# When Can One Test an Explanation? Compare and Contrast Benford's Law and the Fuzzy CLT

David Aldous[*]
Department of Statistics
367 Evans Hall # 3860
U.C. Berkeley CA 94720

Tung Phan

May 11, 2009

**Abstract**

Testing a proposed explanation of a statistical phenomenon is conceptually difficult. This class segment is intended to spotlight the issues.

[*] Research supported by N.S.F Grant DMS-0704159

## 1 Introduction

This material was developed as a class segment for an upper-division course [1] which examines critically the real-world uses of probability theory, and as an illustration of the kind of project (Figure 1 and Table 1) that students can do as part of the course. The goal of the segment is to compare and contrast the "testability" of two explanations of two different phenomena.

**The fuzzy CLT.** The normal approximation for the probability distribution of quantities that are explicitly modeled as sums or averages of random variables is uncontroversial. The normal approximation for observed data has always been a much more delicate issue:

> Everyone believes in the [normal] law of errors: the mathematicians, because they think it is an experimental fact; and the experimenters, because they suppose it is a theorem of mathematics. *Oft-quoted remark, attributed by Poincaré to Gabriel Lippmann.*

As Table 1 and its discussion (relegated to section 3, being peripheral to our main purpose) show, there is a curious disjunction between what many freshman textbooks *say* and what they *do* in exercises and examples regarding this issue. Anyway, one view is expressed in folklore as

> The *"fuzzy" central limit theorem* says that data which are influenced by many small and unrelated random effects are approximately normally distributed.

(This particular phrasing copied from [6].) Suppose we specify some type of data (e.g. biometric or observational errors in astronomy or product quality), examine a large collection of data-sets and find empirically that most such data-sets do follow approximately a normal curve. One could imagine many possible explanations of this finding.
**Question:** Is it possible, in principle and/or in practice, to empirically test whether the fuzzy CLT gives **a correct explanation** of empirically observed approximately normal distributions for specific types of data?

**Benford's Law.** Benford's Law is the assertion that within a large dataset of positive numerical data with large relative spread, the relative frequencies of leading digits $i = 1, 2, \ldots, 9$ will approximately follow the *Benford distribution*

$$b_i := \log_{10}(i+1) - \log_{10} i.$$

Like the birthday paradox, this is memorable because it is initially counter-intuitive. Also like the birthday paradox, an explanation occurs quickly to those with appropriate mathematical background, and this standard *large spread on a logarithmic scale* explanation was expressed succinctly in 1966 in the classic work of Feller [2] (trivially edited).

> The first significant digit of a number taken at random from a large body of physical or observational data may be considered as a random variable $Y > 0$ with some unknown distribution. The first significant digit of $Y$ equals $k$ iff $10^n k \leq Y < 10^n(k+1)$ for some $n$. For the variable $X = \log_{10} Y$ this means
>
> $$n + \log_{10} k \leq X < n + \log(k+1). \qquad (1)$$
>
> If the spread of $Y$ is very large the reduced variable "$X$ modulo 1" will be approximately uniformly distributed on $[0,1)$, and the probability of (1) is then close to $\log_{10}(k+1) - \log_{10} k = b_k$.

Fewster [3] provides a well-written elaboration of this idea in a form more suitable for undergraduate consumption. Again one can imagine other possible explanations, and indeed some are mentioned in [3] and in the Wikipedia article [8]. As before, suppose we specify some type of data (e.g. financial or geophysical or socio-economic), examine a large collection of data-sets and find empirically that most such data-sets do follow approximately Benford's law.

**Question:** Is it possible, in principle and/or in practice, to empirically test whether "large spread on a logarithmic scale" gives **a correct explanation** of empirically observed approximately Benford distributions for specific types of data?

