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

David Aldous* Tung Phan

Department of Statistics

367 Evans Hall # 3860

U.C. Berkeley CA 94720

January 3, 2010

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

The material in this article was developed for a future class segment in an upper-division course (Aldous 2009a) which examines critically the real-world uses of probability theory. This article is addressed to potential instructors of such a course. Because such a course is unusual – perhaps unique – this introduction seeks to outline carefully the style and goals of the course. The style is to treat probability as somewhat analogous to physics – making general predictions about the real world – so analogous to a lab course in physics, the course is a kind of lab course in probability, studying which general predictions are actually verifiable by new data. This differs from typical uses of real data in applied statistics courses. There, typically, the instructor chooses a data set and gives it to the students to analyze, with some "technically correct" analysis in mind. By contrast, in this course students formulate and do a course project (as individuals or in small groups), each project reflecting some different theoretical prediction and chosen where possible to reflect some prior interest or knowledge of the student, and the emphasis is on creative choice of projects for which the student can gather new data and for which the results are not completely predictable, rather than on technical details of statistical analysis of the data.

Now a common reaction to this outline is "OK, but what do you actually do in class?" It's easy to write down a dozen potential projects (see a list at Aldous 2009b), some from textbook introductory statistics (birthday problem; regression effect in sports; empirical accuracy of opinion polls), and

others from more advanced probability (Do prediction market prices behave as martingales? Do intra-day stock prices obey the arc sine laws? Do both halves of the text of *Hamlet* compress equally?). Student projects on such topics can be used and critiqued in future iterations of the course. Indeed an ultimate goal is to built a portfolio of such projects so one could teach the course in seminar style – in each class one could state a theory prediction with brief mathematical background, give analysis of some data, then give some wider-ranging discussion prompted by theory and data. This style is deliberately orthogonal to textbook based courses that systematically develop some particular topic in statistics or probability.

The purpose of this article is to give an example of a class topic. The data was collected by the second (student) author but this write-up is by the first (instructor) author, the *I* in this article. In an attempt to be interesting or even provocative to instructor readers, I have chosen an example for which the wider-ranging discussion leads to conceptual or even philosophical points not usually mentioned in undergraduate statistics courses. The course is a real course (taught 3 times, latterly to 36 students) though the material in this article is, in the spirit of the portfolio above, an illustration of using a recent student project as the basis for planning a new class to be given next time the course is taught.

The central core of the class is the analysis in section 3.1 and the associated conceptual discussion in section 3.2. While the mathematics involved is routine (to mathematical statisticians) and will not be emphasized in class, the conceptual issue is actually quite subtle and easily misunderstood, even by instructors if they think more in terms of mathematics than the

underlying statistical concepts.

Section 2.1 describes the way I would actually frame the topic in class, as a kind of "philosophy of science" issue. How to do this is a matter of taste and other instructors will almost certainly want to do it differently. But some such discussion seems helpful to set the stage – to emphasize we are focussing on concepts rather than mathematics. The style of my course is to talk only about theory where we can get some data to accompany the theory. Thereby eliminating almost all philosophical topics. This particular class is my current attempt to get as close as I can to *some* philosophy issue with a concrete data-set. In practice the majority of students seem neither interested in nor capable of dealing with this type of philosophy, so I would not spend much time on it in class, instead inviting interested students to chat in office hours, and this could lead to formulating some project of interest to the individual student.

In contrast, I do view the conceptual issues in section 3.2 as important – why did we do the analysis this way instead of mindlessly plugging into a chi-squared test of significance? Most undergraduate statistics courses ignore such issues; a few with philosophical bent discuss them but with hypothetical or trite examples. A recurrent theme of the course (done in different contexts in other classes) is to treat such issues seriously within analysis of concrete data. Other classes in the course deal with the widespread misuse of tests of significance and with the justifiability of treating an observed data-set (e.g. exam scores) as if it were i.i.d. samples from a unknown distribution, and this class makes references to such discussions.

In the remainder of the article I distinguish between "what I say in

class" (sections 2.1, 3.1 and 4.1; really "what I plan to say in class", but maintaining a future tense is distracting), and "notes for instructors", the latter including this introductory section. Note that the class ends with a suggested student project, which is very deliberate: I try to talk only about topics which are amenable to student projects.

2 Framing a question

2.1 What I say in class

The goal of today's class 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.*

Indeed there is a curious inconsistency between what many freshman textbooks *say* and what they *do* in exercises and examples regarding this issue, as I'll illustrate at the end of the class (section 4.1 of this article). 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 Triola 1998 p. 260). 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 data-set of positive numerical data which has a *large spread on a logarithmic scale*, 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, there is a simple and standard explanation.

