The Use of Benford’s Law for Evaluation of Quality of Occupational Hygiene Data

FRANK DE VOCHT,1 and HANS KROMHOUT2

1Centre for Occupational and Environmental Health, School of Community Based Medicine, Manchester Academic Health Science Centre, The University of Manchester, Ellen Wilkinson Building, Oxford Road, Manchester M13 9PL, UK; 2Environmental Epidemiology Division, Institute for Risk Assessment Sciences, Utrecht University, Jenalaan 18 D, 3584 CK Utrecht, The Netherlands.

Received 14 June 2012; in final form 15 August 2012; Advance Access publication 20 September 2012

Benford’s law is the contra-intuitive empirical observation that the digits 1–9 are not equally likely to appear as the initial digit in numbers resulting from the same phenomenon. Manipulated, unrelated, or created numbers usually do not follow Benford’s law, and as such this law has been used in the investigation of fraudulent data in, for example, accounting and to identify errors in data sets due to, for example, data transfer. We describe the use of Benford’s law to screen occupational hygiene measurement data sets using exposure data from the European rubber manufacturing industry as an illustration. Two rubber process dust measurement data sets added to the European Union ExAsRub project but initially collected by the UK Health and Safety Executive (HSE) and British Rubber Manufacturers’ Association (BRMA) and one pre- and one post-treatment n-nitrosamines data set collated in the German MEGA database and also added to the ExAsRub database were compared with the expected first-digit (1BL) and second-digit (2BL) Benford distributions. Evaluation indicated only small deviations from the expected 1BL and 2BL distributions for the data sets collated by the UK HSE and industry (BRMA), respectively, while for the MEGA data larger deviations were observed. To a large extent the latter could be attributed to imputation and replacement by a constant of n-nitrosamine measurements below the limit of detection, but further evaluation of these data to determine why other deviations from 1BL and 2BL expected distributions exist may be beneficial. Benford’s law is a straightforward and easy-to-implement analytical tool to evaluate the quality of occupational hygiene data sets, and as such can be used to detect potential problems in large data sets that may be caused by malcontent a priori or a posteriori manipulation of data sets and by issues like treatment of observations below the limit of detection, rounding and transfer of data.

Keywords: Benford’s law; data quality; exposure databases; MEGA; NEDB; rubber manufacturing

INTRODUCTION

Occupational hygiene depends to a large degree on accurate measurements of chemicals to evaluate inhalation, dermal and sometimes ingestion exposure of workers to harmful substances. In principle, occupational exposure measurement can be used for risk analysis by evaluating compliance with exposure limit values but some, especially larger data sets, can also be used in exposure assessment for epidemiological research and for the evaluation of ‘exposure scenarios’ within REACH (European Union regulation concerning the Registration, Evaluation, Authorisation & restriction of Chemicals). There are various ways in which these occupational hygiene measurement databases may be compromised. Most notably, given the potential implications (financial and other) of exceeding limit values for companies, it is not beyond imagination that measurement data are tampered with by, for instance, removal of high measurement values or a posteriori changing of the results prior to inspection. Nonmalicious errors
that could compromise the validity of the data may also occur, such as errors during transfer of data from one electronic medium to another, even if quality-assurance policies are in place. It is generally impossible to post hoc check all measurement data individually, both in terms of resources and in terms of access to data and transparency of the collection and handling processes involved. We describe a method that has been used in other sectors including political sciences, accounting, and banking but, to our knowledge, never in an occupational hygiene setting, which enables easy screening of large data sets to identify potential problems with the data, manipulated or other, for further, more in-depth investigation.