**Analogy with clinical trials.** We will make more comments on the material above in sections 2.3 and 2.4, but let us try to make the issue perfectly clear via an analogy. Statisticians know how to test whether a particular new drug is more effective than a particular old drug in curing a particular disease: set up a randomized clinical trial, assessed double-blind. Such trials answer the empirical question "is it more effective?" After obtaining a positive answer, one might propose several different hypotheses about the physiological process making it more effective. Are any of these hypotheses correct? Well, you would have to do some *different* experiments, tailored to the specific hypotheses, to answer that question.

By analogy, we are assuming a collection of data-sets has passed an empirical test: most are approximately normal/Benford. Are the proposed explanations correct? To answer that question, you need some test tailored to the specific proposed explanation.

Note the analogy breaks down in one respect. With physiology we expect there to be only one correct explanation. With general data there may be several correct explanations applying to partially overlapping types of data. That's OK; one might wear an overcoat because it's cold or because it's raining.

## 2 Our analysis

We presented the two phenomena in parallel to emphasize similarities, but our main purpose is to point out a difference. If "large spread on a logarithmic scale" were a correct explanation of the occurrence of the Benford distribution, one would expect data sets with larger spread to tend to have distributions closer to the Benford distribution, after accounting for finite-sample effects. So it's a "falsifiable hypothesis". Within any given collection of data-sets, for each data-set we can just measure "spread" via some statistic, and measure "closeness to Benford" via some statistic. The hypothesis predicts some substantial association between these two statistics: if we don't see the predicted effect, then the purported explanation is just wrong, for this collection at least. Such a statistical analysis is carried out below.

In contrast, to repeat such analysis for the fuzzy CLT one would need a quantitative measurement of how close a given data-set comes to satisfying the assumption "influenced by many small and unrelated random effects". But this is just impossible to do – at any rate, **we** can't imagine any way of doing this for the kind of data-sets in Table 1.

So that's the point of this paper. Anytime you see a claim that certain types of data typically follow a certain distribution, and a proposed explanation of why, ask yourself

- is there serious evidence that the claim is empirically true (see section 2.3)?
- is the proposed explanation falsifiable? That is, can it be converted into a testable prediction?

More bluntly, though not wishing to enter philosophical debates [9], "is it science or is it just some made-up story"?

### 2.1 Analysis of some Benford distribution data

We study a collection of 18 data-sets (listed after references). As stated above, we will calculate two summary statistics for each data-set.

- $R$ measures spread (on a log scale)
- $D$ measures difference between observed distribution and Benford distribution, allowing for finite-sample effects.

Now (at this point in teaching a class I attempt to lower my voice in a theatrical manner and glance toward the open door of the classroom) academic statisticians with an interest in methodology may spend their careers debating the optimal choice of such statistics, but the boring truth is that if you have a lot of data and you're looking for a substantial effect which actually exists, then any sensible method will find it[1]. So we're just going to make choices of convenience and simplicity. Choose log interquartile range:

$R = \log(Q_3/Q_1)$ where $Q_1$ and $Q_3$ are the lower and upper quartiles.

To measure the distance between a probability distribution $\mathbf{p} = (p_i, 1 \le i \le 9)$ and the Benford distribution $(b_i)$ we use mean-square relative error

$$D(\mathbf{p}) = \sum_{i=1}^{9} b_i \left( \frac{p_i}{b_i} - 1 \right)^2 .$$

Thus if the ratios $p_i/b_i$ were all around 1.2 or 0.8 then $D$ would be around $0.2^2 = 0.04$. We could have used $D^{1/2}$ or any other "distance" measure, but $D$ has a mathematically nice feature we'll see below. If we follow the tradition (see section 2.4) of regarding a data set with $n$ entries as $n$ independent random samples from some unknown distribution $\mathbf{q}$, and calculate $D$ using the observed relative frequencies $\mathbf{p}$, then the observed statistic $D(\mathbf{p})$ will be a biased estimator of the "true" unknown distance $D(\mathbf{q})$. It turns out (the calculations are given in section 2.2, but I don't see any value in doing the algebra in real time in class) that when $\mathbf{q}$ is close to the Benford distribution and $n$ is large then $D(\mathbf{p}) - 8/n$ is an approximately unbiased estimator of $D(\mathbf{q})$ with standard error approximately $4/n$ (the simplicity of these formulas being the mathematically nice feature of the statistic $D$). So in Figure 1 we show, for each data-set, the value of $R$ and the "confidence interval" $[D(\mathbf{p}) - 12/n, D(\mathbf{p}) - 4/n]$ representing approximately $\pm 1$ standard error around the estimate of $D$.

