

Going beyond the book: towards critical reading in statistics teaching

Andrew Gelman

Columbia University, New York, USA.
e-mail: gelman@stat.columbia.edu

Abstract

We consider three examples from our own teaching in which much was learned by critically examining examples from books. Even influential and well-regarded books can have examples where more can be learned with a small amount of additional effort.

Keywords:

Teaching; Categorical variables; Continuous variables; Chi-squared test; Sex ratio.

Introduction

We can improve our teaching of statistical examples from books by collecting further data, reading cited articles and performing further data analysis. This should not come as a surprise, but what might be new is the realization of how close to the surface these research opportunities are: even influential and celebrated books can have examples where more can be learned with a small amount of additional effort.

We discuss three examples that have arisen in our own teaching: an introductory textbook that motivated us to think more carefully about categorical and continuous variables; a book for the lay reader that misreported a study of menstruation and accidents; and a monograph on the foundations of probability that over interpreted statistically insignificant fluctuations in sex ratios.

Categorical or continuous?

The book *Mind on Statistics*, by Jessica Utts and Robert Heckard (2001), is an excellent text that is full of examples for statistics classes at all levels. A fun thing about working from a good textbook is that more can be learned by considering its examples in further depth. For example, early on in the book, the concepts of continuous and categorical variables are introduced, and the following variables are listed as "categorical": Dominant hand (left-handed or right-handed), regular church attendance (yes or no), opinion about marijuana legislation (yes, no, or not

sure), and eye colour (brown, blue, green, or hazel). From another perspective, though, three of these four variables could also be considered as continuous.

The issue is clearest with handedness, which Utts and Heckard categorize as left- or right-handed, but can be also described by a continuous variable, as we illustrate with the left histogram in figure 1, which is based on data we collected from students in a class. (More systematic surveys obtain similar results; see, e.g. Oldfield 1971.) As this histogram shows, many people fall between the two extremes of pure left- and pure right-handedness. But as the right histogram in figure 1 illustrates, students tend to guess the distribution of handedness to be bimodal and thus essentially discrete. This common misconception would make handedness a particularly effective example of a continuous variable that is often summarized discretely.

Similar issues arise for two of the other variables given by Utts and Heckard. Church attendance can be measured by a numerical frequency (e.g. number of times per year), which would be more informative than simply yes/no, or it can be binned in ordered categories. For example, the American National Elections Study (<http://www.electionstudies.org/>) asks, "How often to you attend religious services, not counting weddings or funerals?" and records five sorts of response: More than once per week, once per week, more than once per month, several times per year, and never. Finally, the three options for opinion about marijuana legislation could be coded as 1, 0

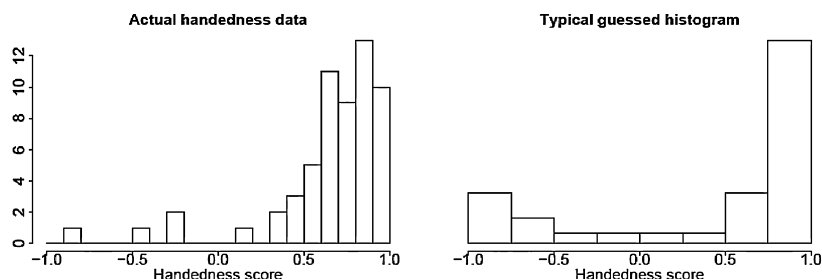


Fig. 1. Handedness can be measured by a 10-item questionnaire to yield an essentially continuous score ranging from -1 (pure left-hander) to $+1$ (pure right-hander). We had the students in an introductory statistics class fill out the questionnaire and also asked them to sketch what they thought the histogram of other students' handedness scores would look like. (a) Data from the class; (b) a guess from a group of students of what they thought the histogram would look like (before seeing the actual data). Bimodality was anticipated but did not actually occur



Fig. 2. Sketch from an example in Zeisel (1985), who writes, "When the frequency of [driving] accidents is plotted against the time of menstruation a surprisingly shaped curve arises [left graph]. Upon investigation, the curve turned out to be the composite of two easily identified separate curves [right graph]; one for parous women (those who had given birth) and one for nonparous women. The one group had the accident peak immediately after their period, the other immediately before it." Compare with the actual data shown in figure 3

and 0.5, and further intermediate preferences could be identified with detailed survey questions asking about medical marijuana, criminal penalties and so forth.

The point of bringing all this up in class is not to lay down the law and say that church attendance, for example, is inherently discrete or continuous. Rather, we want to lead students to think about the ways in which reality is abstracted by numerical measurements. We also find it empowering that we can learn more about the structure of these variables either by collecting our own data (as illustrated in figure 1) or with library research (as by looking at the National Elections Study questions).

