IFA
Institut für Arbeitsschutz der
Deutschen Gesetzlichen Unfallversicherung

Benford's law and the quality of data

Dorothea Koppisch, Rainer Van Gelder, Stefan Gabriel
Institute for Occupational Safety and Health of the German Social Accident Insurance (IFA), Sankt Augustin, Germany

Benford‘s law

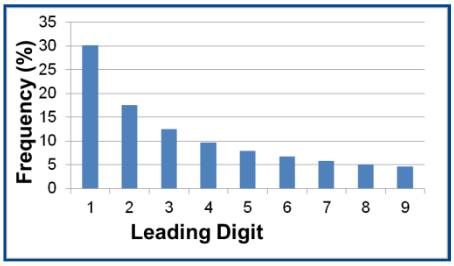

Fig. 1: Frequency of the leading digit for datasets obeying Benford‘s law

According to Benford's law in many empirical datasets the numbers 1-9 are not equally frequent as the leading digit (Fig. 1).

Therefore it has been proposed that deviations from Benford's law can be used to discover data manipulation (Brown 2005, de Vocht, Kromhout 2013).

Benford‘s law and values below the LOD

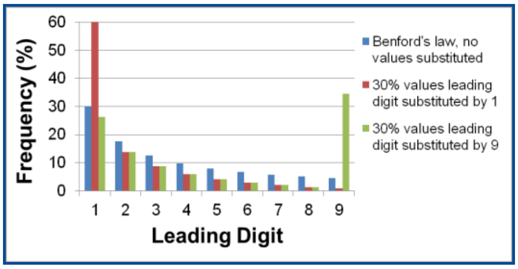

Fig. 2: Frequency of the leading digit for data sets with 0% or 30% leading digit substituted

$$D_{bn} = \sum_{d=1}^{d=9} \left| \frac{Pben(d) - Pobs(d)}{Pben(d)} \right|$$

D_{bn} - normalized deviation from Benford's law
Pben - relative frequency of the leading digit d according to Benford‘s law
Pobs - relative frequency of the leading digit d in the dataset to be tested

Box 1: Calculation of the sum of normalized deviations from Benford's law

The distribution of the leading digits in data sets will be influenced by the method used to substitute values below the limits of detection (LOD). If values below the LOD are substituted by a single value, for example LOD/√(2) (Hornung and Reed 1990), all these data points will have one single leading digit. Fig. 2 shows the influence of such a substitution if 30% of the values are below the LOD.

The degree of the deviation from Benford's law can be calculated as the sum of normalized deviations (Brown 2005, Box 1) and the significance of the deviations can be tested by means of the CHI²-statistic (de Vocht and Kromhout 2013).

Benford‘s law and occupational exposure data

Tab. 1: Deviation from / compliance with Benford's law for exposures in the German rubber industry. Data from the German exposure database MEGA.

	N	% below LOD	D_{bn}	CHI²	p-value (8 dF)
N-nitrosamines	8564	33.5	5.13	5514	< 0.001
Inhalable dust	444	22.7	2.60	61.5	< 0.001
Xylene	665	21.4	2.61	54.9	< 0.001
Trichlorethylene	146	18.5	3.94	55.4	< 0.001
Toluene	1062	10.6	2.20	140	< 0.001
n-heptane	345	2.9	2.03	15.7	> 0.05

To test our theoretical considerations about data sets with values below the LOD we selected six substances with exposure measurements in the rubber industry from the German exposure database MEGA (Gabriel, Range, Koppisch 2010) with different percentages of values below the LOD. For data sets normalized deviations from Benford's law lay between 2.03 and 5.13 if the values below the LOD are substituted by LOD/√(2) (Tab. 1). CHI² is not significant only for the exposure data set on n-heptane, which contains only 2.9% values below the LOD.

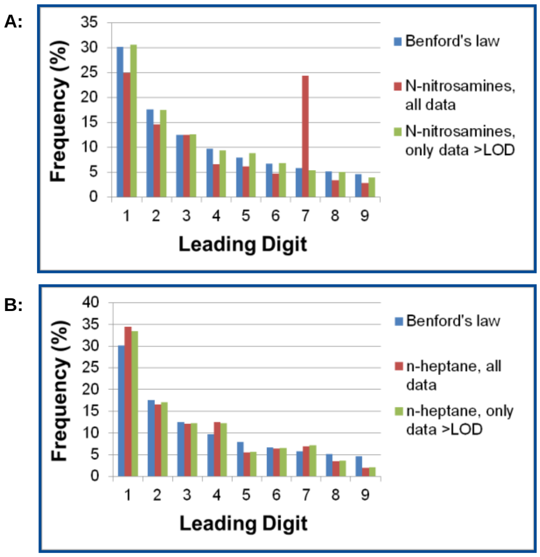

Fig. 3: Frequency of the leading digit for **A:** N-nitrosamines and **B:** n-heptane exposures in the German rubber industry

Conclusion

The authors infer that for data sets with a high percentage of values below the LOD deviations from Benford's law are to be expected and are no indication for data manipulation.

In contrast to de Vocht and Kromhout (2013) we therefore conclude that Benford's law is only suitable for testing the quality of occupational exposure data if the percentage of values below the LOD is sufficiently low. For data sets containing a high percentage of values below the LOD compliance with Benford's law depends on the method how these values are treated.

References

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, Kromhout H. (2013) The use of Benford's Law for evaluation of quality of occupational hygiene data. Ann Occup Hyg; 57: 296-304
Gabriel S, Koppisch D, Range D (2010) The MGU - a monitoring system for the collection and documentation of valid workplace exposure data. Air Quality Control; 70: 43-49
Hornung RW, Reed LD (1990): Estimation of average concentrations in the presence of nondetectable values. Appl Occup Environ Hyg; 5: 46-51