[1] Sometimes a clever student will recognize the tautology: a method which fails to detect a substantial effect with a lot of data is *ipso facto* not very sensible.

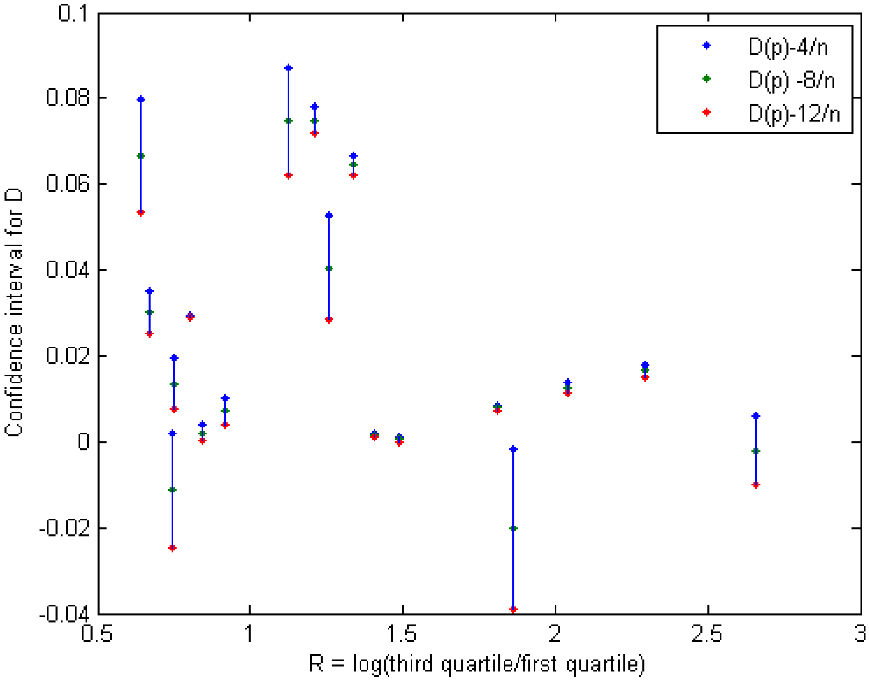

**Figure 1.** The relation between spread and closeness to Benford in 18 data-sets.

Visually, Figure 1 is consistent with the prediction of the "large spread" explanation of Benford's law, and not much would be gained by attempting a more quantitative conclusion, which is hardly justifiable with haphazardly-collected data-sets. Recalling this is an undergraduate course project, we judge it a success. Of course a professional statistician might be expected to do rather more; to follow a stricter protocol (section 2.3) for choosing data-sets, to get more data-sets, to investigate whether different types of data showed better or worse fits, and to think more carefully about possible alternative summary statistics to $R$ and $D$.

### 2.2 The calculation

Write $X_i$ for the number of occurences of $i$ in $n$ independent samples from $\mathbf{q}$. The observed statistic is

$$D := \sum_{i=1}^{K} b_i \left( \frac{X_i}{nb_i} - 1 \right)^2$$

where $K = 9$ (the algebra seems clearer when one writes $K$). Because $EX_i = nq_i$ and $\mathrm{var} X_i = nq_i(1-q_i)$,

$$E \left( \frac{X_i}{nb_i} - 1 \right)^2 = \left( \frac{q_i}{b_i} - 1 \right)^2 + \frac{q_i(1-q_i)}{nb_i^2}$$

