All Data are Wrong, but Some are Useful? Advocating the Need for Data Auditing

All Data are Wrong, but Some are Useful? Advocating the Need for Data Auditing

Sitsofe Tsagbey, Miguel de Carvalho, and Garritt L. Page

Abstract

In a recent paper from the *Annals of Applied Statistics*, Cox (2007) discusses the main phases of applied statistical research ranging from clarifying study objectives to final data analysis and interpreting results. As an incidental remark to these main phases, we advocate that beyond *cleaning* and *preprocessing* the data, it is a good practice to *audit* the data to determine if they can be trusted at all. A case study based on Ghanaian Official Fishery Statistics is used to illustrate this need, with Benford's law being the tool used to carrying out the data audit.

KEY WORDS: Applied statistics; Benford's law; Data quality; Significant digits.

1 INTRODUCTION

Today's growing flood of data, and the increasing demand for working in interdisciplinary environments, has made the field of Statistics more challenging than ever. The deluge of information is such that it has been recently claimed that "90% of the world's data have been created in the last two years" (Horton, 2015). In light of this, the question "Can your data be trusted?" is becoming increasingly more relevant.

In theory, applied statistical research should follow a logical sequence, starting with clarification of objectives, then moving to study design, data collection, analysis, and finally interpretation (Cox, 2007). In a footnote to such main phases, here we advocate that beyond *cleaning* and *preprocessing* the data, it is a good practice *auditing* if the data can be trusted at all. One goal of this article is to illustrate how statistical principles can be used to investigate the validity of reported data and to advocate for adding a data audit to the main phases of a statistical data analysis or even in applied statistical research.

S. Tsagbey is PhD Candidate, Institute of Mathematics and Statistics, Universidade de São Paulo, Brazil (*stsagbey@ime.usp.br*). M. de Carvalho is Assistant Professor, School of Mathematics, University of Edinburgh, Av. Vicuña Mackenna 4860, Santiago, Chile (*miguel.decarvalho@ed.ac.uk*). G. L. Page is Assistant Professor, Brigham Young University, Provo Utah, US (*page@stat.byu.edu*). This research was partially funded by the Chilean National Science Foundation, through Fondecyt grants 11121186 and 11121131.

A famous aphorism in Statistics, one from which the title of this article borrows inspiration, reads: "All models are wrong, but some are useful" (Box, 1979). This aphorism by George Box conveys a simple but powerful idea: Statistical models should only be regarded as an approximation, with some being better suited for that task than others. And what about the data itself? With so many things that can go wrong during data collection, even in carefully designed experiments, is it also the case that "All data are wrong, but some are useful"? Although many reasons could in principle hinder the production of reliable statistics, it is generally agreed that lacking proper data is *the* most difficult obstacle a data scientist can face (Nisbet, Elder and Miner, 2009, chap. 20).

Motivated by a case study in official fishery statistics, in this article we illustrate the need of performing a data audit. Our tool of choice to carry out the audit is Benford's law. What is Benford's law? Roughly speaking, it states that the leading digits of many naturally occurring quantities, are distributed in a nonuniform way. It is a powerful probabilistic result which has been widely used for statistical fraud detection (Hill, 1995a; Nigrini, 1996; Leemis, Schmeiser, and Evans, 2000; Bolton and Hand, 2002; Cho and Gaines, 2007; Diekmann, 2007; de Marchi and Hamilton, 2006; Fewster, 2009; Graham, Hasseldine, and Paton, 2009; Judge and Schechter, 2009; Mebane, 2011) and as a simple, yet effective way to test for erroneous, fraudulent, and fabricated data (Cho and Gaines, 2007; Diekmann, 2007; Leemis, Schmeiser, and Evans, 2000; Ross, 2011). For example, Benford's law has been applied in contexts as diverse as tax auditing (Nigrini, 1996), survival analysis (Leemis, Schmeiser, and Evans, 2000), self-reported toxic emissions data (de Marchi and Hamilton, 2006), numerical analysis (Berger and Hill, 2007), scientific fraud detection (Diekmann, 2007), quality of survey data (Judge and Schechter, 2009), election fraud analysis (Mebane, 2011), and fraud detection in a commercial lobster fishery (Graham, Hasseldine, and Paton, 2009).

