SRRI # 303

DESIGNING DATA ANALYSIS TOOLS FOR STUDENTS

Clifford Konold

Scientific Reasoning Research Institute
Hasbrouck Laboratory
University of Massachusetts
Amherst, MA 01003

and

TERC
2067 Massachusetts Avenue
Cambridge, MA 02140

Draft December 7, 1994

To appear in: *Computer und Unterricht*

I thank Amy Robinson for her comments on an earlier draft and Rolf Biehler for translating the manuscript into German. The software described in this article was developed under a grant from the National Science Foundation (#MDR-8954626) and is being published by Intellimation of Santa Barbara, CA. The opinions expressed here are my own and not necessarily those of the Foundation.

Designing Data Analysis Tools For Students

Few would question the assertion that the computer has changed the practice of statistics. Fewer still would argue with the claim that the computer, so far, has had little influence on the practice of teaching statistics; in fact, many still claim that the computer should play no significant role in introductory statistics courses. In this article, I describe principles that influenced the design of data analysis software we recently developed specifically for teaching students with little or no prior data analysis experience. I focus on software capabilities that should provide compelling reasons to abandon the argument still heard that introductions to statistics are only made more difficult by simultaneously introducing students to the computer and a new software tool. The micro-computer opens up to beginning students means of viewing and querying data that have the potential to get them quickly beyond technical aspects of data analysis to the meat of the activity: asking interesting questions and constructing plausible answers. Although the primary focus of the paper is on how software can facilitate this process, I also attempt to portray various aspects of the iterative nature of data analysis to hint at the difficulties students have maintaining focus and interest in what can be a sea of information.

The software I describe, DataScope®, was designed as part of a 4-year project funded by the National Science Foundation to create materials and software for teaching data analysis at the high-school and introductory college level. We designed DataScope to fill a gap we perceived between professional analysis tools (e.g., StatView and Data Desk), which were too complex for use in introductory courses, and existing educational software (e.g., Statistics Workshop, Data Insights) which were not quite powerful enough to support the kind of analyses we had in mind. Certainly, DataScope is not the educational tool we might dream of having (see Biehler, 1994), but it does give students easy access to considerable analysis power.

Data Analysis Objectives and Software Design

There are a variety of views about what we ought to be teaching in the name of statistics or data analysis (see Gordon & Gordon, 1992). But when it comes time to design or choose software for classroom use, one had better have in mind a clear view of what one wants to accomplish. The overall objective guiding the design of DataScope was to help students quickly to the point where they could use the software

to pursue a series of related questions, possibly involving many variables, with the goal of creating a coherent story. Data analysis is often portrayed as an interactive process in which an individual moves from a question, to collecting and looking at pertinent data, to reformulating and refining the question, to looking at more data, and so on. While this activity depends in part on peoples' perceptive abilities, it is driven by curiosity and by the desire to construct plausible stories which explain or tie together various observations.

Educational software can facilitate this objective by helping students see the "texture" of data, by taking advantage of their visual capabilities to detect trends, patterns, and differences, and by being easy to use so that time otherwise spent learning the software can be devoted to analyzing and thinking about data.

Simplicity

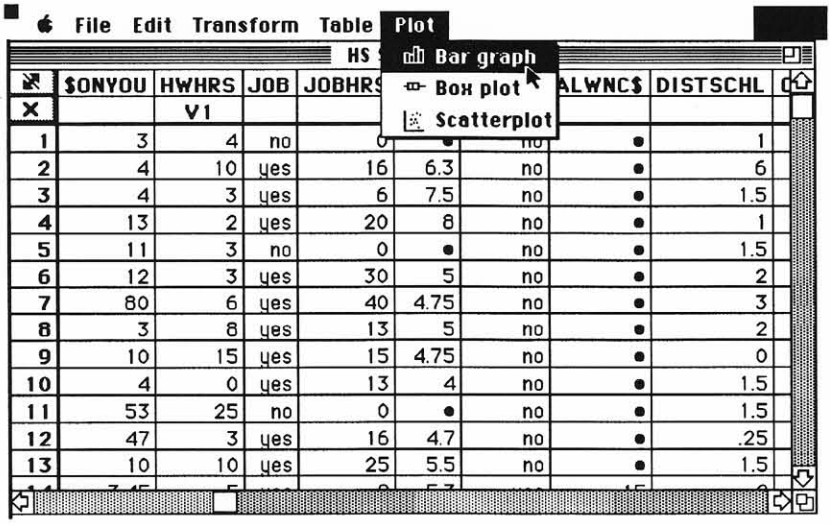
One way we kept DataScope simple was by limiting the number of displays. We had two reasons to keep this number small. Most obviously, the fewer the number of displays and the less complicated the software, the easier to learn. However, one can imagine a tool packed with displays and options but designed so as not to overwhelm the beginner, with the tool revealing its power as the needs and experience of the user develop (see Biehler, 1994). The more important reason to limit the number of displays in an introductory course is that it takes time to acquire expertise "reading" a particular display. For many, interpreting a histogram is now second nature. An expert can look at a histogram, quickly noticing both typical and atypical features, and use this information to guide further exploration. But for many new students, a histogram is still a blur of information. Novices don't know what it is they should look at, cannot perceive the unusual. Limiting the number of plot types in introductory courses allows students to spend sufficient time with each plot to develop the experience that in the end makes a tool an unconscious extension of our ordinary perceptual system (cf. Polanyi, 1969).

