Statistics tend to become interesting to non-methodologists when taught in a research context that is relevant to them. Real data sets supplemented by sufficient background information can provide just such a context. Despite this, many textbook authors and instructors of applied statistics rely on artificial data sets to illustrate statistical techniques. In this paper, it is argued that artificial data sets should be eliminated from the curriculum and that they should be replaced with real data sets. Towards this end, a rationale for using real data sets and the characteristics that make data sets particularly good for instructional use are described. The difficulties encountered when using real data and strategies for compensating for these drawbacks are also discussed. Two authentic data sets and an annotated bibliography of dozens of primary and secondary data sources are included. (Author/PK)
OPENING UP THE BLACK BOX OF RECIPE STATISTICS:
PUTTING THE DATA BACK INTO DATA ANALYSIS

Judith D. Singer
John B. Willett

Harvard University
Graduate School of Education
April 1988

Correspondence: Judith D. Singer, Ph. D.
Assistant Professor
Harvard University
Graduate School of Education
Larsen Hall, Seventh Floor
Cambridge, MA 02138
(617) 495-1961

Running Head: Data Sets to Teach Statistics

Abstract

Statistics becomes interesting to non-methodologists only when taught in a research context that is relevant to them. Real data sets supplemented by sufficient background information provide just such a context. Despite this, many textbook authors and instructors of applied statistics rely on artificial data sets to illustrate statistical techniques. In this paper, we argue that artificial data sets should be eliminated from the curriculum and that they should be replaced with real data sets. Towards this end, we describe the rationale for using real data sets and describe the characteristics that we have found make data sets particularly good for instructional use. Having learned that real data sets can present problems for instructors, we discuss the difficulties that we have encountered when using real data and some of our strategies for compensating for these drawbacks. We conclude by presenting two authentic data sets and an annotated bibliography of dozens of primary and secondary data sources.
Opening up the Black Box of Recipe Statistics:  
Putting the Data Back into Data Analysis

Put yourself in your students shoes. It’s Friday afternoon at 2:00 pm. You’ve survived another week of classes. By next Monday, you have to:

- Observe the interactions between another mother/preschooler dyad and write a three page paper on their use of language;
- Read a chapter from Alan Bloom’s Closing of the American Mind and a chapter from E. D. Hirsch’s Cultural Literacy and be prepared to discuss them during Monday’s section; and,
- Do your statistics homework.

Hoping to complete the worst task first, you turn to your statistics homework and find the following problem.

Here are a set of X and a set of Y scores ...

<table>
<thead>
<tr>
<th>X</th>
<th>2</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>7</th>
<th>6</th>
<th>4</th>
<th>3</th>
<th>6</th>
<th>6</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>9</th>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Calculate:

(a) The means, sums of squares and cross products, standard deviations, and the correlation between X and Y.

(b) The regression of Y on X.

(c) Regression and residual sums of squares.

(d) The F ratio for the test of significance of the regression of Y on X, ...

Pedhazur (1982), p. 43

Would you want to do your statistics homework? Would you learn how regression models can help address interesting research questions? Would you be able to articulate what the F ratio really tells us? Would you learn how to use regression models to analyze data that might be of interest to
you? Would you remember any of this two years from now when you have to analyze your dissertation data?

Now suppose you found the following problem.

How attractive are America’s most prestigious colleges to the high school seniors applying for admission? Is it as difficult as they say to turn Harvard down? How about Princeton and Yale? Is there evidence that some students apply to certain schools just because they are likely to be admitted, when they really have little intention of enrolling?

Table 1 presents the 1986 admissions data for a random sample of 34 private colleges in the northeast. Two variables are given:

\[
\begin{align*}
\text{ACCEPT:} & \quad \text{percent of applicants accepted} \\
\text{YIELD:} & \quad \text{percent of accepted students actually enrolling}
\end{align*}
\]

You are going to examine the variable YIELD (Y), alone, and in relation to ACCEPT (X).

These data are real! Compared to them, Bloom and Hirsch seem abstract and theoretical. You might actually learn something interesting by doing this assignment. Which schools are hot? Which schools are safety schools? You might begin to understand the link between a research question and a statistical model. And you might even begin to think about how to use statistical models to examine the child language data that you have been collecting during the semester.

We believe that data sets of the first type do little to help our students become competent data analysts. Artificial data sets perpetuate the myth that statistics is dry and dull. After "analyzing" the data, students have not experienced the pleasure of doing research to investigate an interesting research question. Nor have they learned how statistical models can represent relationships between variables nor how statistical
models can be interpreted in a real-life context. Most artificial data sets should be eliminated from the applied statistics curriculum.

In their place, we propose that statistics instructors use data sets of the second type, which enable students to learn analytic skills in a realistic research context. Real data sets provide a practical arena for learning how to link research questions to statistical models. Using real data sets helps us show our students how statistical analyses can inform current debates in educational research, thereby teaching students not only how to analyze data, but also why we analyze data. The use of real data sets helps us integrate statistics into the general education curriculum.

The purpose of this paper is to convince textbook authors and instructors of applied statistics to eliminate artificial data sets from the curriculum and replace them with real data sets. Towards this end, we describe the rationale for using real data sets and describe the characteristics that we have found make data sets particularly good for instructional use. Having learned (the hard way) that real data sets can present problems for instructors, we discuss the difficulties that we have encountered when using real data and some of our strategies for compensating for these drawbacks. We conclude by describing two real data sets that we have used in the classroom as well as an annotated bibliography of dozens of primary and secondary data sources.

The Rationale for Using Real Data

As applied statistics instructors, our mission is to teach our students the skills we believe necessary for conducting statistical analyses of high
methodological quality. Although the techniques we cover during our four-semester sequence in quantitative methodology are diverse, our overarching goals are for students to be able to: (1) formulate interesting research questions; (2) select appropriate statistical techniques; (3) conduct all necessary calculations; (4) interpret the results of the analyses; (5) consider rival explanations of the results; and (6) summarize the findings in a cogent and convincing manner. The challenge for us is how best to achieve these goals.