The graph that wasn't there

About fifteen years ago, when preparing to teach an introductory statistics class, I recalled an enthusiastic review I had read (Sills 1986) of the sixth edition of Hans Zeisel's (1985) book, *Say It With Figures*. I bought the book and, flipping through it to find examples for use in class, came across the two sketches reproduced in figure 2. The curves represent data from hospital admissions of premenopausal women who had been involved in traffic accidents, with the left hump

representing accidents that had occurred just before the menstrual period and the right hump showing accidents occurring just after the period.

This seemed like a great example for class. I figured that a graph of the actual data would be even better than a sketch, so I went to the library and found the cited research by Katharina Dalton (1960). The graphs are reproduced in figure 3, and they look nothing like Zeisel's sketches in figure 2! For one thing, the sketched densities show almost all the probability mass just before and after menstruation, but the data show only about half the accidents occurring in these periods. Perhaps more seriously, the sketch shows two modes with a gap in the middle, whereas the data show no evidence for such a gap. Similarly, the two bell-shaped pictures in the right sketch of figure 2 do not match the actual data as shown in the histograms on the right side of figure 3.

Dalton's findings were conveniently summarized by an article in *Time* magazine on November 28, 1960: "In four London general hospitals Dr Dalton questioned 84 female accident victims (age range 15–55), all of whom had normal, 28-day menstrual cycles. Her findings: 52% of the accidents occurred to women who were within four days, either way, of the beginning of menstruation. On a purely random basis, the rate would have been only 28.5% for the same eight days. Childless women, noted Dr Dalton, appear to be abnormally accident-prone just before menstruation, while women who have borne children are vulnerable over the whole premenstrual and menstrual period." What is relevant to our discussion here is that these findings were not accurately described in Zeisel's book. On an unrelated but amusing (from a current perspective) note, the *Time* article quoted Dalton as saying that these

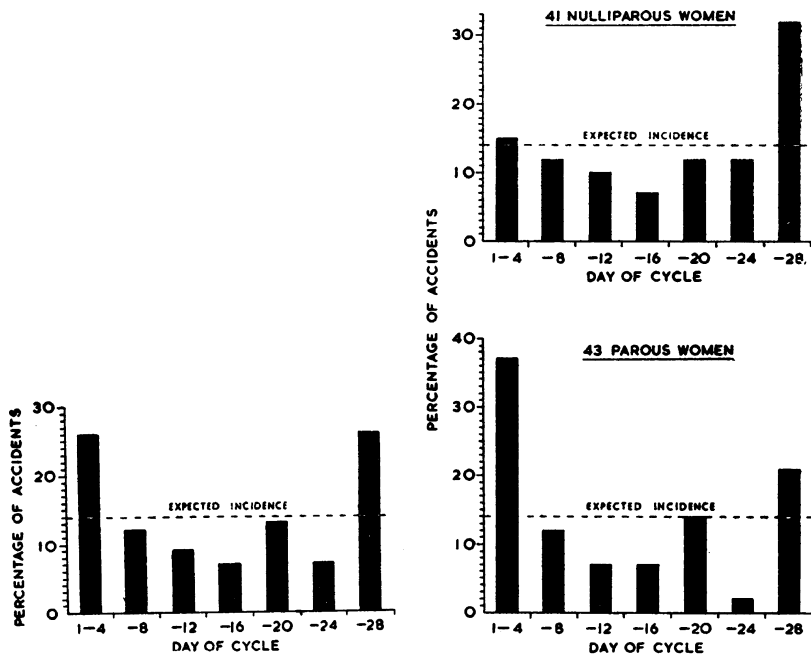


FIG. 1.—Distribution of 84 accidents in the menstrual cycle.

FIG. 2.—Distribution of accidents in nulliparous and parous women.

Fig. 3. Graphs from Dalton (1960) with the raw data on menstruation and accidents. These histograms look almost nothing like the sketches in figure 2 taken from Zeisel's book. Many of the accidents fall outside of the days indicated by the modes in figure 2, and, unlike those sketches, there is no gap between the peaks. The above images, from the *British Medical Journal* (1960) 2, 1425, are reproduced with permission from the BMJ Publishing Group

findings "cause one to consider the wisdom of administering tranquilizers for premenstrual tension."

I suspect that Zeisel heard about the research (perhaps even by reading *Time* magazine), recognized that it would be a good teaching example, and went to the library to read Dalton's original article. He then could have too hastily summarized the data in a sketch, inadvertently knocking out most of the accidents that did not occur just before or after menstruation and mistakenly inserting a gap in his histogram between the two modes. Or perhaps he was looking for an example of a mixture model and didn't look too closely at the data. In any case, this is a benefit to our students, who get a lesson in how easy it is to misread a research report. Had Zeisel's book not been so appealing and well written, we would not have been drawn to the example in the first place.