Benford’s law (or Newcomb-Benford law) is the contra-intuitive empirical observation that the digits 1–9 are not equally likely to appear as the initial digit in numbers resulting from the same phenomenon (Benford, 1938). This law describes that the probability that a multidigit number starts with any particular (nonzero) first digit is (where \( P \) is the probability and \( d_1 \) is the first digit 1–9):

\[
P(d_1) = \log_{10} \left( 1 + \left( \frac{1}{d_1} \right) \right). \quad (1)
\]

This was first observed in 1881 by Simon Newcomb, who observed that the first pages of books of logarithms were much more worn than the last pages. He inferred from this that scientists looked up numbers starting with the numeral one more often than those starting with two, three, and higher numbers. Not until 1938, when Frank Benford published a paper (Benford, 1938) describing this phenomenon together with the analyses of diverse data sets to prove this, did it receive attention and was referred to as ‘Benford’s Law’. In 1995, Hill (1995) provided proof for Benford’s law, showing that it results from the fact that those sets of numbers that conform to the Benford distribution are second-generation distributions (i.e. combinations of other distributions). Benford’s law applies to data from random processes with comparable underlying distributions (Formann, 2010) and as such has been shown to apply to, for example, accounting data (Durtschi et al., 2004) and to air pollution data (Brown, 2005); it is especially useful for skewed distributions like the log-normal distribution generally encountered in occupational hygiene exposure measurement data (Deckert et al., 2010). In contrast, data such as assigned numbers like invoice numbers, numbers influenced by human thought, or numbers with a built-in maximum or minimum (before they are registered) are not expected to conform (Nigrini and Mittermaier, 1997). An important advantage is that Benford’s law is scale invariant so that it can be used for many types of data, even after multiplication by a nonzero constant. For log-normal distributions generally encountered in occupational hygiene, it has further been shown that the fit with the Benford expected distribution generally increases with increased variance of the distribution (Formann, 2010).

It has been shown that manipulated, unrelated, or created numbers usually do not follow Benford’s law, which can be ascribed to most people’s misconceptions of randomness and distributions of real data (Brown, 1995). However, it has also been shown that non-fraudulent causes, such as processing inefficiencies (like a programming error in data manipulation scripts), issue during data transfer (like not all data is transferred or numbers are cutoff to two decimals), or errors (like manual typos during copying of data) may cause data to diverge from the Benford distribution (Nigrini, 1999). As a result Benford’s law can be used as a simple, effective way to evaluate large data sets in wide variety of fields. For example, in accounting, auditors use this to identify operational discrepancies and to uncover fraud (Durtschi et al., 2004), it has been shown to be a robust method to screen air pollution data collected by stationary routine monitors for errors in collection or handling of data (Brown, 2005), it has been used to detect fraud in elections, and it has been used to identify Y2K computer problems (Nigrini, 1999).

If occupational hygiene data sets would be compromised, most importantly if data had been manipulated to adhere to a specific exposure limit value, one would expect the initial digits of measurement values above the limit value to occur less often than expected or lower values to occur more often than expected. For more subtle changes made a posteriori to measured values, this may still be observed in the distributions of the second digits. Specifically for occupational hygiene data sets, in which measurement strategies may not be (entirely) random, resulting in a tendency to be biased toward low values (best-case sampling) or peak exposures (worst-case sampling), respectively, depending on the purpose of the survey (Galea et al., 2009; Van Tongeren et al., 2009; Agostini et al., 2010). Deviations of first digits from the Benford distribution may indicate such nonrandom sampling because biased sampling with the aim of reducing perceived exposure is expected to have a large impact on measured concentrations. Deviation of
the second digits may indicate more subtle changes, originating from a posteriori changes when measured values are known and only small changes may suffice to achieve a target distribution. However, there are also other non-fraudulent ways at which measurement data may be compromised, such as accidental misplacement of digits when copying data from inspector reports or issues of rounding when numeric data are transferred.

Similar to air pollution data from routine measurement programs (Brown, 2005) occupational hygiene measurement data are also expected to adhere to Benford’s law as they are generally best described by the log-normal distribution and measured values do not depend on the input from the occupational hygienist (i.e. within the constraints of the measurement strategy and equipment, they are generally random samples of that distribution), and this principle can be used to screen these data for errors, malicious or other.