Plots That Tell

Strunk and White (1972, p. x) describe a good writer as one who makes "every word tell." Cleveland (1993, p. 1) makes a similar point about tools for visualizing data: "Our tendency is to be misled into thinking we are absorbing relevant information

when we see a lot. But the success of a visualization tool should be based solely on the amount we learn about the phenomenon under study." Since we must limit the number of plots in an educational tool, it is especially important that the plots we include be those that "tell." Such plots are uncluttered by irrelevant information -- what Tufte (1983) calls "chartjunk" -- and are capable of revealing critical features of and relations among data. For displaying non-numeric data, we chose frequency tables and bar graphs; for numeric data, histograms, box plots, and scatterplots. We decided on these by considering not only their power but also their frequency of use. This produced some conflict. Even though they are perhaps the most commonly used plot in the popular press, we didn't include pie charts; they are not well suited either to revealing texture or to making comparisons. To paraphrase Tufte (1983, p. 178), the only thing worse than a pie chart is several of them. And though virtually never seen outside the field of data analysis, we included box plots because they clearly depict features of center and spread and are especially powerful for comparing batches of data.

To demonstrate how students use the software, I will begin with data obtained from a questionnaire administered in 1990 to 84 high-school students in two towns in western Massachusetts: Amherst, a small college town, and Holyoke, a larger, industrial city. This data set includes information on each of the 84 students, including information about their gender, age, birth order, family size, marital status of parents, religious activity, rating of school performance, educational level of parents, curfew times, and a number of other datum, some of which are described below. In DataScope, data are stored in a table with rows holding information about individual cases (students, in this example), and columns holding information on particular variables. In the data table shown in Figure 1, the visible variables include: $ONYOU (dollars the student was carrying when surveyed, HWHRS (hours per week spent doing school work at home), JOB (whether or not the student had a part-time job), JOBHRS (hours worked per week), JOB$ (wages per hour), ALWNC (whether or not the student receives an allowance), ALWNC$ (allowance dollars per week), DISTSCHL (miles from home to school).



	$ONYOU	HWHRS	JOB	JOBHRS			ALWNC$	DISTSCHL
		V1						
1	3	4	no	0	•	no	•	1
2	4	10	yes	16	6.3	no	•	6
3	4	3	yes	6	7.5	no	•	1.5
4	13	2	yes	20	8	no	•	1
5	11	3	no	0	•	no	•	1.5
6	12	3	yes	30	5	no	•	2
7	80	6	yes	40	4.75	no	•	3
8	3	8	yes	13	5	no	•	2
9	10	15	yes	15	4.75	no	•	0
10	4	0	yes	13	4	no	•	1.5
11	53	25	no	0	•	no	•	1.5
12	47	3	yes	16	4.7	no	•	.25
13	10	10	yes	25	5.5	no	•	1.5

Figure 1. Data table showing a subset of information collected from 84 students.

This is one of the data sets we have used in a year-long course at Holyoke High. Specific questions students can pursue using these data include:

- Are girls more likely to have curfews than are boys?
- Are a person's views on abortion related to religious belief?
- Does a person's birth order in the family predict leadership ability?
- Do children of single parents get lower grades than children from two-parent families?
- Is a student's gender related to hourly wage?
- Does holding a part-time job adversely affect school performance?

This last question is one which concerns many parents who wonder whether their children's academic performance will suffer if they try to work a part-time job. While this question might seem to entail looking at only a few variables, other variables in the data set are quickly drawn into the investigation. This is one of the advantages of using data sets with many variables: they invite students to make and test guesses about possible explanations for trends they observe in the data.

Histograms Reveal Texture

Let's begin an investigation of this question by looking at the histogram showing hours per week students reported doing homework. A histogram is gotten simply by designating a column, in this case HWHRS, as a variable and selecting the command Bar Graph from the menu, as is shown in Figure 1. The histogram (Figure 2) appears without the user having first to specify various parameters, such as interval width, as we believe students should not be asked to decide on issues whose implications they do not already know. The program makes guesses about how the data are best displayed. Once displayed, the student is in a good position to decide whether and how the plot should be modified to reveal more about the data.

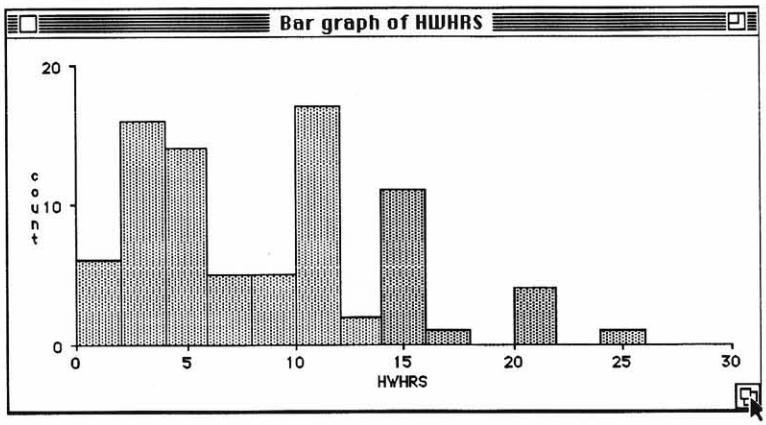


Figure 2. Histogram of hours per week doing homework.

