benford Distributions

Given a density function of first digits, b(p), why should it follow Benford's Law? Beginning with an extended example due to Furry and Hurwitz [8], we review several theoretical explanations for the common empirical occurrence of Benford Law.

Furry and Hurwitz note that x may be expressed as:

$$x = p \cdot 10^{m} = 10^{(m+\log_{10} p)}\,; \tag{6}$$

hence,

$$b(p) = \sum_{m=-\infty}^{\infty} f\left(10^{(m+\log_{10} p)}\right) \cdot \left(\frac{1}{p} \cdot 10^{(m+\log_{10} p)}\right). \tag{7}$$

Factoring out the 1/p and multiplying by ln(10) both inside and outside the summation:

$$b(p) = \frac{1}{p \cdot \ln(10)} \cdot \left\{ \sum_{m=-\infty}^{\infty} f\left(10^{(m+\log_{10} p)}\right) \cdot \left(10^{(m+\log_{10} p)}\right) \cdot \ln(10) \right\}. \tag{8}$$

Denote the factor in braces above as $\Psi$. If $\Psi = 1$, then b(p) is said to follow Benford's distribution, because when $\Psi = 1$:

$$b(p) = \frac{1}{p \cdot \ln(10)}\,, \tag{9}$$

and the fraction of the data with first significant figure between $p_0$ and $p_1$ is given by:

$$\int_{p_0}^{p_1} b(p)dp = \frac{(\ln(p_1)-\ln(p_0))}{\ln(10)} = \log_{10}(p_1/p_0)\,, \tag{10}$$

which is Benford's main result. Equivalently, Benford’s Law describes the frequency distribution of the first digits of the numbers in a data set where those numbers follow a geometric sequence. A key insight is that data describing *growing* processes (e.g., numbers of firms, sizes of firms, values of firms—and associated measures such as stock market values and pollution emissions) often produce first digit frequencies consistent with Benford’s Law because growth is usually a geometric process.

Furry and Hurwitz [8] develop conditions on f(x) that are sufficient for $\Psi = 1$. Suppose f(x) is the *n*<sup></sup>*th iterate* of some density function g(·):

$$f(x) = \int_0^\infty \cdots \int_0^\infty \frac{1}{\alpha_n} \cdot g\left(\frac{x}{\alpha_n}\right) \cdot \frac{1}{\alpha_{n-1}} \cdot g\left(\frac{\alpha_{n-1}}{\alpha_{n-2}}\right) \cdot \cdots \cdot \frac{1}{\alpha_1} \cdot g\left(\frac{\alpha_2}{\alpha_1}\right) d_{\alpha_1} d_{\alpha 2} \cdots d_{\alpha_n}\,. \tag{11}$$

Furry and Hurwitz show via Fourier series analysis that:

$$\lim_{n\to\infty} \Psi = 1\,, \tag{12}$$

hence:

$$\lim_{n\to\infty} b(p) = \frac{1}{p\ln(10)}\,. \tag{13}$$

Thus, if f(x) is the result of a sufficient number of iterations of some density function g(·), then b(p) will follow Benford's Law. In fact, Furry and Hurwitz show numerically for a variety of g(·) (e.g., normal, Cauchy, exponential, etc.) that in practice g(·) need only be iterated a few times to achieve $\psi \approx 1$. For example, suppose the distribution of pollution emissions (not digits) across firms depends on parameter Y, the aggregate production level. Suppose Y is normally distributed and depends on parameter I, aggregate input use level. Suppose I is in turn normally distributed and depends on parameter C, per unit cost of aggregate input I. Finally, suppose C is itself normally distributed and determined exogenously. In this case, the distribution of pollution emissions is the third iterate of input costs. As such, the distribution of the first digits of pollution emissions data should approach Benford's Law. In general, there are many other possible iteration scenarios that might lead to pollution emissions distributed as Benford's Law.

Furthermore, there are other statistical justifications for the common empirical occurrence of Benford's Law. Adhikari and Sarkar [2] show that if a uniform random variable defined on the interval (0,1) is raised to an integer power, then the first digits of the resulting random variable approach Benford's distribution as the integer power increases. They further show that the first digit distribution of the product of many independent random variables each uniformly distributed on (0,1) approaches Benford's distribution as the number of random variables increases. Adhikari and Sarkar show also that if the first digit of a random variable follows Benford's Law, then the first digit of the reciprocal of the random variable and the first digits of the product of the random variable and any constant do as well.

Adhikari [1] considers the product of the reciprocals of independent and identically distributed uniform random variables defined on (0,1). He shows that as the number of factors in the product increases,the distribution of the first digits of the product approaches Benford's distribution. Adhikari proves a similar result for a sequence of quotients of uniformly distributed random variables. Finally, Adhikari shows that if *any* random variable defined on the positive real numbers is divided by the preceding sequence of quotients, then the digits of the resulting random variable approach Benford's distribution.

Lemons [14] explores Benford's distribution from the perspective of physical science. Lemons considers a fixed physical quantity broken into particles of random size (subject to some maximum and minimum values for the particle sizes). He shows that the distribution of first digits of the particle sizes approaches Benford's distribution on average over repeated trials.[3]

Boyle [5] shows that Benford's distribution is the limiting distribution of first digits when *any* continuous, independent and identically distributed random variables are repeatedly multiplied, divided or raised to integer powers. Furthermore, Boyle shows that once the first digits achieve Benford's distribution, then this distribution persists under all further multiplications, divisions and raising to integer powers.

To this point we have considered only first digit frequencies. Hill [10] derived frequency distributions analogous to b(p) for the second significant digit, third significant digit, etc., of a set of data that follow Benford's Law. Indeed, Hill even derived the joint distributions of the digits. Hill's "Generalized Significant Digit Law" is[4]:

$$\mathrm{Pr\,ob}\left(\bigcap_{i=1}^{k}\{D_i = d_i\}\right) = \log_{10}\left[1+\left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)^{-1}\right], \qquad (14)$$

where $D_i$ is the $i^{th}$ significant digit of x, $k \in \text{natural numbers}$, the first significant digit $d_1 \in \{1, 2, . . ., 9\}$, and $d_j \in \{0, 1, . . ., 9\}$, for j = 2, . . . , k.

Hill [11, 12] provides a more rigorous justification for the empirical occurrence of Benford's Law in its full Generalized Significant Digit Law form. Hill proves:

> "If distributions are selected at random (in any 'unbiased' way) and random samples are then taken from each of these distributions, the significant digits of

> the combined sample will converge to the logarithmic (Benford) distribution."
> Hill [11, p. 354]

Hill remarks that Benford's Law is a "limit theorem" for distributions *of digits* of random variables, analogous to the Central Limit Theorem for distributions of random variables themselves. In Hill's words:

> "Justification of the hypothesis . . . is akin to justification of the hypothesis of independence (and identical distribution) in applying the strong law of large numbers or central limit theorem to real-life processes: neither hypothesis can be proved, yet in many real-life sampling procedures they appear to be reasonable assumptions. Conversely, [the result] suggests a straightforward test for unbiasedness of data—simply test goodness-of-fit to the logarithmic distribution."
> Hill [11, p.361]

## 3. REGULATORY APPLICATION OF BENFORD'S LAW

In this section of the paper, we review several statistical tests recently developed by accountants and used to detect fraud in self-reported data. Because the tests are based on examining the frequency of occurrence of digits in a dataset, the tests are known collectively as "Digital Frequency Analysis," or DFA.

### 3.1 Digital Frequency Analysis: Common Digital Tests

Nigrini & Mittermaier [18] and Nigrini [19] describe six digital screening tests used by business accountants when conducting external and internal audits of firms' financial information. Internal audits are conducted typically by a firm's employees to detect data accounting and reporting errors within the firm. Nigrini [19] reviews many case studies where the use of DFA successfully uncovered errors in firms' accounting procedures and outright employee fraud. External audits are conducted typically by third party public accounting firms to validate firms' self-reported financial records. External audits seek to uncover reporting errors and fraud at the firm level.

In an environmental regulation context, internal audits of pollution emissions data would help firms avoid regulatory sanctions through early detection of control system irregularities such as emissions data recording and reporting errors. External audits of emissions data by regulatory agencies would seek to detect suspicious emissions patterns that might indicate pollution control system problems or reporting fraud. The use of DFA as an initial screen for abnormalities in emissions data may help regulatory agencies better target scarce personnel resources used for on-site inspections.

In practice, accountants use several rules of thumb to decide whether a given dataset is likely to conform to Benford's Law under unbiased reporting [19]. A candidate data set should (1) describe a single type of phenomena (e.g., air pollution emissions), (2) have no theoretical maximum or minimum (except zero) values, (3) be expected to have more small numbers than large numbers, (4) not contain systematic number duplication (e.g., it should not be the case that firm X is allowed always to report a value of "12" regardless of the actual data value measured), (5) not consist of systematically-assigned numbers (e.g., social security numbers, bank account numbers, etc.) and (6) be spread across at least one digital order.

Assuming a regulatory situation is consistent with conditions likely to produce an emissions data set conforming to Benford Law under unbiased reporting (as described above), DFA requires that each data value be recorded with sufficient precision (i.e., sufficient number of decimal places) to facilitate the analysis and that the total data set be large enough for valid statistical inference. Assuming these conditions are met, descriptive statistics for the data set should confirm that the mean value of the data is larger than the median and that skewness is positive (necessary conditions for a Benford distribution).