This article aims to describe the use of Benford’s law to screen occupational hygiene measurement data sets. To illustrate the benefits of this screening method, three occupational exposure measurement data sets from the European rubber manufacturing industry collated in the ExAsRub database (de Vocht et al., 2005) that were of special interest will be used, as well as one data set prior to its addition to the ExAsRub database.

MATERIALS AND METHODS

Benford’s law

Equation (1) can be more generally described for the \(k\) first digits (referred to as a 1BL model) as:

\[
P(d_k) = \sum_{d_1}^{9} \log_{10} \left( 1 + \left( \frac{1}{d_1d_k} \right) \right); \quad d_1, 2, 3 \ldots 9 \quad \text{and} \quad d_2 = (0, 1, 2, \ldots 9).
\]

The probability of a digit \(d\) being the first digit is calculated by \(P(d) = \log_{10}(1 + \frac{1}{d})\), and the probability of the second digit being \(d\) is calculated by \(P(d) = \log_{10}(1 + \frac{1}{10d})\).

Hence the probability of observing the first two digits being ‘14’ can be calculated by \(\log_{10}(1 + \frac{1}{1+1/4}) \approx 0.303\), and the probability of the second digit being a ‘4’ \((P \approx 0.100)\) can be obtained by repeating these calculations for first digits 1–9 using equation (3) and adding them together. The expected probabilities, based on equations (2) and (3) for the first and second digits, are shown in Table 1.

Further extensions to the third, fourth, etc. are also possible, but will not be evaluated in this article.

Occupational hygiene data

To support future epidemiological research in the European Rubber Manufacturing Industry an European Union-sponsored Concerted Action was initiated to construct an exposure measurement database combining data from across Europe ['Improved Exposure Assessment for Prospective Cohort Studies and Exposure Control in the Rubber Manufacturing Industry' (ExAsRub)] using a common data-entry protocol (de Vocht et al., 2005).

In the UK, exposure data were provided by two organizations: (i) British Rubber Manufacturers’ Association (BRMA) (currently BTMA; British Tyre Manufacturers’ Association Limited) provided data collected by Dost et al. (2000) to which additional measurement data from BRMA companies,