A histogram can be modified in two general ways: by resizing it using the resize box at the lower right (see arrow in Figure 2) or by changing interval widths. The resize option rescales the plot, based on the rectangular shape to which the window is sized. This allows the user to enhance various features of the histogram. Shortening the length of the histogram, as shown in Figure 3, enhances the spiky appearance of the bars.

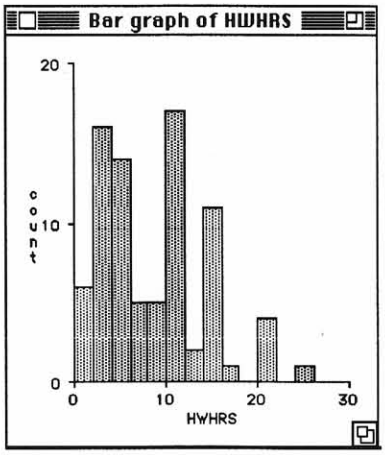


Figure 3. Histogram of homework hours,
resized by shortening the horizontal axis.

One question that arises by looking at this histogram is how to explain the spikes, particularly at 10, 15, and 20 hours. After a little thought it seems reasonably certain that these don't reflect anything about actual study hours, but rather reveals the tendency to round off estimates to some multiple of 5. It is interesting, in this regard, to look at the distributions of HWHRS and JOBHRS together on the same axis (Figure 4). Notice that the spikes occur in similar places in the distribution of JOBHRS. However, this is probably not due in the case of JOBHRS to rounding off, but to the way hours are assigned in the workplace (e.g., a 20 hour per week job).

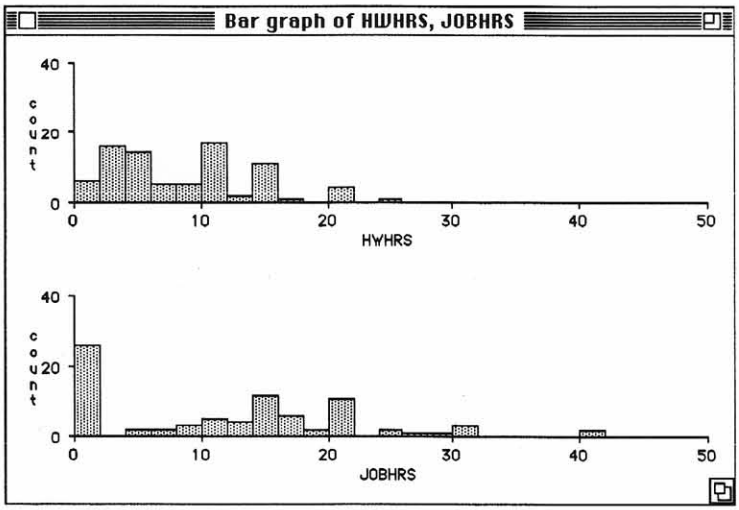


Figure 4. Histograms of homework and job hours displayed
for comparison in the same window over similar axes.

Changing the interval widths allows the user to see how the shape of the distribution appears under finer- or coarser-grained groupings. The spikes in the data of HWHRS disappear in Figure 5 in which the interval width has been changed from 2 to 5, revealing the more general trend in the distribution. As we move from no study time to longer study times, we find fewer and fewer students represented.

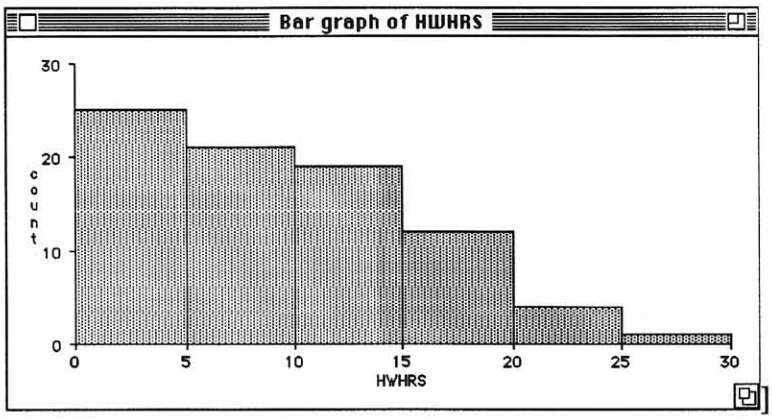


Figure 5. Histogram of homework hours with interval width of 5.

Changing the interval to 1 reveals even more detail (Figure 6), suggesting that there may be two distributions here. Perhaps there is one distribution for Holyoke students which peaks around 5 and then tails off, and another for Amherst students, peaking at 10 and then decreasing. If true, this would fit locally-held stereotypes about students at the two schools.

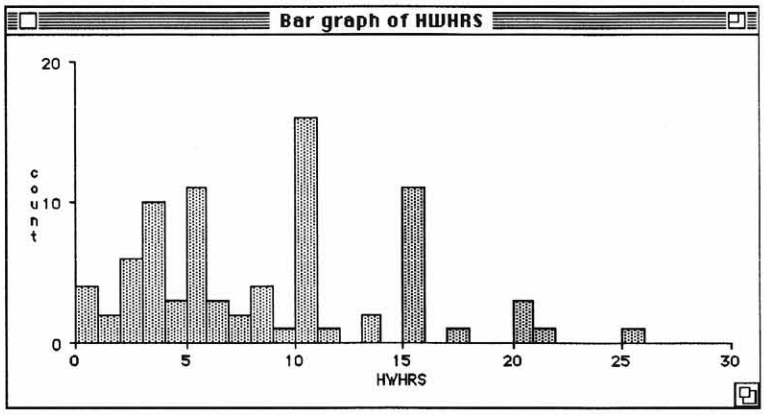


Figure 6. Histogram of home work hours with interval width of 1.