and so

$$ED = D(\mathbf{q}) + n^{-1} \sum_{i=1}^{K} \frac{q_i(1-q_i)}{b_i}.$$

The sum depends on $\mathbf{q}$, but when $\mathbf{q}$ is close to $\mathbf{b} = (b_i)$ we approximate by calculating the sum with $\mathbf{q} = \mathbf{b}$, in which case the sum $= K - 1$. So

$$ED \approx D(\mathbf{q}) - 8/n.$$

Similarly, $\mathrm{var}(D)$ will depend on $\mathbf{q}$, but we approximate by calculating it for $\mathbf{q} = \mathbf{b}$. Assume $n$ large and quote the usual multivariate Normal approximation for empirical frequencies:

$$\frac{X_i}{n} - b_i \approx n^{-1/2} G_i$$

for multivariate Normal $(G_i)$ with mean zero and

$$\mathrm{var}(G_i) = b_i(1-b_i); \quad \mathrm{cov}(G_i, G_j) = -b_i b_j.$$

We also use the fact that for mean-zero bivariate Nornal $(Z_1, Z_2)$

$$\mathrm{var}(Z_1^2) = 2\ [\mathrm{var} Z_1]^2; \quad \mathrm{cov}(Z_1^2, Z_2^2) = 2\ [\mathrm{cov}(Z_1, Z_2)]^2.$$

From the definition of $D$ we see

$$nD \approx \sum_i \frac{G_i^2}{b_i}.$$

Use the variance-covariance formula for sums to get

$$\begin{aligned}
\lim_n n^2 \mathrm{var}(D) &= \sum_i \frac{2[b_i(1-b_i)]^2}{b_i^2} + \sum_i \sum_{j \neq i} \frac{2[b_i b_j]^2}{b_i b_j} \\
&= 2\left( \sum_i (1-b_i)^2 + \sum_i \sum_{j \neq i} b_i b_j \right) \\
&= 2\left( K - 2 + \sum_i b_i^2 + 1 - \sum_i b_i^2 \right) \\
&= 2(K-1).
\end{aligned}$$

So $\mathrm{var}(D) \approx 16/n^2$.

## 2.3 Are such approximations empirically true?

We have deliberately side-stepped the controversial issue of to what extent the often-asserted general fit of particular kinds of data to particular theoretical distributions is empirically true – in this context, the extensive interest in power-law distributions over the last 20 years comes to mind. Such assertions are typically justified by exhibiting examples where the fit is good, but this is rather like claiming general effectiveness of a medical procedure by merely exhibiting a set of patients on whom the procedure was successful. It seems to us that, as a bare minimum requirement for scientific seriousness, one should formulate and follow some protocol analogous to clinical trials:

- one person states a prediction that data of type [verbal description A] will usually fit distribution B;
- another person, given only [verbal description A] and not the prediction, gathers a large number of data-sets satisfying the description;
- a third person, given the data-sets and told distribution B, analyzes the data-sets and reports how well they do fit distribution B.

Has this ever been done?

## 2.4 Conceptual remarks

**1.** Pedantically, Feller's explanation is not really an "explanation" in the logical sense: it merely reduces a non-intuitive phenomenon to a more intuitive one (that when $X$ has substantial spread, "$X$ modulo 1" will be approximately uniformly distributed on $[0, 1)$) without justifying the latter.

**2.** To a mathematical statistician, Feller's paragraph says all there is to say, so it's not surprising there is little further discussion of the argument in the research literature, other than technical work quantifying bounds. Just as mathematicians like to discover and explore the applicability of different proofs of the CLT, there is subsequent work such as Hill [5] giving derivations of Benford's law based on more structural assumptions. Fewster [3] gives the impression that the law is regarded as mysterious and that Hill's is the standard mathematical derivation, but in our experience this impression is just wrong; Feller's derivation has been common knowledge in the academic community throughout the last 40 years.

**3.** As discussed elsewhere in the course, the tradition (invoked to calculate a confidence interval) of regarding a data set with $n$ entries as $n$ independent random samples from some unknown distribution is hard to justify *a priori*, though is itself a testable hypothesis.