<table>
<thead>
<tr>
<th>Digit</th>
<th>First digit</th>
<th>Second digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>—</td>
<td>0.120</td>
</tr>
<tr>
<td>1</td>
<td>0.301</td>
<td>0.114</td>
</tr>
<tr>
<td>2</td>
<td>0.176</td>
<td>0.109</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>0.104</td>
</tr>
<tr>
<td>4</td>
<td>0.097</td>
<td>0.100</td>
</tr>
<tr>
<td>5</td>
<td>0.079</td>
<td>0.097</td>
</tr>
<tr>
<td>6</td>
<td>0.067</td>
<td>0.093</td>
</tr>
<tr>
<td>7</td>
<td>0.058</td>
<td>0.090</td>
</tr>
<tr>
<td>8</td>
<td>0.051</td>
<td>0.088</td>
</tr>
<tr>
<td>9</td>
<td>0.046</td>
<td>0.085</td>
</tr>
</tbody>
</table>
were added and (ii) UK Health and Safety Executive (HSE) provided data from the HSE National Exposure Data Base (Burns and Beaumont, 1989). The majority of the exposure data from the 'BRMA' were collected from available occupational hygiene reports provided by the factories, and as such will have contained data collected with different sampling strategies and of varying quality. The exposure data from the NEDB database were obtained from HSE inspectors' surveys directly, and although from varying quality, these have all been collected by trained inspectors expected to have used a more uniform measurement strategy. The analyses described in this article will use the rubber process dust measurements. These have been previously used to evaluate historical time trends in exposure levels in Europe (de Vocht et al., 2008) and included personal, stationary, and source-oriented measurements. Exposure to rubber process dust, which is measured as inhalable dust, primarily occurs in the first stages of the production of rubber goods where ingredients are handled, weighed, or mixed with uncured materials or synthetic elastomers (Dost et al., 2000) and is a suspected human carcinogen (Kogevinas et al., 1998). A specific occupational exposure limits (OELs) for rubber process dust exists in the UK only and has been 8 mg m$^{-3}$ [8-h time-weighted average (TWA)] from 1987 until 1995, when it was lowered to 6 mg m$^{-3}$ (8-h TWA). Recent work (Agostini et al., 2010) has shown that exposure levels as well as annual time trends differed between both sources of data, with average personal exposures about a factor 2 higher in the NEDB data compared to data collected by the industry (BRMA) and annual trends indicated an average decrease of about 2% (95% CI 5.2–0.7%) and 4% (95% CI 4.7–3.6%) for NEDB and BRMA data, respectively. Both data sets aimed to measure occupational exposure in the same industry although anecdotal information indicates industry data primarily included measurements from larger carefully managed plants whereas the HSE inspector data were collected from small- and medium-sized companies. The different aims with which these data were collected by both organizations provide an interesting illustration of how Benford’s law can be used for screening of errors—fraudulent or other. Because we are not interested in the actual measured levels in these analyses and Benford’s law is scale invariant, all UK rubber process dust data available in ExAsRub ($N = 4310$) are used irrespective of type of measurement (personal or stationary) or suspected errors (this after all is one of the potential uses of screening data using Benford’s law); hence these data are not directly comparable to those used by de Vocht et al. (2008), which excluded data with obvious erroneous values or Agostini et al. (2010), which was based on personally monitored exposure data only and for which additional exclusion criteria were used.

Similarly, the ExAsRub database also contains measurement data extracted from the German MEGA Exposure Database. The MEGA database is the chemical workplace exposure database of the German Berufsgenossenschaften (BG) and contains exposure measurement data collected for risk assessment, prevention, epidemiological questions, and investigations of occupational diseases (Stamm, 2001). It is maintained by the Institute for Occupational Safety and Health of the German Social Accident Insurance (IFA). In 1999, the database included about 31 000 samples in approximately 4000 companies obtained by 300 measurement technicians from 55 different institutions (Stamm, 2001), and by the end of 2005, the MEGA contained 1629 million measured values for 760 chemical and 330 biological agents from about 47 000 companies (Gabriel, 2006). With such large numbers of individuals and companies involved again, the quality of the data will by definition be variable. On the legal basis of the social insurance law the inspectorates of the BGs conducted workplace measurement surveys. As such, measurement locations are not randomly selected, but are based on criteria such as supposed critical exposure situations or testing the efficiency of exposure reducing measures. The data drawn from these measuring and inspection activities are subsequently stored in the MEGA database. For the purpose of the ExAsRub project all measurement data from the rubber manufacturing industry within the MEGA database were transferred to the ExAsRub database (de Vocht et al., 2005). However, for subsequent analyses it was realized that although the same set of $n$-nitrosamines was analyzed for all German measurements in most, but not all, situations only the measurements of specific $n$-nitrosamines above the limit of detection (LOD) were added to the MEGA database. To analyze data more representative of true exposure levels all ‘missing’ measurements were added to the ExAsRub database (de Vocht et al., 2007). These values were subsequently, but prior to analyses, substituted by a constant of $\text{LOD}/\sqrt{2}$ (Hornung and Reed, 1990). Similar to the rubber process dust data all values have been used for these analyses and as such the data set included personal, stationary, and source-oriented measurements. However, 310 measurements were removed from the original data set because they were only added as ‘0’ and with no
available data it was unknown whether these were actual performed measurements at all.