This is the kind of conjecture that we hope students make as they inspect data displays. It is one of the reasons that students should analyze data they already know something about. Background knowledge provides the basis for asking interesting questions and interpreting findings. And the software should make it easy to explore such hypotheses. In DataScope this is done using a generalized "grouping" capability. Students can form subgroups of one variable for each level of another variable. In this way, they can explore relationships among variables by comparing the distribution of one variable across the various levels of another variable. To produce the histograms in Figure 7, I left HWHRS as the main variable and selected SCHOOL as a grouping variable. This produced two histograms of HWHRS, one for Holyoke students and one for Amherst students. (Note that the distributions are labeled with the levels of the grouping variable rather than with the main variable, HWHRS.)

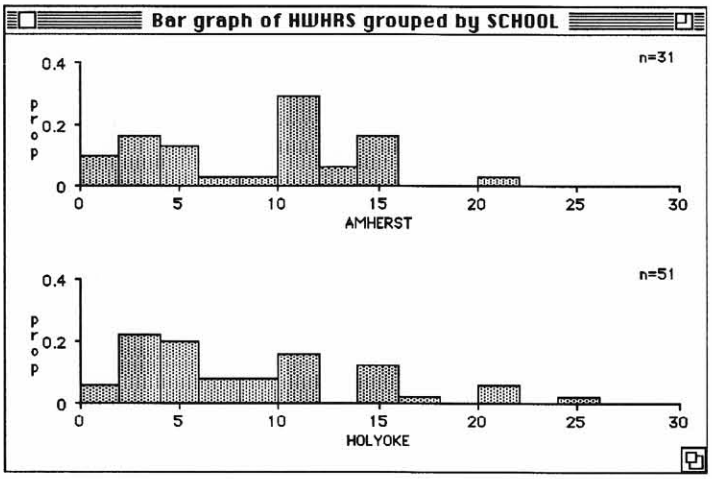


Figure 7. Separate histograms of HWHRS for each level of the variable SCHOOL (Amherst/Holyoke).

The two sub distributions for homework hours are displayed one above the other to make visual comparison easy. Because there are different numbers of students surveyed from each school, I have changed the values on the y-axis to proportions. The number in each subgroup is displayed to the right of the histograms. The option of switching from frequencies to proportions makes it easier to visually compare the two distributions, which do appear to be somewhat different. However, histograms may not be the best choice for exploring this question. We are getting more information in the details of the histograms than we want, and it is difficult to judge where the distributions are centered.

Comparing Distributions with Box Plots

Figure 8 are box plots of the same data displayed as histograms in Figure 7. These show a median of 10 for the Amherst students compared to 6 for the Holyoke students. (The small dot over the 25 in the Holyoke box plot is an "outlier," a value far enough from the bulk of the data so as to warrant special consideration.) Note how easy it is to compare these two box plots when they are placed one above the other on a common axis.

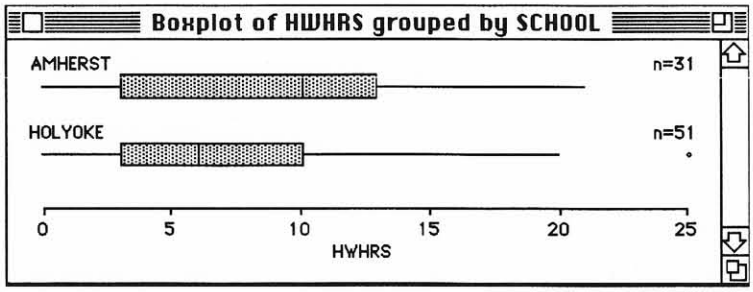


Figure 8. Separate box plots of HWHRS for each level of the variable SCHOOL (Amherst/Holyoke).

What I have been doing up to now might be considered a preliminary investigation, becoming familiar with the "texture" of individual variables before addressing the question of the relationship between them. Let's return to our question and refine it somewhat: "Do those holding a part-time job spend less time on school work than those not working?" Figure 9 is a box plot of HWHRS grouped by the variable JOB. Surprisingly, the 56 students who work ("yes") have a higher median for homework hours than the 26 students who do not work ("no").

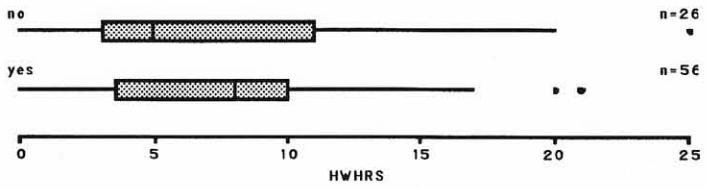


Figure 9. Box plot of HWHRS grouped by JOB (yes/no).

