UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Do people fit to Benford’s law, or do they have a Benford bias?

Permalink

https://escholarship.org/uc/item/2hx5f2rx

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 42(0)

Author

Burns, Bruce D.

Publication Date

2020

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Do people fit to Benford's law, or do they have a Benford bias?

Bruce D. Burns (bruce.burns@sydney.edu.au)
University of Sydney, School of Psychology, Brennan MacCallum Building (A18)
Camperdown, NSW 2006, Australia

Abstract

Smith (2015) describes an explosion of interest in Benford's law, that for data from many domains the first digits have a log distribution. Few studies have similarly asked whether the numbers people generate fit to Benford's law, but recent data show a reasonable fit. This paper argues that testing for fit to Benford's law is the wrong question for behavioural data, instead we should think in terms of a "Benford bias" in which the first-digit distribution is distorted towards Benford's law. We propose calculating the effect size of this bias by testing a linear contrast weighted by Benford's law. Analyses of existing data sets yielded effect sizes of 0.43-0.52. Applying this approach to a new task extended the scope of Benford bias to predicting outputs of a linear system and found an effect size of .40. Benford bias may be a ubiquitous influence on judgments and decisions based on numbers.

Keywords: Benford's law, decision making, judgment

Introduction

People frequently estimate numbers: How long is the Nile River? What is an item worth? How long will this project take? We use our knowledge to help us make such numerical estimates but often there is a gap between what we know and the precise number we must propose. How then do we generate estimates that are beyond the limitations of our knowledge? Benford's law provides a possible window into this process and suggests the existence of a previously unrecognized bias in our estimates.

Benford (1938) collected 20,229 data points from 22 unrelated domains (e.g., length of rivers, newspaper circulation and physical constants). He found that the first digit of those numbers, independent of magnitude, had a log distribution, as shown in Table 1. This distribution has become known as Benford's law.

Table 1: Percentage frequency of each first digit from theory (Benford's law) and observation (Benford, 1938).

	1	2	3	4	5	6	7	8	9
Benford's law	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Benford's data	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7

In recent years there has been an explosion of interest in Benford's law as a property of data with the publication of an edited book (Smith, 2015) and an online Benford's law bibliography (http://www.benfordonline.net/) that has grown to contain more than 500 papers published in mathematics and statistics; 400 in Finance and Accounting; 130 in science and psychology; 90 in computer and digital science, 60 in politics and economics, and 15 in clinical or medical settings (estimated by Chi, 2020, in August 2019). Whereas these papers have shown that a lot of data about humans fit Benford's law, whether the data people generate fits to Benford's law has only been examined by a handful of studies. The early studies found no evidence of this pattern in human generated numbers which led to experts on numerical cognition, such as Dehaene (1997), to conclude that Benford's law was not a psychological phenomenon. However, these studies had asked people to generate arbitrary, random numbers. When Diekmann (2007) and Burns & Krieger (2015) asked people to generate meaningful, nonarbitrary numbers then their first digits were a reasonably good fit to Benford's law. They do not perfectly fit Benford's law because no single phenomena could explain everything about number generation, but they suggest people can have a strong bias towards Benford's law. Understanding the extent of this "Benford bias" and why it occurs can provide a window into how people estimate numbers and what errors they make, which would also improve its effectiveness as a tool of fraud detection.

This paper will develop the idea of Benford bias and how we can estimate the size of it. It will then apply this analysis to a study extending Benford bias to a prediction task.

