



LETTERS TO THE EDITOR

Editorial Note. Letters to the Editor are peer reviewed to ensure that the arguments are reasonable and clearly expressed. However, letters may express a particular opinion rather than a balanced interpretation. Authors of papers commented on are invited to reply, but neither the journal nor peer reviewers should be assumed to support the arguments made.

Benford's Law and the Quality of Occupational Hygiene Data

Dorothea Koppisch¹, Rainer Van Gelder¹, Stefan Gabriel¹ and Roger Stamm²

1. Unit 1.3 Monitoring of Working Conditions, Institute for Occupational Safety and Health of the German Social Accident Insurance (IFA), Alte Heerstraße 111, 53757 Sankt Augustin, Germany

2. Division 1 Information technology, Risk management, Institute for Occupational Safety and Health of the German Social Accident Insurance (IFA), Alte Heerstraße 111, 53757 Sankt Augustin, Germany
E-mail: Dorothea.Koppisch@dguv.de

Submitted 14 August 2013; revised version accepted 26 December 2013.

In their paper, [de Vocht and Kromhout \(2013\)](#) conclude that Benford's law can be used to evaluate the quality of data on occupational exposure. We agree with this conclusion in general but would like to stress that this is true only if the data set does not contain too many imputed values in the case of measurements below the limit of detection (LOD).

In this letter, we will first derive theoretically why compliance with Benford's law cannot be expected for such data sets. Secondly, we will show some examples of exposure data from the MEGA database that confirm our theoretical considerations. Thirdly, we want to explain why we think that our considerations about imputed values could be a reasonable explanation for the deviations from Benford's law found by [de Vocht and Kromhout \(2013\)](#) in

the MEGA *n*-nitrosamine data. In the fourth place, we have a question regarding the selection of data sets from the ExAsRub study. Finally, we have some minor remarks regarding the publication of [de Vocht and Kromhout \(2013\)](#).

Benford's law is the name for the observation that in many empirical data sets the numbers 1 to 9 are not equally frequent as the leading digit. One is the most frequent leading digit with a frequency of 30.1% and 9 is the least frequent leading digit with a frequency of 4.6%. This fact was first published as a mere observation by [Benford \(1938\)](#). Later, it was used to detect manipulated data or errors in data handling in various scientific disciplines ([Brown, 2005](#); [Nigrini and Miller, 2007](#)).

The fact that the percentage of measurements below the LOD influences the distribution of numbers in a

data set is straightforward for a data set which contains a single LOD. The distribution of first digits in such a data set will be influenced by the methods used to substitute values below the LOD. If values below the LOD are substituted by a single value, e.g. $\text{LOD}/\sqrt{2}$ (Hornung and Reed, 1990, the method that was used by de Vocht and Kromhout, 2013), all measurements below the LOD will be substituted by one value with a single first digit. The percentage of this specific first digit will therefore be higher than would be predicted by Benford's law, and the other eight digits will show a lower percentage.

How great will this effect be in terms of the normalized deviations Δ_{bf} from Benford's law for first digits [Formula (5) in de Vocht and Kromhout, 2013]? Since, for data obeying Benford's law, 1 is the most frequent first digit and 9 the least frequent, let us assume two cases and a proportion of values below the LOD of 10% in order to obtain an estimate of the effect. If 1 is the leading digit of the imputed value, this number alone will contribute to Δ_{bf} with $(10/30.1) = 0.33$. If 9 is the leading digit, the contribution to Δ_{bf} of this digit alone will be $(10/4.6) = 2.2$. As the higher frequency of one number leads to lower frequencies of the other eight possible numbers and therefore to deviations from Benford's law for these numbers, Δ_{bf} will be even higher than 0.33 and 2.2.

If we assume that 30% of values are below the LOD, a proportion which can easily be found in occupational health data, the contribution of the increased frequency of 1 or 9 to Δ_{bf} increases to 1.0 and 6.5. This proportion of Δ_{bf} of one digit is already more than the sum of normalized deviations from the rubber dust data sets and the MEGA data set in table 2 of de Vocht and Kromhout (2013). That means that if the three data sets would contain 30% of imputed values in the case of measurements below the LOD this fact would be sufficient to explain the deviation from Benford's law found by de Vocht and Kromhout (2013).

Since occupational hygiene data very often contain many values below the LOD, the estimate of the influence of imputed values as described above shows that Benford's law for theoretical reasons cannot be used to detect manipulations in such data sets. The question now is, whether this theoretical consideration is confirmed by real data sets with different percentages of measurements below the LOD.

