Applications of the generalized law of Benford to informetric data

Applications of the generalized law of Benford to informetric data

Leo Egghe[1,2] and Raf Guns[2]

1. Universiteit Hasselt (Uhasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium(*)

2. Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

(*) Permanent address

leo.egghe@uhasselt.be
raf.guns@ua.ac.be

Key words and phrases: Benford's law, Zipf's law

Abstract

In previous work of the first author, one could show that Benford's law (describing the logarithmic distribution of the numbers 1, 2, ..., 9 as first digits of data in decimal form) is related to the classical law of Zipf with exponent 1. Work of Campanario and Coslado (Scientometrics 88, 421-432, 2011) however shows that Benford's law does not always fit practical data in a statistical sense. In this article we use a generalization of Benford's law related to the general law of Zipf with exponent $\beta > 0$. Using data from Campanario and Coslado, we apply nonlinear least squares to determine the optimal β and show that this generalized law of Benford fits the data better than the classical law of Benford.

Introduction

Benford's law (Benford ,1938), originating from Newcomb (1881), determines the probability distribution of the first digits of numbers in decimal form. It reads as follows. The distribution $P(d)$ of the numbers $d = 1, 2, ..., 9$ as first digits of numbers in decimal form is

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \qquad (1)$$

Acknowledgement: The authors are grateful to an anonymous referee for drawing our attention to three references.

It is easily seen that $P(d)$ is indeed a distribution since the numbers $P(d)$ for $d = 1$, 2, ..., 9 add up to 1. As a consequence of this, the numbers 1, 2, ..., 9 as first digits of data in decimal form are not uniformly distributed (contrary to what one might think): (1) shows that smaller digits are favored since (1) is a decreasing function of d .

There are not many references to Benford's law in the informetric literature: Brookes and Griffiths (1978) and Brookes (1984) do not use the name Benford but call it the "anomalous law of numbers". In another paper – which does not mention Benford's name either –, Brookes (1980) describes the logarithmic structure of information, i.e. that information quantities should be measured logarithmically, as is the case in Benford's law. Brookes's (1980) reasoning to a large extent builds on the law of Bradford, but his quest for the foundations of information science may also help to clarify why Benford's law (as well as generalizations like the one used in this paper) seems applicable to informetric data. Quite recently we have the paper by Campanario and Coslado (2011) on which the present paper is based.

Most explanations of Benford's law are mathematical (based on probability theory or combinatorics). Egghe (2011) showed that there also exists a direct relation between Benford's law and the simple law of Zipf

$$g(r) = \frac{B}{r} \tag{2}$$

where $g(r)$ is the number of item densities in the source on rank density $r \in [1, T]$ (T = total number of sources). This result constitutes a link between the law of Benford and the informetric laws.

On a referee's request we further explain the argument in Egghe (2011). For each number in a data set we consider it as being between 1 (included) and 10 (not included) by putting the decimal dot '.' behind the first digit. This yields a set of numbers in the interval [1, 10[(closed in 1 and open in 10). We assume (as was also done by Pietronero et al. (2001), as remarked to us by the second referee) that for these numbers Zipf's law (2) is valid. Since all numbers with first digit r (r = 1, …, 9) belong to the interval $[r, r + 1[$ (closed in r and open in $r + 1$), the area under the curve (2) for abscissae between r and $r + 1$ is the frequency that the digit r occurs. Since the highest used abscissa is 9 + 1 = 10, we have here that $T = 10$. This explains the use of Zipf's law (2).

Campanario and Coslado (2011) remark that Benford's law does not always fit practical data. They examine the frequency of occurrence of first digits in articles, citations and impact factors. In some cases, Benford's law fits and in some cases it does not (according to a χ^2 goodness-of-fit test). This finding gave us the idea of deriving a generalized form of Benford's law from the generalized form of the law of Zipf:

$$g(r) = \frac{B}{r^{\beta}} \tag{3}$$

for $\beta > 0$ (see e.g. Egghe (2005)). Although we were not the first to do this (see Pietronero, Tosatti, Tosatti and Vespignani, 2001; Nigrini and Miller, 2007; Luque and Lacasa, 2009; see also Tao, 2009) we represent our reasoning for the sake of completeness and since it is short. Of course the 'generalized law of Benford' depends on the parameter β. In the third section we use the data found in Campanario and Coslado (2011); we apply nonlinear least squares to obtain the optimal β which gives us the best fit of the generalized law of Benford. In this way we obtain improvements of the fitting exercise in (Campanario and Coslado, 2011). This is also illustrated graphically. The conclusions are presented in the final section.