Before the widespread availability of high-speed computing and pre-packaged statistics programs, the third goal—the computational aspects of data analysis—assumed priority over the other goals. After all, the success of any statistical analysis hinged upon the analyst's ability to perform the requisite calculations. Recognizing that the calculations could be time-consuming and tedious, many instructors and textbook authors tried to reduce student burden by using artificial data sets, constructed so as to simplify the arithmetic. For example, the observations in such data sets usually were integers, often chosen so that summary statistics, such as means, standard deviations and regression coefficients, also were integers. The American Statistician periodically published articles that described methods for constructing artificial data sets with specific characteristics (see, e.g., Edwards, 1959; Carmer & Cady, 1969; Dayton, 1972; Searle & Firey, 1980; Read & Riley, 1983; and Read, 1985) and artificial data sets were common fare in many popular applied statistics textbooks in education and the behavioral and social sciences (see, e.g., Hays, 1973, 1981; Winer 1962, 1973; and McCall, 1970, 1977).
Although the use of artificial data sets decreased the number of hours that students spent squaring and summing columns of numbers, it did not completely eliminate the drudgery of hand computation. The calculations were easier, but they still had to be performed. In the hope of keeping student attention focused on statistical concepts, not arithmetic details, many textbook authors provided step-by-step formulas ("recipes") designed to decrease the computational burden. By their very nature, these textbooks focused on confirmatory analyses, because only confirmatory analyses could be written as a sequence of specific steps, followed as one might follow a cookbook.

Unfortunately, although the rationale for using artificial data sets and cookbook strategies came from the desire to improve the quality of statistics instruction, the result usually fell far short of that goal. Artificial data sets perpetuated the myth that statistics is boring and unrelated to the students' substantive interests. Cookbook approaches to data analysis seduced students into believing that statistical analysis could be reduced to a set of predefined steps, conducted by a robot.

The widespread availability of high-speed computing has allowed statistics instructors to change the way in which they teach data analysis. Computers have eliminated the need for simplified arithmetic; the computer does not care if the observations are integers or if the summary statistics are integers. Most tedious calculations now can be relegated to a machine. No longer do students need to learn (let alone memorize) formulas whose sole purpose was to simplify computation. Exploratory and descriptive analyses, which previously were avoided, due in part to the time required to conduct
them, can now be easily incorporated into the data analyst’s tool kit. Just as computers have revolutionized the way in which we analyze data, so too should they revolutionize the way in which we teach how to analyze data.

One positive step in this direction is the increasing presence of computer output in statistics textbooks. In a review of 16 introductory statistics texts, Cobb (1987) noted that 8 included some computer output. But to our mind, the simple inclusion of computer output is not enough; it is now time for teachers of applied statistics to change the data sets used to illustrate statistical techniques.

What difference does the authenticity of a data set make? We have found that using real data sets has been a major factor in keeping our students—masters and doctoral candidates in education—motivated to learn statistical techniques. Although students report a diverse set of reasons for preferring authentic data sets, the major reason we hear is that the students find the real data we use to be intrinsically interesting—for example, most of them enjoy comparing the acceptance and yield rates at their school to the rates at other institutions. With real data, their efforts are rewarded not only with information on how to use statistics to conduct research, but also with information on an interesting research question. And our students report that some of the data sets themselves are memorable, thereby becoming mnemonics for recalling statistical techniques.

Over and above capturing the students’ interest, we find real data sets to be particularly helpful instructional aids. Real data sets allow a student to assume the role of researcher, exploring data in the hopes of addressing a specific set of research questions. Class examples and
Using Real Data to Teach Statistics

homework exercises become "trial" runs for data analysis problems that students encounter later in their own research. In essence, real data sets bring students as close as possible to an actual research experience.

But real data sets are helpful for another reason as well: They provide us an opportunity to teach students how to cope with many of the common problems that arise in real data, such as non-linearity, outliers, and missing values. These non-standard problems remind students of the need to investigate the tenability of assumptions; and the students respond by becoming interested in learning what they should do when the standard assumptions do not hold. Thus, the use of real data sets shows students that exploratory data analysis is an essential component of all statistical investigations.

Desired Characteristics of Real Data Sets

Not all real data sets are equally effective vehicles for teaching applied statistics. In this section, we discuss seven attributes that make a data set particularly well-suited for instructional use.

Authenticity

First, and foremost, a real data set must be authentic. The data given must be actual measurements taken on an actual sample of cases. Attaching life-like variable names to artificial data is not an acceptable substitute.

Consider the following exercise from Hays (1981):

An experimenter was interested in the possible linear relationship between the measure of finger dexterity \( X \), and another measure representing general muscular coordination \( Y \). A random sample of 25 persons showed the following scores: ... Compute the correlation coefficient, and test its significance. (p. 490).
Why should a student believe that these data are real? How were finger dexterity and general muscular coordination measured? From what population was a random sample chosen? Is the sample homogeneous with respect to age, a factor that might influence general muscular coordination and perhaps the relationship between coordination and finger dexterity? Is the experimenter only interested in a linear relationship?

The problem with "life-like" data is that most students can easily see through the artifice. As a result, most students would not bother asking the questions raised above, because why should they care about how the data were "collected." Yet these questions reflect the very issues we would like students to raise when reviewing other people's research and when conducting their own. Because students can see through the ruse of "life like" data, we should not demean them by attempting to fool them.

**Background information**

A real data set should be accompanied by background information on the purpose and design of the research, the source of the data, measurement techniques, variable definition and so on. It is the provision of this information that allows students to fully assume the role of researcher.

As Cobb (1987) wrote when assessing the data examples used in 16 introductory textbooks:

A data set is no longer alive if it is uprooted from its context like a pulled tooth. (What would you think of a dental school whose students only practiced drilling individual teeth that their instructor had already extracted?) To make a data set feel alive, the author must tell enough about what the numbers mean so that analysis is a search for meaning, not just an exercise in arithmetic. (pp. 331-332).
If the data come from a published paper or published tabulations, students should be given access to the original document. If the data are extracted from another source, the instructor must provide the background information. Many of the sources listed in the appendix are published papers; our students have found it interesting to read the original papers while conducting their analyses.

Interest and Relevance

Some of the best-selling statistics texts are filled with real data, but on topics of little interest to students in education and the social sciences. Snedecor and Cochran (1980) make ample use of real data, but on topics such as the calcium concentration in turnip greens (p. 239) and the average daily weight gain of swine (p. 303). Draper and Smith (1981) also use real data, but on topics such as the viscosity of filled and plasticized elastomer compounds (p. 228) and the effects of temperature on the growth rates of ice crystals (p. 66). Most of the classic statistics data sets, such as Fisher’s iris data (1936) and Brownlee’s stack loss data (1965) also fail to inspire our students.

Intrinsic interest is obviously in the eye of the beholder, but we can go a long way towards ensuring it by using data sets from our discipline. For example, the annual salary survey conducted by the American Association of University Professors (published annually in Academe), includes data of interest to most students: the average salaries of faculty members by institution and academic rank. The survey of school districts conducted by Education Resources Corporation is another useful source; it provides information on teacher's and administrator's salaries by district for a
nationwide stratified random sample of districts. (The full citations for these sources are given in the appendix.)

Topicality often provokes student interest. Our students became very engaged in a data set recently reported in Chance (1988) on the relationship between race of victim, race of defendant and whether the defendant was given the death penalty. Although these data had little to do with education, students perceived it as quite relevant, especially in the current climate of racial tension on our nation’s college campuses.