This article's principal goal is to encourage the regular practice of performing a data audit when analyzing data or carrying out applied statistical research. In our particular case study assessing conformance to Benford's law was a fairly obvious option for carrying out an audit. In other contexts carrying out a data audit by employing Benford's law may not be reasonable. What options for data auditing exist in these settings? Prado and Sansó (2011) provide a recent example of a statistical analysis of fraud that might be applicable. Additionally, depending on the applied context, there is a wealth of other methods for statistical analysis of fraud (e.g., mixture models, classification trees, and other unsupervised learning statistical techniques) which are surveyed by Bolton and Hand (2002), and that can potentially be applied for data auditing purposes.

The rest of paper is organized as follows. In the next section we provide more background associated with Benford's law. In Section 3 we formally consider the case study on Official Fishery Statistics. We conclude in Section 4 with some final remarks.

2 BENFORD'S LAW

2.1 Background

Benford's law (also known as the 'first-digit law') was discovered by Newcomb (1881) and Benford (1938) who suggested that for many naturally occurring quantities, the probability of the first significant digit being d is a logarithmic distribution given by

$$\pi_d := P(\text{first digit} = d) = \log_{10}(1 + d^{-1}), \tag{1}$$

for $d = 1, \ldots, 9$. Benford's law thus predicts that the probability of observing a leading digit equal to 1 is 0.301, and hence almost seven times larger than the probability of observing a leading digit equal to 9. Feller (1971) briefly describes the law in his seminal monograph:

> "A distinguished applied mathematician was extremely successful in bets that a number chosen at random in the *Farmer's Almanac* or the *Census Report* or similar compendium, would have the first significant digit less than 5. One expects naively that all 9 digits are equally likely, in which case the probability of a digit $\leqslant 4$ would be 4/9. In practice it is close to 0.7."

The mathematical and probabilistic roots of Benford's law can be found in Hill (1995a,b); simplified yet instructive descriptions are given for instance in Raimi (1976), Fewster (2009), Formann (2010), and Ross (2011).

Benford's law is employed in fraud detection by comparing the empirical first digit relative frequencies with those in Eq. (1). There are a number of statistical procedures that can be used to formally test this comparison. In this article we employ the χ^2 goodness-of-fit test and briefly comment on alternatives in Section 3.4. Datasets for which the χ^2 test indicates a deviation from Eq. (1) are thought to be fabricated or at least suspicious motivating the need for further investigation. However, before making such conclusions one must be reasonably confident that absent fraud, the data should follow Benford's law.

Two general justifications for assuming that Benford's law applies in a particular context are as follows: Hill (1995a,b) shows that Benford's law can be understood as a central-limit type theorem

for significant digits, and Leemis, Schmeiser, and Evans (2000) show that Benford's law applies to a rich class of mixture models. Hence, in contexts where we expect the data to have originated from a diversity of populations, Benford's law should be regarded as a theoretically grounded working assumption. In practice, a remarkable number of naturally occurring first digit distributions have been shown to follow Benford's law. In fact, Fewster (2009) states that the leading digits of numerical values generated from any distribution that is relatively smooth will at least approximately follow Benford's law if the sample size is large enough (>100) and if the recorded values span approximately 4 orders of magnitude. In the absence of such criteria, Morrow (2010) provides a transformation technique (similar in spirit to the Box–Cox transformations in linear models) such that the first digits of the transformed response variable will approximately follow Benford's law.

3 CASE STUDY: OFFICIAL FISHERY STATISTICS

3.1 Motivation for the Analysis

The global scale of the recent Chinese scandal regarding excessive fishery harvests has put the ocean policy community in disarray (Pauly et al., 2014). The illegal exploitation of marine resources is particularly troublesome in West Africa, where it poses serious threats to livelihoods and marine ecosystems (Pala, 2013). This scandal has highlighted the need of designing and implementing enhanced evidence-based fishery policies which can only be carried out if reliable fishery data are collected. Unfortunately, over the past few years, the confidence in official fishery statistics has been widely questioned (see, for instance, Watson and Pauly, 2001; Tesfamichael and Pauly, 2011; Pala, 2013; Pham et al., 2013; Pauly et al., 2014). Sustainable marine policies certainly need to take into account the impact of fisheries on marine ecosystems and biodiversity. For this to be possible, regular stock assessment must be carried out; as put simply by Daniel Pauly, a widely cited fisheries scientist (Pala, 2013): "We can't assess the state of the oceans without knowing what's being taken out of them." Data regarding stock assessment are collected regularly, including catch, fishing effort, vessel type, location, type of fishing gear, and temperature. As can be understood from the recent scandal on excessive catches by China, *the* challenge however is on collecting accurate data, particularly catch data.