Don't Report Findings: Tell Stories

While it might be tempting to stop here thinking we have answered the question, there are a large number of possible explanations for the observed difference, and many possible interpretations. First, we could be seeing a chance difference which occurred when we drew our samples. Every once in a while (about two times in a thousand, in fact) you can deal yourself a five card poker hand of all one suit even when the deck has been well shuffled. If your "flush" were in spades, someone unfamiliar with poker cards might look at your hand and infer that all the cards in the deck were spades. You could prove this untrue by showing the person the entire deck, explaining that what happened to you doesn't happen very often, at least among honest players. Similarly, if we looked at the whole student body at each school, perhaps we would find students putting in about the same time on homework whether they lived in Holyoke or Amherst and whether they worked or not. The question is, how hard is it to draw a sample from the school "decks" and get a result like the one we got? This question is addressed with inferential statistics and is handled in DataScope via a randomization procedure (see Konold, 1993; Noreen, 1989). We will not explore the "chance explanation" here, but it is an important consideration for students at higher grade levels to keep in mind when analyzing data.

Let's suppose that the differences in homework hours are in fact characteristic of all students at Holyoke and Amherst. There is still a large number of possible explanations for these differences. One of the challenges in teaching data analysis is getting students beyond reporting a "finding" to creating and testing the plausibility of "stories." Stories provide possible explanations for our observations, and it is stories about data that ultimately matter. Here are some simple stories that might explain the finding that students who work tend to study more and which could be pursued by exploring relationships among variables in this data set:

- More of the students at Amherst are planning to attend college, and thus they are both studying harder and working to save money for college.
- Some students are more motivated than others, and therefore more likely to hold a job and be diligent about their school work.
- Students who work are older than students who don't work and as students advance into higher grades, they get more homework assigned.

When data sets are both long (with many cases) and wide (including many variables), explanations students formulate can often be "tested." In this case, because the data set includes students' ages, school, plans to attend college, and self ratings on motivation, students can explore the possibilities above. Space doesn't permit exploring all these explanations here, and, in fact, none of the above explanations are supported by the data. But I want to pursue one of the possibilities (that Amherst students are both studying and working more than Holyoke students) to stress that analyzing data is not a linear, straightforward process.

In exploring the possibility that students at Amherst are both working and studying more than those at Holyoke, students discover a large discrepancy between the percentage of students at the two schools who work. But it is the Holyoke students who are working more, as shown in the two-way table in Figure 10. Two-way tables are the most commonly used display for exploring dependencies between two non-numeric variables, but they are notoriously difficult for beginning students to learn to interpret. There are almost too many numbers to look at, which belies the fact that the relationship can be described by choosing a pair of them. In this case, we could summarize the results by noting that 82% of the Holyoke students hold a job compared to 45% of the Amherst students.

	JOB		
SCHOOL	no	yes	total
AMHERST	17 (0.55)	14 (0.45)	31
HOLYOKE	9 (0.18)	42 (0.82)	51
total	26 (0.32)	56 (0.68)	82

Figure 10. Two-way frequency table of JOB (yes/no) grouped by SCHOOL (Amherst/Holyoke).

Grouped Bar Graphs Are Often Better Than Two-way Tables

Figure 11 shows the same data displayed as a grouped bar graph. The upper graph shows the proportion of Amherst students who don't ("no") and do ("yes") work. It is much easier in this graph than it is in the two-way table to see the difference in the rate of working between the two schools. If you want to know precise values, the

two-way table is what you want. But if you are interested in seeing whether there is a relationship between two non-numeric variables, look at the grouped bar graph.

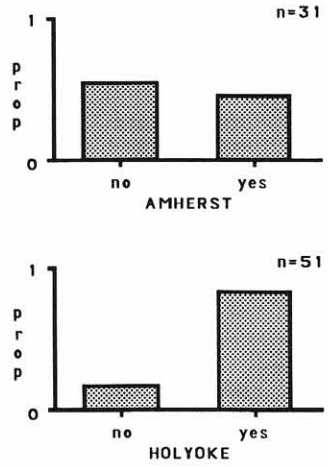


Figure 11. Bar Graph of JOB (yes/no) grouped by SCHOOL (Amherst/Holyoke).

This collection of findings might lead us to conclude that work does adversely affect study time: Holyoke students are twice as likely to hold a part time job as are students at Amherst, and since we also know that Amherst students study on average 4 more hours a week than those at Holyoke (see Figure 8), this suggests that students who work spend less time studying than those who don't work. However, this conclusion is not supported by the box plots in Figure 12 which show homework hours separately for students at each school, and within school, for those who do and do not work (the upper plot show the 17 Amherst students who don't work). In our samples from each schools, those who work spend more time studying than those who don't.

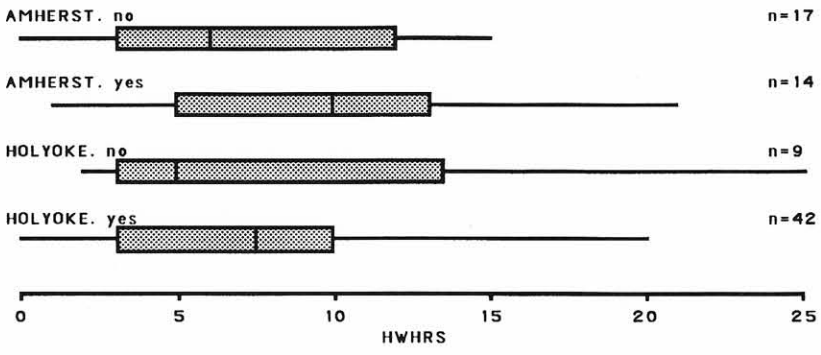


Figure 12. Box plots of HWHRS grouped by SCHOOL (Amherst/Holyoke) and by JOB (yes/no).