The first DFA test considered is the "First Digits Test." This test compares the frequencies of the first digits in a data set to the Benford Law first digit frequencies (Table 1). If the frequencies of the smaller digits in the data set are larger (smaller) than the corresponding Benford frequencies, then the data values may have been biased (i.e., "fudged") downward (upward). Z-statistics are used to test for significant differences between actual and expected (Benford) frequencies and to construct confidence intervals. Figure 1 is an example of the type of graph typically produced when conducting first digits tests. The graph shows empirical first digit frequencies for a hypothetical data set, the corresponding Benford frequencies and 95% confidence limits. If an empirical frequency lies outside the confidence limits, then the null hypothesis that the empirical frequency is identical to the corresponding Benford frequency at the 5% confidence level is rejected.

The first digits test is typically used simply as a general test of conformity of the data set with Benford's Law; i.e., if several digits show massive deviation from Benford frequencies, then the maintained assumption that the unbiased data follow Benford's Law may not be appropriate for the given data set (or, of course, fraud may be *very* widespread in the data; but if this is so, it should be relatively obvious). Similar tests could be conducted for the second or any other single digit. Of course, chi-square or Kolmogorov-Smirnoff tests could be used to test the conformity of all digital frequencies *as a group* with the corresponding Benford Law frequencies.

The "First Two Digits Test" is a more precise test that compares the frequencies of the first two digit *combinations* in the data with the frequencies of the first two digit *combinations* consistent with Benford's Law. There are ninety possible first two-digit combinations (10 through 99 inclusive). Again, Z-statistics and confidence intervals may be calculated to investigate significant differences in digital combinations. Consider a graph of (hypothetical) empirical first two digit frequencies and associated confidence intervals (Figure 2). An empirical frequency extending above the upper confidence interval line is termed a "spike" in the accounting literature. Spikes indicate unusual presence of a first two-digit combination in the data set. (Similarly, an empirical frequency less than the lower confidence limit indicates unusual absence of the corresponding combination from the data.) Of course, some spikes are expected to occur due to chance alone ("false positives"), but the First Two Digits test has proved useful in practice. Spikes have been found to signal systematic system (engineering or accounting) malfunctions, data reporting or recording errors, and fraud [19]. Each of these possible sources of spikes would be of interest to either the reporting firm, the auditor or both. Two particular spike patterns deserve emphasis. The first consists of one or several significant positive spike followed by one or several significant negative spikes. This pattern typically indicates a *threshold* value that is being avoided by the data reporter. The second pattern consists of spikes at multiples of ten and five, indicating the potential of excessive rounding in the data. Of course, "First Three Digit Tests," "First Four Digit Tests," etc., may be conducted also. However, data set size may not be sufficient to achieve statistically valid distributions over the larger supports required for these higher-precision tests.

The "Number Duplication Test" investigates number duplication as one possible source of positive spikes. A positive spike occurring at "25" on the First Two Digits Test may represent many values of 25 or an assortment of the values 25, 250, 2500, 252, etc. The Number Duplication Test is simply a list of each number in the data set and the frequency of occurrence of each number in the data set, sorted by decreasing frequency. If the positive spike at “25” is due to many values of 25 in the data set, the value 25 will occur high in the list of duplicated numbers. If the spike is due to an assortment of “25, 250, 2502, 253, etc.” values, then none of the values will appear high on the list. High frequency duplication may indicate systematic errors in emissions monitoring equipment, data entry errors, or errors in data “cleaning” and data analysis conducted by regulatory agencies.

The “Last Two Digits Test” is similar to the First Two Digits Test except that the distribution of the last two digit combinations of the data is examined. This test is useful because the distribution of the last two digits of a data set conforming to Benford’s Law is quite different from the distribution of the first two digits. Notice in Table 1 that the distributions of the succeeding digits become more and more uniform. The distribution of the last two-digit combinations of Benford Law data (of sufficient precision) is essentially *uniform*. Assuming a uniform distribution for the last two digits combinations, Z-statistics and confidence intervals are calculated, and a graph of the empirical digit frequencies is checked for spikes. Not only is this test useful as an additional indicator of excessive data rounding, the test is used also to detect *less* than expected rounding. For example, when choosing fictitious data, an evader may shy away from choosing round numbers, as they may appear “made up.” However, we would *expect* an unbiased data set to include numbers ending in “00” fully one percent of the time. Furthermore, numbers ending in “x0” would be expected to appear *ten percent* of the time. Hence, a *lack* of round numbers may indicate fraud. Of course, this test is not possible when the data do not possess a sufficient number of significant figures to ensure that the distribution of the last two digits approximates a uniform distribution.

The “Round Numbers Test” calculates the frequencies of multiples of round numbers such as 25, 100 and 1000 in the data set. The calculated frequencies are compared with expected frequencies (via Z-statistics) based on the assumption that the last two digits of the numbers in the data set follow a uniform distribution. The Round Numbers Test is useful for identifying excessive estimation and its order of magnitude. (In data sets where rounding is expected, other digital tests derived from the Last Two Digits test can be used to determine whether rounding is unbiased.)

### 3.2 Nigrini’s Distortion Factor Model

Suppose the common digital tests indicate that distortion is present in a data set. Nigrini [16, 17] develops a simple measure of the *direction and average magnitude* of the distortion in a data set that follows Benford’s Law under unbiased reporting. Nigrini’s measure is called the Distortion Factor Model (DFM) and depends on two assumptions. First, any data manipulations do not change the order of magnitude of the manipulated data values. This assumption is based on psychological evidence that people use orders of magnitude as reference points, that data manipulators are aware of this tendency, and that manipulators will therefore avoid conspicuous order of magnitude changes when altering data. Second, the model assumes that the relative magnitude of data manipulation is similar across orders of magnitude (i.e., that average percentage manipulation is equal across orders of magnitude). This assumption is consistent with a manipulator choosing to alter data such that the “level of significance” of the alteration is similar across orders of magnitude.

The DFM assumes that the unmanipulated data set follows Benford's Law and spans the range [10, 100). If the data span a greater range, they are *collapsed or expanded* to the assumed range by moving the decimal point via:

$$X_{collapsed} = \frac{10 \cdot X}{10^{\mathrm{int}(\log_{10}(X))}}, \qquad (15)$$

where X is an uncollapsed (raw) data value, $X_{\mathrm{collapsed}}$ is the corresponding collapsed (or expanded) data value, and "int" is the integer function, which removes digits to the right of the decimal. Because (1) the unbiased data are assumed to follow Benford's distribution, (2) Benford's distribution is invariant to changes in scale and (3) any data manipulation is assumed proportional to order of magnitude, collapsing/expanding the data does not distort any percentage manipulation present in the data. Numbers with less than two significant figures after collapsing/expanding are deleted from the data set.

The DFM compares the mean of the collapsed data set with the mean of an unbiased data set that contains the same number of observations over the same range and that follows Benford's Law. The actual mean, AM, of the collapsed data set is:

$$AM = \frac{\sum X_{collapsed}}{N}, \qquad (16)$$

where N is the number of observations. Nigrini [17] shows that the expected mean, EM, of an unbiased Benford data set with N observations over interval [10,100) is:

$$EM = \frac{90}{N \cdot (10^{1/N} - 1)} . \tag{17}$$

The Distortion Factor, DF, is calculated as:

$$DF = \frac{100 \cdot (AM - EM)}{EM} . \tag{18}$$

DF gives the (signed) average percentage manipulation of the data. Nigrini shows that the expected value of DF is zero and that the standard deviation of DF, STD(DF), is:

$$STD(DF) = \frac{\left[11 \cdot N \cdot (10^{1/N} - 1)\right] - \left[9 \cdot (10^{1/N} + 1)\right]}{9 \cdot N \cdot (10^{1/N} + 1)} . \tag{19}$$

Since AM is the mean of N random variables, by the central limit theorem the distribution of DF approaches a normal distribution with mean zero and variance $\text{STD(DF)}^2$as N increases. As a result, Z-test statistics may be computed for DF for relatively large N.

## 4. EMPIRICAL APPLICATION: NORTH CAROLINA VOC AIR POLLUTION DATA

In this section we provide an example of how DFA might be applied by environmental regulators.

### 4.1 Data

We consider the most recent (1996-1998, depending on each firm’s audit schedule) data on annual volatile organic compounds (VOC) air emissions for all permitted North Carolina firms [21]. The data are self-reported by firms to the North Carolina Divisions of Air Quality (NCDAQ). Firms are classified by NCDAQ into three categories: Title V[5] Facilities, Synthetic Minor Facilities, and Small Facilities. Title V facilities emit 100 or more tons/year of at least one criteria air pollutant[6], or 10 or more tons/year of at least one hazardous air pollutant, or 25 or more tons/year of all hazardous air pollutants combined. Synthetic Minor facilities “would be minor facilities except that the potential emissions are reduced below the thresholds by one or more physical or operational limitations on the capacity of the facility to emit an air pollutant. Such limitations must be enforceable by the EPA . . . ” [21]. Minor, or “small,” facilities are all facilities other than Title V or Synthetic Minor. All facilities must pay both an initial emissions program application fee and annual permit fees to NCDAQ. Title V facilities (only) must pay an additional, annual fee per ton of air emissions on all air emissions (both criteria and hazardous).

Is it plausible to assume that unbiased pollution emissions data should follow Benford's Rule? The data set meets the practical requirements for conformance to Benford’s Law: (1) the data describe a single type of phenomena (e.g., air pollution emissions), (2) they have no theoretical maximum or minimum (except zero) values, (3) they are expected to contain more small numbers than large numbers, (4) they are not expected to contain systematic number duplication, (5) they do not consist of systematically-assigned numbers and (6) they are spread across several digital orders (from 0.01 to 100,000 tons/yr). However, perhaps the best justification for the Benford Law assumption is Hill’s [11] result that a dataset consisting of random samples from a random collection of distributions will converge to the Benford Law distribution. If the data generating mechanism can be characterized as random samples from a random collection of distributions, then digital frequencies should follow Benford’s Rule (Hill's Generalized Significant Digit Law). Assuming such a data generating mechanism applies in the present case, if the distributions of significant figures do not follow Benford's Rule, then there is reason to suspect that the data have been manipulated ex post.