Controversy also has provoked student interest. Cyril Burt’s data on the IQs of identical twins is interesting (Jensen, 1974), especially when analyzed in the context of Burt’s views on the nature/nurture debate and Dorfman’s (1979) and Kamin’s (1976) evidence that Burt falsified data to support the nature argument. Powell and Steelman’s (1984) analysis of the relationship between state SAT scores and the percent of students taking the test also arouses interest, especially when accompanied by newspaper accounts of Secretary Bennett’s wall chart that ranks states according to these scores and critiques of all state comparisons of SAT scores given by Wainer, Holland, Swinton and Wang (1985), Rosenbaum and Rubin (1985), and Wainer (1986). By analyzing controversial data sets, students learn not just statistical techniques, but also how these techniques can support or undermine a hypothesis.

Historical data sets also have been an effective motivator for many of our students. The early volumes of journals such as Child Development, Journal of Educational Psychology, and Journal of Genetic Psychology, are filled with individual data. Although their topics are not always
fascinating (e.g., four different types of math tests), their age often overcomes this gap. Moreover, it is interesting to compare modern’s statistics more sophisticated analyses to the older tabular presentations given in the original sources.

Substantive learning

Empirical researchers analyze data because they want to learn something about the way the world works, not because they want to conduct statistical analyses for their own sake. When students learn something from their data analyses that they did not know before, they discover just how useful statistical analysis can be. The substantive learning does not have to be on a grand scale relating to fundamental theories of education, but it should be real.

One of our most popular data examples is not from the research literature but from a local magazine. Every few years, Boston magazine conducts a survey of school districts in the local area and publishes data for each district on per-pupil expenditures, teacher salaries, student demographics and so on. The Boston Globe publishes similar data sets on a regular basis. When students analyze these data, they discover how their home town compares to others in the area and how district characteristics are related to each other. They gain new insight into the on-going political debate as to why some school districts are reported to be "better" than others. Substantive learning reinforces the reasons for conducting statistical analyses in the first place.

Availability of multiple analyses

As practicing statisticians, we often use more than one type of
analysis to address a given research question. Different analyses provide different insights into the measures under study and when a data set is used in multiple analyses, the students learn that there may be more than one way to investigate a research question.

No experience reinforces the importance of multiple analyses as much as the discovery of previously unknown findings. For example, Scarcella (1984) published an analysis of the influence of language background and proficiency on choice of writing device (repetition, paraphrase, explanation). A re-analysis using log-linear modeling (given as a homework assignment to students) revealed previously unsuspected effects. In particular, it was possible to demonstrate that language proficiency, not language background, was a significant predictor of writing device.

The importance of raw data

We feel very strongly that the data must be given in raw form, not summarized using means and variance-covariance matrices. Rich information is lost when raw data are replaced by sufficient statistics and because students are one-step removed from the data they are seduced back into the cookbook approach fostered by hypothetical data. When raw data are available, students are free to adopt the data-analytic approach preferred by many practicing statisticians, be it the exploratory approach advocated by Tukey (1977) or the initial data examination approach advocated by Chatfield (1985). This allows students to look for high-leverage cases, heteroscedasticity, non-linearity and other non-standard problems that all too often arise in real data. The use of summary statistics may fool students into believing that such problems do not exist, or if they do, they
Case identifiers

Many published data sets have case identifiers which allow students to bring background knowledge to their data analyses. State, school district and school identifiers have meaning for our students. If such identifiers are available, they should be provided with the data set so that students can use their background information about the cases to inform their statistical analyses. Case identifiers are particularly helpful for identifying outliers and high-leverage observations. When students analyze data on the citation frequencies of prominent researchers, for example, their knowledge about the researchers being studied helps them understand why Sigmund Freud and Jean Piaget might be outliers (Gordon, Nucci, West et al., 1984).

Drawbacks to Using Real Data

Using real data sets to teach statistics is not without shortcomings. Below we describe four problems that we have encountered and offer some remedies for overcoming these problems.

The Workload of Finding Data Sets

A major motivation for using artificial data is that an instructor can readily create any number of data sets with specific characteristics. For example, Dayton (1960) presented a simple method for constructing a data set illustrating the effects of suppressor variables. Searle and Firey (1980) suggested that an instructor could reduce student plagiarism by generating dozens of data sets and giving each student a different data set to
"analyze." Producing a variable that is normally distributed, but with an outlier or two, is indeed a simple programming problem; identifying a real data set with the same features can take hours.

Having used real data sets for several years, we can clearly state that they increase the amount of time required to prepare classes, homeworks and exams. To identify a single data set to illustrate a specific technique we spend a great deal of time analyzing several data sets, some of which will not reveal interesting findings, others of which will present difficult analytic problems. This is especially true when developing materials for lower level courses, when students are still learning basic skills before learning to cope with non-standard problems.

These difficulties are, in fact, the major reason that we have written this paper. By including references to dozens of data sets that we have used, we hope that statistics instructors can gain access to a wide array of data sets that have many of the desired characteristics described in the previous section. Although it is still necessary to examine the data sets to determine which is most appropriate for introducing a specific concept, the availability of annotated bibliographies should facilitate the process.

**Small Data Sets and Statistical Power**

In our introductory and intermediate courses, we prefer small data sets, with sample sizes in the 35-75 range. Small data sets encourage the students to become intimately acquainted with each case, thereby fostering a more detailed understanding of the relationship between the data and the analyses. Once students have developed these skills, we introduce larger data sets in more advanced courses.
Unfortunately, small data sets create a false impression of how big effect sizes actually are in the real world. After all, null findings are not terribly interesting, so we tend to present data sets in which the effect sizes are large, yielding "statistically significant" results despite the small sample size. Although we, as instructors, know that these large effect sizes are not common in practice, the students do not see much evidence of this in their class problems or their homeworks. Thus, when we give them journal articles to read that report R²'s of 9%, many students conclude that the effect size is small, and it is, relative to their experience.

This problem is not unique to real data sets; most artificial data sets presented in applied statistics textbooks are also relatively small. The difference is that real data sets seem to reflect the larger class of statistical problems that arise in the real world. Because we see little means of eliminating this problem, we have chosen to specifically focus our students attention on it by discussing the concepts of statistical power, effect size, and the distinction between statistical significance and practical significance.

