Teaching the Craft of Data Analysis

Daniel W. Schafer and Fred L. Ramsey
Oregon State University

Journal of Statistics Education Volume 11, Number 1 (2003), ww2.amstat.org/publications/jse/v11n1/schafer.html

Copyright © 2003 by Daniel W. Schafer and Fred L. Ramsey, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Applied statistics; Case studies; Real data; Statistics major; Undergraduate.

Abstract

We describe our experiences and express our opinions about a non-introductory statistics course covering data analysis. In addition to the methods of statistics, the course emphasizes the process of data analysis, the communication of results, and the role of statistics in the accumulation of scientific evidence. Since it is impossible to provide explicit instructions for all data analytic situations, the course attempts to impart a body of tools, a spirit of approach, and enough thoroughly covered case studies to give students the skills and confidence to apply this craft on their own.

1. Introduction

We teach a one-year sequence that serves both as a service course for graduate students in various departments and as a serious data analysis course for undergraduate mathematical sciences majors. In this paper we discuss the undergraduate audience, we further advocate the applied side of statistics in the undergraduate program, and we convey our opinions about covering the process of analyzing data more deeply in a statistics course beyond “Stat 101.”

Our attention to the graduate audience has been instrumental in our thinking about instruction for both groups. Most importantly, since we serve on many graduate committees, we have come to see how students who are one or two years removed from the data analysis course attack their “outside-the-classroom” data problems. The bungling of data analysis, often by students who received grades of A, indicated to us that we weren’t training the students in the full range of skills necessary for successful application of statistics. It was a picture of more capable data analysts that motivated us to restructure the course 14 years ago.

Our goal was to graduate students from the class who could successfully apply the spirit and techniques of statistical data analysis to a wide variety of data problems, who could successfully communicate the results to their intended audience, who could recognize problems they were incapable of solving on their
own, and who had a fairly clear picture of the types of mental activities associated with applied statistics.

Our restructuring strategy to achieve this goal involved two elements. The first was adding a prerequisite (since the course previously had none) so that a suitable starting point could be achieved. The second was to center the instruction on case studies. By these, we mean data problems - with context - that are considered in advance of the presentation of the statistical tools they demonstrate, and which have a clear question of interest. We attempt to linger longer on the case studies and take advantage of whatever lessons they offer. By including a summary of statistical findings and a discussion of the scope of inference for every case study, we try to demonstrate good techniques for the communication of statistical results and to illustrate the role of individual studies within a broader scientific enterprise.

In Section 2 we offer some observations and opinions about the continuing evolution of applied statistics in undergraduate statistical education. Section 3 further details our approach for centering the instruction on case studies. We provide an assessment in Section 4, and Section 5 contains comments on prerequisites and the role of mathematics.

2. The Continuing Evolution of Undergraduate Statistical Education

We suspect there is general agreement that the academic discipline of statistics can benefit from a better self-definition. Much of this definition involves a pronounced divorce from mathematics. Much of this definition also involves a pronounced usefulness of statisticians. In addition, we ought to do a better job at attracting students to the field. Most importantly, we believe that a deeper attention to data analysis at the undergraduate level is key.

The balance of theoretical and applied training in undergraduate programs or tracks in statistics is variable, but is evolving from a starting point in which the theoretical side was dominant. We are advocating, as others have already done, more respect for the applied part. It should be possible to better communicate what it is that professional applied statisticians do, what type of mental process is associated with attacking a data analysis problem from start to finish, and what is exciting about the field of applied statistics.

There are, of course, questions about how far undergraduates can be taken in data analysis. While serious undergraduate statistical training has traditionally served to prepare students for graduate studies, leaving the instruction of real data analysis to the graduate program, there are some changes that now make the communication of the data analysis process more accessible to these students.

First, students are entering college with a better sense of what the field of statistics is about because of exposure to statistical thinking in the K-12 curriculum and the development of Advanced Placement (AP) Statistics courses. Second, statistical software is easier to introduce, both because students are already familiar with computers and software, and because statistical software is now more mainstream in its user interface. Finally, since so many statistical educators are searching for and disseminating good “data and story” problems, and since so much data is now accessible over the Internet, good data examples - which are digestible and relevant to undergraduates — are readily available.