Descriptive statistics on VOC emissions for the Title V and small facility categories are presented in Table 2. Facilities with less than 1 ton/yr of VOC emissions were excluded from the analysis, as the data for such facilities would have too few significant figures for analysis. This reduced the number of small facility observations from 1993 to 631 and the number of Title V facility observations from 431 to 380. For each data set, the mean is greater than the median and the skewness is positive, necessary conditions for unbiased Benford data sets. Figures 3 and 4 present the VOC emissions data (uncollapsed) distributions by facility size sorted in ascending order of emissions level. These distributions have the general form of geometric sequences, further supporting the assumption that the unbiased data approximate Benford’s Law.

## 4.2 DFA and DFM Test Results

Assuming the unbiased data follow Benford's Law, we apply the standard DFA tests to the VOC data. All DFA tests are conducted using the DATAS® 2000 digital analysis software package [27]. We begin with the First Digits Test. Figures 5 and 6 present the distributions of first digits of the VOC data and associated confidence intervals (5% confidence level) by facility size category. The first digits graph for Title V firms (Figure 5) indicates that the data generally conform to a Benford distribution, although the upper digits appear to be somewhat underrepresented and digits 1 and 2 appear overrepresented. The overrepresentation of first digit 1 is statistically significant, as is the underrepresentation of first digit six. This pattern is consistent with downward bias in the data.

For small size category firms (Figure 6), digits one and two again appear overrepresented, and the larger digits appear underrepresented. Only digit nine departs significantly from Benford's distribution. The underrepresentation of digit nine is somewhat suspicious, as numbers with leading digit nine lie just below the Title V emissions threshold of one hundred tons/yr. Annual permit fees jump by an order of magnitude at this threshold and annual emissions fees are not required for firms below the threshold. Note further that digit eight is relatively well represented among the upper digits. Firms with emissions in the nineties may be fudging their numbers to lie within the eighties to avoid appearing "close" to the emissions threshold and attracting regulatory attention.

To investigate the digital frequencies with more precision, we consider the distributions of the first two-digit combinations of the VOC data and associated confidence intervals (5% confidence level) by facility size category. Figure 7 shows that the Title V facility data exhibit relatively large spikes at emissions levels 15, 27, 43 and 50, though only the latter two spikes are statistically significant. Although a few significant spikes may occur due to chance alone, it typically would not be a large task for regulators to investigate possible explanations for these few overrepresented combinations. In some cases, spikes may be easily explained by factors other than fraud or evasion. For example, perhaps an unusually common (in the statistical sense) boiler type produces 430 tons/year of emissions when used at capacity, causing “43” to appear more frequently than expected. On the other hand, if a regulatory threshold were 440 tons/year, then a spike at 43 might raise suspicion. If so, the names of firms with emissions levels beginning with digits “43” could be extracted from the database and perhaps inspected sooner or with a greater frequency until an explanation for the unusual occurrence surfaced. The lack of digit combinations in the high 50s and low 60s in the Title V data is also unexpected, though an explanation for this observation is not apparent to the authors.

The first two-digit combination data for the small size category firms (Figure 8) show significant positive spikes at 10, 22 and 44. Again, a few spikes would be expected due to chance alone. Whether these spikes should be investigated would depend on additional knowledge of the regulatory environment. Of greater interest in the small facility data is the underrepresentation of digit combinations in the 90s. Firms may be lowering emissions data to avoid the 100 ton/year threshold for Title V classification. This lack of small facility digit combinations in the 90s is more suspicious since Title V two-digit combinations in the 10s, values just above the small facility 90s combinations, are well represented. In contrast, the 90s combinations are well represented in the Title V data.

Table 3 presents the results of the Number Duplication Test by facility size category. The ten numbers in each data set that occur with highest frequency are listed together with their respective frequencies.  All (uncollapsed)[7] data values greater than 1 ton/yr. are considered. Recall that the purpose of this test is to investigate simple number duplication as a possible cause of distortions in the digit frequency data.  For Title V facilities, there is not unusual duplication of any data value; in fact, no data value appears more than twice in the data set.  However, duplication of numbers as specific as "434.32" arouse curiosity.  In fact, the first three numbers on the duplication list indicate problems in the data set.  Numbers 434.32 and 332 appear twice in the data set because the data records from which they are drawn were apparently keyed in twice by NCDAQ by mistake.  Number 311.94 appears twice in the data set because the data records for a particular firm for two different years are included in the data set, even though the data should include only the data from the most recent inventory year for each firm.  Aside from the fact that two years of data for a given firm should not appear in the data set, if the data for the firm are correct, then the firm is reporting the same *exact* emissions values year-on-year.  If reported values represent field measurements typically subject to variation, then these repeated values raise suspicion.  Hence, the Number Duplication Test can reveal abnormalities in the data set as well as simple number duplication.  However, number duplication may occur due to chance alone and does not necessarily indicate a data problem.  For example, in the small facility data, the value "10.2" is duplicated six times.  When the corresponding data records were investigated, no irregularities were discovered.

Figures 9 and 10 present results for the Last Two Digits Test.  The distributions of the last two-digit combinations (in the uncollapsed data) and associated confidence intervals (5% confidence level) by facility size category are shown.  The significant spikes above multiples of 10 are clear evidence of rounding in the last two digits.  However, this finding is likely of little regulatory concern in the present case, as it concerns only fractions of a ton/yr per source.

Table 3 presents the results of the Round Numbers Test by facility size category. The test looks for excessive rounding in reported data. In contrast to the Last Two Digits Test, the Round Numbers Test considers rounding in the integer (left of the decimal point) digits only. The Round Numbers data for the Title V facilities indicate no more rounding than expected is occurring to the left of the decimal point. That is, facilities do not appear to be rounding to the nearest 5, 10, 25, 100, etc. tons/yr. when reporting emissions levels. The interpretation of results for the small facility firms is the same. However, the small facilities data indicate also that firms may be *avoiding* round numbers, as the observed proportions of round numbers in the data set are significantly less than the expected proportions for several round numbers values.

Nigrini's DFM test determines the direction, magnitude and significance of average distortion in a data set. The DFM compares the actual mean (AM) of the collapsed data with the expected mean (EM) of a data set with the same number of observations distributed according to Benford's Law. Consider Title V facilities and small facilities. Title V facilities have an incentive to distort reported emissions downward, as they must pay emissions fees per ton of emissions. Small facilities have an incentive to distort reported emissions downward, at least at higher emissions levels, in order to avoid classification as a (fee paying) Title V facility. Although both types of facilities have incentives to underreport emissions, they may not do so. If underreporting does occur, its relative magnitude may differ for the two facility categories. We test two null hypotheses using the DFM test:

> Null Hypothesis 1: For each facility class, the average percentage distortion in reported emissions values is zero (i.e., $DF_{\text{Title V facilities}} = 0$, $DF_{\text{small facilities}} = 0$).
>
> Null Hypothesis 2: The difference across facility classes in average percentage distortion in reported emissions values is zero. (i.e., $DF_{\text{Title V facilities}} = DF_{\text{small facilities}}$)

Table 4 presents test results (5% confidence level) for Null Hypothesis 1 by facility size category. For the Title V facilities, AM is 9.97% lower than EM, a result that is significant at the 5% level of confidence. This means that the average of all the numbers in the Title V data set is 9.97% lower than expected, or that the numbers in the Title V data set appear to be distorted downward by 9.97%, on average. Similarly, DFM test results for the small facility category indicate that actual mean emissions are less than expected mean emissions by 9.45%, a result that is significant at the 5% level.

The second null hypothesis concerning significant difference between Title V facility and small facility DF's is tested via a Z test of differences in means. Because the DF variances are significantly different (at the 1 % level of significance) across facility size categories, we use the following large sample Z test that allows for differences in category variances:

$$Z = \frac{\left(DF_{Title\ V} - DF_{Small}\right) - 0}{\sqrt{\frac{s^2_{Title\ V}}{N_{Title\ V}} + \frac{s^2_{Small}}{N_{Small}}}}, \tag{20}$$

where $N_{\text{Title V}}$ and $N_{\text{Small}}$ denote sample sizes and $s^2_{\text{Title V}}$ and $s^2_{\text{Small}}$ denote DF variances for Title V and Small facility categories, respectively. The calculated Z statistic of –1.7328 does not exceed the critical Z value of –1.96 (two-tailed test, 5% significance level). Hence, we do not reject the second null hypothesis that the degree of distortion in the data as measured by the category DF's is the same (at a 5% significance level) across facility size categories.

## 5. SUMMARY AND DISCUSSION

This paper explores the use of Digital Frequency Analysis (DFA) based on Benford's Law to detect evidence of non-compliance in self-reported pollution emissions data. The theory of Benford's Law is reviewed, statistical justifications for its wide applicability to empirical data sets are presented, and several tests for dataset irregularities based on Benford's Law are described. These tests are being adopted by public accounting firms for use in detecting fraud in financial data. We argue that these techniques can be employed by environmental regulators when attempting to detect fraud in self-reported pollution emissions data.