At least one new question has emerged from this investigation: Why are Holyoke students working more? Again, several explanations are possible. One is that work is more valued in Holyoke. The grouped bar graphs in Figure 13, however, suggests another explanation: The sample of students from Holyoke are older than those at Amherst.

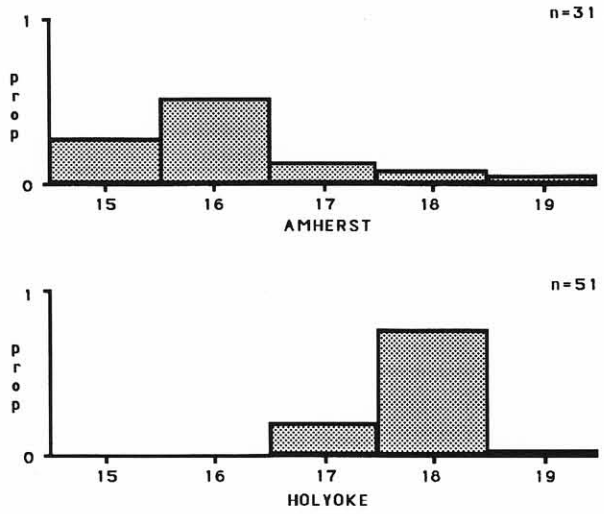


Figure 13. Separate histograms showing proportion of students of various ages at Holyoke and Amherst.

The next logical step in the investigation would be to see whether age and job status are related (and indeed they are). However, what I wanted to point out is that while large data sets give students the opportunity to pursue a family of questions, the results of the analyses are usually complex and often inconclusive. This can frustrate students whose interest in and knowledge about the questions they are pursuing are fragile. I've had students throw up their hands in despair after looking at a few ambiguous or contradictory results and claim that they weren't that interested in the question in the first place. This is one of the biggest challenges to confront in using large data sets. The computer can exacerbate this problem in that it makes it easy for students to look at plot after plot without giving serious thought to any of them, and without taking time to organize and reflect on what they have found. This is one of the areas in which teacher guidance and wise selection of tasks is critical.

Sometimes a Story Can be Woven Around a Single Plot

It does happen, however, that some findings lend themselves to clear explanations. Figure 14 is a scatterplot from a data set which includes information on 104 countries obtained largely from an almanac. Most of the information is 1990 vintage, which means that this data set is already history. In 1990, the USSR is still an entity and Germanys in East and West are still separated by a wall. (One of the problems building large data sets with up-to-date information is that you never finish building the data set.)

The scatterplot shows birth rate (births per thousand population) on the x-axis and deaths rate on the y-axis. In addition, the regression line has been added, with its formula displayed in the upper left, and the value of r, the regression coefficient, shown in the upper right.

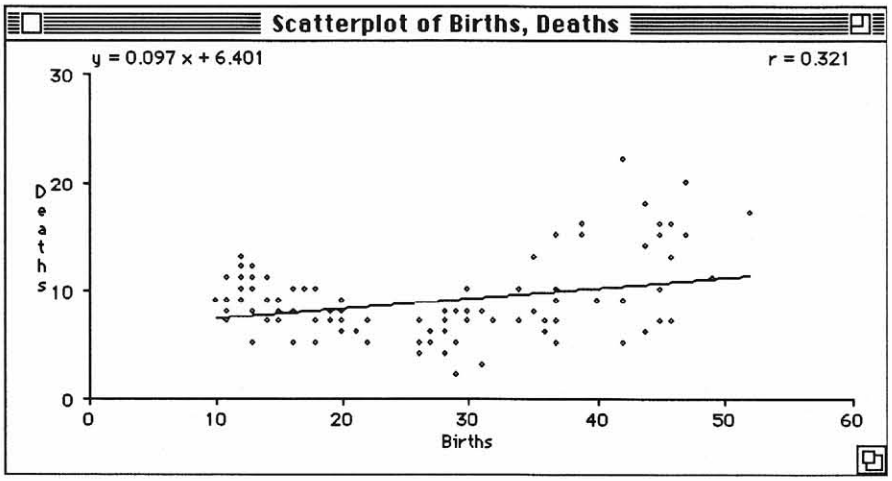


Figure 14. Scatterplot with regression line of birth rate and death rate of 104 countries.

In statistics classes I took as an undergraduate, plots of this type were shown soon after we'd been introduced to the idea of fitting straight lines through bivariate data. The instructor would draw a straight line through a U-shaped scatterplot and ask "What's the matter?" One of the better students would point out that the data weren't linear. Then the instructor would inform us that r measures goodness of fit only around straight lines, and is therefore a poor estimate of the degree of relationship when the data aren't linear. Lesson learned, the instructor would proceed to another scatterplot.

Because of the nature of the lesson, the plot was usually shown with axes unlabeled. The lesson as taught, after all, is a general one that applies to any non-linear scatterplot. The information about birth and death rates in the world's countries might focus attention on other than the statistical properties the instructor is hoping students learn.

Of course, limitations of linear regression are important statistical lessons to learn. However, if one wants to show students how they can learn something about the world by looking at and thinking about data, a great opportunity is lost when the axes are left unlabeled, and when the students' knowledge of and curiosity about the world are left untapped. The interesting question in this case is not "What's the matter with fitting a straight line?" but "What's going on here?"