Evaluation of data sets

Following Brown (2005), these data sets can be characterized, in addition to standard geometric means and standard deviations, by their numerical range ($R$), where $x_{\text{max}}$ is the highest measured concentration and $x_{\text{min}}$ the lowest, as:

$$R = \log \left( \frac{x_{\text{max}}}{x_{\text{min}}} \right). \tag{4}$$

While the extent to which any of these data sets diverges from Benford’s law can be characterized by the sum of normalized deviations, $\Delta_{bl}$, for the first digits as (with $P_{\text{obs}}(d)$ the normalized observed frequency):

$$\Delta_{bl} = \sum_{d=1}^{9} \left| \frac{P(d_i) - P_{\text{obs}}(d_i)}{P(d_i)} \right|. \tag{5}$$

Normalized frequencies of the first and second digits will be shown graphically for comparison with Benford’s expected frequencies, while in addition compliance with Benford’s law will be formally tested using classical chi-squared ($\chi^2$) goodness-of-fit tests (Jewell, 2003). All statistics are done using R version 2.14.2.

RESULTS AND DISCUSSION

Characteristics of the rubber process dust measurements provided by the UK HSE (the NEDB data set) and those by the BRMA are shown in Table 2. As shown, the BRMA data set is about 10 times larger than the NEDB data set, but the two data sets approximately covers the same 20–25 years time period. Although the variability, expressed as the geometric standard deviation (GSD) of both data sets, is similar, the average measured concentration is lower in the BRMA data set (GM $\sim$ 0.59 mg m$^{-3}$) compared to the NEDB data (GM $\sim$ 1.34 mg m$^{-3}$). There are a number of reasons for these differences that have been described in more detail previously (Agostini et al., 2010), but these are not important in the context of comparison to the Benford distribution. The numerical range ($R$) of the data is slightly larger in the BRMA data set compared to the NEDB data set (11.30 versus 8.47, respectively). Overall, the summary absolute normalized deviation ($\Delta_{bl}$) from the expected Benford distribution (1BL) is almost twice as large for the NEDB data compared to the BRMA data, although in absolute terms the difference is marginal. Chi-square test indicates that the 1BL distribution in the BRMA data set ($P \sim 0.03$), but not the NEDB data set ($P \sim 0.92$), differs significantly from the Benford distribution. For the second digits, both data sets show small, but statistically significant deviations from the 2BL distribution. Nonetheless, deviations from the expected 1BL distributions ($\Delta_{bl}$) were relatively small for both data sets, 0.90 and 0.53 for the NEDB and BRMA sets, respectively. Comparisons of the data sets with the expected Benford distributions are shown graphically in Fig. 1 in order to further explore these deviations. Both data sets show a relatively good agreement with the expected Benford distribution for the first digit (1BL). Evaluation of the second-digit (2BL) distributions does not indicate large deviation compared with the expected distribution for the NEDB data other than what may be attributed to natural variation in a relatively small data set. The BRMA-collected data, on the other hand, shows a clear deviation from the Benford second-digit distribution in that ‘0’ occurs about twice as much as expected (and consequently all other numbers less than expected).