(Note to instructors. At this point I relate what I regard as the "simple and standard explanation", indicated in section 3.3. This explanation is expressed very clearly and with very helpful graphics in Fewster (2009) and

in the Wikipedia article Wikipedia:Benford's law (2009), and in class I show the graphics from one of those sources).

This explanation explicitly uses the *large spread on a logarithmic scale* assumption, of which a quantification will be given later in this class (section 3.1 of this article). As before (with the fuzzy CLT), 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. As before one can imagine other possible explanations (indeed, some are mentioned in Fewster (2009) and Wikipedia:Benford's law (2009)).

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. 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.

As a rough analogy, we are assuming a collection of data-sets has passed an empirical test: most are approximately Normal/Benford. Are the pro-

posed explanations correct? To answer that question, you need some test tailored to the specific proposed explanation.

How can we try to answer the questions? I presented the two phenomena in parallel to emphasize similarities, but my main purpose in this class 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. That's the kind of a statistical analysis we shall do in a moment (section 3.1 of this article).

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, I can't imagine any way of doing this for the kind of data-sets typically presumed to approximately Normal, as we will see later. But let's first study some data in the Benford's law setting.

2.2 Note to instructors on preceding material

Obviously I am touching upon a topic in the philosophy of science – in saying the phrase "falsifiable hypothesis" I am referring to the popularized Popperian view that to count as "scientific" a proposed theory must be falsifiable. I am well aware that most academic philosophers of science regard the popularized Popperian view as naively simplistic. But my purpose is not to summarize a debate within the philosophy of science, but simply to choose *some* philosophical methodology I can work with, and can get to a bottom line. One explanation is falsifiable and the other is not.

The analogy with clinical trials just a superficial analogy to illustrate the distinction between studying "what happens" and "why it happens". It gets used later (section 4.2) in the context of a *protocol*; students know that clinical trials follow some prespecified protocol, so should one analogously prespecify a protocol to gather data to test some general theory prediction?

3 Analysis of some Benford distribution data

3.1 What I say in class

As a student project, Tung Phan studied a collection of 18 data-sets (listed after references). As just outlined, 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.

Now (note to instructors: 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. For spread let's 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 (note to instructor: see section 3.2 for what I say here) 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 $D(\mathbf{q})$. A calculation of the kind you learn in

[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.

mathematical statistics courses (see section 3.4) shows 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 notional "confidence interval" $[D(\mathbf{p}) - 12/n, D(\mathbf{p}) - 4/n]$ representing approximately ± 1 standard error around the estimate of D.

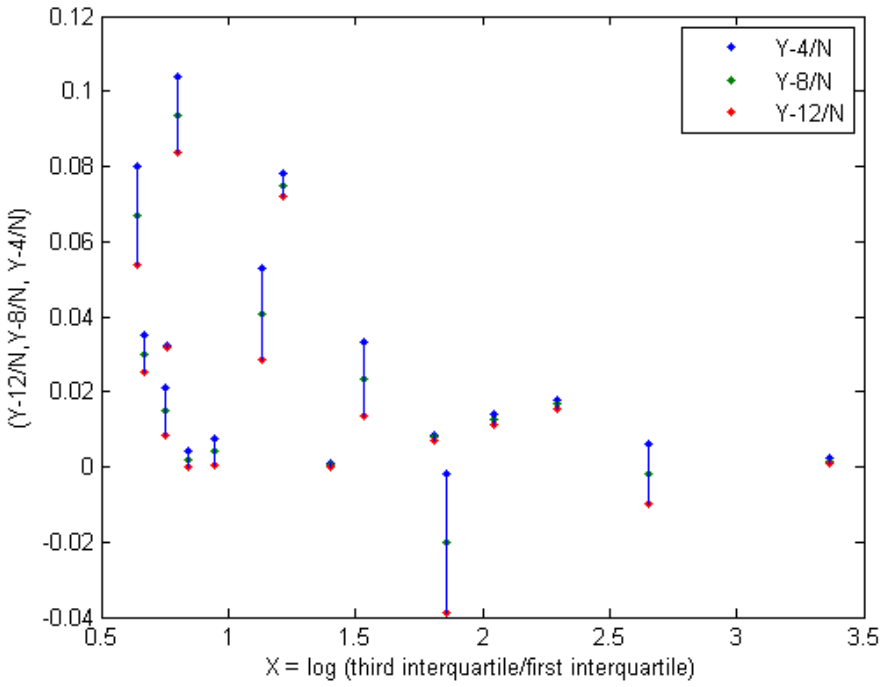

Figure 1. The relation between spread and closeness to Benford in 18 data-sets.