**4.** Our $\pm 1$ S.E. confidence intervals are not 68% confidence intervals (the null distribution is not Normal) but the numerical value is hardly important. The fact that 3 of the 18 confidence intervals include (impossible) negative values might be embarrassing to a professional statistician, but amounts to an inefficiency rather than an error in our methodology.

## 3 Textbook examples of the Normal distribution for data

Almost all textbooks on introductory statistics have a chapter on the Normal distribution. What do they say about the general question

> For what kinds of data is it empirically true, or reasonable to suppose, that histograms are approximately Normal?

We are talking about observational data, not sample averages. We read the relevant chapter in three textbooks: [7] Chapter 6; [6] Chapter 5; [4] Chapter 9. What they explicitly say in the chapter is brief and unobjectionable. [7] says

> The normal distribution is used as a model for a variety of physical measurements since it has been discovered that many such measurements have distributions that are normally distributed or at least approximately so

[6] gives the "fuzzy CLT" quote we copied in section 1; and [4] says nothing.

| cited | asserted | assumed | example | type |
|---|---|---|---|---|
| √ | √ |  | height | human physiology |
|  |  | √ | weight |  |
| √ |  |  | cholesterol level |  |
| √ |  |  | blood pressure |  |
|  | √ |  | gestation time |  |
|  |  | √ | body temperature |  |
| √ |  |  | brain weight |  |
| √ |  |  | skull breadth |  |
| √ |  |  | eye-contact time |  |
|  |  | √ | reduction in oxygen consumption |  |
|  |  |  | . . . during transcendental meditation |  |
|  |  | √ | SAT (and similar exam) scores | human behavior |
|  | √ |  | farm laborer wages |  |
|  |  | √ | family food expenses |  |
|  |  | √ | recreational shopping expenses (teenage) |  |
|  |  | √ | TV hours watched (child) |  |
|  |  | √ | time in shower |  |
| √ |  |  | 10k race times |  |
| √ |  |  | baseball batting ave |  |
| √ |  |  | household paper recycling quantity |  |
|  |  | √ | in-home furniture assembly time |  |
|  |  | √ | military service point scores |  |
|  |  | √ | store complaints per day |  |
|  | √ |  | horse gestation time | non-human physiology |
|  | √ |  | rattlesnake length |  |
|  |  | √ | scorpion length |  |
|  |  | √ | grapefruit weight |  |
| √ |  |  | TV (and similar) lifetimes | product quality |
|  | √ |  | electrical resistance of product |  |
|  |  | √ | auto tire life (miles) |  |
|  |  | √ | weight in package |  |
|  | √ |  | thermometer inaccuracy |  |
|  |  | √ | coil springs strength |  |
|  |  | √ | weight of quarters (25c) |  |
| √ |  |  | yearly major earthquakes | geophysics |
| √ |  |  | annual rainfall Iowa |  |
| √ |  |  | inter-eruption times, Old Faithful |  |
|  |  | √ | inflight radiation exposure | miscellaneous |
|  |  | √ | mice: number of fights |  |
|  |  | √ | repeated weighings |  |

**Table 1.** Textbook [4, 6, 7] examples of the Normal distribution for data.

Having said this (and no more) about the general question, each proceeds to in-text examples and student exercises, many of which involve data having a Normal distribution. The subjects of the examples and exercises in those three chapters are listed (essentially completely) in Table 1.

A reasonable supposition is that, to a student, the memory of the one (or zero!) explicit sentence is crowded out by many examples and exercises; so the student comes away, consciously or unconsciously, with the perception that data on these kind of subjects follows the Normal distribution. Is this perception correct? Well, we don't really know: as implied in section 2.3, scientifically serious studies are hard to find, to put it mildly.

*Notes on Table 1.* Examples categorized as follows, even though the author's intention is often unclear.