Previous Research

Mathematical research. The idea that first digits follow a log distribution was first proposed by Newcombe (1881) who noticed the wear pattern in tables of logarithms. Benford (1938) was the first to empirically confirm this speculation, although he did not name it Benford's law himself. In its most general form Benford's law describes the leading digit d ($d \in \{1, \ldots, b-1\}$) in base b ($b \geq 2$) as occurring with probability $P(d)=\log_b(d + 1) - \log_b d = \log_b((d + 1)/d)$. For a base 10 number system this gives the proportions in Table 1. Mathematicians have tried to derive Benford's law from the general properties of numbers, but had limited success (see Raimi, 1976). Hill (1995) then proved an important result by deriving Benford's law from the assumption that data that fits to it will be scale invariant. Furthermore, Hill (1998) argued that because a distribution seems to fit better the more it arises from completely unrelated data (e.g., baseball averages, areas of rivers); the critical point may be that the data is a combination of different distributions. Hill therefore proposed a theorem, "If distributions are selected at random (in any 'unbiased' way) and random samples are taken from each of these distributions, then the significant-digit frequencies of the combined sample will converge to Benford's distribution, even though the individual distributions selected may not closely follow the law." (Hill,

©2020 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

1998, p. 361). Benford's law has a tendency to attract informal attempts at explanation by people first encountering it, but these do not work. This led Berger & Hill (2011) to entitle their paper "Benford's law strikes back: No simple explanation in sight for mathematical gem". The current paper will not focus on the mathematics of Benford's law, so the interested reader is directed to Berger & Hill (2015) for a good summary of its mathematics.

Empirical research. Several attempts were made to test whether people generate numbers that conform to Benford's law numbers. Hsü (1948) asked 1044 participants to "write a 4-digit number that must be original, i.e., created in your own mind". Kubovy (1977) found priming effects but no fit to Benford's law when people generated random numbers. Later Hill (1988) asked mathematics students to generate a 6-digit number "out of their heads" with similar results: There was no fit of first digits to Benford's law. Thus, the consensus was that the numbers produced bear no relation to Benford's law.

That consensus held until Diekmann (2007) challenged it. He first showed that unstandardized regression coefficients reported in journals were a good fit to Benford's law. He then asked students in sociology or economics to fabricate multiple four-digit "plausible values" of regression coefficients that would support a hypothesis, and found that the generated first-digits were a good fit to Benford's law. However the samples were small (10 or 13 participants) and the pattern could be due to knowledge about regression coefficients (i.e. they tend to be low for data in social science).

Burns & Krygier (2015) suggested that perhaps the critical difference between earlier studies and Diekmann (2007) was that the earlier studies had no meaningful context, instead they explicitly asked participants to produce arbitrary, random numbers. Burns & Krygier pointed out that there is some evidence for a bias towards small first digits in random number generation but the effects sizes are much smaller than Benford's law would predict (for example, Loetscher & Brugger, 2007, aggregated across experiment using random generation of single digits and found a mean 0.02% increase over expected for digits 1, 2, and 3). To test this, Burns & Krygier asked students to estimate quantities for domains similar to those for which Benford collected data.

In Burns & Krygier (2015) Study 1, a set of nine questions was given to 127 psychology students. The questions were selected so that one had a correct answer with each of the first-digits "1" through "9". Thus, either correct or random answers would yield a flat distribution of first-digits. Participants were asked to "Please try to estimate the following values. Even if you have no idea, just guess." Answers were recorded by a computer program that required them to enter a valid number for each question. The selected questions were as follows, with correct answers (not shown to subjects) in brackets:

1. US gross national debt: $ [9] trillion
2. The number 2 raised to the power of 33: [8,589,934,592]
3. The peak summer electricity consumption of Melbourne: [7000] MV
4. Atomic weight of zinc: [65.39]
5. Population of the urban area of Philadelphia, USA: [5,330,000]
6. Area drained by the Pearl (Xi Jiang) river: [437,000] km2
7. Length of the Indus river: [3,180] km
8. Daily circulation of UK newspaper The Daily Mail: [2,340,255]
9. Infant mortality rate of Afghanistan: [157.43] deaths per 1000 live births

The first digits of each participant's answers were extracted and the percentage of their nine answers using each digit was calculated. Figure 1 shows the mean percentages for each first-digit together with columns that represent Benford's law and a line representing the correct answers.