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.

Conclusion. On a small stage (18 data-sets) we have checked a theoretical prediction. Not just the literal assertion of Benford's law – that in a data-set with large spread on a logarithmic scale, the relative frequencies of leading digits will approximately follow the Benford distribution – but the rather more specific prediction that "distance from Benford" should decrease as that spread increases.

In one sense it's not surprising this works out. But the background point is that we *don't* know how to formulate and check any analogous prediction for the Normal approximation for data, and in that light our Benford analysis is perhaps noteworthy.

Recalling the data was collected as 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 4.2) 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.

3.2 Notes to instructors.

Implicit in the above are conceptual issues that are actually quite subtle and easily misunderstood, even by instructors. I explain here for instructors,

relating where relevant what I say in class.

1. Of course, what to say in class depends on what you think the students understand and misunderstand from previous courses. A disheartening way to find out is to ask the question

(*) Given one sample U_1 from Uniform$[0, T]$ with unknown T, write down a 50% CI for T

which few students can answer with any self-assurance, because (I guess) repeated exposure to symbolism of the type $\hat{\mu} \pm 2\hat{\sigma}/\sqrt{n}$ beloved by authors of mathematical statistics textbooks has overwritten the actual concept of a CI in their minds. The instructor needs to remind them what a CI is, and that it depends on assumptions whose validity one needs to consider (recall (*) arises in the Doomsday argument, discussed in another class, an extreme case of dubious validity of assumptions!)

2. The tradition (which we invoked to calculate a confidence interval for each data-set) of regarding a data set with n entries as n independent random samples from some unknown distribution is, in many contexts, hard to justify. This is the topic of another class in the course; let me summarize here what I say in the other class.

(i) Some statisticians act (without explicitly asserting this as a general principle) as if any data-set (e.g. exam scores from a class of students) can be analyzed as if it were i.i.d. from some unknown distribution, unless there's some particular reason not to do so. Others statisticians more explicitly assert that one should not do so without positive justification.

(ii) I take an intermediate view: that without positive justification
(a) it's incorrect to suggest that calculated CIs or P-values are "objectively

true" in the conclusion of a study;

(b) but having some indication of the uncertainty attached to an estimate is in practice often preferable to the alternatives, which are to regard the estimate as exact or to ignore it completely – just remember that the measure of uncertainty is itself uncertain. Everyday life requires decisions under uncertainties which are almost never mathematically quantifiable, but we survive using more intuitive degrees of confidence.

In this class, I remind students of that discussion, show the titles of the 18 data-sets in the study to illustrate we're in the "no positive justification" context, so that the "notional CIs" shown in Figure 1 should be regarded only as rough indications of the distance between the distribution in each data-set and the Benford distribution.

3. Another class discusses the misuse and misinterpretation of tests of significance – it's fun to do this with reference to the recent polemical book of Ziliak and McCloskey (2008). Again, let me summarize here what I say in the other class (and textbooks say, but too often forgotten by students).

(i) There is certainly a list of textbook settings where tests of significance are appropriate (randomized controlled experiments, etc).

(ii) A test of significance addresses some very specific question such as "can we be sure beyond reasonable doubt that the size of a certain effect is not exactly zero"; this is often not the relevant question. It's often more relevant to study the size of effect.

(iii) Often the null hypothesis is conceptually not plausible; in which case a test of significance is merely an exercise is finding out whether you can "knock down a straw man".

(iv) A test of significance assumes some particular probability model for the data under the null hypothesis; often the particular probability model is not plausible.

(v) With a small amount of data a big effect will not be detected as statistically significant; with a large amount of data a small effect will be detected as statistically significant.

In class I point out how much of this applies to the present setting. There is no reason to believe Benford is exactly true for any particular data set. The entries in our data-sets are complete lists, neither random samples nor arising from some underlying random mechanism we might plausibly model. A list of 18 P-values, from the chi-squared test of fit to the Benford distribution for each data-set, would be essentially meaningless – the widely varying width of the CIs in Figure 1 reflects widely varying sizes of the data-sets, so the implied statistical significance or non-significance of P-values would mostly reflect size of data set.