Aggregate data and self-selected samples

Easy access to information has often led us to use aggregate data or data on self-selected samples, such as mean SAT scores by state for the high school seniors who chose to take the test. In some of these data sets, the variables actually are measured at the aggregate level--for example, college tuition, student/faculty ratio, number of students enrolled--but many other data sets involve aggregate data, with all its attendant problems.
The question is whether the gains are worth the drawbacks, and in most instances, we believe they are. Aggregate data sets are some of the most readily available, intrinsically interesting data sets we use. The observations contained in aggregate data sets often have meaningful identifiers--names of towns, cities, counties, school districts, states or countries--thus enabling students to become more intimately associated with each individual data point. Nevertheless, in our more advanced classes we use these data sets to illustrate some of the problems involved in analyzing aggregate summaries or data on self-selected samples.

In-class testing

It is difficult, although not impossible, to test students in-class using real data. We do not use in-class exams, but rather multiple homework assignments and take-home exams. If you prefer to give an in-class exam, however, the solution is probably to hand out computer output and have the students interpret it. In doing so, though, note that the students are not choosing the analyses to be conducted--they are simply interpreting the output--and thus such an in-class exam may not be testing all of their analytic skills.

Two Examples

Perhaps the best way of discovering the advantages of authentic data sets is to try them in your classes. To assist in the search for real data, the appendix presents an annotated bibliography of primary and secondary sources. As illustrations of what you are likely to find in these sources, we present below two real data sets.
What does college tuition buy?

The cost of a college education has been rising rapidly during the past decade; at many schools in the northeast, it costs over $10,000 in tuition alone for a year at a private college. David W. Breneman, president of Kalamazoo College, has suggested that some colleges are simply raising their tuition to increase their prestige (President Says, 1988). With tuition at an all-time high, the question arises as to what tuition actually buy. Better trained faculty? Better student/faculty ratios? Better students?

Table 1 presents data on tuition and selected characteristics of the faculty and student body at 34 private colleges in the northeast, including:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUITION</td>
<td>Total tuition for 1986-87 academic year</td>
</tr>
<tr>
<td>NAPPLY</td>
<td>Number of freshman applicants in Fall 1985</td>
</tr>
<tr>
<td>PCTADMIT</td>
<td>Percent of freshman applicants admitted in Spring 1986</td>
</tr>
<tr>
<td>PCTYIELD</td>
<td>Percent of admitted applicants who matriculated in Fall 1986</td>
</tr>
<tr>
<td>PCTDOC</td>
<td>Percent of faculty holding a doctorate or the highest degree in their field</td>
</tr>
<tr>
<td>PCTFIFTH</td>
<td>Percent of matriculating freshmen in 1986 who were in the top fifth of their high school graduating class</td>
</tr>
<tr>
<td>PT_FAC</td>
<td>Number of part-time faculty members</td>
</tr>
<tr>
<td>FT_FAC</td>
<td>Number of full-time faculty members</td>
</tr>
<tr>
<td>SFRATIO</td>
<td>Student/faculty ratio</td>
</tr>
<tr>
<td>PROFSAL</td>
<td>Average salary of full professors</td>
</tr>
<tr>
<td>ASSTSAL</td>
<td>Average salary of assistant professors</td>
</tr>
</tbody>
</table>

---

insert Table 1 here

---
We have used this data set to illustrate regression model building. Additional variables such as the school's endowment, mean financial aid award, and mean SAT scores of entering freshman could be easily added, as could additional colleges. In particular, we strongly recommend adding the school at which you teach so that the students can compare their institution to other schools.

What are the ratings rating?

In 1982, the National Academy of Sciences published a report rating "the scholarly quality" of research programs in the humanities, physical sciences and social sciences. The ratings were based upon rankings of quality and reputation made by senior faculty in the field who taught at institutions other than the one being rated. The report stirred much controversy, as most published ratings do. Critics argued that peer rating scales were not necessarily indices of quality, but may instead reflect the institution's prestige, reputation, productivity, or perhaps size. Thus, the question arises as to what the ratings rate.

Table 2 presents the quality ratings of 46 research doctorate programs in psychology, as well as six potential correlates of the quality ratings. The variables are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUALITY</td>
<td>Mean rating of scholarly quality of program faculty</td>
</tr>
<tr>
<td>NFACULTY</td>
<td>Number of faculty members in program as of December 1980</td>
</tr>
<tr>
<td>NGRADS</td>
<td>Number of program graduates from 1975 through 1980</td>
</tr>
<tr>
<td>PCTSUPP</td>
<td>Percentage of program graduates from 1975-1979 that received fellowships or training grant support during their graduate education</td>
</tr>
<tr>
<td>PCTGRANT</td>
<td>Percent of faculty members holding research grants from the Alcohol, Drug Abuse and Mental Health Administration,</td>
</tr>
</tbody>
</table>
We have used this data set for teaching how to build regression models, but smaller portions of the data set could easily be used to illustrate other techniques. The data set could be modified by adding additional schools, additional predictors, or by choosing a sample from a different subject specialty.

Conclusion

Real data sets can be a statistics instructors strongest ally in motivating students to learn how to analyze data. Although the use of real data sets is not without problems, the strengths far outweigh the weaknesses. Moreover, the biggest drawback—the amount of time needed to identify data sets exhibiting specific statistical patterns and problems—can be overcome by communications in statistical journals that identify where such data sets can be found. The annotated bibliography in this paper is a first step. As more individuals identify data sources, we should be able to eliminate most artificial data sets from the applied statistics curriculum.
References


President says 100 private colleges follow crowd: the higher their prices, the more students apply. *The Chronicle of Higher Education*, 2 March 1988, p. A29.