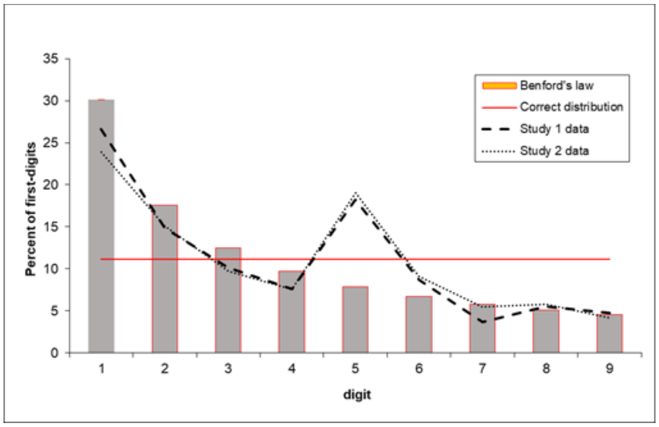

Figure 1: Distribution of first digits in Burns & Krygier (2015) Studies 1 and 2 data. The columns show Benford's law and the straight line is the correct distribution.

Figure 1 also shows the result of Burns & Krygier's (2015) Study 2 which dealt with the possibility that the pattern of results from Study 1 was the result of some peculiarity of the specific questions asked. Participants were again presented with questions from nine different domains but now with nine different possible questions for each domain, each with a true answer starting with a different first digit. Each participant received nine of the possible 81 questions, selected so that each participant received one question from each domain and such that the correct answers started with each of the first digits 1-9. Figure 1 shows that the resulting distribution of first digits had almost the identical distribution to Study 1.

Visually, the data in Figure 1 suggest that Benford's law fits somewhat for numbers generated by people, but not perfectly. The pattern shown in Figure 1 was repeated in studies by Tripodi (2016) and Chi (2020), who consistently found the peak at Digit-5, though it varied in size. So, although human data somewhat approximates Benford's law, it is not a perfect fit. This is not surprising; rather it would be a shock if the only influence on what numbers people generate was Benford's law. Any particular set of questions will have its own set of characteristics that may have an impact on fit to Benford's law, and there may be other general heuristics that have an impact. For example, the peak at Digit-5 may be due to people trying to estimate magnitudes when they don't know the answer to a question, and then sometimes settling on a number halfway between two magnitudes.

So perhaps rather than trying to test if people fit to Benford's law, we should be trying to estimate the size of a "Benford bias" in the numbers produced. This Benford bias may represent a previously unrecognized bias in those numbers.

Analysing the size of Benford bias

Many of the papers that claim to report data that fits to Benford's law do so by using chi-square tests or z-tests and when such tests fail to reject the null hypothesis of fit, they conclude that their data's distribution fits to Benford's law. Problematically such analysis relies on drawing a conclusion based on failure to reject a null hypothesis. An alternative has been to use Bayesian analysis to measure the degree to which the hypothesis of fit is supported (Geyer & Williamson, 2004). However, any test of the hypothesis that data fit to Benford's law suffers from the fact that it is almost certainly untrue because any specific data set will have characteristics or influences other than that which produces Benford's law. Benford (1938) recognized this when he labelled what he had discovered "the law of anomalous data" from his observation that the best fit to Benford's law was the aggregation across all his data sets. Aggregation makes the signal of Benford's law stand out most clearly from the noise of individual data sets' characteristics. So there is something of a paradox in testing fit to Benford's law in that evidence for it should be strongest when aggerating across a large set of diverse data, but Benford's law is most interesting when it is tested for specific data sets.