4. The real logical point (too hard to convey effectively in class, I suspect) is that there's no direct connection between the two issues

(a) given two probability distributions (either theoretical, or empirical data without any consideration of randomness), calculate some measure of distance between the distributions;

(b) given an i.i.d. sample from some distribution, might this have come from a particular specified distribution ?

other than the fact it's mathematically convenient to use "sum of squares of differences" in both contexts. We are interested in (a) – that's part of the "explanation" we are studying – – but we also take into account notional

sampling variability, expressing distance via a notional CI rather than a point estimate.

5. Note that our ± 1 S.E. confidence intervals are not 68% confidence intervals (the null distribution is not Normal). I don't say this in class (unless it arises from a student question) because it's a bit of a distraction – the numerical value is hardly important. Similarly, a student might be puzzled by the fact that 3 of the 18 confidence intervals include (impossible) negative values. If asked, I explain that this might be embarrassing to a professional statistician, but is not an actual error; rather, one could devise a technically better rule for constructing the CI so this didn't happen.

3.3 Note to instructors: background to Benford's law

I don't talk in class about the history of Benford's law, and say only three sentences (see below) about rigorous hypothesis-conclusion formulations, but let me write a few words here because I find some current literature unsatisfactory. Feller (1966, pp. 62–63, trivially edited) derives Benford's formula as follows.

> 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 $10^n k \le Y < 10^n(k+1)$ for some n. For the variable $X = \log_{10} Y$ this means
>
> $$n + \log_{10} k \le X < n + \log_{10}(k+1). \qquad (1)$$

> 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 $\log_{10}(k+1) - \log_{10} k = b_k$.

Now Feller's particular phrase "if the spread of Y is very large" is a poor choice of words because if Y is Uniform$[0, T]$ and T is very large then we would normally think of Y as having large spread, whereas what the argument implies and what Feller obviously *meant* is that $X = \log_{10} Y$ should have large spread. The contemporary phrase "large spread on a logarithmic scale" is better. With this emendment[2] Feller's explanation strikes me as the same as that in the previously cited modern sources (Fewster (2009); Wikipedia:Benford's law (2009)) though their graphics make the argument much easier for students to understand.

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" should be approximately uniformly distributed on $[0, 1)$) without justifying the latter. And this intuitive assertion does not correspond directly to any rigorous theorem, in that it is easy to write examples where the asserted conclusion is wrong.

There are several ways to obtain Benford's distribution as a rigorous limit under precisely specified hypotheses. Feller (1966), in text preceeding that cited above, implicitly outlines a theorem under hypotheses to the effect that the Y is a product of an increasing number of independent ran-

[2] *emend* – "to improve (a published sentence) by critical editing" – is an obscure but useful word for academics to know. Alas, Google and many online dictionaries wrongly treat *emendment* as a mis-spelling of *amendment*.

dom variables. Hill (1995) gives derivations of Benford's law based on more structural assumptions such as scale-invariance, and further technical work can be found in the online bibliography at http://www.benfordonline.net/.

A recurring theme in the course is to emphasize that a logically correct mathematical theorem is only relevant to the real world to the extent that you can check its assumptions in the particular real-world setting of interest. As I don't see how to check the assumptions of the Benford theorems for a particular kind of data-sets we studied.

Anyway, what I say in class is a 3 sentence summary.

> Just as there are many versions of the CLT, proving a Normal limit under specified hypotheses, so there are several theorems in which Benford's law appears as the distributional limit under specified hypotheses. However, as with the fuzzy CLT it seems impossible to check such hypothesis for a typical data-set of interest. The significance of the associated "explanation" in the Benford case is that its assumption is readily checked.

3.4 The calculation

(Note to instructor: I see little value in doing algebra in real time in class; this calculation would be posted on the course web site, which contains extensive course-related material. One could give a shorter derivation of the variance by quoting results about the chi-squared distribution, but this leads to possible confusion with the chi-squared test of significance, and the few calculations I do in class are intended to remind students of basic

mathematical probability techniques.)

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 $\text{var}X_i = nq_i(1 - q_i)$,

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

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, $\text{var}(D)$ 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

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

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

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

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

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

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

4 Back to a bigger picture

4.1 What I say in class

The general theme of the course is to examine critically what parts of the standard theoretical treatment of probability and statistics, as presented in textbooks, correspond to empirically verifiable features of the real world. Of course textbooks are rarely wrong in what they say explicitly; it's more a case of emphases, implicit impressions and omissions. Recall, for instance, that in the recent best-seller *The Black Swan*, Taleb (2007) argues forcefully that basic statistics textbooks place too much emphasis on Normal curve behavior and too little on power-law tail behavior.