Table 1. Characteristics of 34 private northeast colleges

<table>
<thead>
<tr>
<th>College</th>
<th>N</th>
<th>A</th>
<th>A</th>
<th>Y</th>
<th>F</th>
<th>T</th>
<th>T</th>
<th>S</th>
<th>P</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelphi</td>
<td>5990</td>
<td>2855</td>
<td>66</td>
<td>39</td>
<td>75</td>
<td>91</td>
<td>500</td>
<td>350</td>
<td>12</td>
<td>39</td>
</tr>
<tr>
<td>Albright</td>
<td>1000</td>
<td>1237</td>
<td>60</td>
<td>45</td>
<td>75</td>
<td>69</td>
<td>35</td>
<td>93</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>Babson</td>
<td>9152</td>
<td>1723</td>
<td>42</td>
<td>48</td>
<td>91</td>
<td>37</td>
<td>50</td>
<td>94</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>Bard</td>
<td>11450</td>
<td>1000</td>
<td>60</td>
<td>33</td>
<td>80</td>
<td>36</td>
<td>20</td>
<td>65</td>
<td>9</td>
<td>34</td>
</tr>
<tr>
<td>Bates</td>
<td>11835</td>
<td>2808</td>
<td>42</td>
<td>37</td>
<td>85</td>
<td>75</td>
<td>18</td>
<td>122</td>
<td>13</td>
<td>39</td>
</tr>
<tr>
<td>Bentley</td>
<td>7300</td>
<td>3945</td>
<td>54</td>
<td>41</td>
<td>74</td>
<td>58</td>
<td>150</td>
<td>197</td>
<td>20</td>
<td>41</td>
</tr>
<tr>
<td>Boston College</td>
<td>9120</td>
<td>16163</td>
<td>31</td>
<td>48</td>
<td>83</td>
<td>73</td>
<td>138</td>
<td>566</td>
<td>16</td>
<td>43</td>
</tr>
<tr>
<td>Boston Univ</td>
<td>10950</td>
<td>18476</td>
<td>60</td>
<td>33</td>
<td>94</td>
<td>70</td>
<td>920</td>
<td>1511</td>
<td>15</td>
<td>42</td>
</tr>
<tr>
<td>Bowdoin</td>
<td>10760</td>
<td>3555</td>
<td>23</td>
<td>50</td>
<td>78</td>
<td>90*</td>
<td>0</td>
<td>112</td>
<td>12</td>
<td>42</td>
</tr>
<tr>
<td>Brandeis</td>
<td>10950</td>
<td>3684</td>
<td>63</td>
<td>31</td>
<td>87</td>
<td>79</td>
<td>105</td>
<td>357</td>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>Brown</td>
<td>11690</td>
<td>13630</td>
<td>19</td>
<td>52</td>
<td>95</td>
<td>93</td>
<td>60</td>
<td>475</td>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td>Bucknell</td>
<td>10640</td>
<td>6279</td>
<td>43</td>
<td>33</td>
<td>90</td>
<td>74</td>
<td>20</td>
<td>213</td>
<td>14</td>
<td>41</td>
</tr>
<tr>
<td>Clark</td>
<td>10200</td>
<td>2969</td>
<td>58</td>
<td>37</td>
<td>95</td>
<td>65</td>
<td>50</td>
<td>155</td>
<td>14</td>
<td>39</td>
</tr>
<tr>
<td>Conn. College</td>
<td>11700</td>
<td>3456</td>
<td>46</td>
<td>28</td>
<td>75</td>
<td>56</td>
<td>94</td>
<td>136</td>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>Franklin/Marsh</td>
<td>10200</td>
<td>3350</td>
<td>62</td>
<td>26</td>
<td>92</td>
<td>68</td>
<td>25</td>
<td>134</td>
<td>14</td>
<td>41</td>
</tr>
<tr>
<td>Gettysburg</td>
<td>9425</td>
<td>2640</td>
<td>66</td>
<td>32</td>
<td>78</td>
<td>53</td>
<td>36</td>
<td>132</td>
<td>13</td>
<td>37</td>
</tr>
<tr>
<td>Harvard</td>
<td>11390</td>
<td>13614</td>
<td>16</td>
<td>73</td>
<td>99</td>
<td>98</td>
<td>198</td>
<td>674</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>Hofstra</td>
<td>6400</td>
<td>6229</td>
<td>68</td>
<td>39</td>
<td>85</td>
<td>55</td>
<td>325</td>
<td>400</td>
<td>18</td>
<td>39</td>
</tr>
<tr>
<td>Iona</td>
<td>5520</td>
<td>2940</td>
<td>74</td>
<td>39</td>
<td>56</td>
<td>13</td>
<td>130</td>
<td>247</td>
<td>22</td>
<td>48</td>
</tr>
<tr>
<td>Ithaca</td>
<td>7646</td>
<td>7306</td>
<td>65</td>
<td>31</td>
<td>73</td>
<td>14</td>
<td>104</td>
<td>362</td>
<td>15</td>
<td>33</td>
</tr>
<tr>
<td>Long Isl Univ</td>
<td>6350</td>
<td>4323</td>
<td>72</td>
<td>35</td>
<td>70</td>
<td>4</td>
<td>345</td>
<td>382</td>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>Marymount</td>
<td>6710</td>
<td>631</td>
<td>81</td>
<td>45</td>
<td>55</td>
<td>16</td>
<td>82</td>
<td>67</td>
<td>9</td>
<td>29</td>
</tr>
<tr>
<td>Middlebury</td>
<td>14500</td>
<td>3890</td>
<td>30</td>
<td>42</td>
<td>75</td>
<td>80</td>
<td>25</td>
<td>178</td>
<td>13</td>
<td>40</td>
</tr>
<tr>
<td>Monmouth</td>
<td>6662</td>
<td>2254</td>
<td>77</td>
<td>43</td>
<td>60</td>
<td>29</td>
<td>126</td>
<td>160</td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td>Mt St Vincent</td>
<td>6150</td>
<td>617</td>
<td>76</td>
<td>35</td>
<td>75</td>
<td>27</td>
<td>10</td>
<td>64</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>Northeastern</td>
<td>6753</td>
<td>20901</td>
<td>57</td>
<td>33</td>
<td>80</td>
<td>30*</td>
<td>2108</td>
<td>784</td>
<td>20</td>
<td>43</td>
</tr>
<tr>
<td>Ravier</td>
<td>5280</td>
<td>385</td>
<td>82</td>
<td>68</td>
<td>30</td>
<td>36</td>
<td>135</td>
<td>54</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>St Rose Coll</td>
<td>5570</td>
<td>672</td>
<td>73</td>
<td>47</td>
<td>60</td>
<td>45</td>
<td>100</td>
<td>108</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>Simmons</td>
<td>8992</td>
<td>1156</td>
<td>80</td>
<td>40</td>
<td>78</td>
<td>39</td>
<td>172</td>
<td>177</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>Swarthmore</td>
<td>11200</td>
<td>2590</td>
<td>32</td>
<td>41</td>
<td>90</td>
<td>92</td>
<td>15</td>
<td>135</td>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>Syracuse</td>
<td>7550</td>
<td>13654</td>
<td>69</td>
<td>32</td>
<td>71</td>
<td>43</td>
<td>405</td>
<td>898</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>Trinity</td>
<td>10355</td>
<td>3388</td>
<td>40</td>
<td>33</td>
<td>90</td>
<td>46</td>
<td>14</td>
<td>121</td>
<td>15</td>
<td>42</td>
</tr>
<tr>
<td>Villanova</td>
<td>6560</td>
<td>8488</td>
<td>50</td>
<td>39</td>
<td>73</td>
<td>61</td>
<td>306</td>
<td>518</td>
<td>14</td>
<td>40</td>
</tr>
<tr>
<td>Wesleyan</td>
<td>10860</td>
<td>4365</td>
<td>38</td>
<td>42</td>
<td>84</td>
<td>82</td>
<td>62</td>
<td>251</td>
<td>11</td>
<td>43</td>
</tr>
</tbody>
</table>