A better question to ask than whether the data support the hypothesis of fit to Benford's law may instead be how much of the variance in the data is explained by Benford's law? This question can be addressed by calculating the effect size ($\eta 2$) for the linear contrast weighted by Benford's law for the proportions of digits 1 through 9 produced by participants. A difficulty posed by such an analysis is that given that participants produce only nine data points no single participant can closely approximate Benford's law, so it would impossible for such a contrast to explain all the variances. Simulations of samples of 300 participants found that the maximum $\eta 2$ for such data is approximately $\eta 2$=0.73. Applying this analysis to Burns & Krygier (2015) Study 1 found that for the weighted-linear contrast, $F(1,126)=137.17$, $p < .001$, $\eta 2=.521$ and for Study 2, $F(1,289)=215.72$, $p < .001$, $\eta 2=.427$. Analysis of Tripodi's (2016) nonarbitary data found, $F(1,381)=335.06$, $p < .001$, $\eta 2=.468$. The $\eta 2$ statistic represents the proportion of variance accounted for by the contrast, therefore higher values for the proposed contrast are evidence of a stronger Benford bias.

Thus, analyzing the effect sizes for the weighted-linear contrast yielded reasonably consistent results indicated a moderate to large effect size, especially in the context that 0.73 was the maximum possible. Therefore, such analysis appears to be useful tool for gauging the size of Benford bias. Note that although it is novel to use contrast effect sizes to examine Benford's law, there is nothing statistical novel about this. Contrasts are simply being applied to means that happen to be predicted by a distribution, they are not testing the distribution in any other way.

Extending Benford Bias to Prediction

In Burns & Krygier (2015) participants were given knowledge questions, which leads to the question of whether the evidence for Benford bias is restricted to such questions. If this was the case, then it would be hard to argue that it is telling us anything about judgement or numerical cognition. A situation in which the numbers we generate can have important impacts on judgement are when we need to make predictions, so this paper will extend the investigation of Benford bias to a prediction task.

The prediction task chosen was the type of complex problem solving task used by Vollmeyer, Burns & Holyoak (1996) in which participants could manipulate the inputs to a linear system and then observe the outputs in order to learn over a series of trials how to control the system. Participants were shown a computer display listing what the current output values were, and they could enter new values for each of the inputs. Once this was done the computer would display the new outputs as well as the history of previous manipulations. However, before participants see the result of the changes to the inputs it is possible to ask them to predict the outputs. One of the key reasons researchers have used this task is that it generates metrics for how people go about learning about the system by observing the way they manipulated the inputs; however, this is a limitation when using it to observe predictions. The best learning strategy is to change one input at a time by small amounts, which can leave multiple outputs unchanged and the ones that do change may be only changed by an amount that does not change the first digit. It also means that first digit of outputs will be heavily influenced by the first digit of the initial values of the outputs, and therefore the first digits of predictions will be biased.

To overcome these limitations instead of participants choosing the inputs they were presented with inputs, each one of which was changed on every trial. Therefore, participants observed every output change on every trial, and sets of inputs were created that led participants to observe outputs that started equally often with each first digit.

It was predicted that the distribution of participants' first digits should be similar to that shown in Figure 1, and that analyzing weighted contrasts should produce at least moderate effect sizes.

All previous studies of the psychology of Benford's law have used samples drawn from university undergraduates, which raises the question of how universal is Benford bias? So, in this study participants were recruited using Amazon's Mechanical Turk. Amazon originally created this so that people could request and pay workers to complete online tasks. It has become extensively used for conducting online behavioral experiments and research suggest that it allows recruitment of a more diverse sample than university undergraduates and can produce results of similar reliability (Buhrmester, Kwang, & Gosling, 2011).

Method

Participants

A total of 317 participants were recruited via mechanical Turk. The only recruitment restriction was that they be from the USA. This was chosen so that the sample would be English speaking, and this was the cheapest way to achieve this (Mechanical Turk can charge for recruitment restrictions). Mechanical Turk should enforce this restriction, but when asked what country they were in 6 participants indicated that they were outside the USA. Although we did not deliberately try to collect information on participants' actual location, because our program time stamped every event with the time on the participant's computer, we were able to observe that some machines were using time zones outside the USA. There were 35 participants whose computers indicated a time zone outside of the USA, the 18 with an Indian time zone being the largest single group.

It was decided not to eliminate any participants from the sample based on location information. Time zone settings are not an absolutely reliable indicator of location, and setting the restriction to the USA was for convenience rather than being critical to the experiment. It is slightly worrying though that about 9% of participants may have inaccurately answered a direct question.