On the theme of implicit impressions, let's return to an opening comment. There is a curious inconsistency between what many freshman textbooks *say* and what they *do* in exercises and examples regarding the issue of the validity of the Normal approximation for observed data. Table 1 shows the contexts, in three textbooks, where in-text examples and student exercises use the Normal distribution.

(Note to instructor: I show the table as a slide, while saying that I categorized the examples 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.

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 (Freund 2001, Triola 1998, Weiss 1999) examples of the Normal distribution for data.

It seems a reasonable presumption that, because of such examples, students comes away from introductory statistics courses with the perception that data on these kind of subjects follows the Normal distribution. Is this perception correct? Well, we don't really know: scientifically serious studies are hard to find, to put it mildly.

Now I am not suggesting that you students enter a centuries-old debate concerning the Normal. But an interesting future student project is to collect data (analogous to that in Table 1) on asserted instances of power-law distributions, and *The Black Swan* would be one place to start. It is also interesting to think about protocols for choosing data-sets, and some guidance is given at (point 4 of section 4.2; the protocol could be said briefly in class and would be posted on the course web site along with many other possible student projects).

4.2 Notes to instructors.

1. *Some details relating to Table 1.* 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.

2. What these three texts explicitly 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) is brief and unobjectionable. Weiss (1999) 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

Triola (1998) gives the "fuzzy CLT" quote we copied in section 1; and Freund (2001) says nothing.

But 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 concerning variables of the type in Table 1.

3. Almost any statistics course can be enlivened by discussing quotes from *The Black Swan*, which I find often insightful and equally often misleading; I mention it frequently in this course. Aside from any page-by-page analysis, the main moral I draw for statisticians from *The Black Swan* is that one could not write such a book on a typical scientific topic, only on one which has been inadequately explained by its practitioners.

4. In class I only briefly touch upon, in the Normal context, the broader issue of to what extent the often-asserted general fit of particular kinds of data to particular theoretical distributions is empirically true. Let me here address the issue of how to gather data-sets to test such assertions, in order to formulate a future student project on occurrence of power law distributions.

The data we studied for Benford's law resulted from instructing a student

go on the Internet and find authoritative data-sets with large but varying spreads

and we used every such data-set collected; call this *haphazard* data-set collection. In contrast, authors who choose examples which fit the theoretical distribution and reject other examples are using *selective* data-set collection; this may be fine in textbook discussion of an uncontroversial theory but is hardly convincing as evidence for a general theory.

Now haphazard and selective data collection could be viewed as the "stamp collecting" phase of scientific enquiry – if Table 1 had arisen from studying actual data-sets and distinguishing the (approximately) Normal from the non-Normal, and if we noticed that most of the "human physiology" examples in our collection were Normal, then we could hypothesize that most "human physiology" examples are indeed Normal. How to test such a hypothesis? It seems to me that to avoid selection and classification bias one should formulate and follow some protocol (just as clinical trials are required to follow some prespecified protocol). A very strict protocol would be

- 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.

This is perhaps over-elaborate in practice, certainly for my students. But this is what it would take to convince me of the correctness of some gener-

alization such as "most human physiology data is approximately Normal" — what about you?

Acknowledgements. We thank two referees and the Editors for thoughtful critiques.

5 References

Aldous, D.J. (2009a), "Overview of Probability in the Real World project", http://www.stat.berkeley.edu/users/aldous/Real-World/cover.html.

Aldous, D.J. (2009b), " Which mathematical probability predictions are actually verifiable?", http://www.stat.berkeley.edu/ aldous/Real-World/bet.html.

Feller, W. (1966), *An Introduction to Probability Theory and its Applications* (Vol. II), New York: John Wiley.

Fewster, R.M. (2009), "A simple explanation of Benford's law", *The American Statistician*, 63, 26–32.

Freund, J.E. (2001), *Modern Elementary Statistics* (10th edition), Englewood Cliffs, NJ: Prentice Hall.

Hill, T.P. (1995), "A statistical derivation of the significant-digit law", *Statistical Science*, 10, 354–363.

Taleb, N.N. (2007), "The Black Swan: The Impact of the Highly Improbable", New York, NY: Random House.

Triola, M. F. (1998), *Elementary Statistics* (7th edition), Reading MA:

Addison-Wesley.

Weiss, N.A. (1999), *Elementary Statistics* (4th edition), Reading MA: Addison-Wesley.

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

Ziliak, S, and McCloskey, D. (2008), *The Cult of Statistical Significance*, Ann Arbor MI: Unversity of Michigan Press.

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