One way to draw students into the question using DataScope is by clicking the cursor over one of the points to see the name of a country (see Figure 15). After some thought, they can give a verbal description of the scatterplot, which highlights a puzzling question. Why is it that for a while, as birth rates decline, so do death rates, but when births reach about 25 per thousand, further decreases in birth rate are associated with increases in the death rate?

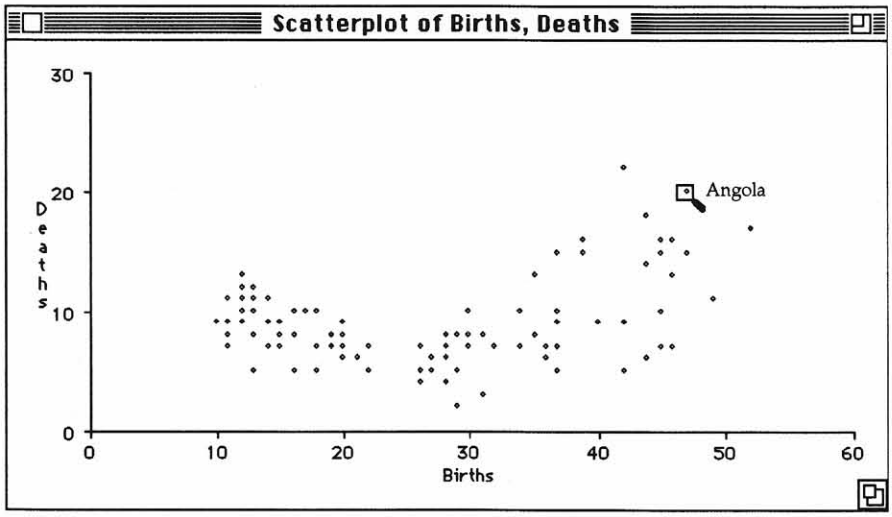


Figure 15. Scatterplot of birth and death rate showing the location of Angola.

Some possible explanations are suggested by grouping on the variable "Labor" which classifies the countries according to the predominate type of labor (Industrial,

Agricultural, Service, Government). In this case the grouping feature replaces each point in the scatterplot with the first letter of the appropriate labor type (Figure 16). Someone weary of the modern age might reason on the basis of these data that in the evolution from Agricultural to Service economies, the death rate eventually begins to increase because of stress.

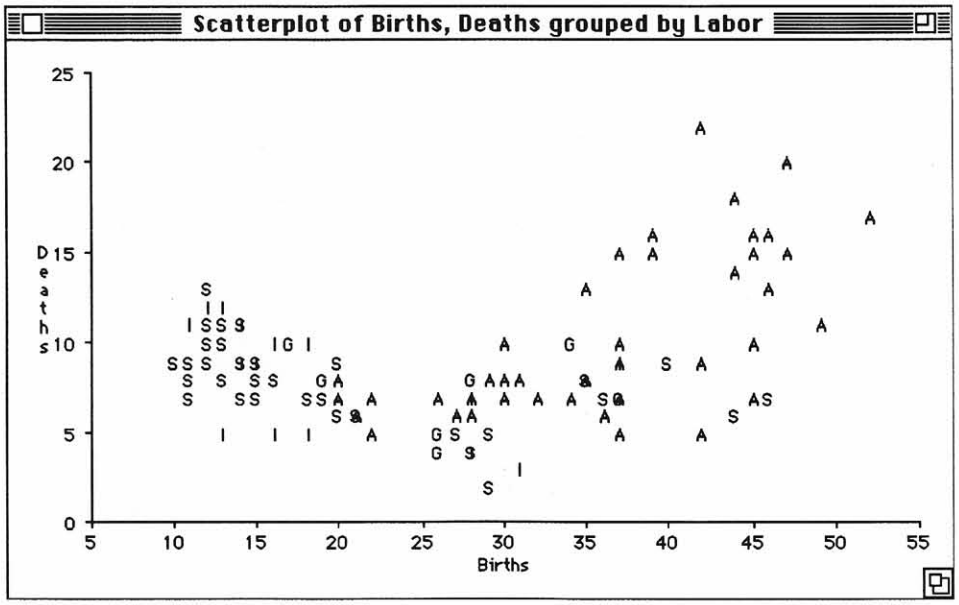


Figure 16. Scatterplot of birth- and death rates grouped on type of Labor (I=Industrial, A=Agricultural, S=Service, G=Government).

Looking at the names of other countries on the plot allows students to generate other possible explanations (Figure 17). Here's one: Countries who have successfully maintained low birth rates for a number of years have older populations. Thus death rates are higher in these countries than they are in countries such as China who have only recently brought down their birth rate but still have relatively young populations.