We don’t mean to imply that the master’s degree in statistics will become obsolete. There are limits to how much undergraduates can be taught professional-quality data analysis and the important reasoning behind it. Further, despite the improved conditions, we still find that some undergraduates are data-mature enough to appreciate the big picture and some aren’t. Those that aren’t won’t get much out of an
undergraduate statistics program. But many are mature enough and can be taught, at least, an appreciation for data analysis.

Our point is that conveying a more thorough understanding and appreciation of data analysis should be central in our thinking about the undergraduate statistical curriculum. Such focus will help not only to communicate a more complete image of the field of statistics but also to encourage those students who like "making sense out of what we observe" to consider careers in the field. Further, the task of teaching data analysis at the undergraduate level is not as unrealistic as it once seemed.

3. The Case Study Approach

There now seems to be a near universal recognition of the need for real data problems in applied statistics classes. What we wish to add, though, is that real data problems are necessary but not sufficient. It is not enough to have “data examples.” Considerable care and some skill are needed to use the full data problems to communicate the entire process of data analysis and the role of statistics in scientific learning.

By "case studies" we mean data sets with accompanying context that are considered thoroughly and that occupy a central position in the course structure. Our approach is to present the case studies first, as an introduction to a data structure, then use them to demonstrate the methods. For each case study we also include a summary of statistical findings to illustrate statistical communication, discuss the scope of inference as it relates to the study’s design, and talk about any additional, broader issues of data analysis that arise in the analysis and interpretation.

3.1. The Broader Set of Data Analytic Skills

To begin data analysis we ask questions such as: "What exactly is being asked here?" "How were the data gathered and what do they look like?" "What options do I have for analysis?" "What can I do to get some initial ideas of answers to the questions of interest and possibilities for models?" "Can I start with a simplification?" Then, during the course of the analysis: "If I transform variables will a convenient tool be appropriate and still permit answers to the question of interest?" "What do I know about the importance of the various assumptions for the tool I am considering?" "Do different tools give different answers to the questions of interest and, if so, why?" "Are there influential observations, what do I know about resistance, and what action should be taken?" And, in the end: "How can I best communicate the statistical evidence in answer to the question of interest?" "What kind of graph best highlights the results, without too much clutter?" "Do I need to caution the audience of my report on the limitations of interpretation here?" "If the questions of interest remain unresolved, what type of study would overcome the obstacles?"

Each data problem requires unique consideration and action. In this sense, statistical data analysis is a craft. We cannot simply teach students the calculations and interpretation of the one-way analysis of variance table, for example, and expect them to know how to deal with all data problems having the "several sample" structure. We cannot possibly present templates for all possible situations that will arise, so we must attempt to focus on the spirit of what we're trying to do and provide enough examples to instill confidence in applying this craft to new situations.

The following two examples have nearly identical "structures" but different questions of interest.

**Respiratory Rates Example.** A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly "high," however, a physician must have
a clear picture of the distribution of typical respiratory rates in healthy children. To this end, Italian researchers measured the respiratory rates of 618 children between the ages of 15 days and 3 years. The data are displayed in Figure 1 below (read from a graph in Rusconi, et al. 1994).

![Figure 1](image)

**Figure 1.** Respiration rate versus age for 618 children.

Notice the well-defined purpose of this data study. Since a normal, straight-line regression model fits well for log respiration rate on age, it is straightforward to estimate the percentiles of the distribution of respiration rate for various ages. A report of the findings might include a graphical display of some of the percentiles as in Figure 1. A look-up table listing some important percentiles by month of age, such as that displayed in Table 1, (with careful attention to the appropriate number of presented digits), would be useful to pediatricians in their offices.