In response to demographic questions, 60% indicated that they were male, and 91% indicated that their first language was English. Their age range was 18-92 with a mean of 31.7 years and standard deviation of 8.5. In terms of education, 171 indicated that they had a bachelor's degree and a further 78 had an advanced degree. Thirty-nine indicated that they had some college but no degree at this time, leaving only 20 participants with only a high school level education. So, in terms of education the Mechanical Turk sample was not too different from a university sample, but their age spread was much greater.

Materials

Linear systems. In Vollmeyer et al (1996) participants learnt about a single system with three inputs and three outputs over three or four rounds, with each round containing 6 trials. Each trial was a chance to change the inputs and observe the resulting outputs. Such conditions enabled many participants to gain at least partial knowledge of the system. However, in the current study it was better if participants didn't learn the system, so that their predictions were not biased too much by the correct answer. It was also better to have multiple systems so that predictions were less likely to be biased by the characteristics of a particular system. So, each participant received three systems but only four trials for each. Thus, each participant made a total of 36 predictions after starting with all outputs set to zero.

Randomly, the frequency of first digits would be expected to come out to be unequal, and it is possible that observing such unequal frequencies could influence people's predictions. Therefore, once the values for the weights on the links between inputs and outputs were decided, the inputs were selected systematically. This selection was constrained so that across the first three trials for each system with a total of nine outputs, the first-digits 1-9 each appeared once. Therefore, participants observed each first digit value equally often. Furthermore, across the three systems, in each trial the first-digits 1-9 each appeared once. Thus, across trials there was no bias towards seeing any digits more often earlier in the task. The order of presentation of the three systems was also controlled such that each of the six possible orders occurred equally often.

Creating sets of 27 inputs that would produce outputs with these constraints would be hard by hand, so appropriate sets of inputs were created by simulation. For each simulation the 27 inputs were randomly selected and if the resulting outputs did not fit the constraints then a new set of inputs was generated. Fortunately, the code to do this is simple and desktop computers can generate billions of such simulations in an hour.

Four sets of inputs were generated by simulation, but another aspect of them was varied. For two sets the range from which random inputs were chosen was 1-100 each time, but for two sets the range for inputs accelerated from 1-10 on Trial 1, to 1-100 on Trial 2, then 1-1000 for Trial 3. By accelerating the inputs range it was more likely that an output would increase in magnitude from one trial to the next. Therefore, by testing for an effect of input set the design allowed us to test the effect on Benford bias of increasing uncertainty regarding magnitudes. The sets which more often led to an increase in magnitude should have stronger Benford bias.

The inputs for the first three trial for each system were determined for each of the systems but it was decided to provide participants with a fourth trial for each system. This reduced the chance that participants might notice a pattern that the first digits were being presented equally often, and reduced a concern that participants might not pay as much notice to the outputs of the last trial they were given. In addition, it generated more data from each participant. The inputs for the fourth trials were randomly generated for each participant using the input range used on the third trial.

Catch page. A risk for any online study is that participants may not be able to understand the instructions given or may not be motivated enough to try to understand. One way of mitigating this risk is to include "catch pages" in which it is only possible to responding correctly if the text on the page is read and understood. Often such pages contain a large amount of text in order to catch out participants who might skip lengthy instructions.

A catch page was designed for this study which instructed participants that they were being paid for 15 minutes so they should try to go through the experiment briskly and not pull out a calculator to try to find precise answers. We wanted their best guess at predicting what would happen. Due to the difficulty of the task a participant could take a long time, so the instructions were intended to give participants permission to not have to get things right. To make sure that participants read these instructions the end of the instructions told participants to click on the title at the top of the page rather than the prominent button below the text. Participants who clicked on the button saw a message telling them to read the page carefully before proceeding. If they clicked on the button again then they saw this message again with a count of how often they had acted incorrectly. Participants could not advance in the experiment until they had clicked on the title of the page.