The $n$-nitrosamines data sets from the MEGA database (Table 2) show much larger deviations ($\Delta_{bl}$ = 9.88 and 3.75, respectively) from the expected 1BL distribution than the two rubber process dust data sets, even though the numerical ranges ($R$) and GSDs are comparable. Both MEGA data sets, but especially the MEGA–ExAsRub data set with imputed values, are up to 50 times larger data set than the rubber process dust data sets, which should generally improves the comparison with the Benford distribution because of attenuation effects. Nonetheless, significant deviations from the Benford distributions are observed for both the first and second digits ($P < 0.01$) for both data sets. Graphical exploration of the observed and expected 1BL and 2BL distributions (Fig. 2) indicates an over-representation of 7s as first digits for MEGA–ExAsRub data set which is not present in the original MEGA data. As described above, this can be ascribed to imputation of ‘missing’ specific $n$-nitrosamine measurements by LOD/$\sqrt{2}$, or 0.10/$\sqrt{2} \sim 0.07$ mg m$^{-3}$. Furthermore, both data sets have more ‘1’s than expected (and consequently too few of the other numbers). Evaluation of the 2BL distributions indicates too many zeroes are present combined with too few of the other numbers (aside from 3s and 7s in the MEGA–ExAsRub data but not in the original data). This, upon further inspection, can be ascribed to issues with rounding. As
indicated in the original paper describing the database, these data do not come from random sampling surveys in the first place (Stamm, 2001), and based on this information and observed deviation from the expected Benford distributions additional evaluation of these data may be warranted. In the case where it is determined that the measurement data are valid, more sophisticated treatment of values below the LOD in the MEGA–ExAsRub data set, such as described by Hewett and Ganser (2007), may be beneficial. Explanation of the deviation of the second-digit distribution of the MEGA data compared to the Benford distribution (i.e. under-representation of 2s, 4s–6s, 8s, and 9s but not 3s and 7s) is not straightforward, and further investigation of possible causes, for example data transfer across various reports and databases and manipulation of (strata of) the data prior to inclusion in MEGA, may be worthwhile.

### CONCLUSIONS

This article describes the use of Benford’s law to screen occupational hygiene exposure measurement data set. Additionally, we evaluated four data sets from the European Union ExAsRub project as examples of the use of Benford’s law in this context. Previous sensitivity studies, based on data sets of pollutant concentrations in ambient air, which generally follow a similar log-normal distribution to occupational hygiene data sets and in which first digits were omitted as an example of ‘data mishandling’, indicated that \( \Delta_{bl} \) is very sensitive to any degree of data mishandling with percentage changes of about 9% for as little as 1% mishandled data. The arithmetic mean and standard deviation of the data set (other summary methods used to screen data for errors), in contrast, do not become useful criteria until about 25% of first digits are omitted (Brown, 2005).

Including the NEDB and BRMA rubber process dust data sets was of additional interest since previous work (Agostini et al., 2010) showed that even though both data sets supposedly measured exposures in the same industry in the same country, their results differed. Agostini et al. (2010) argued that the lack of auxiliary data on company size, reasons for sampling, measurement strategy, and other potentially important determinants prevented an explanation for the observed differences in both data sets. Because the use of Benford’s law has been shown to be more sensitive to data mishandling (Brown, 2005) than evaluation of geometric means and standard deviations we decided to evaluate this further. Nonetheless, although \( \chi^2 \) goodness-of-fit tests indicated some deviation from the expected distributions, these were relatively minor. Over-representation of ‘0’ in the second digits was most notable in the data collected by the BRMA, and further evaluation of these data may be done to evaluate whether these deviations may originate from when the data were initially entered or subsequently in transfer of the data across various databases.

The MEGA \( n \)-nitrosamines data, both the original and the treated ExAsRub data in contrast, showed much larger deviations from the expected 1BL and 2BL distributions. Further inspection indicated this may, to some extent, be attributed to issues of rounding and in the ExAsRub data set to the large number of added measurements below the LOD that were replaced by the same value (i.e. 0.07). Reanalysis of the ExAsRub–MEGA data after removal of all values below the LOD (\( N = 6739 \)) indeed indicated that although the goodness-of-fit test still indicated statistically

<table>
<thead>
<tr>
<th>Characteristics data sets</th>
<th>1BL</th>
<th>2BL</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( \chi^2 ) (8 df)</td>
<td>( \chi^2 ) (9 df)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Rubber process dust</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEDB</td>
<td>358</td>
<td>1980–2002</td>
</tr>
<tr>
<td>BRMA</td>
<td>3952</td>
<td>1977–2002</td>
</tr>
<tr>
<td>( n )-Nitrosamines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEGA–ExAsRub</td>
<td>18 619</td>
<td>1974–2003</td>
</tr>
<tr>
<td>MEGA</td>
<td>5243a</td>
<td>1974–2003</td>
</tr>
</tbody>
</table>