In a case study of volatile organic compound air pollution emissions data in North Carolina, DFA tests indicate that the data appear to contain distortions that reduce mean emissions by about 9.5-10% below expected levels. This *relative* distortion is similar across facility size categories, although the distortion in absolute emissions levels would be larger for the larger Title V facilities. While the Last Two Digits Test indicates that firms are rounding emissions numbers, the Round Numbers Test shows that rounding is not occurring in the larger digit positions to the left of the decimal point. Hence, rounding is not the source of the sizeable 9.5-10% distortion in mean emissions. First Digit and First Two-Digit Tests indicate that the Title V facility data exhibit unusually high occurrences of the digit combinations 15, 27, 43 and 50 and an unusually low proportion of data values beginning with digits 5x and 6x. These same tests indicate that the small facility data exhibit unusually high occurrences of the digit combinations 10, 22 and 44 and an unusually low proportion of data values beginning with digits 9x. Hence, similar distortions in average emissions levels across facility size categories may have different causes—a lack of 5x and 6x numbers in the Title V data and a lack of 9x numbers in the small facility data. Given that an emissions level of 100 tons/yr. represents the regulatory threshold for categorization as a Title V firm, which entails significant increases in permit and emissions fees, small firms with emissions of 9x tons/yr. may be distorting emissions numbers downward to avoid Title V classification.

DFA may be used to determine the relative likelihood of fraudulent underreporting across other pollution source categories, such as across industry types, across geographic regions, across number of pollutants emitted per source, across urban vs. rural source location, etc. If evidence of underreporting is found to vary by industry type, for example, then the efficiency of regulatory auditing might be improved by reallocating agency resources toward industries exhibiting higher evidence of underreporting. In addition to further applications within the field of pollution control, other potential regulatory applications involving natural resources include detecting cheating in fishery landings data, hunting data and cattle grazing data.

DFA in its current form faces several limitations. First, DFA will not detect an equal percentage multiplication of all elements in a dataset (due to the scale invariance property of Benford's Law). Similarly, DFA will not detect systematic multiplication by random numbers drawn from a closed interval, nor will it detect systematic addition or subtraction of a constant, if the constant is sufficiently small to leave first few digits unaffected by the manipulation. Second, the data values in a dataset must span at least one digital order (i.e., 1-10, 10-100, 100-1000, etc.) for DFA to be useful. Third, if regulatory agencies adopt DFA as an auditing tool, we would expect sophisticated non-compliant firms to adjust self-reported data in ways that avoid detection. In particular, we would expect polluting firms to restrict the types of self-reported data manipulations to those that would be consistent with Benford's Law or some analogous law implied by the relevant data generating mechanism. However, although firms may still find it possible to cheat, application of DFA as an audit tool places additional restrictions on self-reported data, reducing the “degrees of freedom” in cheating activity. For example, although a firm may still be able to cheat on self-reported data by multiplying each value by a given percentage, the firm may not be able to subtract a given amount from each reported value, or reduce each reported value to some relevant threshold, without detection. Furthermore, in cases where multiple firms are analyzed together, an individual firm would need to know the data values reported by other firms in order to pick a fraudulent data value that would “fit” the expected digit distribution across firms. In effect, the use of DFA makes it more difficult to cheat, raising the cost to the firm of cheating activity. A useful extension of the analysis would be to model the potential economic welfare gains from reduced cheating activity due to DFA implementation. For example, one might investigate the incorporation of adherence to Benford's Law as a constraint on strategic emissions reporting behavior in the context of regulatory mechanism design models.

From a more general statistical viewpoint, any *data generating mechanism* [13] implies patterns of digital frequencies in generated data. Although Benford’s Law may well describe the predicted digital frequency patterns associated with many data generating mechanisms (for reasons discussed in section 3), the expected digital frequencies in some empirical situations may follow some other type of distribution. Nonetheless, generalized DFA is still useful in such cases, as it enables comparison of observed frequencies with the expected digital frequencies implied by the maintained data generating mechanism, whatever its structure may be. Future work might develop digital frequency tests applicable to alternative data generating mechanisms. If we trust our statistical model specification, then deviation of observed from predicted frequencies is a sign that data may be manipulated. Of course, such deviation may simply signal misspecification of the statistical model rather that data manipulation, but identification of model misspecification is useful also for pointing out situations where our understanding of firms' pollution behavior (or pollution reporting behavior) may be poor.

## REFERENCES

1. A. K. Adhikari, Some results on the distribution of the most significant digit, *Sankhya Series B* **31**, 413-42 (1969).

2. A. K. Adhikari and B. P. Sarkar, Distribution of most significant digit in certain functions whose arguments are random variables, *Sankhya Series B* **30**, 47-58 (1968).

3. R. U. Ayers and A. V. Kneese, Production, consumption and externalities, *American Economic Review* **59**(3), 282-297 (1969).

4. F. Benford, The law of anomalous numbers, *Proceedings of the American Philosophical Society* **78**(4), 551-572 (1938).

5. J. Boyle, An application of fourier series to the most significant digit problem, *American Mathematical Monthly* **101**(9), 879-886 (1994).

6. C. Carslaw, Anomalies in income numbers: Evidence of goal oriented behavior, *The Accounting Review* **63**, 321-327 (1988).

7. C. Christian and S. Gupta, New evidence on "secondary evasion," *The Journal of the American Taxation Association* **15**, 72-92 (1993).

8. W. H. Furry and H. Hurwitz, Distribution of numbers and distribution of significant figures, *Nature* **155**, 52-53 (1945).

9. S. A. Goudsmit and W.H. Furry, Significant figures of numbers in statistical tables, *Nature* **154**, 800-801 (1944).

10. T. P. Hill, Base-invariance implies Benford's law, *Proceedings of the American Mathematical Society* **123**(3), 887-895 (1995).

11. T. P. Hill, A statistical derivation of the significant-digit law, *Statistical Science* **10**(4), 354-363 (1995).

12. T. P. Hill, The first digit phenomenon, *American Scientist* **86**, 358-363 (1998).

13. G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lutkepohl and T-C Lee, "Introduction to the Theory and Practice of Econometrics, Second Edition," John Wiley & Sons, Inc., New York, NY (1988).

14. D. S. Lemons, On the numbers of things and the distribution of first digits, *American Journal of Physics* **54**, 816-817 (1986).

15. S. Newcomb, Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics* **4**, 39-40 (1881).

16. M. J. Nigrini, "The Detection of Income Tax Evasion Through An Analysis of Digital Distributions," Ph.D. dissertation, Department of Accounting and Business Law, University of Cincinnati. Cincinnati, Ohio. (1992).

17. M. J. Nigrini, A taxpayer compliance application of Benford's law, *Journal of the American Taxation Association* **18**(1), 72-91 (1996).

18. M. J. Nigrini and L. J. Mittermaier, The use of Benford's law as an aid in analytical procedures, *Auditing: A Journal of Practice & Theory* **16**(2), 52-67 (1997).

19. M. J. Nigrini, "Digital Analysis Using Benford's Law: Tests and Statistics for Auditors," Global Audit Publications, Vancouver, Canada. (2000).

20. North Carolina Department of Environment and Natural Resources (NCDENR). Personal communication. Mr. Steven Boone, Regional Air Quality Supervisor, North Carolina Division of Air Quality. Wilmington, NC. July 11, 2000.

21. North Carolina Division of Air Quality (NCDAQ). NCDAQ web site, Rules and Regulations. http://daq.state.nc.us/Rules/. (2000).

22. R. A. Raimi, The Peculiar Distribution of First Digits, *Science* **221**, 109-120 (1969).

23. R. A. Raimi, The first digit problem, *American Mathematical Monthly* **83**, 521-538 (1976).

24. J. K. Thomas, Unusual patterns in reported earnings, *The Accounting Review* **64**, 773-787 (1989).

25. U.S. Environmental Protection Agency, About the Toxics Release Inventory (TRI) Data Collection. U.S. Environmental Protection Agency internet web site address http://www.epa.gov/opptintr/tri/triwww.htm. February 10, 1999.

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

27. M. J. Nigrini. DATAS® 2000 for EXCEL 97. Allen, TX.

## ENDNOTES

[1] Nigrini [19] provides an extensive bibliography on the theoretical development and empirical application of Benford's Law in the field of accounting.

[2] Simon Newcomb [15] provides the earliest known description of Benford's Law. It appears that Benford [4] discovered the phenomenon independently, and it is Benford's paper that motivates the current interest in these issues.

[3] If we consider pollution emissions as "pieces" of original production inputs (as we might if we take Ayers and Kneese's (1967) "materials balance," or conservation of matter, approach to the study of pollution), and if we consider each firm's pollution emissions as a trial, then by Lemon's argument the distribution of pollution emissions across firms might well exhibit first digits that follow Benford's Law.

[4] Hill's Generalized Significant Digit Law holds for data measured in any base. The version of the law appropriate for base 10 data is presented here.

[5] Title V refers to Title V of the federal Clean Air Act, which specifies minimum regulations for state air pollution permit programs and fees.

[6] The criteria air pollutant data are: volatile organic compounds (VOC), nitrogen oxides (NOX), carbon monoxide (CO), fine particulate matter (PM 10), total suspended particulates (TSP) and sulfur dioxide (SO2).

[7] Uncollapsed data are considered because we are looking for duplication in the data values themselves, rather than duplication in the digits of the data values.

Table 1.

Benford's Distribution[a]

| | Relative Frequency | Relative Frequency | Relative Frequency | Relative Frequency |
|---|---|---|---|---|
| Digit Value | in 1st Digit Position | in 2nd Digit Position | in 3rd Digit Position | in 4th Digit Position |
| 0 | ---------- | 0.11968 | 0.10178 | 0.10018 |
| 1 | 0.30103 | 0.11389 | 0.10138 | 0.10014 |
| 2 | 0.17609 | 0.10882 | 0.10097 | 0.10010 |
| 3 | 0.12494 | 0.10433 | 0.10057 | 0.10006 |
| 4 | 0.09691 | 0.10031 | 0.10018 | 0.10002 |
| 5 | 0.07918 | 0.09668 | 0.09979 | 0.09998 |
| 6 | 0.06695 | 0.09337 | 0.09940 | 0.09994 |
| 7 | 0.05799 | 0.09035 | 0.09902 | 0.09990 |
| 8 | 0.05115 | 0.08757 | 0.09864 | 0.09986 |
| 9 | 0.04576 | 0.08500 | 0.09827 | 0.09982 |

## Table 1, Footnote a

The number "238" has three digits, with "2" as the first digit, "3" as the second digit, and "8" as the third digit.