Procedure

All participants completed the experiment online at a time of their choosing. The experiment was presented as a set of webpages controlled by javascript through any standard web browser. They were first given demographic questions regarding gender, age, first language and highest education level. They then read the instructions and the catch page. Participants then gave predictions for each of the three linear systems, as described above. The three systems could be presented in six different possible orders and each order was used equally often.

Participants took about 15 minutes to complete the task, but it varied because no aspect of the experiment was time limited. Participants who completed the experiment were paid US$2.

Results

Catch page

The majority of participants either never clicked on the incorrect button on the catch page (35.8%) or did so only once (32.3%). A further 12.3% clicked twice and 12.9% between three and eight times. A total of 5.2% clicked between twelve and sixty-nine times, while a further four participants clicked more than 100 times. One of these was recorded as clicking 6914 times which would only be possible if they ran some sort of computer script. Failing to read the instruction correctly 10 or more times seems a reasonable criterion for distrusting a participant's data, so participants who did this were eliminated for the sample. This reduced the sample to 289 participants.

A total of 160 participants started the task without completing it. Of these the most common point to stop was on the catch page, a total of 63. Perhaps because they lacked the motivation or the language skills to get past this page.

Across the four trials for the three systems the participants generated 36 predictions. The first digits for these predictions were counted and frequencies for digits 1 to 9 converted into proportions of total digits (0 could be a legitimate prediction, especially in the first trial, so such responses were ignored for the purpose of calculating proportions). Figure 2 shows the proportions for these digits.

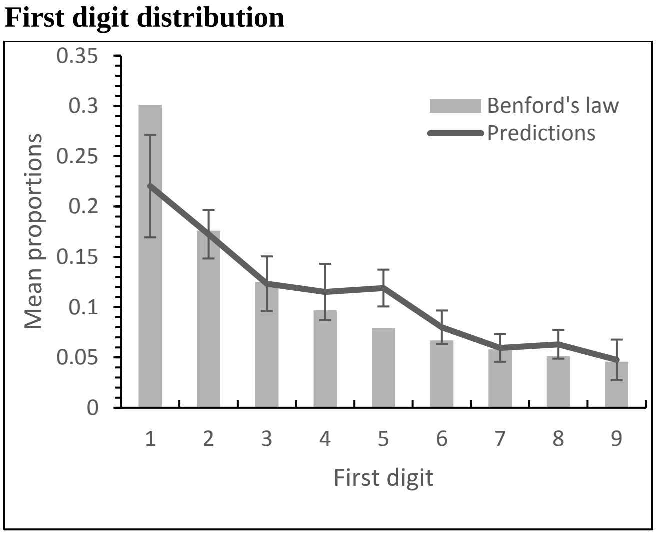

Figure 2: Mean proportions of first digits for prediction with 95% confidence intervals. Benford's law proportions are in columns. The correct distribution for the answers would be a straight line across the graph at 11.1%.

Figure 2 shows a similar pattern to Figure 1 except that Digit-1 and Digit-5 are a little lower. These are the only digits whose mean frequencies' 95% confidence intervals do not encompass Benford's law.

A 4x9 mixed design analysis of variance was run with a linear contrast weighted by the digit proportions for Benford's law. The between factor was input set (4 different sets) and the within factor the nine first digit proportions. The contrast was statistically significant, $F(1,285) = 190.4$, $p < .001$ with effect size $\eta2 = .400$. The contrast did not interact with input set, $F(3,285) = 0.56$, $p = .645$, $\eta2 = .004$. The lack of any effect inputs set suggests that the frequency with which outputs changed magnitudes was not critical.