Motivated by the fallout of excessive catches by China in West Africa, in our case study we work with tuna catch data from the Gulf of Guinea, and we focus on questioning and examining the

reliability of reported catches. In the Gulf of Guinea, tuna is one of the most valuable fish, being the number one species exported with value of USD 200,604,000 as of 2011 (FishStatJ, 2014).

Although it is widely recognized that official fishery statistics often fail to account for what has been actually harvested (Pham et al., 2013), our analysis is based on Benford's law and thus allows us to assess the extent to which the raw data used to produce such statistics, are themselves trustworthy.

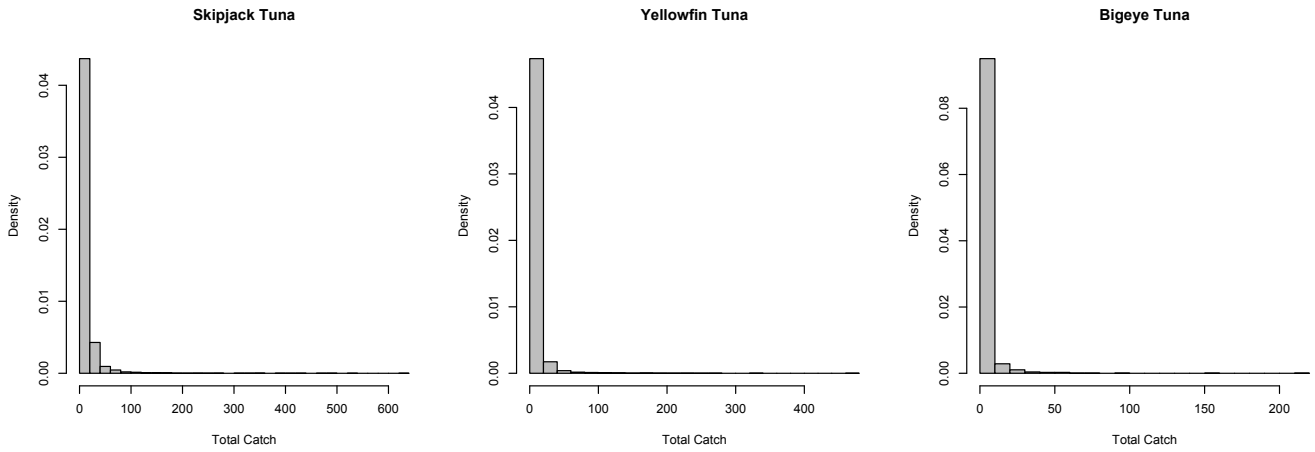

Figure 1: Frequency histograms of total catch weight (in metric tons) for each of the three species considered regardless of location, time, or boat type. The horizontal axis represents the percentage of observations that fall in each bin. Notice the extreme right skew of the data resulting in a difference between minimum and maximum that is a few orders of magnitude.

3.2 Data Description

We analyze tuna catch data of the three main species caught from the Gulf of Guinea, and landed in Ghana, West Africa. The data includes the subset of Chinese vessels which have been recently accused of misreporting which are registered in Ghana. The species considered are skipjack tuna (*Katsuwonus pelamis*), yellowfin tuna (*Thunnus albacares*), and bigeye tuna (*Thunnus obesus*). Data per species include the catch weight (in metric tons, mt), time (day, month, and year), location of the catch (latitude and longitude), and vessel type (baitboat, or purse seine). The Ghanaian Marine Fisheries Research Division of the Ministry of Fisheries and Aquaculture Development (*www.ghana.gov.gh*), granted access to the data except for the following years 1993, 2000, and 2002. In addition, we were not given access to part of 2008. We are thus led to believe that all the data

to which we were given access, corresponds to the part of data which the Marine Fisheries Research Division most trusts.