* estimated from available data

### Table 2. Ratings of 46 research doctorate programs in psychology

<table>
<thead>
<tr>
<th>Rank</th>
<th>Program</th>
<th>Q</th>
<th>F</th>
<th>P</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AuELPHI</td>
<td>12</td>
<td></td>
<td></td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>ARIZONA--TUCSON</td>
<td>23</td>
<td>29</td>
<td>72</td>
<td>67</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>BOSTON UNIV</td>
<td>29</td>
<td>38</td>
<td>111</td>
<td>66</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>BROWN</td>
<td>36</td>
<td>16</td>
<td>28</td>
<td>52</td>
<td>63</td>
</tr>
<tr>
<td>5</td>
<td>U C BERKELEY</td>
<td>44</td>
<td>40</td>
<td>104</td>
<td>64</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td>U C RIVERSIDE</td>
<td>21</td>
<td>14</td>
<td>28</td>
<td>59</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>CARNEGIE MELLON</td>
<td>40</td>
<td>44</td>
<td>16</td>
<td>81</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>UNIV OF CHICAGO</td>
<td>42</td>
<td>60</td>
<td>57</td>
<td>65</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>CLARK UNIV</td>
<td>24</td>
<td>16</td>
<td>18</td>
<td>87</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>COLUMBIA TEACHERS</td>
<td>30</td>
<td>37</td>
<td>41</td>
<td>43</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>DELAWARE, UNIV OF</td>
<td>20</td>
<td>20</td>
<td>45</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>DETROIT, UNIV OF</td>
<td>8</td>
<td>11</td>
<td>27</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>FLORIDA ST--TALAH</td>
<td>28</td>
<td>29</td>
<td>112</td>
<td>64</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>FULLER THEOL SEMIN</td>
<td>14</td>
<td>14</td>
<td>57</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>UNIV OF GEORGIA</td>
<td>27</td>
<td>38</td>
<td>167</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>16</td>
<td>HARVARD</td>
<td>46</td>
<td>27</td>
<td>113</td>
<td>62</td>
<td>52</td>
</tr>
<tr>
<td>17</td>
<td>HOUSTON, UNIV OF</td>
<td>29</td>
<td>32</td>
<td>122</td>
<td>51</td>
<td>19</td>
</tr>
<tr>
<td>18</td>
<td>UNIV ILLINOIS-CHAMP</td>
<td>42</td>
<td>56</td>
<td>116</td>
<td>56</td>
<td>32</td>
</tr>
<tr>
<td>19</td>
<td>IOWA, UNIV OF</td>
<td>33</td>
<td>32</td>
<td>54</td>
<td>49</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>KANSAS, UNIV OF</td>
<td>31</td>
<td>42</td>
<td>79</td>
<td>41</td>
<td>14</td>
</tr>
<tr>
<td>21</td>
<td>KENT STATE UNIV</td>
<td>23</td>
<td>30</td>
<td>76</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>22</td>
<td>LOUISIANA STATE</td>
<td>18</td>
<td>18</td>
<td>62</td>
<td>39</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>UNIV OF MARYLAND</td>
<td>29</td>
<td>41</td>
<td>98</td>
<td>41</td>
<td>12</td>
</tr>
<tr>
<td>24</td>
<td>MIAMI UNIV</td>
<td>21</td>
<td>23</td>
<td>52</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>U MICH--ANN ARB</td>
<td>45</td>
<td>111</td>
<td>222</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>26</td>
<td>U MISSOURI</td>
<td>25</td>
<td>26</td>
<td>63</td>
<td>39</td>
<td>23</td>
</tr>
<tr>
<td>27</td>
<td>U NEW HAMPSHIRE</td>
<td>18</td>
<td>16</td>
<td>24</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>28</td>
<td>NEW YORK UNIV</td>
<td>33</td>
<td>38</td>
<td>154</td>
<td>55</td>
<td>34</td>
</tr>
<tr>
<td>29</td>
<td>U NC -- GREENSBORO</td>
<td>21</td>
<td>19</td>
<td>40</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>NORTHEASTERN</td>
<td>24</td>
<td>16</td>
<td>18</td>
<td>25</td>
<td>63</td>
</tr>
<tr>
<td>31</td>
<td>NOTRE DAME</td>
<td>15</td>
<td>13</td>
<td>29</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>32</td>
<td>OKLA ST--STILLWATER</td>
<td>15</td>
<td>23</td>
<td>41</td>
<td>51</td>
<td>4</td>
</tr>
<tr>
<td>33</td>
<td>PENN STATE</td>
<td>36</td>
<td>32</td>
<td>69</td>
<td>65</td>
<td>16</td>
</tr>
<tr>
<td>34</td>
<td>PRINCETON</td>
<td>38</td>
<td>21</td>
<td>38</td>
<td>28</td>
<td>48</td>
</tr>
<tr>
<td>35</td>
<td>UNIV OF ROCHESTER</td>
<td>32</td>
<td>28</td>
<td>90</td>
<td>70</td>
<td>36</td>
</tr>
<tr>
<td>36</td>
<td>SUNY ALBANY</td>
<td>27</td>
<td>22</td>
<td>52</td>
<td>10</td>
<td>27</td>
</tr>
</tbody>
</table>

continued
Table 2. continued

|    | N  | P  | C  | A  | U  | A  | N  | C  | T  | R  | P  | C  | A  | N  | P  | C  | A  | U  | A  | N  | C  | T  | R  | P  | C  | A  |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 37 | ST LOUIS UNIVERSITY | 16 | 20 | 80 | 46 | 10 | 19 | 40 |
| 38 | UNIV SOUTH FLORIDA  | 26 | 32 | 41 | 13 |  6 | 64 | 56 |
| 39 | STANFORD           | 48 | 26 | 81 | 70 | 58 | 155| 100|
| 40 | TEMPLE             | 26 | 40 | 81 | 42 | 10 | 70 | 68 |
| 41 | TEXAS TECH LUBBOCK| 14 | 19 | 87 | 15 |  5 | 72 | 79 |
| 42 | UNIV OF TOLEDO     | 12 | 17 | 26 |  9 |  6 | 15 | 59 |
| 43 | UNIV OF UTAH, SALT L| 29 | 29 | 71 | 74 | 17 | 85 | 76 |
| 44 | VIRGINIA POLYTECH  | 34 | 27 | 20 |  0 | 29 | 79 | 57 |
| 45 | WASHINGTON UNIV-ST.L| 28 | 26 | 70 | 68 | 27 | 84 | 73 |
| 46 | UNIV WISC--MADISON | 39 | 36 | 59 | 57 | 67 | 172| 83 |

APPENDIX

Annotated Bibliography of Published Data Sets


Several extensive data sets describing the blood chemistry (cholesterol, blood pressure, etc.), cardiovascular state, socioeconomic status, and year of death. Some censored cases, could be used in the teaching of survival analysis. Other datasets include body flexibility, diet, testosterone levels in right and left testes of mice (!), weaning of rats. Some educational data sets on infant cognitive development.