For teachers, the most important lesson is that going to the source of the data turned up a better example to use in class. For students, the lesson is to be sceptical when seeing second-hand reports of data, even when coming from a credible-seeming source.

Finding patterns in noise

The book *Probability, Statistics, and Truth* by Richard Von Mises (1957) is an important text in the foundations of probability, laying out a deri-

vation of the axioms of probability theory from the concept of infinite random sequences. This work has long been influential in statistics (see, e.g. Wald (1939) and Good (1958) for classical frequentist and Bayesian reactions) and in philosophy (e.g. Gillies (2000) connects von Mises's ideas to those of Karl Popper and others).

I bought the book several years ago, and, in skimming it, alighted on the chapter on "Applications in Statistics," within which von Mises uses the sex ratio of births to illustrate the binomial distribution. He reports the proportion of boys born in each of the 24 months of 1907–1908 in Vienna and found less variation than expected. In his words: "The average of these 24 values is 0.51433; the dispersion [n -weighted variance] . . . is 0.0000533." He computes the expected dispersion as $(23/24)(p(1-p)/n) = 0.0000613$ (here, n is about 3900 per month, and p is taken to be 0.514) and then writes, "The actual dispersion is smaller than the theoretical one. In other investigations of the proportion of male births, a value of Lexis's ratio closer to 1 is obtained. We must therefore look for an explanation of the slightly subnormal dispersion found in this special case." He goes on to attribute this lower variance to different sex ratios in different racial or socio-economic groups.

However, while the variance is less than expected by chance under the assumption of a constant sex ratio, is not at all statistically

significantly less. We can see this by using the chi-squared test for overdispersion – a topic that is not typically covered in a secondary-school statistics course, but is conceptually similar to other statistical tests. The null hypothesis is derived from the assumption that the number of births each month is binomially distributed with a constant probability, and the mathematical form of the test is similar to that of a normal population having a specified variance based on a sample from it using the statistic $(n - 1)s^2/\sigma^2$.

With 24 months, the sample variance $s^2 = (24/23)(0.0000533)$ is estimated based on 23 degrees of freedom, and we can use the chi-squared test to compare it with the theoretical variance $\sigma^2 = (24/23)(0.0000613)$ from the model that assumes a constant sex ratio. As $23s^2/\sigma^2$ follows a chi-squared distribution with 23 degrees of freedom, the observed ratio of 0.869 has a p -value of 0.36; i.e. one would observe a ratio this extreme or smaller more than a third of the time, just by chance.

Thus, it is unnecessary to search for an explanation for the discrepancy, especially given that, as von Mises notes, birth numbers of boys or girls are among the rare data that actually do generally follow the binomial distribution. In addition, irrelevantly for the technical point but of interest when teaching, von Mises makes a presentational lapse by summarizing dispersion with variances rather than standard deviations, which are more interpretable on the original scale of the data.

Von Mises is hardly alone in over-interpreting birth data: there is a long tradition of looking for patterns in birth data, despite that there is no convincing evidence that boys or girls run in families or that sex ratios vary much at all except under extraordinary conditions. (See Freese and Powell (2001), Das Gupta (2005) and Gelman (2007) for more on the over-interpretation of statistical fluctuations in sex ratios.) Thus, in addition to illustrating the important technical point of assessing statistical significance of a variance ratio, this example opens the door to a more general discussion of how and why statistics can be misread. (Note: For the analysis above, the calculations of von Mises are taken at face value. He, in fact, computes the dispersion incorrectly from the 24 observations he lists. We invite the reader to go to the original source to compute a dispersion of 0.0000394, which still gives a fairly healthy p -value of 0.10 compared with the value of 0.36 stated above.)

That this occurred in an influential book merely underscores that even a standard chi-

squared test for overdispersion cannot be taken for granted. In a similar vein, finding that the great Francis Galton performed inaccurate calculations with the normal distribution (mistakenly predicting that there were nine-foot-tall men in Britain; see Gelman (2006) and Wainer (2007)) gives us a new respect for the pioneers who worked out the mathematical property of that model.

Discussion

Individually, these examples are of little importance. After all, one does not go to a statistics textbook to learn about handedness, menstruation, and sex ratios. It is striking, however, that the very first examples I looked at in the Zeisel and von Mises books – the examples with interesting data patterns – collapsed upon further inspection. In the Zeisel example, we went to the secondary source and found that his sketch was not actually a graph of any data, and that he in fact misinterpreted the results of the study. In the von Mises example, we reanalysed the data and found his result to be not statistically significant, thus casting doubt on his already doubtful story about ethnic differences in sex ratios. In the Utts and Heckard example, we were inspired to collect data on handedness and look at survey questions on religious attendance to find underlying continuous structures.