As a means to visualize the total catch amounts for each species we provide histograms in Figure 1. Notice that the shape for each species is heavily right skewed and as a result the difference between the minimum and maximum is a few orders of magnitude providing some justification for employing Benford's law for these data. Additionally, it can be argued that the catch data originate from a mixture of processes (e.g., fishing methods, time, ocean temperature and other characteristics). This consideration along with the relatively large sample sizes (13,735 skipjack, 11,702 yellowfin, and 2,969 bigeye) provide more justification to believe that these data if absent fraud should follow Benford's law.

3.3 Benford's Law-Based Data Analysis

We start with a separate analysis for each tuna species and then conduct a pooled analysis of all species. (We also explored the influence that other factors such as location have on conformance to Benford's law. These results are provided in an one-line supplementary material report.) In Figure 2 we depict some outputs from our Benford's law-based data analysis. In the first row of Figure 2 we plot the histogram of the proportion of observations having d as the leading digit ($p_d = n^{-1}O_d$), and compare it with the proportion predicted by Benford's law (π_d). Notice, for example, that for yellowfin the observed proportion of first digits equal to 9 is almost half of Benford's law predicts. Similar deviations can be observed for skipjack and bigeye, but to formally test whether such deviations are significant, we use χ^2 goodness-of-fit tests. The Pearson residuals are plotted in the lower row of Figure 2, and they can be used to visualize the degree of the deviation from Benford's law in a sample size corrected scale. A χ^2 goodness-of-fit test of π_d against p_d for each species rejects the null hypothesis with p-values close to 0 for each species. The results were the same for the pooled species, and thus it is evident that there is a general lack of conformity to Benford's Law. As it can be observed from the Pearson residual plots in Figure 2, the significant deviation is mostly driven by the higher-than-expected occurrence of digit '5', and by the lower-than-expected occurrence of digit '9'; the only exception is bigeye tuna for which, the lower-than-expected occurrence of digit '7' plays an even more important role than the one of the lower-than-expected occurrence of the digit '9.'

Another analysis was carried out for each species based on the type of fishing vessel. The

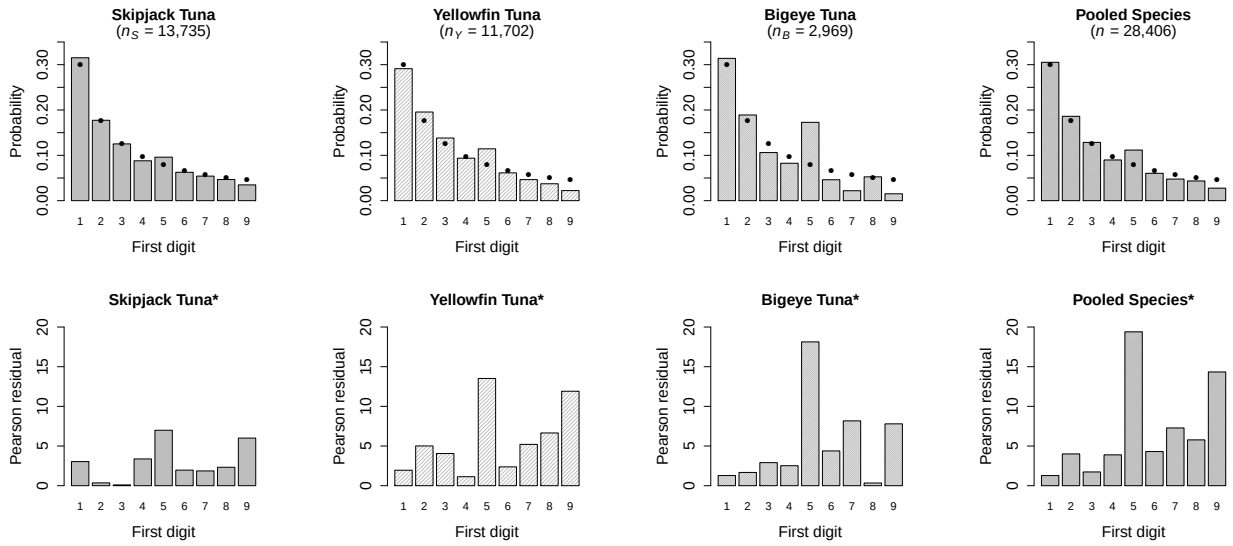

Figure 2: Benford's law-based analysis by species and for pooled species. Above: Density of first digits of catch weight; the solid dots (•) represent Benford's first digits probability law in Eq. (1). Below: Pearson residuals; the asterisks ($\star$) in the title of the Pearson residuals graphs represent rejections at 5% significance level.