The table indicates that under Benford's distribution the expected proportion of numbers with first digit "2" is 0.17609.

Similarly, the expected proportion of numbers with third digit "8" is 0.09864.

Table 2.

Descriptive Statistics for VOC Data by Facility Class

(all VOC data values >= 1 ton/yr)

| | Title V Facilities | Small Facilities |
|---|---:|---:|
| Mean | 162.4824 | 12.6268 |
| Standard Error | 12.69531 | 0.606101 |
| Median | 85.055 | 7.18 |
| Mode | 1.6 | 10.2 |
| Standard Deviation | 247.4772 | 15.22508 |
| Sample Variance | 61244.96 | 231.803 |
| Kurtosis | 17.3961 | 11.95194 |
| Skewness | 3.623078 | 2.83997 |
| Range | 1990.84 | 137.41 |
| Minimum | 1.06 | 1 |
| Maximum | 1991.9 | 138.41 |
| Sum | 61743.33 | 7967.51 |
| Count | 380 | 631 |

Table 3.

Number Duplication Test Results

| Title V Facilities (380 Obs.) |  | Small Facilities (631 Obs.) |  |
|---|---|---|---|
| (VOC data >=1 ton/yr) |  | (VOC data >=1 ton/yr) |  |
| Value | Frequency | Value | Frequency |
| 434.32 | 2 | 10.2 | 6 |
| 332 | 2 | 1 | 6 |
| 311.94 | 2 | 3 | 5 |
| 11.3 | 2 | 5.5 | 4 |
| 5 | 2 | 5.2 | 4 |
| 3.96 | 2 | 2.45 | 4 |
| 1.6 | 2 | 2 | 4 |
| 1991.9 | 1 | 1.6 | 4 |
| 1631.96 | 1 | 1.45 | 4 |
| 1618.3 | 1 | 5.7 | 3 |

Round Numbers Test Results

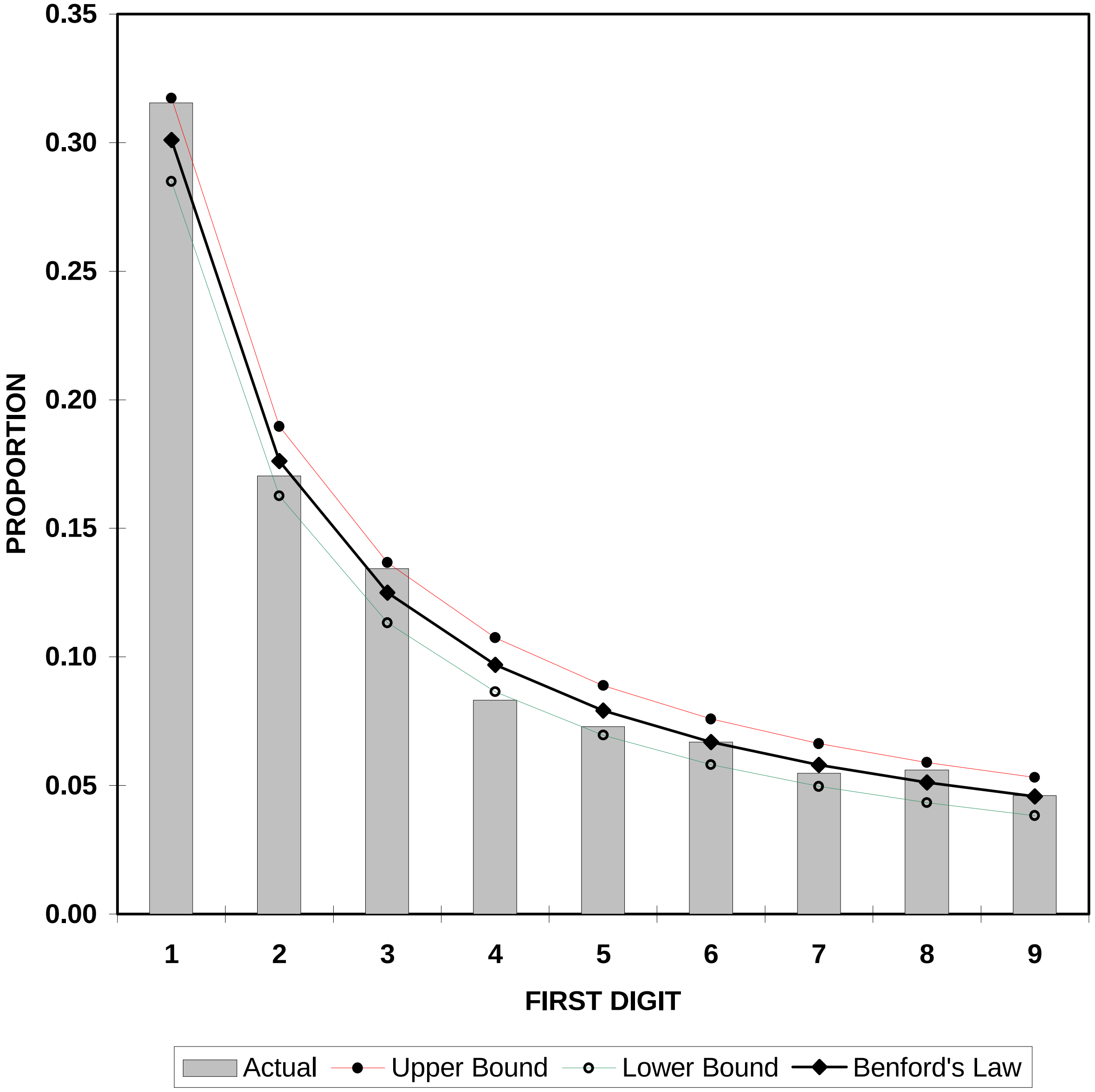
| Title V Facilities (380 Obs.) |  |  |  | Small Facilities (631 Obs.) |  |  |  |
|---|---|---|---|---|---|---|---|
| (VOC data >=1 ton/yr) |  |  |  | (VOC data >=1 ton/yr) |  |  |  |
| Multiples | Freq. | Proportion | Expected | Multiples | Freq. | Proportion | Expected |
| 5= | 75 | 0.197368 | 0.2 | 5= | 97 | 0.153724 | 0.2 |
| 25= | 9 | 0.023684 | 0.04 | 25= | 8 | 0.012678 | 0.04 |
| 100= | 2 | 0.005263 | 0.01 | 100= | 0 | 0 | 0.01 |
| 1000= | 0 | 0 | 0.001 | 1000= | 0 | 0 | 0.001 |

Table 4.

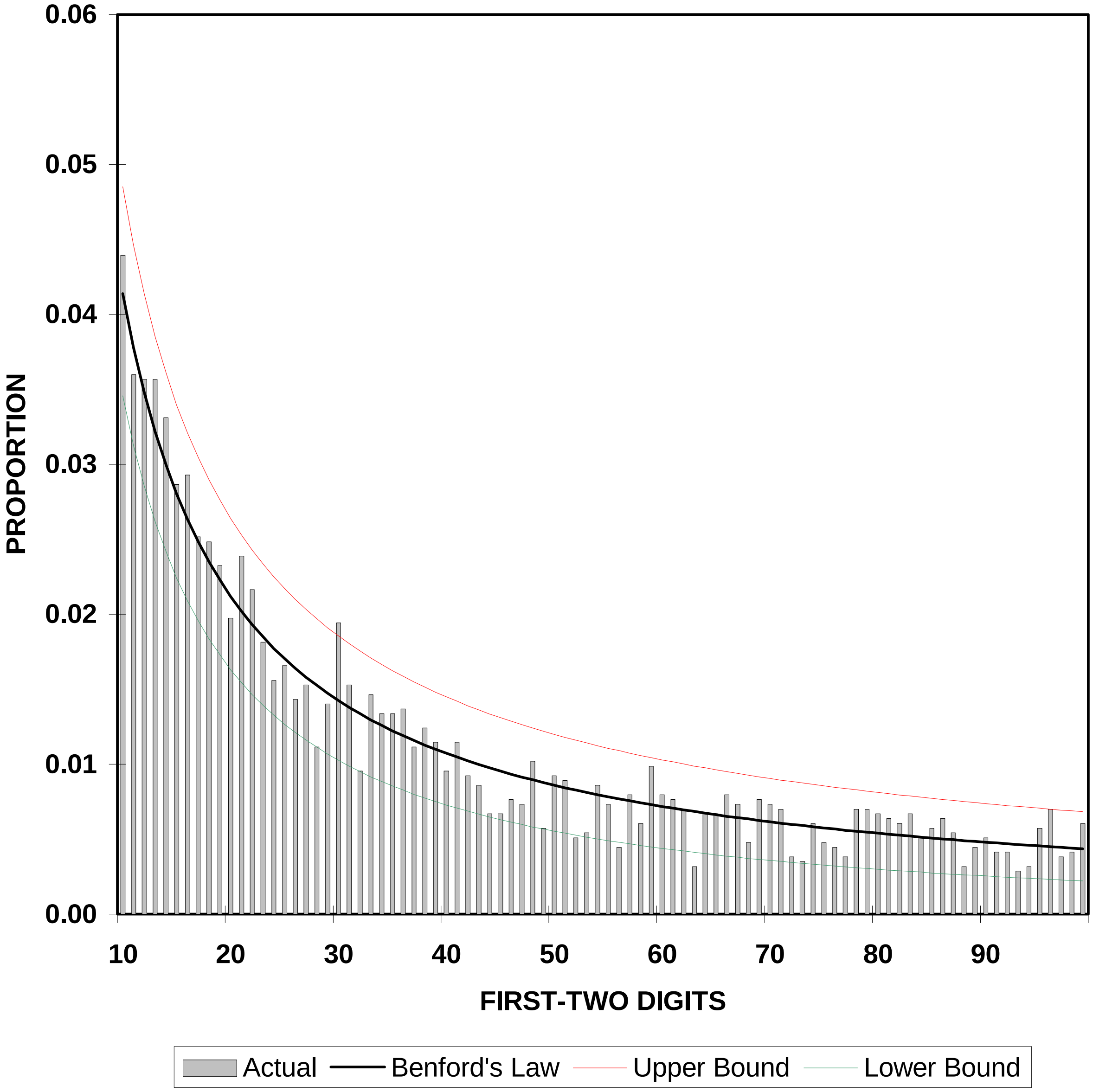
Distortion Factor (DF) Model Results

| Title V Facilities | | Small Facilities | |
|---|---|---|---|
| AM | 35.15278 | AM | 35.32775 |
| EM | 38.9682 | EM | 39.01523 |
| DF | -0.09791 | DF | -0.09451 |
| %distortion | -9.79113 | %distortion | -9.45139 |
| Std.Dev.(DF) | 0.032742 | Std.Dev.(DF) | 0.025408 |
| Zstat(DF) | -2.99041 | Zstat(DF) | -3.71978 |