The generalized law of Benford

We use (3) (excluding $\beta = 1$ since this yields the classical law of Benford, see Egghe (2011)) and interpret the interval $r \in [d, d+1[$ for $d = 1, 2, ..., 9$ in (3) as the range where the digit d occurs. Hence $T = 10$ here. The same technique was used by Egghe (2011) in the proof of the classical law of Benford, using the simple law of Zipf (2). We first normalize (3) so that it becomes a distribution:

$$\int_1^{10} g(r)dr = 1 \tag{4}$$

Using (3) this yields

$$\frac{B}{1-\beta}(10^{1-\beta} - 1) = 1 \tag{5}$$

from which we find

$$B = \frac{1-\beta}{10^{1-\beta} - 1} \tag{6}$$

and hence our Zipfian distribution equals

$$g(r) = \frac{1-\beta}{(10^{1-\beta} - 1)r^{\beta}} \tag{7}$$

With distribution (7) we can now calculate the generalized law of Benford as follows. The probability $P(d)$ for digit d to be the first digit of a number in decimal form is

$$P(d) = \int_d^{d+1} g(r)dr \tag{8}$$

We obtain

$$P(d) = \frac{1}{10^{1-\beta} - 1}((d+1)^{1-\beta} - d^{1-\beta}) \tag{9}$$

being the generalized law of Benford where $d = 1, 2, ..., 9$ and where $\beta > 0$ but $\beta \neq 1$. Note that (9) is again a distribution since

$$\sum_{d=1}^{9} P(d) = \frac{(10^{1-\beta} - 9^{1-\beta}) + \ldots + (2^{1-\beta} - 1^{1-\beta})}{10^{1-\beta} - 1} = 1 \qquad (10)$$

Since the law of Zipf (3) is explained (Egghe, 2005, 2010), the above yields an explanation of the generalized law of Benford and an introduction of it into the informetrics field.

Note. Shi (2009) and Fu, Shi and Su (2007) propose the following generalization of Benford's law (in the notation of $P(d)$ above):

$$P(d) = N \log_{10}\left(1 + \frac{1}{s + d^{q}}\right) \qquad (11)$$

where N is a normalization factor; s and q are extra parameters that do not occur in Benford's original law. In the original law, $s = 0$ and $q = 1$, yielding $N = 1$. The main difference between (9) and distribution (11) is that the latter is neither explained nor related to existing literature. Furthermore, it contains two parameters, instead of only one. For these reasons, we consider the law in (9) to be more important from – at least – a model-theoretic view.

Practical results

This section is based on the data collected by Campanario and Coslada (2011), as shown in their Tables 1–3. The data consists of the number of articles published, citations received and impact factors of all journals in the Science Citation Index between 1998 and 2007. We refer to each of these in a specific year (e.g., the number of articles published in 2005) as a case study.

We employed the following method. All data was read into the R software for statistical computing (*http://www.r-project.org*), along with the relevant formula (9). We then used the *nls* function to perform a nonlinear least squares estimate of the parameter β for each case study. Filling in this value in (9) allowed us to create plots to visually inspect the results, as well as to determine the predicted number of items $N(d)$ for $d = 1, 2, \ldots, 9$. Finally, we determined how well the predictions fitted the empirical data according to the χ^2 goodness-of-fit test.

Figure 1 clearly illustrates that, for the case of articles published in 2003, the generalized law of Benford provides a better fit than the original one. In some cases, the differences are much smaller and visual inspection is not sufficient to state which of the two provides a better fit.

The results are summarized in Table 1. As can be seen from the Table, optimal values for β range from 0.8838 to 1.113. In four cases, the nonlinear regression could not improve on the original of Benford – in these cases the optimal β is 1. We note that there is a remarkable difference between the three kinds of case studies. All article case studies have optimal β smaller than 1, whereas all impact factor cases have optimal β greater than or equal to 1. The citation case studies are a mixed group.