Depression scores and selected covariates for 294 participants in the Los Angeles Depression Study. Data set includes individual item responses for a 20 question depression scale, person background characteristics and selected health variables.


A large variety of categorical data sets including: tenure in American universities, dolphin sightings, transitions between Piagetian stages, college expectations and participation in high school athletics, political preferences, religion and marijuana, sudden infant death.


Data concerning the effect of the "Personalized System of Instruction" on course grades in an intermediate macroeconomics course, useful for logit analysis and log-linear modeling.


Average brain weights and body weights for 62 species of mammals. Both variables are very skewed, but logarithmic transformations alleviate the skewness and improve the linearity of the scatterplot.

Raw data for 71 data sets. Many substantive areas are included, but the emphasis is generally on the physical and natural sciences. Several interesting social science examples are given, including: unemployment statistics, insurance rate information, literary data sets (The Federalist papers data set and another on Platonic prose rhythm) and the birthday/deathday problem.


Measures of the moral integration, ethnic heterogeneity, crime, welfare effort, integration and mobility of residents in 43 American cities.


Salary data by rank, sex, and tenure status for faculty at 1,901 colleges and universities. Institutions are categorized according to Carnegie classifications.


Contingency table of the relationship between gestational age and neurological status for 505 babies. Also see detailed log-linear analysis of these data in Green, J. A. (1988). Loglinear analysis of cross-classified ordinal data: Applications in developmental research, Child Development, 59, 1-25.


One of many sources describing the more than 1,500 four-year colleges in this country. Relevant data include: number of applicants, number of students accepted, number of students enrolling, mean SAT scores of incoming freshman, mean class rank of incoming freshmen, faculty/student ratios, financial aid available, number of part-time students and faculty, percent of faculty with doctorates, sex composition of student body. Can be supplemented with information from American Association of University Professors salary survey and endowment data given in the Digest of Education Statistics.


Individual data for 25 college sophomores at the University of Texas
on the speed and accuracy with which they solved four types of arithmetic problems (addition, subtraction, multiplication and division).


Scores on nine tests for 37 of the "best" students and 37 of the "worst" students, with notations of class rank, designed "to be of assistance to college authorities in aiding freshmen to adjust themselves to their environment" (p. 381).


A large variety of non-educational data sets (lawn service, real estate market, professional sports, foreign food), with some educational data sets scattered here and there: categorical data on health issues in children by graduating class of pediatrician, starting salaries of MBA graduates, etc.


A variety of educational growth data sets suitable for repeated measures/MANOVA analysis, including data on responses to inkblot plates by grade and IQ over time, longitudinal (4 grades) data on scaled vocabulary scores for boys and girls, and so on. (Data repeated in Finn, J. D., & Mattsson, I. (1978). *Multivariate Analysis in Educational Research*. (Chicago, IL: National Educational Resources).


Perceptions of course performance among high ability men and women in physics and chemistry courses at Stanford.


Raw data for 46 children divided into four groups--stutterers, well-adjusted, medium adjusted and poorly adjusted. Measures include age, achievement, receptivity to education, physical condition, social-personality traits, insightfulness, family background, somotype and anthropometrics.


Raw data for 33 data sets. Many substantive areas are included, and many of these data sets are just plain interesting, such as the ages of signers of the Declaration of Independence, murder/suicides by crashing private airplanes, heights of singers in the New York Choral Society.


Six to ten longitudinal data sets (10 waves) on measures of computation, color-naming and opposites-naming for 22 college-age males in New York at the turn of the century. Suitable for growth curve analysis.


Data for on the mother, father and children in eight families, with the age of each family member and their scores on five tests (addition, subtraction, multiplication, division and copying figures). The author concludes that "it is difficult to avoid the conclusion that ... likeness is due to heredity" (p.16).


Two large data sets: (1) a 20 variable subset of the PROJECT TALENT data (234 males, 271 females); and (2) the RECTANGLES data set on the physical dimensions of 100 rectangles, useful for factor analysis and principal components analysis.


Raw data for 39 data sets. Relevant examples include: educational plans of Wisconsin school boys, statistical aspects of literary style, satisfaction with housing conditions.


Educational and demographic descriptors for 32 large urban school districts, including data on how these characteristics have changed over time.

The data sets tend to be small, and many are from the sciences, but there are dozens of them. One interesting example is the movie production and promotion costs for "Dumb Movies," such as Revenge of the Nerds and Police Academy.


Some educational data sets are submerged among the many others, including: sex differentials in teacher pay, aptitude and age of first word, nutrition of preschoolers, ailments of university alumni.


Data on the eating habits of 37 children, including age, sex, means (and standard deviation) for total amount of calories eaten and minutes spent at table.


Data for 300 participants in the National Longitudinal Study on reading, math, gender, race, college status, SES, High school program, High school grades, creativity, stress avoidance, etc.


Raw data for 1,031 school districts on enrollment, per pupil expenditures, and salaries for superintendents, central office administrators, principals, teachers, staff and support personnel.


An introductory textbook that melds together Tukey’s exploratory data analysis and the more traditional confirmatory approaches. Many interesting data sets, including frequency of teacher criticism by student IQ, sex differences in reactions to hostile treatment by an experimenter, experimenter artifacts in social psychology research, characteristics of social networks.


Age, IQ scores and activity ratings for 16 boys and 16 girls. Two
activity ratings are available for each child.


Raw data for four studies of relevance to education: Creativity and achievement, memory for words, essay grading practices and effects of programmed instruction.


Several interesting data sets including: relationship between status, authoritarianism and conformity, methods to enhance recall of words, causes of the 1907 Romanian Peasant rebellion.


Aggregate cigarette smoking and cancer death rates, by type of cancer and state.


Sociodemographic characteristics for 353 Massachusetts towns and cities, including data on age, race, sex, income, labor force participation, voter registration, police, fire, crime, taxation, libraries and schools. The company publishes similar books for other states; write to them at Box 356, Wellesley Hills, MA 02181.


State level data on percentage of children in each category of special education and sociodemographic composition of the states. Washington, DC is a high-leverage outlier for the relationship between percent of students classified as educably mentally retarded and percent of population that is black.


Raw data for 61 maladjusted children on two IQ tests, as well as information on their sex, age, parentage (both Foreign, both American, Mixed) and maladjustment type (aggressive or asocial). Authors explore relationship between IQ and all these predictors.

Several multivariate data sets from a variety of disciplines including engineering, manufacturing, biology and mining. The volume includes several well-known data sets such as Fisher’s iris data (1936) and Rothkopf’s Morse-code confusion data (1957) which have utility for the teaching of principal components analysis, factor analysis, multidimensional scaling and cluster analysis.