Discussion

The title of this paper asked a question, do people fit to Benford's law or do they have a Benford bias? It was phrased this way because papers about Benford's law usually pose the question simply as whether the data set fit Benford's law, however this paper argues that a better way to pose the question for human behavior is to look for the degree of Benford bias. Whether people generate data that fits to Benford's law has a muddy answer, but if the question is whether their data shows evidence of a Benford bias then the empirical evidence strongly answers in the affirmative. The current experiment extended the scope of Benford bias by finding for a prediction task a similar pattern for first digits as found by Burns & Krygier (2015). Although the peak for Digit-5 was lower than in the previous studies, Digit-5 was the only digit for which its mean proportion exceeded that predicted by Benford's law by more than the 95% confidence interval. It also found a similar effect size for the Benford-weighted contrast (0.40). The pattern found here and in Tripodi (2016) and Chi (2020) has been consistent: Digit-1 is most common and then there is a monotonic decline in frequencies to Digit-4, then an upward spike for Digit-5, followed by a decline for Digit-6 and then Digits 7-9 have the lowest (though similar) frequencies. The current study also extended this finding to a new population, online Mechanical Turk workers.

This paper represents a conceptual and analytic advance over the previous behavioral studies of Benford's law. Both the studies that failed to find evidence for Benford's law when they asked people to generate random numbers and the more recent studies that found evidence when people generate nonarbitrary numbers, focused on testing the hypothesis that people exactly fit to Benford's law. The current paper argues that this hypothesis is not supported by the data and that conceptually it is the wrong hypothesis to test because it implies a too reductive view of the process of generating numbers. Instead we should be thinking in terms of a Benford bias, meaning that like any heuristic or bias there is a distortion of the data towards the bias, but the bias rarely completely determines the data. Such a view seems consistent with the behavioral data and brings the conceptualization of how to understand the relationship between Benford's law and human behavior more in line with how heuristics and bias affect judgement in the framework of Tversky & Kahneman (1974).

Evidence of systematic effects other than Benford bias on generation of this data is suggested by the consistent peak at Digit-5. Such a peak was also found by Scott, Barnard & May (2001) who had people provide numbers under various constraints in order to test hypotheses about executive function. However, they found that the peak at Digit-5 was present in unelaborated numbers (those consisting of one nonzero digit followed by zeros) rather than elaborated ones (those with more than one nonzero digit followed by zero). They interpreted elaboration as indicating greater involvement of executive functions. Burns & Krygier (2015) saw a similar reduction in the Digit-5 peak for elaborated numbers. Therefore, as suggested already, the Digit-5 peak could be due to some sort of heuristic that would decrease the extent to which Benford's bias explains the data.

If there is a Benford bias, then the question is how do we measure its size? An important innovation of this paper is to propose that this can be done by calculating the effect size of a linear contrast weighted by the proportions proposed by Benford's law. This approach was applied to the data from Burns & Krygier (2015) and then successfully applied to a new data set. Therefore, it appears to be a useful analytic tool for asking new questions about Benford's law and human behavior.

Note that the claim is not that the weighted linear contrast is the only possible description of the first digit data, or even necessarily the best. For example, an unweighted linear contrast would also yield a substantial effect size for the data in Figure 2. The weighted linear contrast is useful because it allows a focused question to be asked: to what extend does the Benford's law pattern explain people's first digit data?

There has been so little research on the psychology of Benford's law that there are many open questions for new research to address. We now have the conceptual and analytic tools to explore this topic and doing so is important for the following reasons.

First, Benford's law has practical consequences because fit to it is being used as a way of detecting fraud, first in financial data but more recently in many types of data (Nigrini, 2015). Using it as a tool relies on the assumption that deviation from it in data can be evidence that a human hand has distorted the data. However, such tests would benefit from a good model of human generated data. A better understanding of what the first digits of human generated data looks like would allow the development of more sensitive tests of fraud and fewer false positives. For example, our tests of Benford bias suggest that elevation of Digit-5 in data may be a particularly strong indicator of fraud.

A second reason for exploring Benford bias is that it may be distorting human judgment in ways that have not been recognized before. In particular, by extending Benford bias to prediction, the current study opens up the possibility that the many decision we make based on predictions may be being distorted. How much of an impact Benford bias has on decision making will depend on how robust its effects are. In particular, determining how knowledge mediates Benford bias is an important target for future research.