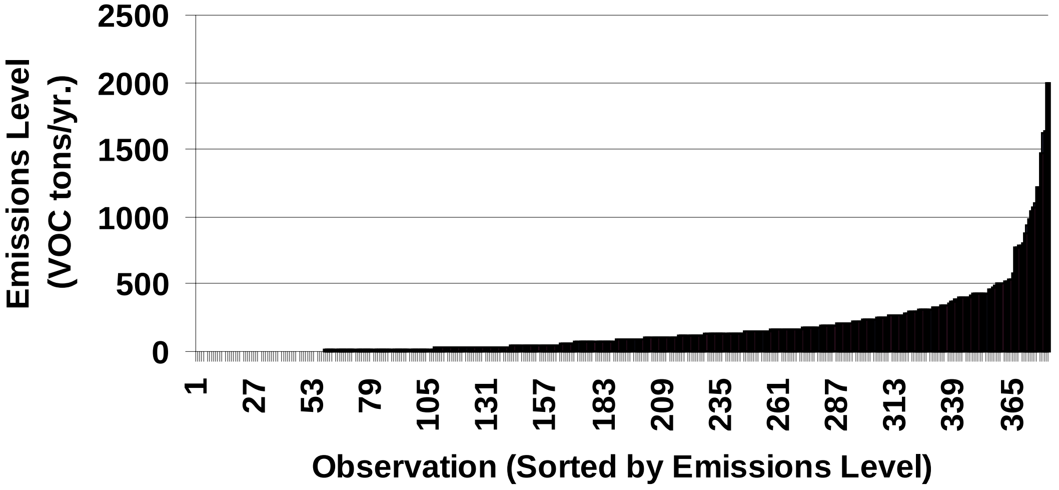
## Figure 1.
## FIRST DIGIT DISTRIBUTION
## Hypothetical Data Example



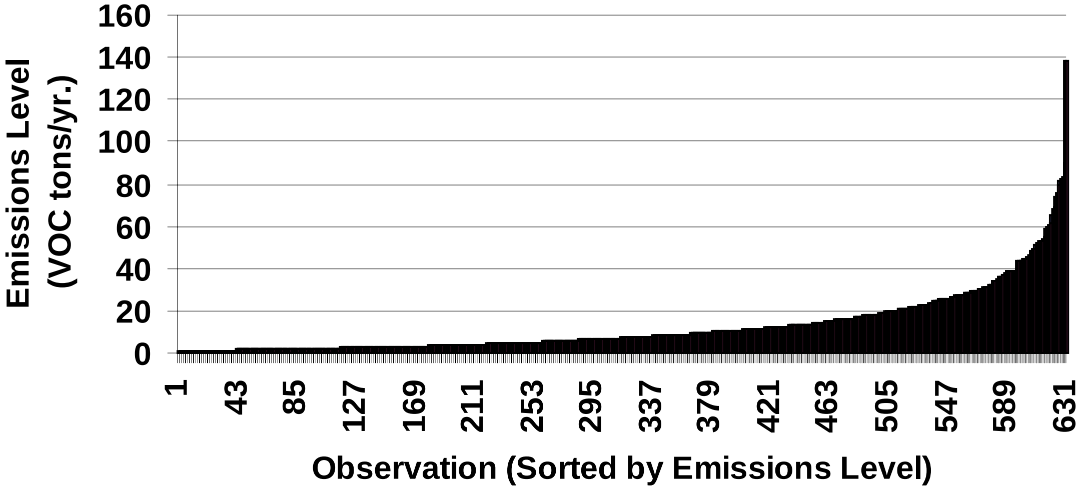
## Figure 2.
## FIRST-TWO DIGITS DISTRIBUTION
## Hypothetical Data Example



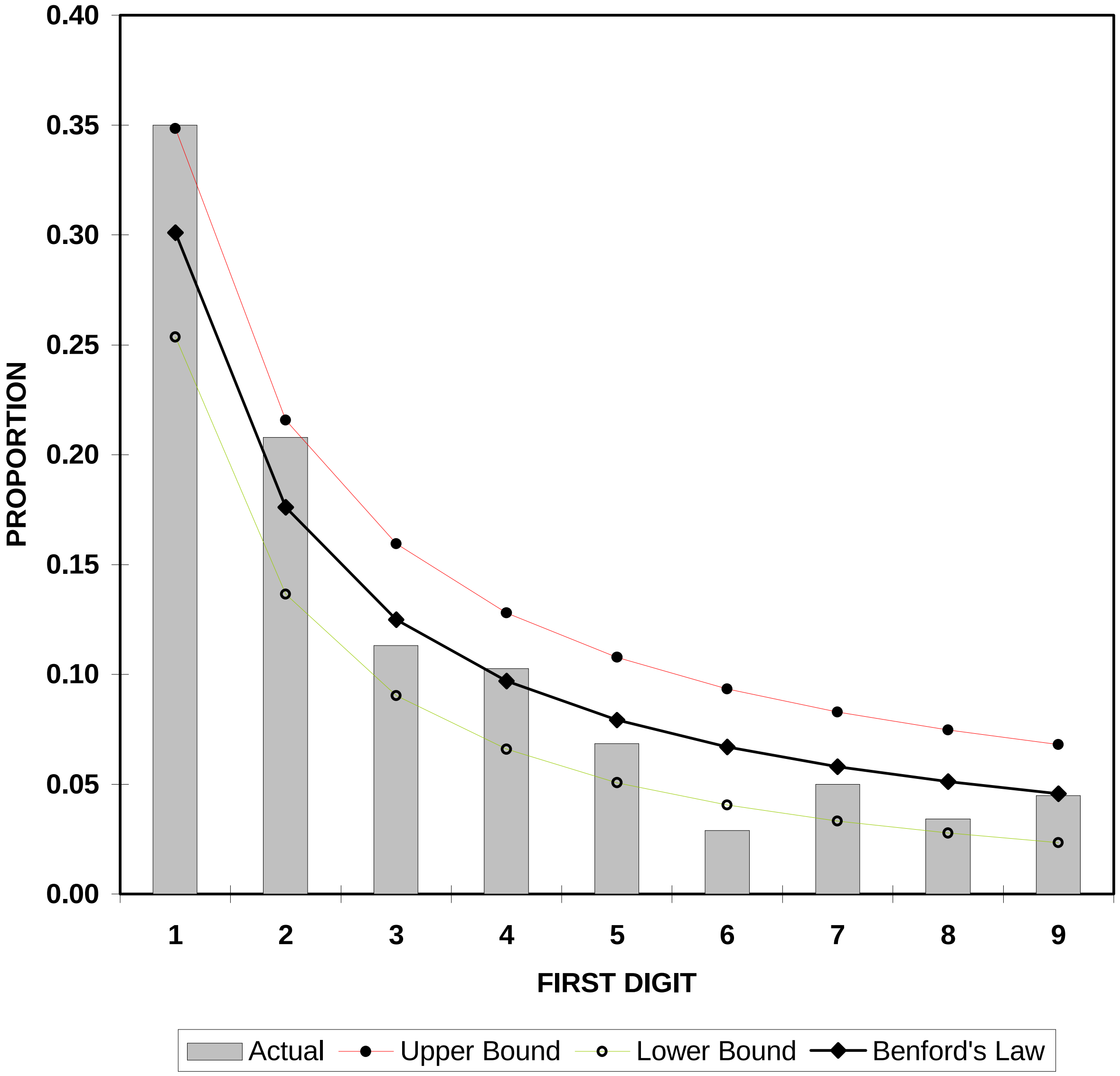
## Figure 3.
## Title V Facility VOC Emissions
## (Uncollapsed Data, Sorted By Emissions Level)