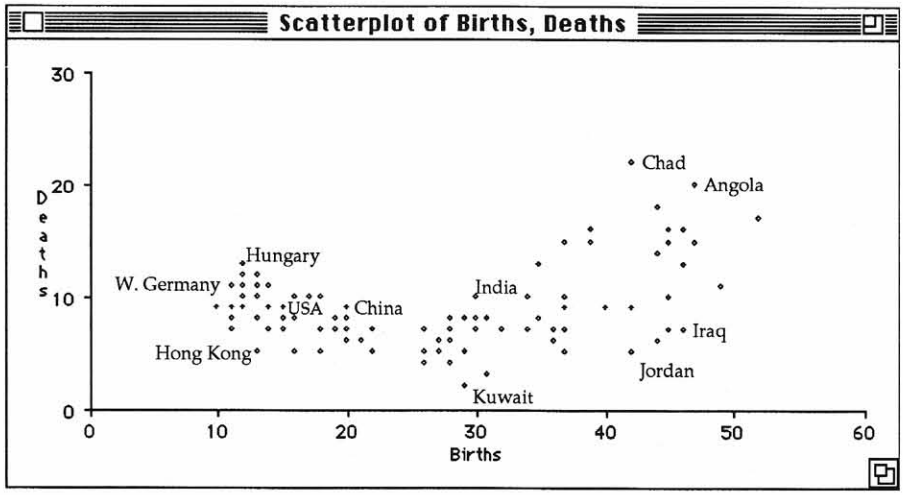


Figure 17. Scatterplot of birth and death rate showing the identity of several of the countries.

The scatterplot in Figure 18 shows the relation between birth rate and the percentage of the population over 65. This plot gives strong support to one piece of our explanation -- that as birth rate is decreased, the percentage of the older segment of the population increases. The accelerated rate of this increase as the birth rate decreases could help explain the U-shaped relation between birth and death rate.

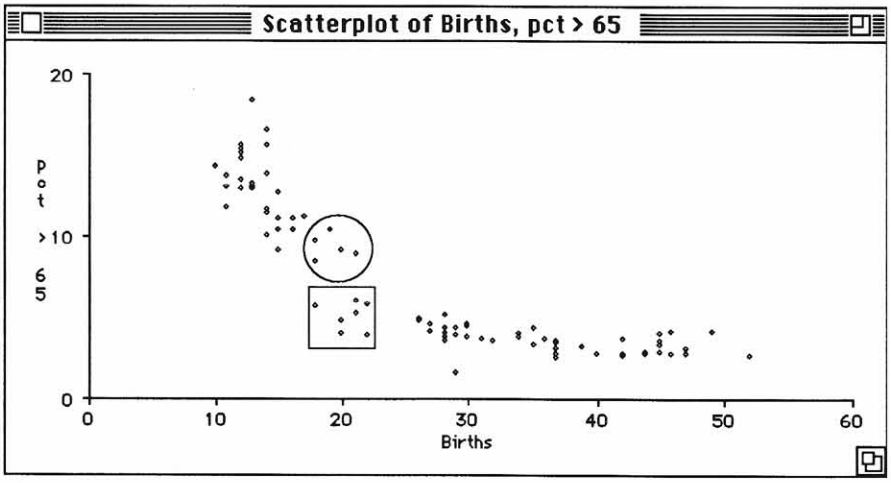


Figure 18. Scatterplot of birth rate and percentage of population over 65 with two groups of countries marked for comparison.

The other part of our argument was that the increase in death rate would only be seen some time after a country had maintained a relatively low birth rate. Some

support for this argument comes from comparing the five counties enclosed in the circle (which include Argentina, Israel, and the former USSR) with the seven countries in the square (which include China, Sri Lanka, and Thailand). The birth rates in these two groups are fairly close compared to the difference in the percentage over 65. However, over the 20-year period from 1970 to 1990, the group in the circle lowered their birth rates by an average of 5.3 per 1,000, while the group in the square lowered theirs by an average of 12. Even though the birth rates in these two groups are nearly equal, because the countries in the circle have maintained a relatively low birth rate for longer, their populations are now older on average, and we could therefore expect them to have somewhat higher death rates as a result.

Summary

My intention in this article is to show how data analysis software designed specifically for educational use can help students get beyond rudimentary skills to the interesting and more difficult tasks of learning how to probe a set of data, and build from pieces of information, stories that tie these pieces together into understandable wholes. I certainly do not believe that the computer is a panacea, or that it isn't important to introduce students to various graphing and statistical conventions off the computer. There are many activities I find better done off the computer. My students, for example, draw their first box plots by hand using modest amounts of data. But when it comes time to have students go through cycles characteristic of exploratory data analysis, I want them at the computer. I have in this article only touched on the problems students still experience trying to do exploratory data analysis using even simple software, and these don't appear to me to be simple problems. Finding our way through these difficulties will be an interesting part of the educational success story we should be able to tell ten years from now about how the computer changed the teaching of data analysis.

References

Biehler, R. (1994). Cognitive technologies for statistics education: Relating the perspective of tools for learning and of tools for doing statistics. In L. Brunelli & G. Cicchitelli (Eds.), Proceedings of the First Scientific Meeting of the International Association for Statistics Education (pp. 173-190). Università di Perugia, Italy.

Cleveland, W. S. (1993). Visualizing data. Summit, NJ.: Hobart Press.

Gordon F. S., and Gordon S. P. (1992). Statistics for the twenty-first century. MAA Notes, #26. Mathematical Association of America.

Konold, C. (1994). Understanding probability and statistical inference through resampling. In L. Brunelli & G. Cicchitelli (Eds.), Proceedings of the First Scientific Meeting of the International Association for Statistics Education (pp. 199-211). Università di Perugia, Italy.

Noreen, W. (1989). Computer intensive methods for testing hypotheses. New York: John Wily & Sons.

Polanyi, M. (1969). Knowing and being. Chicago: University of Chicago Press.

Strunk, W. Jr., and White, E. B. (1972). The elements of style. New York: Macmillan.

Tufte, E. R. (1983). The visual display of quantitative information. Cheshire, Conn.: Graphics Press.