Citation frequencies and dates of birth for 187 prominent educational researchers. Four sources of citations are given: AERJ & JEP, Review of Research in Education, selected Educational Psychology Texts, and Social Science Citation Index.


Several categorical data sets of wide interest: Suicides by day of the week, homicides by month, stressful events, etc.


Individual data from an experiment on the effects of time lapsed between pre-tests and post-tests for 167 students in grades 4, 5, 6, and 7.


Raw data for 8 studies in psychology and psychiatry, on topics as diverse as headaches, smoking and Alzheimer’s disease.


Many unique and interesting data sets on the link between physical and psychological characteristics. Do blonds have more fun? Do lunatics eyebrows join together in the middle? Are manic depressives thin or fat?

Harris, S., & Harris, L. B. (1986). *The Teacher’s Almanac* (New York: Facts
Assorted education data by state and school district, including teacher salaries, high school graduation rates, functional illiteracy rates, and presence of computers in schools.


Ages and IQ scores of siblings in 44 families.


IQ scores for a six year period for 343 children.


A host of data sets of different sizes on many different topics including: blood pressure and obesity of Mexican-Americans, baseball data, 1970/1 draft lottery, promotion rates among male and female pharmacists, leisure time companions of black women, ranking of rum brands by different nationalities, preference for Charlie's Angel's actors, longevity and environment, color of canned tuna, etc.


Productivity, reputation and size of psychology departments at 75 universities.


Body length of crickets as a function of geographical location and weather throughout the USA.


Twenty-five children were asked to build towers on each of two occasions. Each time they were given: (a) a set of cubes; and (b) a set of cylinders. Raw data are given on the number of blocks of each type used each time, and how many minutes it took to construct the tower.


"Quality" rankings and characteristics of university departments in the social sciences, by discipline. Data include number of faculty, number of students, productivity of faculty, number of grants awarded, follow-up placement of doctoral students.


Data for 38 infants on their crying activity in early infancy and later measures of IQ.


Popularity scores for 17 children: percent each child was named first, percent each child name last, effects of ordering and the effects of sex.

Three measures of reading instruction for a sample of 53 learning disabled students, by curricular approach and school.


Life expectancy and per capita income for 105 nations divided into five national wealth classifications (industrialized, petroleum exporting, higher, middle and lower).


Individual data for 80 children on the sizes of 10 bones, measured by both x-rays and anthropometry at each of 3, 4, or 5 occasions, by sex. The authors construct lots of individual growth curves.


Lung cancer mortality by degree of urbanization and gender, in Louisiana.


Gesell adaptive scores and age at first word (in months) for 21 children with cyanotic heart disease. The data set contains some interesting outliers and high leverage cases.


Raw data for 13 data sets across several disciplines. Relevant examples include a subset of 20 from the Coleman Report, educational expenditures for Massachusetts school districts, municipal bond data for 20 US cities.


Annual reports issued by the Department of Education providing descriptive information on education, often over time, sometimes by state, occasionally by school district. The data on university endowments can be used in conjunction with other university level
data, such as that given in Barron's (1987).


Data for 61 school districts on educational policies and practices, as well as selected education and economic descriptors.


Selected results from a study of opening lines used in singles bars in the St. Louis area. Two-way contingency table describing the relationship between type of opening line (compliments, propositions, etc.) and time of evening.


David Phillips has made a cottage industry of looking at what many might term coincidences--birthdays and deathdays and cc ycat suicides after popularized accounts in the media. These are but a handful of articles, each listing the detailed raw data on deaths following these events that led him to his conclusions.


A large variety of categorical data sets including: fingerprints, family size, work conditions and work quality, behavioral problems and birth order, high school rank by gender and socioeconomic status.


Mean SAT scores and percent of high school seniors taking the SAT, by state for 1982. For additional data, and a critique of their analyses, see: Wainer, H., Holland, P. W., Swinton, S., & Wang, M. H. (1985). On "State Education Statistics". *Journal of Educational Statistics, 10*, 293-325. Also see: Rosenbaum, P. R. & Rubin, D. B.
Using Real Data to Teach Statistics


Number of meals served, breads baked and ale brewed at the de Bryene household from October 1412-September 1413, by month. That's right, the fifteenth century. These data have nothing to do with education, but their age makes them intrinsically interesting.


Thirty data sets of small to moderate sizes, on topics ranging from education to cartoons. The educational data sets include information on school strikes and freshman SAT verbal and math scores.


Categorical data on the language background and language proficiency of native and non-native speakers and how this influences their choice of writing device.


Rankings for 17 first ladies from Florence Harding through Nancy Reagan on 10 dimensions ranging from integrity, leadership and accomplishments.


Raw data for 100 children who were adopted at birth. Measures include: natural mother's IQ and education level, foster mother's IQ and education level, foster father's occupation and child's IQ on each of 5 occasions, from infancy through pre-adolescence.


Approximately 20 interesting small to moderately sized educational (and other) data sets, including: pre/post data on the influence of
Sesame Street, risk of reading problems among kindergartners, behavior reversal, programmed music instruction of elementary school children, IQ testing, etc.


Data for 28 boys given including: 2 IQ scores (Terman and Stanford-Binet), 2 socioeconomic status measures (parents education and father's occupation), somotypes (endomorph, mesomorph, ectomorph), and drawing type.

Supreme court ruling on death penalty. *Chance, 1*, 7-8.

Three-way contingency table on the relationship between race of victim, race of defendant and use of the death penalty, showing that the death penalty is not uniformly applied.


Raw data for a handful of data sets gathered in educational settings, including: effects of delay in oral practice on second language learning (pp. 228-229), relationship between recall and sentence structure (p. 233), predictors of student performance on the Peabody Picture Vocabulary Test (p. 281).


Sociodemographic, education, health and economic indicators for 130 countries.


The authors analyze data for the 50 states on the relationship between failure on the selective service exam administered during 1969-1970 and contextual and education descriptors of the states. The selection bias inherent in analyses of state level SAT scores are also present here, but it does make an interesting example.


Raw data for several interesting data sets including Cyril Burt’s IQ data, Allison and Cicchetti’s brain weight and body weight data, three time points for 26 boys and 32 girls who participated in the Berkeley Guidance study (anthropometric information only, however.)


Many characteristics in dozens of societies around the world, including age at weaning, toilet training, fear of ghosts, rituals, etc.


Age of person at death (in years) and the length of the person’s lifeline (in centimeters) for 50 individuals. Not surprisingly, the test of H₀: r=0 cannot be rejected.


Individual data for 100 students on these two IQ tests, with information on student sex, age and native language.