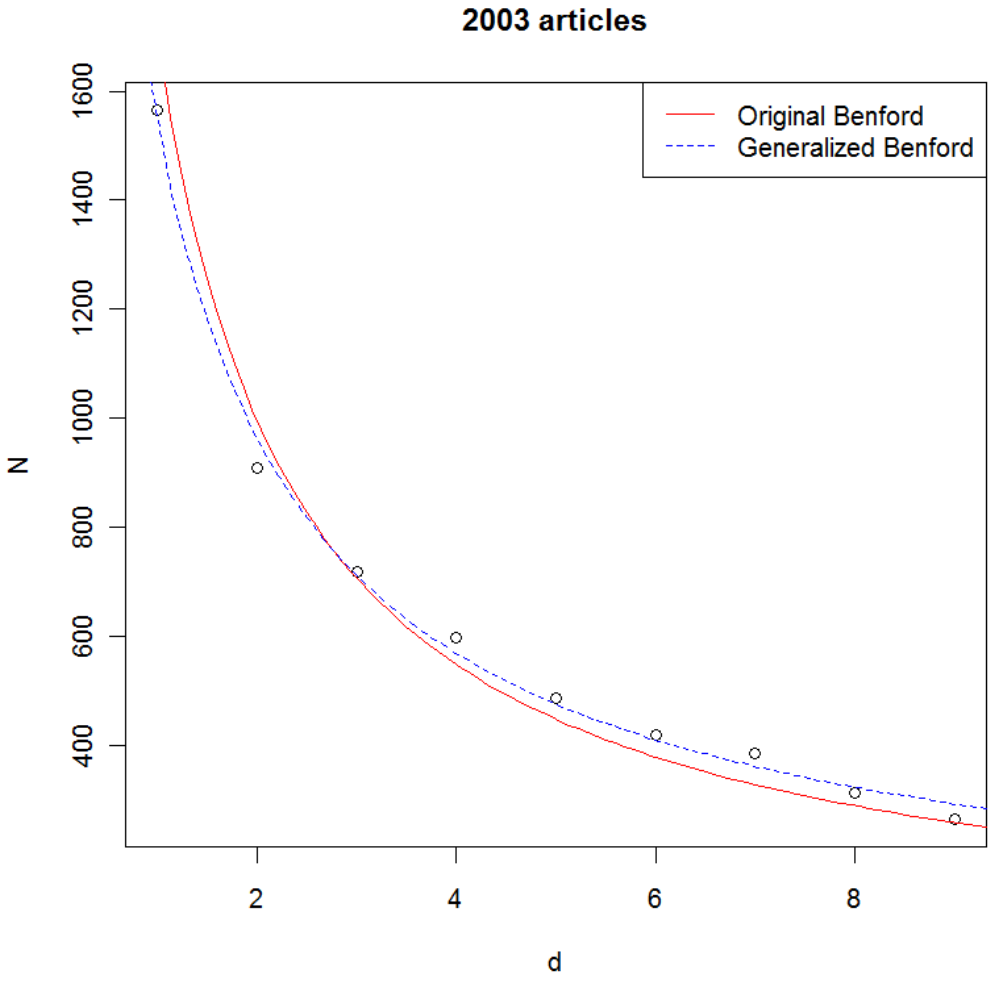

Figure 1. Comparison between original and generalized Benford for articles published in 2003

The last two columns of Table 1 compare the results of the chi-square test for the original and the generalized Benford law, showing both χ^2 and p values. Note that the χ^2 distribution for the original law of Benford has 8 degrees of freedom, whereas we use the χ^2 distribution with 7 degrees of freedom for the generalized law of Benford, because the latter contains one estimated parameter β. It can be seen that the number of statistically significant differences is much lower for generalized Benford. For instance, while all nine of the articles case studies have significantly different values ($p < 0.01$) compared to the ones predicted by Benford's original law, this is only the case for one ($p < 0.01$) or four ($p < 0.05$) when compared to the values predicted by the generalized law of Benford.

In some cases (most of them impact factors) we still find significant differences between empirical data and generalized Benford. Also, in some cases the generalized law of Benford yields predictions with slightly higher χ^2 values than the original law of Benford. The Gauss–Newton method used by R performs nonlinear regression by

optimizing the sum of least squares. Although this is highly correlated with chi-square, it is possible that optimizing for one has a negative effect on the other. This is the reason for this effect.

Table 1. Results per case study

Case study	Optimal β	Original Benford χ^2 (p)	Generalized Benford χ^2 (p)
1998 articles	0.9294	27.8 (0.001) **	14.8 (0.039) *
1999 articles	0.9279	27.4 (0.001) **	17.1 (0.017) *
2000 articles	0.9344	16.2 (0.039) *	6.5 (0.480)
2001 articles	0.8962	38.1 (0.000) **	10.7 (0.151)
2002 articles	0.8974	57.9 (0.000) **	28.9 (0.000) **
2003 articles	0.8838	43.5 (0.000) **	9.6 (0.210)
2004 articles	0.9103	31.3 (0.000) **	10.7 (0.152)
2005 articles	0.9037	41.5 (0.000) **	16.8 (0.019) *
2006 articles	0.8961	27.8 (0.001) **	5.6 (0.588)
2007 articles	0.8941	31.3 (0.000) **	5.1 (0.645)
1998 citations	0.9916	15.1 (0.056)	15.3 (0.033) *
1999 citations	0.9627	7.1 (0.524)	4.1 (0.771)
2000 citations	0.9798	4.5 (0.807)	3.3 (0.852)
2001 citations	1	5.2 (0.732)	5.2 (0.632)
2002 citations	1.003	3.1 (0.925)	3.2 (0.864)
2003 citations	0.9783	3.5 (0.897)	2.3 (0.943)
2004 citations	0.9858	3.0 (0.933)	2.0 (0.957)
2005 citations	1	11.2 (0.192)	11.2 (0.131)
2006 citations	1.019	9.7 (0.290)	9.4 (0.227)
2007 citations	1.019	8.4 (0.397)	8.6 (0.281)
1998 impact factor	1.016	6.6 (0.586)	6.9 (0.443)
1999 impact factor	1.006	11.3 (0.184)	11.9 (0.103)
2000 impact factor	1	22.2 (0.004) **	22.2 (0.002) **
2001 impact factor	1	20.2 (0.010) **	20.2 (0.005) **
2002 impact factor	1.003	24.9 (0.002) **	25.3 (0.001) **
2003 impact factor	1.001	12.5 (0.130)	12.6 (0.082)
2004 impact factor	1.059	16.7 (0.033) *	13.8 (0.054)
2005 impact factor	1.035	16.3 (0.038) *	17.0 (0.017) *
2006 impact factor	1.085	39.3 (0.000) **	28.7 (0.000) **
2007 impact factor	1.113	40.4 (0.000) **	14.7 (0.040) *