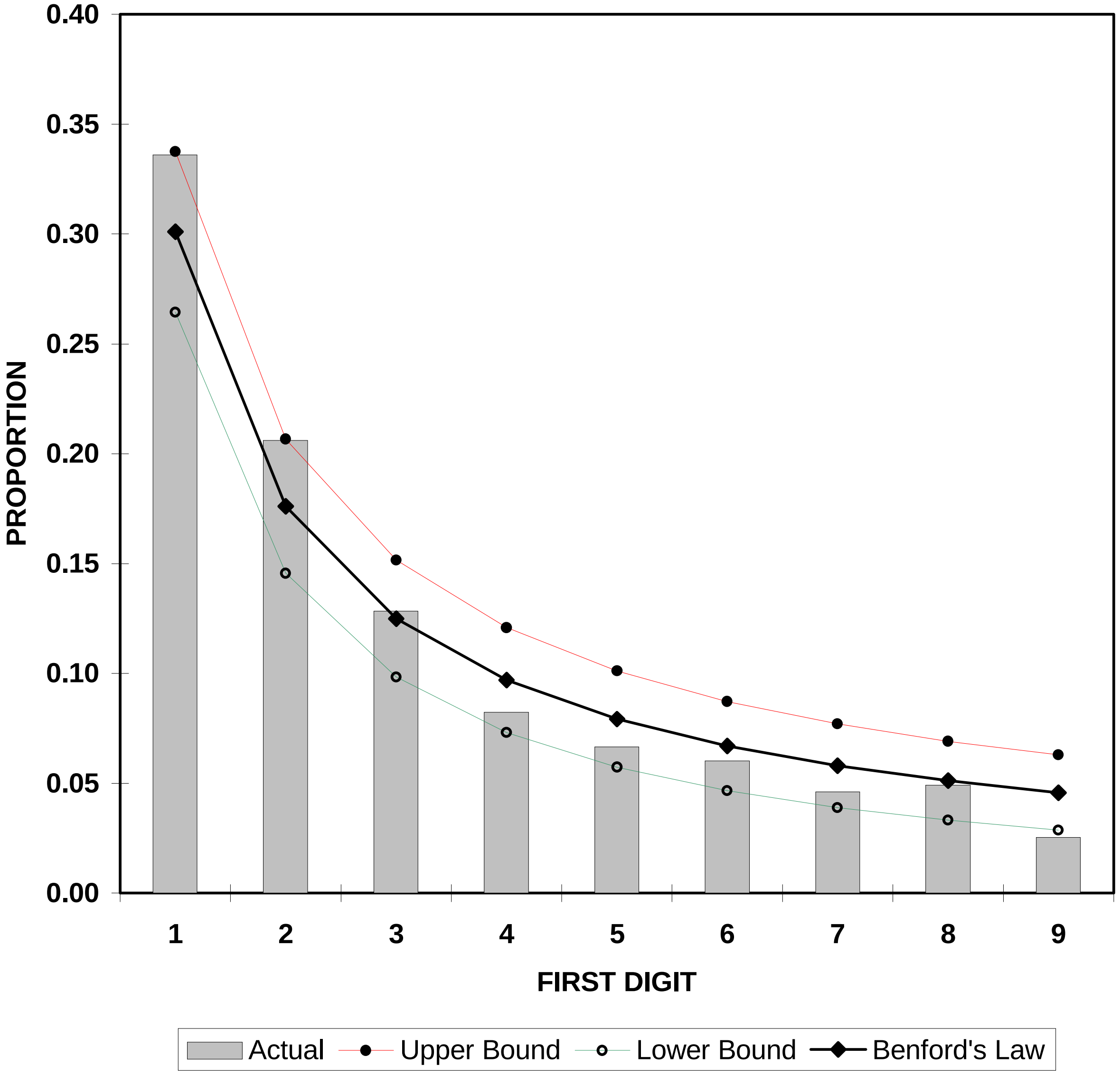
## Figure 4.
## Small Facility VOC Emissions
## (Uncollapsed Data, Sorted By Emissions Level)



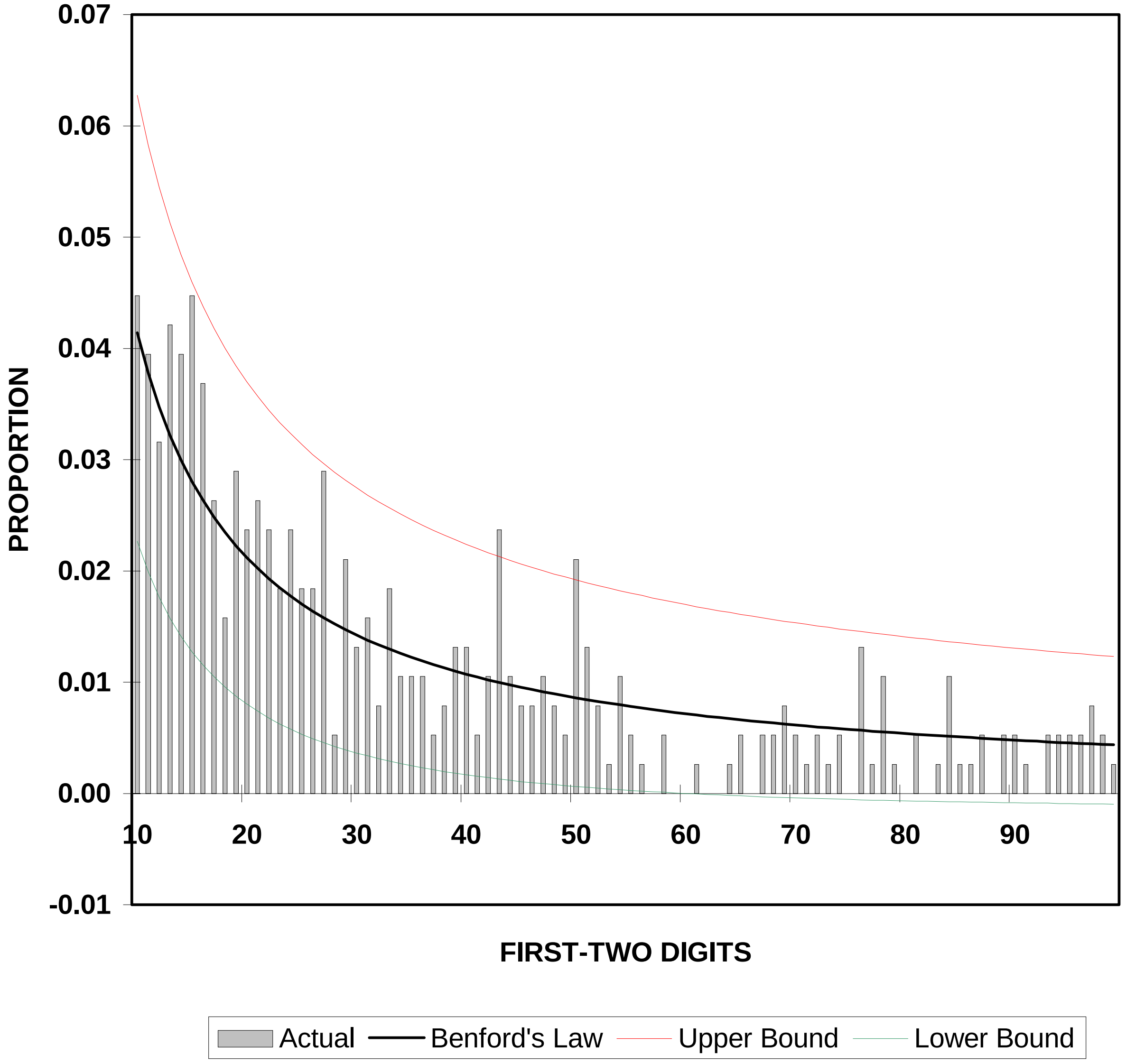
## Figure 5.
## Title V Facilities
## FIRST DIGIT DISTRIBUTION
## (Collapsed Data)



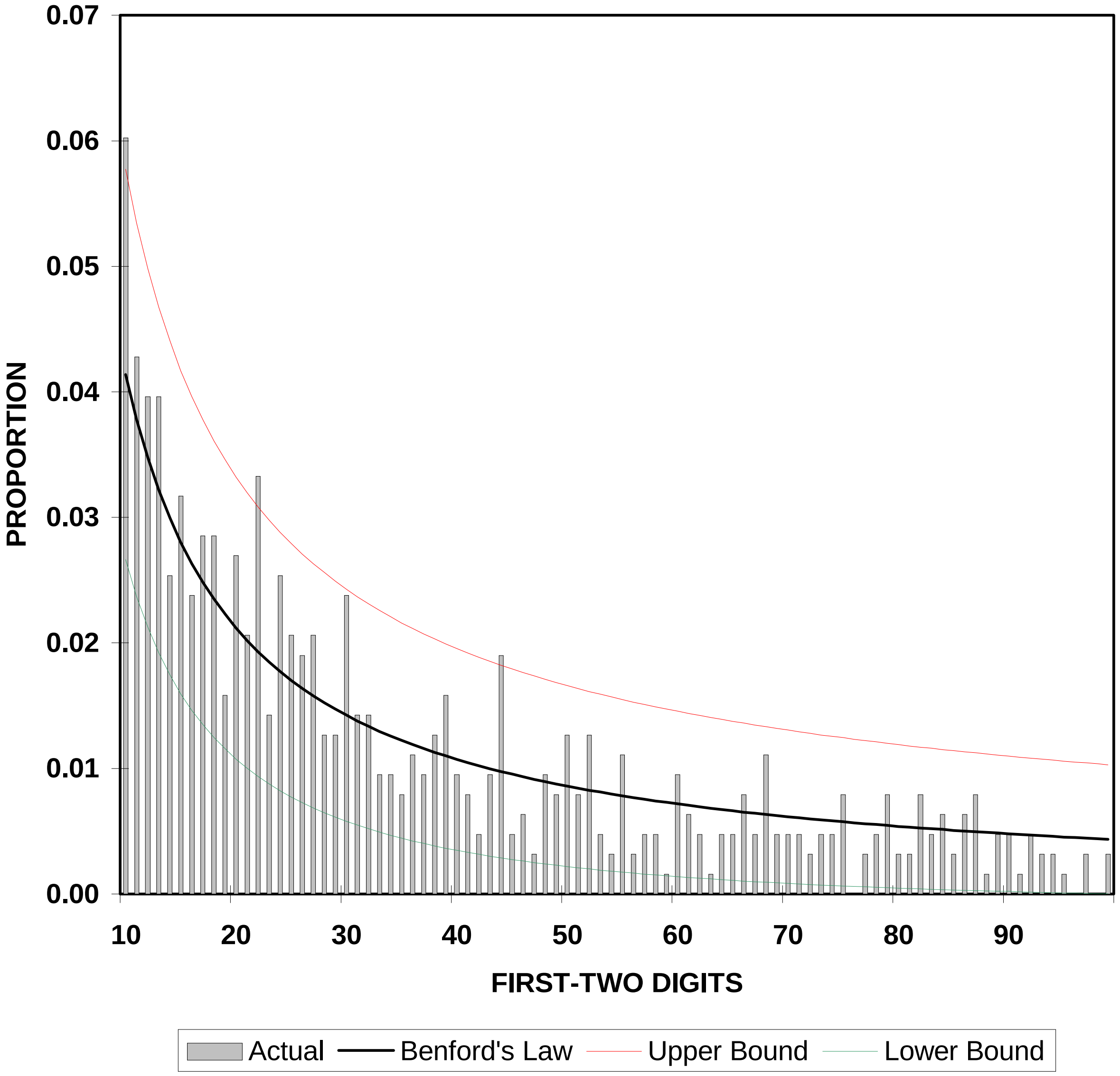
## Figure 6.
## Small Facilities
## FIRST DIGIT DISTRIBUTION
## (Collapsed Data)