- **Cited**: the author appears to have some specific data set in mind, though typically cited in a vague way (e.g. "the National Health Survey") so that the actual data cannot be readily located.
- **Asserted**: the author appears to be saying that, as a known general fact, data of this type is approximately normally distributed.
- **Assumed**: the author has either completely hypothetical data, or real data with given mean and s.d. but no given distribution, and tells the reader to assume normal distribution for the purpose of doing a calculation.

A few examples were excluded: quantities standardized to Normal (IQ scores); classical data (Quételet); numerical data intended for student computer-aided analysis. Much of the human data is "broken down by age and sex", in the classic phrase.

## References

[1] David Aldous. Overview of Probability in the Real World project. http://www.stat.berkeley.edu/users/aldous/Real-World/cover.html, 2009.

[2] W. Feller. *An Introduction to Probability Theory and its Applications. Vol. II.* John Wiley & Sons Inc., New York, 1966.

[3] R.M. Fewster. A simple explanation of Benford's law. *American Statistician*, 63:26–32, 2009.

[4] J. E. Freund. *Modern Elementary Statistics.* Prentice Hall, 10th edition, 2001.

[5] T. P. Hill. A statistical derivation of the significant-digit law. *Statist. Sci.*, 10(4):354–363, 1995.

[6] M. F. Triola. *Elementary Statistics.* Addison-Wesley, 7th edition, 1998.

[7] N. A. Weiss. *Elementary Statistics.* Addison-Wesley, 4th edition, 1999.

[8] Wikipedia. Benford's law — wikipedia, the free encyclopedia, 2009. [Online; accessed 22-April-2009].

[9] Wikipedia. Falsifiability — wikipedia, the free encyclopedia, 2009. [Online; accessed 22-April-2009].

**Datasets**

(1) Annual Sunspots (1700-2008, yearly):
ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/YEARLY
(2) Employment (Jan 1939 to Aug 2006):
http://www.swivel.com/data_sets/spreadsheet/1000209?page=1
(3) Energy Expenditure by State 2005, Table 889:
http://www.census.gov/compendia/statab/cats/energy_utilities/production_consumption_trade.html
(4) US exports and imports for consumption of merchandise by custom districts, Table 1262:
http://www.census.gov/compendia/statab/cats/foreign_commerce_aid/exports_and_imports.html
(5) US CO2 Emissions (1949-2004): http://www.swivel.com/data_sets/spreadsheet/1000222
(6) California Timber production:
http://www.dof.ca.gov/HTML/FS_DATA/STAT-ABS/Toc_xls.htm
(7) World Paper Consumption by Countries: http://www.swivel.com/data_sets/spreadsheet/1000520
(8) Farm Income-Farm Marketings , 2005 and 2006, and Principal Commodities, 2006 by State, Table 805:
http://www.census.gov/compendia/statab/cats/agriculture/farm_income_and_balance_sheet.html
(9) Timber product, Production, Foreign trade, and consumption by type of product: http://www.census.gov/compendia/statab/cats/natural_resources/timber-based_manufacturing.html
(10) Number of Tickets Sold in 2006: http://www.swivel.com/data_columns/show/1007788
(11) Retail Trade and Food Services–Sales by Kind of Business,Table 1009:
http://www.census.gov/compendia/statab/cats/wholesale_retail_trade.html
(12) Oil Reserve, extracted from workbook which was downloaded in historical data section:
http://www.bp.com/multipleimagesection.do?categoryId=9023755&contentId=7044552

(13) Foreign Exchange Rates (table 1352):
http://www.census.gov/compendia/statab/cats/international_statistics.html
(14) World Coal Consumption: http://www.eia.doe.gov/iea/coal.html
(15) World CO2 emissions from consumptions and flaring of fossil fuels, table H.1co2: http://www.eia.doe.gov/iea/carbon.html
(16) Electricity Consumption by country:
http://www.swivel.com/data_columns/spreadsheet/1532471
(17) US Aid by country: http://www.swivel.com/data_sets/show/1002196
(18) Farm Income–Cash Receipts From Farm Marketings ,Table 803:
http://www.census.gov/compendia/statab/cats/agriculture/farm_income_and_balance_sheet.html