Teaching activities already exist in which students apply critical reading skills to news reports and scientific articles with statistical content (Gelman and Nolan 2002); here, the recommendation is to have an inquiring eye when reading books that we teach from as well. Much can be learned by redoing analyses and going to the primary and secondary sources to look at data more carefully, and this can help us improve our teaching, even from our favourite books.

Acknowledgements

We thank Ji Meng Loh, Martin Lindquist, Roger Johnson, and an anonymous reviewer for helpful comments and the U.S. National Science Foundation, National Institutes of Health, and Columbia University Applied Statistics Center for financial support.

References

- Dalton, K. (1960). Menstruation and accidents. *British Medical Journal*, **2**, 1425–1426.

- Das Gupta, M. (2005). Explaining Asia's "missing women": A new look at the data. *Population and Development Review*, **31**(3), 529–535.
- Freese, J. and Powell, B. (2001). Making love out of nothing at all? Null findings and the Trivers-Willard hypothesis. *American Journal of Sociology*, **106**(6), 1776–1788.
- Gelman, A. (2006). Galton was a hero to most. *Statistical Modeling, Causal Inference, and Social Science* blog, 23 October 2009. <http://www.stat.columbia.edu/~gelman/blog> (accessed 11 October 2010).
- Gelman, A. (2007). Letter to the editor regarding some papers of Dr. Satoshi Kanazawa. *Journal of Theoretical Biology*, **245**(3), 597–599.
- Gelman, A. and Nolan, D. (2002). *Teaching Statistics: A Bag of Tricks*. Oxford: Oxford University Press.
- Gillies, D. (2000). *Philosophical Theories of Probability*. London: Routledge.
- Good, I.J. (1958). Review of *Probability, Statistics and Truth*, by Richard von Mises. *Journal of the Royal Statistical Society, Series A*, **121**(2), 238–240.
- Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, **9**(1), 97–113.
- Sills, D.L. (1986). Review of *Say it with Figures*, by Hans Zeisel. *Journal of the American Statistical Association*, **81**(393), 257.
- Utts, J.M. and Heckard, R.F. (2001). *Mind on Statistics*. Pacific Grove, CA: Duxbury.
- Von Mises, R. (1957). *Probability, Statistics, and Truth* (2nd edn). New York: Dover. Reprint.
- Wainer, H. (2007). Galton's normal is too platykurtic. *Chance*, **20**(2), 57–58.
- Wald, A. (1939). Review of *Probability, Statistics and Truth*, by Richard von Mises. *Journal of the American Statistical Association*, **34**(207), 591–592.
- Zeisel, H. (1985). *Say it with Figures* (6th edn). New York: Harper and Row.

Technology tip

Randomisation Tests in R.

Kabacoff R (2011) "Data Analysis and Graphics with R" is one recent example of a textbook suitable for use at university level which features a solid section on Bootstrapping and Randomisation tests. However, a basic two sample randomisation test is very simple to implement in R.

First we type in Student's sleep data (see ?sleep for full attribution)

```
x1 ← c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0, 2)
```

```
x2 ← c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4)
```

Then we measure the observed test statistic, in this case the absolute difference between sample means:

```
observed.test.statistic ← abs(mean(x1)-mean(x2))
```

Next we need to generate the null hypothesis distribution for this test statistic. We create a storage vector, pool all the data and then use a simple loop to shuffle and cut (generate two data vectors by sampling with replacement),

each time calculating the absolute difference in means:

```
null.hypothesis.dist ← vector("numeric", 1000)
pooled ← c(x1, x2) for (i in 1:1000){ pooled.
shuffle ← sample(pooled) x1.shuffle ← pooled.
shuffle[c(1:10)] x2.shuffle ← pooled.shuffle
[c(11:20)] null.hypothesis.dist[i] ← abs(mean
(x1.shuffle)-mean(x2.shuffle))
}
```

Having simulated an approximation to the null hypothesis distribution we just need to find out how unusual the observed test statistic was, given this null.

```
hist(null.hypothesis.dist, freq = FALSE, xlab =
"Test statistic", main = "Null hypothesis distribu-
tion") abline(v=observed.test.statistic, col =
"red")
```

```
length(null.hypothesis.dist[null.hypothesis.dist >
observed.test.statistic]) / 1000
```

Simple extensions to this are to develop a one sided test, and to alter the test statistic that is used.