Characteristics of data set: number of measurements (\( N \)), years, geometric mean (GM), geometric standard deviation (GSD), minimum and maximum measured values (min–max), and numerical range of measurements [equation (4)] (\( R \)). For observed and expected Benford distributions of first digits (1BL) and second digits (2BL): summary absolute normalized differences of first digits [equation (5)] (\( \Delta_{bl} \)), \( \chi^2 \) statistics of goodness-of-fit for 8 and 9 degrees of freedom (df), and \( P \)-values of \( \chi^2 \) tests. a310 values removed because of missing values.
significant differences between the observed and the expected frequencies of the first and second digits, the differences, expressed as $\chi^2$-statistics were much smaller (3 301.80 and 6 368.68 compared to 36 890.99 and 33 287.76 for the 1BL and 2BL tests, respectively).

Unexplained deviation from the expected Benford distributions remains to exist, even after removal of the values below the LOD. In general there are at least four possible explanations for these deviations:

1. The data do not follow Benford’s law;
2. The data in general really do follow Benford’s law, but due to random chance this particular set of observations does not (type I error);
3. There is a reasonable explanation for the deviation from Benford’s law;
4. Some of the measurements in the data set are fraudulent, or others have been omitted.

When the Null hypothesis is rejected, the presence of fraudulent entries is just one of the possible explanations. For example, alternative explanation may be that this could have been caused by copying of data from inspector’s reports into digital format, subsequent transfer to the MEGA databases, and finally transfer to the ExAsRub database (and subsequent data treatment). However, these analyses do indicate that additional inspection of the MEGA $n$-nitrosamines data is warranted to further investigate whether there is a reasonable explanation for this
deviation from Benford’s law (explanation 3 above) or whether errors may have occurred in the various processes involved from taking measurements to their inclusion in a digital database used in research; for example, when large volumes of data are copied from paper reports into digital format by hand.

As such, there is no easy ‘test’ to determine what may have occurred. This may involve thorough re-examination of the measurement surveys, the data-entry process, and the final data set after statistical manipulation for use in research. An important disadvantage of Benford’s law is that it is in fact an observation rather than a law, and that although Hill (1995) has provided an explanation there is no theory describing why manipulated data would lead away from the predicted Benford distribution (Deckert et al., 2010). There are various reasons why occupational hygiene data sets may diverge from Benford’s law as outlined above, and as such it is important to acknowledge that just because a data set diverges, this does not indicate fraudulent attempts to alter data per se. Moreover, although it is a relatively simple and straightforward method Benford’s law should be just one of the tools employed to evaluate the quality of measurement data sets (Deckert et al., 2010). Other tools, for example, manual rechecking of the data set and/or the data-entry process however, are generally very time consuming (and hence costly). Benford’s law in these cases will provide a cheap

---

**Fig. 2.** First- and second-digit distributions relative to Benford distribution for original MEGA n-nitrosamines data and imputed data set from the ExAsRub database.
and easy-to-implement method for routine checking of occupational hygiene data sets so that resources can be better directed toward those data sets where some errors may be expected, as well as provide some hints as to what may have caused these deviations (for example, in this particular case study, the imputation of values below the LOD with a constant).

In conclusion, Benford’s law is a straightforward and easy-to-implement analytical tool to evaluate the quality of occupational hygiene data sets, and as such can be used to detect potential problems in large data sets that may be caused by malcontent a priori or a posteriori manipulation of data sets and by issues like treatment of observations below the LOD, rounding, and transfer of data. As such, it can be used to identify ‘strange’ issues in data sets, which subsequently need additional and more detailed evaluation to determine whether data are valid to be used for occupational hygiene and/or epidemiological purposes.

**FUNDING**

This study was funded by a grant of the Quality of Life and Management of Living Resources Programme in the European Union 6th framework (contract nos QLK4-CT-2001-00160 and QLRT-2001–02786).

**REFERENCES**