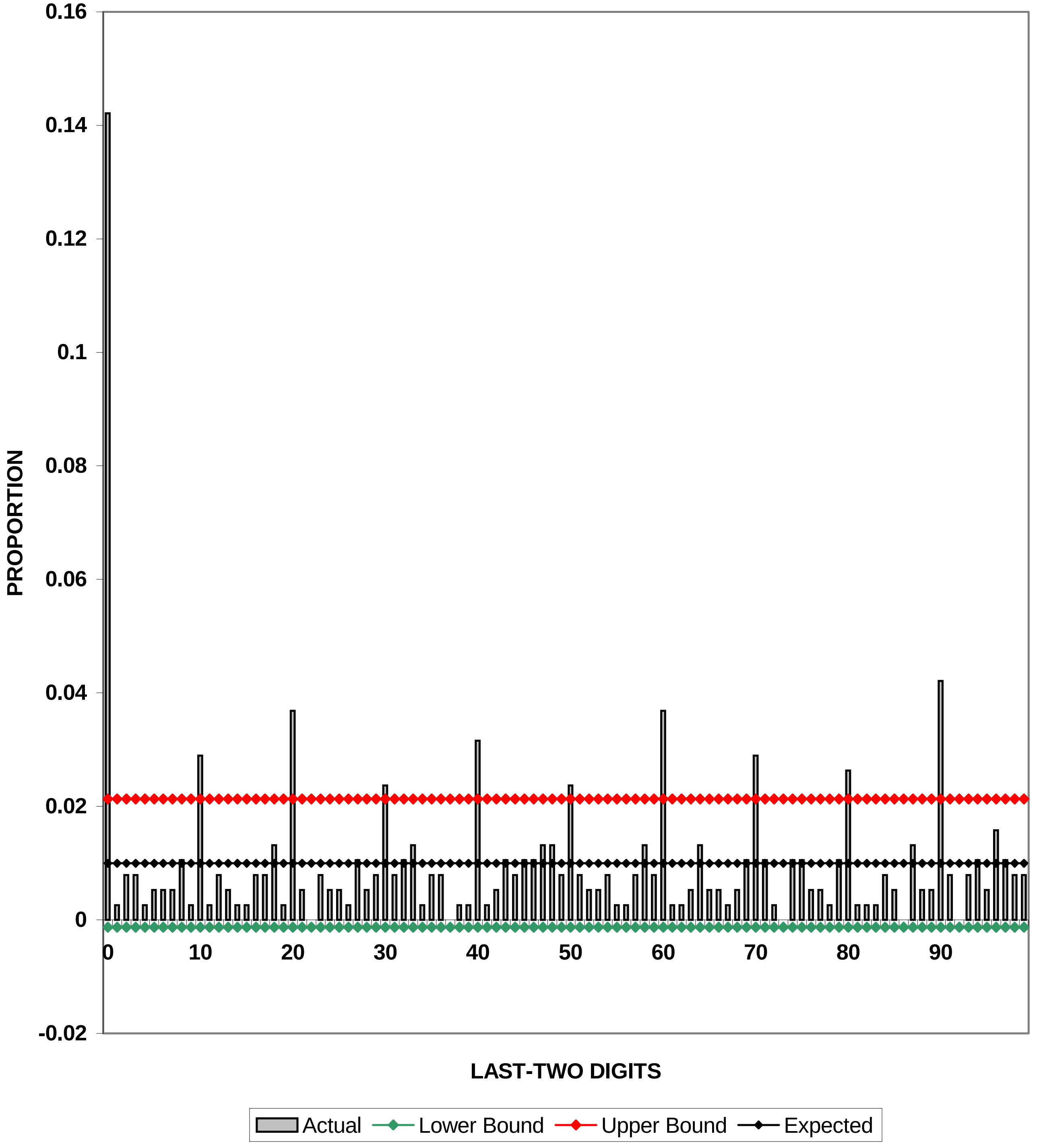
## Figure 7.
## Title V Facilities
## FIRST-TWO DIGITS DISTRIBUTION
## (Collapsed Data)



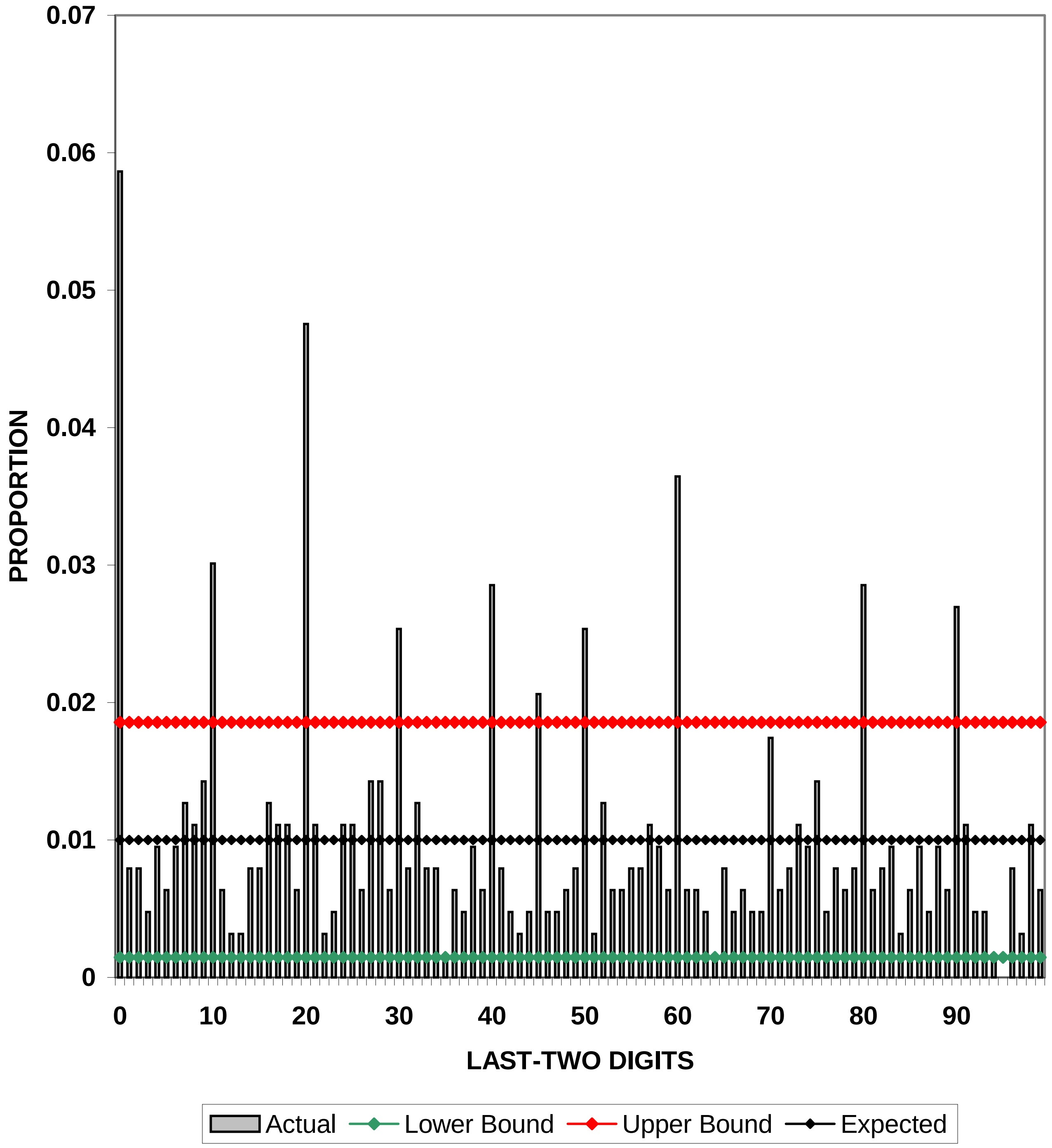
## Figure 8.
## Small Facilities
## FIRST-TWO DIGITS DISTRIBUTION
## (Collapsed Data)



## Figure 9.
## Title V Facilities
## LAST-TWO DIGITS
## (Uncollapsed Data)



## Figure 10.
## Small Facilities
## LAST-TWO DIGITS
## (Uncollapsed Data)


PROPORTION
LAST-TWO DIGITS
Actual
Lower Bound
Upper Bound
Expected