This paper has not proposed a model of why people's data shows Benford bias. The goal of this paper and of Burns & Krygier (2015) has been to establish empirically that Benford bias is a real and reliable behavioral phenomenon. This is not an uncommon approach in research into decision making biases, for example, the anchoring bias was established as an empirical phenomenon for number generation (Tversky & Khaneman, 1974) long before the multiple cognitive mechanism for it were understood (see Epley, 2004). Strong and consistent regularities in human behavior can be illuminating, but under-constrained attempts to explain them before there is an understanding of the nature of the empirical phenomenon can lead to a lot of wasted research effort. The current research

So, a final reason for examining Benford bias is that it is a surprising finding about human behavior. Like any surprising consistency in human behavior, understanding it potentially offers a unique window into the processes of cognition. We believe that it is now has a solid enough empirical basis that understanding why Benford bias exists is now an important question for future research.

References

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

Berger, A., & Hill, T. P. (2011). Benford's law strikes back: No simple explanation in sight for mathematical gem. *Mathematical Intelligencer, 33*, 85-91.

Berger, A., & Hill, T. P. (2015). A short introduction to the mathematics of Benford's law. In S. J. Miller (ed.), *The Theory and Applications of Benford's Law* (pp 23-67). Princeton, NJ: Princeton University Press.

Buhrmester, M., Kwang, T., & Gosling, S. L. (2011) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3-5.

Burns, B. D., & Krygier, J. (2015). Psychology and Benford's Law. In S. J. Miller (ed.), *The Theory and Applications of Benford's Law* (pp 267-274). Princeton, NJ: Princeton University Press.

Chi, D. (2020). *First Digit Phenomenon in Number Generation Under Uncertainty: Through the Lens of Benford's Law*. Unpublished master's thesis, The University of Sydney, New South Wales, Australia.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* Oxford, UK: Oxford University Press.

Diekmann, A. (2007). Not the first digit! Using Benford's law to detect fraudulent scientific data. *Journal of Applied Statistics, 34*, 321-329.

Epley, N. (2004). *A tale of tuned decks? Anchoring as accessibility and anchoring as adjustment.* In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (p. 240–257). Blackwell Publishing.

Geyer, C. L., & Williamson, P. P. (2004). Detecting fraud in data sets using Benford's Law. *Communications in Statistics: Simulation & Computation, 33*, 229-246.

Hill, T.P. (1988). Random-number guessing and the first digit phenomenon. *Psychological Reports, 6*, 967-971.

Hill, T.P. (1995). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society, 123*, 887-895.

Hill, T.P. (1998). The first digit phenomenon. *The American Scientist, 10*, 354-363.

Hsü, E. H. (1948). An experimental study on "mental numbers" and a new application. *Journal of General Psychology, 38*, 57-67.

Kubovy, M. (1977). Response availability and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Performance and Performance, 2*, 359-364.

Loetscher, T., & Brugger, P. (2007). Exploring number space by random digit generation. *Experimental Brain Research, 180*, 655-665.

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

Nigrini, M. J. (2015). Detecting fraud and errors using Benford's law. In S. J. Miller (ed.), *The Theory and Applications of Benford's Law* (pp 191-211). Princeton, NJ: Princeton University Press.

Miller S. J., (ed.) (2015). The Theory and Applications of Benford's Law. Princeton, NJ: Princeton University Press.

Raimi, R.A. (1976). The first digit problem. *American Mathematical Monthly, 83*, 521-538.

Scott, S.K., Barnard, P.J., & May, J. (2001). Specifying executive representations and processes in number generation tasks. *The Quarterly Journal of Experimental Psychology: Section A, 54*, 641-664.

Tripodi, M. (2016). *Unknowingly influenced by statistics in the environment. Why do people produce Benford's law?* Unpublished honours thesis, The University of Sydney, New South Wales, Australia.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185,* 1124-1130.\

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity and systematicity of strategies on the acquisition of problem structure. *Cognitive Science, 20,* 75 - 100.