**Table 1.** Selected percentiles of respiratory rate distribution.

| Age (months) | 1% | 5% | 50% | 95% | 99% |
|--------------|----|----|-----|-----|-----|
| 1            | 29 | 33 | 46  | 63  | 72  |
| 2            | 29 | 33 | 45  | 62  | 71  |
| 3            | 28 | 32 | 44  | 61  | 70  |
| 4            | 27 | 31 | 43  | 60  | 68  |
| ...          |    |    |     |     |     |
**Insulating Fluid Example.** Electrical engineers interested in the sensitivity of an insulating fluid to changes in voltages applied one of seven voltages to each of 76 batches of the fluid then observed the length of time that the fluid retained its insulating properties (data from Nelson 1970). The time to breakdown is plotted against the amount of applied voltage in Figure 2.

![Figure 2](image-url)

**Figure 2.** Breakdown time versus voltage for 76 batches of an insulating fluid.

As with the first example, a simple linear regression model with constant variance fits after log transformation of the response. In this case, the engineering question is best answered with the following statement of conclusion: "It is estimated that the median breakdown time decreases by 40% for each 1kV increase in voltage (95% confidence interval: 32% to 46%), for voltages between 26 and 38 kV."

While these examples have the same structure, what we end up doing with them and how we communicate the results is quite different. For example, normality is important to check in the first example but not the second. An interpretation of the regression coefficient (in a meaningful way) is required in the second, but not the first. Although it seems possible to roughly devise a list of the various purposes that might arise, what can only be conveyed through repeated serious attention to real examples is that the process of data analysis involves applying available statistical tools in ways that are customized to the individual problems. There is a need to focus on the questions of interest that is best communicated not with any instructive statement but with repeated demonstration.

This is difficult for the students at first. They aren’t able to write down exactly what it is they are supposed to be learning. But by the end of the one-year course they will have seen about 50 case studies
treated seriously by the instructor and will have analyzed and written summaries of statistical findings for another 25 on their own. By the end, the strategies, tools, and styles of presentation take form.

We spend the first four weeks of the course on the following topics: drawing conclusions, confounding variables, the connection between statistical inference and probability models for sample selection or treatment assignment, robustness, transformation, a strategy for dealing with outliers, and the communication of statistical results. We then revisit these topics, opportunistically and repeatedly, in discussions of the case studies in subsequent weeks. After the initial general issues we follow a typical outline of topics based on data structures: several samples, simple regression, multiple regression, and so on; and introduce two new case studies each week. These are selected to demonstrate the particular data structure, but are also used, whenever possible, to highlight the broader issues.

3.2. Statistical Ideas and the Accumulation of Scientific Evidence

By paying a bit more attention to where a study falls in the development of understanding of some topic we are also better able to convey the value of statistics in scientific investigation. The following is a simple example.

**Cardiovascular disease and obesity example.** There are two schools of thought about the relationship between obesity and heart disease. Proponents of a physiological connection cite several studies in North America and Europe that estimate higher risks of heart problems for obese persons. Opponents argue that it is the strain of social stigma brought on by obesity that is to blame. In a clever but simple approach for shedding light on this controversy, researchers investigated the association between obesity and heart disease in American Samoa, where obesity is socially acceptable and even desirable. The data in Table 2 are from a 5-year prospective study on Samoan women (from Crews 1988).

| Obese | Yes | No  |
|-------|-----|-----|
| 16    | 2045|
| 7     | 1044|

The following is a possible summary of statistical findings: "It is estimated that the odds of 5-year heart disease mortality in the obese sub-population is 16% greater than the odds in the non-obese sub-population (95% confidence interval: 52% less than to 182% greater than). The data are consistent with the theory that the odds of 5-year cardiovascular disease mortality does not depend on obesity (one-sided p-value = 0.37) but, as evident in the confidence interval, the data are consistent with a wide range of other values of the odds ratio as well."

The interesting feature of this example is the identification by the researchers of a special population as a way of avoiding a difficult confounding variable. While the choice of the Samoan population is quite clever, the resolution provided by the 5-year study of 3000 individuals is disappointing. The following
are possible discussion questions: In what way is the resolution disappointing? Would reinvestigating the mortality rates on the same individuals after 10 or 20 years help? Would increased sample size help? What value would there be in a similar study in a distinct population for which, as in Samoa, obesity is acceptable? What tools of study design and analysis do we now know of for eliminating the effects of known confounding variables?