* (**) denotes significant difference between observed and predicted values at $p = 0.05$ ($p = 0.01$) level.

There is a declining tendency for β for article-based samples from 1998 to 2007 and a less definite increase for impact-based samples. Whatever the statistical significance of this may be, we can comment on this as follows. As proven in (Egghe, 2005, Corrolary IV.3.2.1.5, p. 204–205) increasing values of β stand for an increase of the inequality (concentration) of the $g(r)$ values, as expressed by the increasing Lorenz curves $L(g)$ of the rank-frequency distribution g. In the article-based examples this would mean a

decrease in the inequality between the $g(r)$ values over the years 1998–2007. We do not have an informetric explanation for this.
Using the same argument we can derive – since the article-based samples have a lower β value than the impact-based samples – that the distribution of the first digits in the latter case is more skewed (less equal, more concentrated) than in the former case. We consider this an interesting finding, but we lack an informetric explanation for it.

Conclusions

In this article, we discussed and applied the generalized law of Benford, which can be derived from Zipf's law. This new formula has the same parameter β as Zipf's law. Since the law of Zipf has been explained in an informetric setting (Egghe, 2005, 2010), the generalized law of Benford is also relevant to informetrics and related fields.

Empirical testing shows that the generalized law of Benford yields improved fits to practical data on the distribution of the digits $d = 1, 2, ..., 9$ as first digits in numbers in decimal form. To a certain extent, this is to be expected, since we have an extra parameter.

Our results suggest that the generalized law of Benford could be used heuristically to determine changes and differences in inequality. It would take, however, more research and much more data to see if this is really the case or just an accidental regularity in the data.

References

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

B.C. Brookes (1980). The foundations of information science. Part III. Quantitative aspects: objective maps and subjective landscapes. *Journal of Information Science* 2, 269–275.

B.C. Brookes (1984). Ranking techniques and the empirical log law. *Information Processing and Management* 20(1-2), 37-46.

B.C. Brookes and J.M. Griffiths (1978). Frequency-rank distributions. *Journal of the American Society for Information Science* 29, 5-13.

J.M. Campanario and M.A. Coslado (2011). Benford's law and citations, articles and impact factors of scientific journals. *Scientometrics* 88, 421-432.

L. Egghe (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Elsevier, Oxford, UK.

L. Egghe (2010). A new short proof of the Theorem of Naranan, explaining the laws of Lotka and Zipf. *Journal of the American Society for Information Science and Technology* 61(12), 2581-2583.

L. Egghe (2011). Benford's law is a simple consequence of Zipf's law. *ISSI Newsletter*, 7(3), 55-56.

D. Fu, Y. Q. Shi and W. Su (2007). A generalized Benford's law for JPEG coefficients and its applications in image forensics. *Proceedings of SPIE* 6505, 65051L. doi:10.1117/12.704723 (retrieved on October 27, 2011)

B. Luque and L. Lacasa (2009). The first digit frequencies of prime numbers and Riemann zeta zeros. *Proceedings of the Royal Society A* 465, 2197-2216.

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

M.J. Nigrini and S.J. Miller (2007). Benford's law applied to hydrology data – Results and relevance to other geophysical data. *Mathematical Geology* 39, 469-490.

L. Pietronero, E Tosatti, V. Tosatti and A. Vespignani (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A* 293, 297-304.

Tao, T. (2009). Benford's law, Zipf's law, and the Pareto distribution. Retrieved March 2, 2012 from http://terrytao.wordpress.com/2009/07/03/benfords-law-zipfs-law-and-the-pareto-distribution/

Y. Q. Shi (2009). First digit law and its application to digital forensics. In: H. J. Kim et al. (eds.), *Digital Watermarking: 7th International Workshop (IWDW2008)*, 444–456.