Unfortunately, de Vocht and Kromhout (2013) do not state the percentage of values below the LOD in the data sets in table 2 of their paper, so we cannot use their data sets to answer this question. We therefore checked

selected data sets with exposure values in the German rubber industry taken directly from the MEGA exposure database (Table 1). Values below the LOD were substituted by $\text{LOD}/\sqrt{2}$ in accordance with the procedure described by de Vocht and Kromhout (2013). It can be seen that for *N*-nitrosamines with 33.5% of values below the LOD, inhalable dust with 22.7% of values below the LOD and toluene with 10.6% of values below the LOD, the deviation from Benford's law for the first digit is highly significant ($P < 0.001$), while for the data sets with exposures to *n*-heptane, containing only 2.9% of values below the LOD, the normalized deviations from Benford's law are lower and the X^2 statistic is just slightly significant with $P = 0.047$.

Indeed, de Vocht and Kromhout (2013) refer to the influence of values below the LOD, stating in their conclusion (p. 301) that for '*the ExAsRub-MEGA data after removal of all values below the LOD (N=6739) ...*'. Regarding this statement, we would like to ask if the number $n = 6739$ is the number of all values below the LOD within the ExAsRub MEGA data set?

According to table 2 in de Vocht and Kromhout (2013), the ExAsRub-MEGA data set contains 18 619 values and the MEGA data set contains 5243 values. In 'Materials and methods', de Vocht and Kromhout (2013) state: '*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 $\text{LOD}/\sqrt{2}$.*' We infer from this explanation that all added values which are contained in the ExAsRub data set but not in the original MEGA data set are values below the LOD. We therefore calculate that the ExAsRub-MEGA data set contains at least $18\,619 - 5243 = 13\,376$ values below the LOD and we would like to ask the authors to explain the number of 6739.

Furthermore, de Vocht and Kromhout (2013) state in their conclusions (p. 301 f.) that for '*the ExAsRub-MEGA data after removal of all values below the LOD ... although the goodness-of-fit test still indicated statistically significant differences ... the differences, expressed as X^2 -statistics were much smaller.*' We cannot verify the still significant deviation from Benford's law after removing all values below the LOD in the data sets shown in Table 1. Instead for all four substances it is shown in Table 1, that after removal of the values below the LOD, deviations from Benford's law are no longer significant. In contrast to de Vocht and Kromhout (2013), we therefore

Table 1. Compliance with Benford's law of different occupational exposure datasets from the German MEGA database (German rubber industry, 1974–2011)

	Whole dataset						Data above LOD			
	<i>n</i>	% <LOD	<i>R</i> ^a	Δ_{bf}	X^2 (8 df)	<i>P</i>	<i>n</i>	Δ_{bf}	X^2 (8 df)	<i>P</i>
<i>N</i> -nitrosamines	8564	33.5	4.4	5.13	5514	<0.001	5695	0.43	14.5	=0.070
Inhalable dust	444	22.7	3.2	2.60	61.5	<0.001	343	1.46	10.8	=0.215
Toluene	1062	10.6	5.1	2.20	140	<0.001	949	0.78	6.1	=0.639
<i>n</i> -heptane	345	2.9	4.8	2.03	15.7	=0.047	335	1.81	12.8	=0.117

^aNumerical range $R = \log_{10}(\text{maximum/minimum})$.

think that the non-compliance with Benford's law for their two data sets MEGA and MEGA-ExAsRub can be explained by the high percentage of values below the LOD and there is no need to search for data manipulations as proposed in the conclusions of [de Vocht and Kromhout \(2013\)](#).

In addition, this observation means that some types of data manipulation (removal of all data points below the LOD) can lead to a better compliance with Benford's law. Compliance or non-compliance with Benford's law can therefore in our opinion not be used as a criterion for data manipulations in data sets with a high percentage of values below the LOD.

The data examined by [de Vocht and Kromhout \(2013\)](#) were taken from a study on exposure in the rubber industry conducted by the authors and their coworkers ([de Vocht et al., 2005](#); [de Vocht et al., 2007](#)). According to the 'Materials and methods' section, the question stemming from [Agostini et al. \(2010\)](#) whether the two rubber process dust data sets from the British Rubber Manufacturers' Association (BRMA) and the UK Health and Safety Executive National Exposure Data Base (NEDB) are comparable and 'how Benford's law can be used for screening of errors – fraudulent or other' (p. 299, left column) were the starting point for the publication. According to the paper by [de Vocht et al. \(2005\)](#) within the ExAsRub project, in addition to the data on inhalable dust from the UK, data sets were also available on inhalable dust from the Netherlands ($n = 2307$), Germany ($n = 188$), Poland ($n = 6407$), and Sweden ($n = 443$). We question why the authors did not take these data sets to compare the degree of compliance with Benford's law with the two data sets from the UK. We would be interested to know

how many values below the LOD they contain and how the omitted data sets perform in relation to Benford's law.