economic motivation for this inquiry is as follows: Vessels with different capacities can in principle be owned by business players with different utility functions, and who can have different incentives to misreport catch. The null hypothesis is rejected in both type of fishing vessel for all three species. Pooling all the data together based on the type of vessel tells the same story. In Figure 3 we plot the Pearson residuals corresponding to this analysis. Similarly to the analysis by species, the significant deviation is mostly explained by an unexpected frequency of digits '5' and '9.'

3.4 Monitoring Reliability of Inferences

In Section 3.3 we assessed conformance to Benford's law through a χ^2 goodness-of-fit test. This procedure is simple and can be easily implemented to assess conformity of the observed first digits of total catch to Benford's Law. That said, as with all statistical tests, valid inference depends on assumptions and power. Among the concerns that accompany the χ^2 goodness-of-fit test is that in order to invoke asymptotic results a large sample size is needed. However, it is known that a large sample size also increases power leading to a decision to reject H_0 even for small insignificant deviations from Benford's (practical vs statistical significance). In light of this, a number of other tests with varying assumptions and power have been developed (Nigrini , 2012) many of which are included in the `R` package `BenfordTests` (Joenssen, 2015). To assess the reliability of the inferences

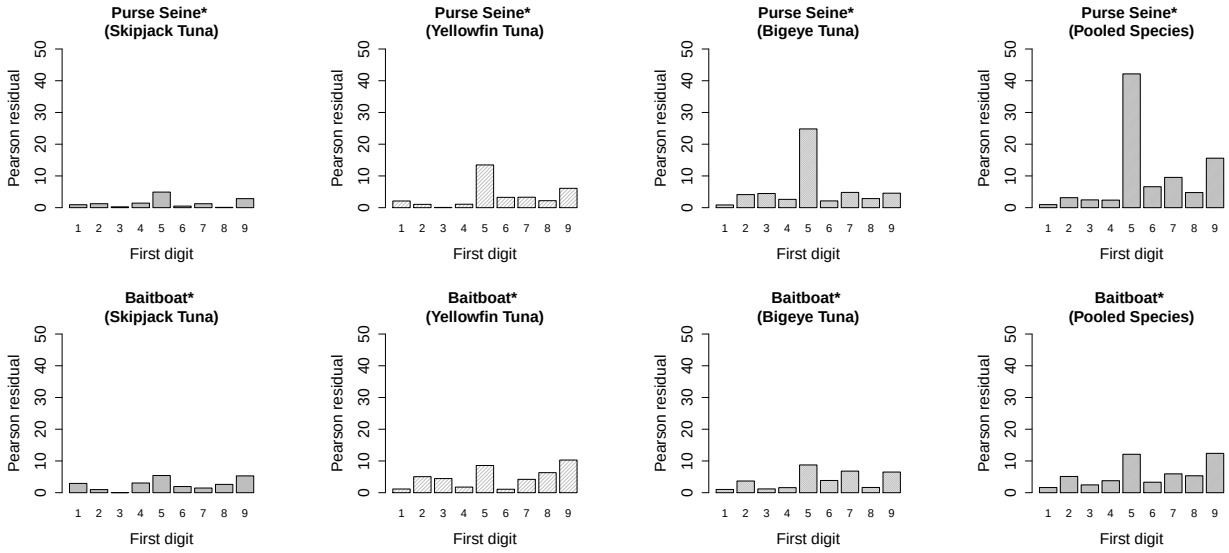

Figure 3: Benford's law-based analysis by vessel; the asterisks ($\star$) in the titles of the graphs represent rejections at 5% significance level.

reported above, we employ a few of these tests to conduct a battery of additional experiments. In particular, we assess conformance to Benford's law using a discrete Kolmogorov–Smirnov statistic a version of the Cho–Gaines Euclidean distance statistic (Cho and Gaines, 2007, p. 221), and a variant of the Leemis–Schmeiser–Evans statistic (Leemis, Schmeiser, and Evans, 2000, p. 237). The later two tests are discussed in Morrow (2010). Peculiarities with the Kolmogorov–Smirnov test for discrete data are discussed in Conover (1972) and Wood and Altavela (1978).