It is the consideration of examples within a broader scientific investigation - such as the study of the causal relationship between obesity and heart disease - that makes the value of statistics evident. In addition, the problematic issues associated with the interpretation of $p$-values and confidence intervals become easier to accept when data problems are considered as a part of a scientific process for learning, rather than in isolation.

3.3. The Poetry of Good Data Examples

We have advocated the central role of (many) data examples in the data analysis course. The successful execution of this idea is not trivial. It is easy to get bogged down by the cognitive load involved in understanding data problems. As mentioned in Section 2, there are now more data problems to choose from.

By "poetry of good data examples" we mean those that convey several lessons, but with little extra clutter. We would like to have a relatively simple data problem having as many as possible of the following characteristics: It is interesting. It involves good science. There is an understandable context. There is a well-defined question of interest. The statistical tool being demonstrated is the one that is needed for addressing the question of interest. The statistical analysis is relevant. The data problem demonstrates some lesson about broader issues of data analysis.

**Interesting.** Naturally we wish to find examples that are interesting to students; but it might be even more important that they are interesting to the teacher - whose enthusiasm can be particularly contagious.

**Good science.** There are scientists who, through passion, understanding, and dedication have produced major scientific advances. Their clever choices for overcoming scientific obstacles in pursuing a theory (as in the obesity example) are often simply a result of persistence and a keen desire to find answers, yet often involve underlying statistical principles. We wish to show, wherever possible, this connection between good science and good statistics.

**Understandable context and well-defined question of interest.** There is a need to avoid too much extra cognitive load in presenting scientific examples. We believe rather strongly, though, that there must be a well-defined question of interest. There are three bad consequences of examples without well-defined questions of interest: (a) we miss out on any opportunity to provide additional lessons about the process of data analysis, (b) we foster the idea that statistics is an end rather than a means, and (c) we fail to give important practice on how data analytic choices are dictated by both the data structure and the questions of interest.

**The demonstrated statistical tool is the correct statistical tool.** It is similarly undesirable to demonstrate a statistical tool on an example for which some other tool is more appropriate. Often there is more than one way to analyze a particular dataset and it doesn't hurt to demonstrate one method and point out that there are others. What we wish to avoid is a vacuous use of statistics for the same reasons discussed in the previous paragraph.
The statistical analysis is relevant. Statisticians sometimes go to a lot of effort to report the obvious. Ideally, we like to find data problems for which someone who knows how to do good statistical data analysis can provide better answers to the questions of interest than someone who doesn't; that is, where the value of good statistical data analysis is clearly evident.

The case studies we use can be seen in our textbook (Ramsey and Schafer 2002). Some additional information is also provided on the Web site www.statisticalslneuth.com, including links to other data and stories pages. The Data and Stories Library (DASL, lib.stat.cmu.edu/DASL) and the Journal of Statistics Education Data Archive (ww2.amstat.org/publications/jse/jse_data_archive.html) are particularly worthy of attention.

4. Assessment

We believe that by focusing more on data analysis as a craft and highlighting the role of statistical methods as a broader enterprise, students see why the field of statistics - as a means of scientific exploration and discovery - is relevant and interesting. The success for the graduate students in our class is evident, in our opinion, on the graduate committees we attend. What works for the graduate students does not necessarily work for undergraduates, though, and it is more difficult to judge the latter. For what it’s worth, about half of the undergraduates score near the top of the class overall, and the remainder score from mediocre to poor. Aside from course performance, though, it is difficult to know whether undergraduates can grasp the big picture of the process of data analysis, since they don’t have the same scientific context to judge the usefulness of applied statistics as graduate students do. A major result, at least, is that some of the undergraduates decide, as a result of this class, to pursue further training in statistics.