It would also have been possible to compare the data sets from BRMA and NEDB with other data from the rubber industry in the UK. [de Vocht et al. \(2005\)](#) list data on *N*-nitrosamines ($n = 595$), rubber fumes ($n = 3965$), and solvents ($n = 1533$). It is not clear to us whether the authors considered these data sets and why in the end they decided to work only with the two data sets on rubber process dust from the UK referred to in [Agostini et al. \(2010\)](#) and exposure data on *N*-nitrosamines from Germany.

Finally, we have some minor remarks. We cannot verify the values for the numerical range *R* in table 2 of [de Vocht and Kromhout \(2013\)](#). From the minimum to maximum values in that table 2 and using \log_{10} in the formula in [Brown \(2005\)](#), we calculate the following *R* values: 3.2, 4.5, 3.4, and 4.3. In consequence when using \log_{10} the numerical range is below 4 for two data sets in contrast to the values in table 2 of [de Vocht and Kromhout \(2013\)](#). Therefore, according to [Brown \(2005\)](#) one would not expect a good compliance with Benford's law for the data sets from NEDB and for the MEGA-ExAsRub data set.

Nor can we verify the *P* value for 1BL of the BRMA data that is given in table 2 of [de Vocht and Kromhout \(2013\)](#). For $df = 8$ and $X^2 = 47.92$, we obtained a *P* value of <0.001 and not of 0.03. This would mean that the BRMA data shows highly significant deviations from Benford's law. This should have been considered in the conclusions.

In summary, our main concern with the paper of [de Vocht and Kromhout \(2013\)](#) is that the authors

did not consider in depth the role of high percentages of imputed values in case of measurements below the LOD for the usefulness of Benford's law in evaluating occupational hygiene data and that they did not explain the rationale behind their selection of data sets from the vast range of substances and countries that supplied data for the ExAsRub project.

REFERENCES

- Agostini M, de Vocht F, van Tongeren M *et al.* (2010) Exposure to rubber process dust and fume since 1970s in the United Kingdom; influence of origin of measurement data. *J Environ Monit*; **12**: 1170–8.
- Benford F. (1938) The law of anomalous numbers. *Proc Am Philos Soc*; **78**: 551–72.
- Brown RJ. (2005) Benford's Law and the screening of analytical data: the case of pollutant concentrations in ambient air. *Analyst*; **130**: 1280–5.
- De Vocht F, Burstyn I, Straif K *et al.* (2007) Occupational exposure to NDMA and NMor in the European rubber industry. *J Environ Monit*; **9**: 253–9.
- De Vocht F, Straif K, Szeszenia-Dabrowska N *et al.* (2005) A database of exposures in the rubber manufacturing industry: design and quality control. *Ann Occup Hyg*; **49**: 691–701.
- De Vocht F, Kromhout H. (2013) The use of Benford's law for evaluation of quality of occupational hygiene data. *Ann Occup Hyg*; **57**: 296–304.
- Hornung RW, Reed LD. (1990) Estimation of average concentrations in the presence of nondetectable values. *Appl Occup Environ Hyg*; **5**: 46–51.
- Nigrini MJ, Miller SJ. (2007) Benford's law applied to hydrology data - results and relevance to other geophysical data. *Math Geol*; **39**: 469–90.

doi:10.1093/annhyg/meu009

Advance Access publication 18 February 2014

REPLY

Author's Reply to Koppisch *et al.* 2014

Frank de Vocht

Centre for Occupational and Environmental Health
The University of Manchester

Hans Kromhout

Institute for Risk Assessment Sciences
Utrecht University

E-mail: Frank.deVocht@manchester.ac.uk

Submitted 15 January 2014; revised version accepted 15 January 2014.

We would like to thank Dr Koppisch and co-authors for their interest in our paper in which we described the use of Benford's Law (BL) for occupational hygiene data, using two data sets from the EXASRUB database (de Vocht *et al.*, 2005) as illustrative examples (de Vocht and Kromhout, 2013). We could indeed have used other data sets collated within the ExAsRub data set but included the *n*-Nitrosamines data set from the

MEGA database because of its interesting feature that although a standard number of specific *n*-Nitrosamine-swere measured only those below the limit of detection (>LOD) were reported in the MEGA database. To make the *n*-Nitrosamines data useful for exposure assessment within the EXASRUB project, we had to 'manipulate the data' by adding the missing (<LOD) concentrations for each measurement. Moreover, because of the