In all cases we obtained critical values through Monte Carlo simulation, and in terms of our data analysis the main conclusions are as follows. Apart from some minor exceptions, the inference remains unchanged regardless of the test used. The results are reported in Supporting Information. Lastly, we consider the mean absolute deviation test (MAD) which (Nigrini , 2012, chap. 7) argues is not overly sensitive to huge sample sizes. Conclusions based on this test are the same as those arrived at using the previously mentioned tests. Namely that the first digits associated with total tuna catch off the coast of Ghana do not conform to Benford's law.

4 DISCUSSION

In applied work we argue that more emphasis needs to be put on academic data auditing, in the sense that datasets used in research should be more carefully audited for accuracy and dependability. We provided a case study illustrating this need. In our particular case study assessing conformance

to Benford's law was a fairly obvious option for carrying out an audit. In other contexts carrying out a data audit by employing Benford's law may not be reasonable. Even though the statistician's toolbox contains a variety of solutions—many of which discussed by Bolton and Hand (2002)—the need for developing statistical methodology for spatial, functional, and biometric data auditing is of the utmost interest in practice.

The findings presented in this article suggest anomalies in reported weights of tuna harvest, and raise serious suspicion of misreporting. While the reliability of official fishery statistics has been widely questioned in the fisheries literature (see Watson and Pauly, 2001; Tesfamichael and Pauly, 2011; Pala, 2013; Pham et al., 2013; Pauly et al., 2014, and the references therein), our analysis provides further statistical evidence supporting the common belief that the reliability of the raw data themselves is the fundamental problem. In common with other Benford's law applications mentioned earlier (Nigrini, 1996; Leemis, Schmeiser, and Evans, 2000; Bolton and Hand, 2002; Cho and Gaines, 2007; Diekmann, 2007; de Marchi and Hamilton, 2006; Fewster, 2009; Graham, Hasseldine, and Paton, 2009; Judge and Schechter, 2009; Mebane, 2011), we emphasize however that this study does not provide conclusive evidence for the existence of fraud or misreporting, it only suggests that the anomalies identified in the data are compatible with their occurrence and further investigation is required.

ACKNOWLEDGEMENTS

We thank the Editor, Associate Editor, and two anonymous Reviewers for their careful reading of this paper and many useful suggestions for improvements. We also thank Paul Bannerman, Director of the Marine Fisheries Research Division, for providing them with the authorization to access the data, and to Vanda Inácio de Carvalho for helpful discussions and recommendations.

REFERENCES

Benford, R. J. (1938), "The Law of Anomalous Numbers," *Proceedings of the American Philosophical Society*, 78, 551–572.

Berger, A., and Hill, T. P. (2007), "Newton's Method Obeys Benford's Law," *American Mathematical Monthly*, 114, 588–601.

Bolton, R. J., and Hand, D. J. (2002), "Statistical Fraud Detection: A Review," *Statistical Science*, 17, 235–255.

Box, G. E. P. (1979), "Some Problems of Statistics and Everyday Life," *Journal of the American Statistical Association*, 74, 1–4.

Cho, W. K. T., and Gaines, B. J. (2007), "Breaking the (Benford) Law," *The American Statistician*, 61, 218–223.

Conover, W. J. (1972), "A Kolmogorv Goodness-of-Fit Test for Discontinuous Distributions," *Journal of the American Statistical Association*, 67, 591–596.

Cox, D. R. (2007), "Applied Statistics: A Review," *Annals of Applied Statistics*, 1, 1–16.

de Marchi, S., and Hamilton, J. T. (2006), "Assessing the Accuracy of Self-Reported Data: An Evaluation of the Toxics Release Inventory," *Journal of Risk and Uncertainty*, 32, 57–76.

Diekmann, A. (2007), "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data," *Journal of Applied Statistics*, 34, 321–329.

Feller, W. (1971), *An Introduction to Probability Theory and its Applications*, Vol. 2, 2nd ed, New York: Wiley.

Fewster, R. M. (2009), "A Simple Explanation of Benford's Law," *The American Statistician*, 63, 26–32.