5. Prerequisites and Mathematical Level

The official prerequisite for our course is either a class in statistical thinking or an introductory class in statistical methods. This vagueness is a result of the graduate student audience, the majority of whom have had their first course in statistics at some other college or university. Our attitude in first proposing a prerequisite was to take advantage of whatever we could. Simply avoiding initial discussions about means, standard deviations, sampling distributions, standard errors, the normal distribution, and so on, saves time. Since there were a wide variety of backgrounds, however, we were forced to choose a relatively low-level starting point: the two-sample problem. In the end, though, the result of this constraint has been quite satisfactory. By starting with the structure of the two-sample problem we are able to simultaneously review this elementary structure and, more importantly, discuss the broader issues mentioned in Section 3.1 above. (While we assume that the one-sample problem has already been covered, we review it for the special case of paired data.) In our experience, students who have had a previous statistics course understand very little, if anything, about such things as robustness, interpretation after a log transformation, and writing statements of statistical conclusions; and often very little about the extent to which conclusions are tied to the study design.

Our preference for the prerequisite is a course in statistical literacy. We believe there is a great deal about consuming statistical information and statistical thinking that ought to be covered before actually attempting to do statistics. Such a course would prepare the students sufficiently for our starting point and, in our opinion, is more likely to spawn an interest in the field than is a course that introduces methods of statistics at a shallow level.

Although the sophistication of our course is high when it comes to statistical models and modeling, the mathematical content in the course is low. It has been said that our approach would be good for non-
mathematically inclined students. While this is true, we would take issue with the implication that it is not appropriate for more mathematically inclined students. If students are serious about statistics they deserve a deeper treatment of the process of data analysis. A data analysis course that attempts to delve into very much mathematical rationale behind the statistical methods involved in the analysis of data might easily be un successful at both.

We do not avoid mathematics altogether, but try to restrict its use to those areas where it can help the applied statistician. For example, we find it useful and important to explain the statistical and algebraic steps to show the interpretations after back-transformation when logarithms have been used; but we don’t feel there is much merit in deriving the formulas for least squares estimators.

Ideally, one can imagine parallel courses in data analysis and in the theory involved in the methods used. Logistically, however, the steps involved in doing both things right would make it quite difficult to cover the same topics at the same pace.

A related issue is the difficulty that mathematics departments have in justifying a course for its majors with very little mathematical content. We have no answer for this, except to point out that many graduate programs in statistics, like ours, do not see strength in mathematics as necessarily the best indicator of success of a potential graduate student. In addition to mathematical training, we wish to see applicants who know what they’re getting into and who have a genuine appreciation for the use of statistics.

Regarding the question of where this course should fall within the entire undergraduate curriculum, our opinion is that because it will be much more successful with data-mature students, that later is better. Having said that, though, we think the inclusion of a course that goes deeply into data analysis is the important point, and that how it is included will depend a lot on the size and structure of the particular department.

**6. Conclusions**

We have expressed our opinions about the need for a deeper treatment of data analysis in the undergraduate curriculum, about the improving atmosphere for doing so, and about the choices we have made in our data analysis class. We think there is merit in treating applied statistics as a craft and in conveying the skills of the craft to undergraduates who are serious about statistics. To do this requires a suitable starting point, hence the requirement of some statistics prerequisite, accompanied by the broad incorporation of relevant case studies.

**References**

Crews, D. E. (1988), "Cardiovascular Mortality in American Samoa," *Human Biology*, 60, 417-433.

Nelson, W. B. (1970), *General Electric Company Technical Report 71-C-011*, Schenectady, NY.

Ramsey, F. L., and Schafer, D. W. (2002), *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed.), Belmont, CA: Duxbury Press.

Rusconi, F., Castagneto, M., Gagliardi, L., Leo, G., Pellegatta, A., Porta, N., and Razon, S., (1994),
"Reference values for respiratory rate in the first 3 years of life," *Pediatrics*, 94, 350-355.

Daniel W. Schafer  
Department of Statistics  
Oregon State University  
Corvallis, OR 97331  
USA  
schafer@stat.orst.edu

Fred L. Ramsey  
Department of Statistics  
Oregon State University  
Corvallis, OR 97331  
USA  
ramsey@stat.orst.edu