Food and Agriculture Organization (1995), "The Coordinating Working Party on Fishery Statistics: Its Origin, Role and Structure," *FAO Fisheries Circular*, 903.

FishStatJ (2014), "Fisheries and Aquaculture Software. `FishStatJ`—Software for Fishery Statistical Time Series," In: *FAO Fisheries and Aquaculture Department* [online]. Rome. Updated 3 July 2014. [Cited 15 July 2014]. `www.fao.org/fishery/statistics/software/fishstatj`.

Formann, A. K. (2010), "The Newcomb–Benford Law in its Relation to Some Common Distributions," *PLoS ONE*, 5, e10541.

Graham, S. D. J., Hasseldine J., and Paton, D. (2009), "Statistical Fraud Detection in a Commercial Lobster Fishery," *New Zealand Journal of Marine and Freshwater Research*, 43, 457–463.

Joenssen, W. D. (2015), "BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law," *R package version 1.2.0*, `https://CRAN.R-project.org/package=BenfordTests`.

Judge, G., and Schechter, L. (2009), "Detecting Problems in Survey Data using Benford's Law," *Journal of Human Resources*, 44, 1–24.

Hill T. (1995a), "A Statistical Derivation of the Significant Digit Law," *Statistical Science*, 10, 354–363.

Hill T. (1995b), "Base-Invariance Implies Benford's Law," *Proceedings of the American Mathematical Society*, 123, 887–895.

Horton, N. J. (2015), "Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward," *The American Statistician*, 69, 138–145.

Leemis, L. M., Schmeiser, B. W., and Evans, D. L. (2000), "Survival Distributions Satisfying Benford's Law," *The American Statistician*, 54, 236–241.

Mebane, W. R. (2011), Comment on "Benford's Law and the Detection of Election Fraud." *Political Analysis*, 19, 269–272.

Morrow, J. (2010), "Benford's Law, Families of Distributions and a Test Basis," Technical Report, University of Wisconsin–Madison.

Newcomb, S. (1881), "Note on the Frequency of Use of the Different Digits in Natural Numbers," *Amer. J. Math.*,

4, 39–40.

Nigrini, M. (1996), "A Taxpayer Compliance Application of Benford's Law," *Journal of the American Taxation Association*, 18, 72–91.

Nigrini, M. (2012), *Benford's Law Applications for Forensics Accounting, Auditing, and Fraud Detection*, New Jersey: John Wiley & Sons.

Nisbet, R., Elder iv, J., and Miner G. (2009), *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier, Canada.

Pala, C. (2013), "Detective Work Uncovers Under-Reported Overfishing: Excessive Catches by Chinese Vessels Threaten Livelihoods and Ecosystems in West Africa," *Nature*, 496, 18.

Pauly, D., Belhabib, D., Blomeyer, R., Cheung, W. L., Cisneros-Montemayor, A. M., Copeland, D., Harper, S., et al. (2014), "China's Distant-Water Fisheries in the 21st Century," *Fish and Fisheries*, 15, 474–488.

Pham, C. K., Canha, A., Diogo, H., Pereira, J. G., Prieto R., and Morato T. (2013), "Total Marine Fishery Catch for Azores (1950–2010)," *ICES Journal of Marine Science*, 13, 564–577.

Prado, R., and Sansó, B. (2013), "The 2004 Venezuelan Presidential Recall Referendum: Discrepancies between two exit polls and official results," *Statistical Science*, 26, 517–527.

R Development Core Team (2011), R: *A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing.

Raimi, R. A. (1976), "The First Digit Problem," *American Mathematical Monthly*, 83, 521–538.

Ross, K. A. (2011), "Benford's Law: A Growth Industry," *American Mathematical Monthly*, 118, 571–583.

Tesfamichael, D., and Pauly, D. (2011), "Learning from the Past for Future Policy: Approaches to Time Series Catch Data Reconstruction," *Western Indian Ocean Journal of Marine Science*, 1, 99–106.

Watson, R., and Pauly D. (2001), "Systematic Distortions in World Fisheries Catch Trends," *Nature*, 414, 534–536.

Wood, C. L., and Altavela, M. M. (1978), "Large-Sample Results for Kolmogorov–Smirnov Statistics for Discrete Distributions," *Biometrika*, 65, 235–239.