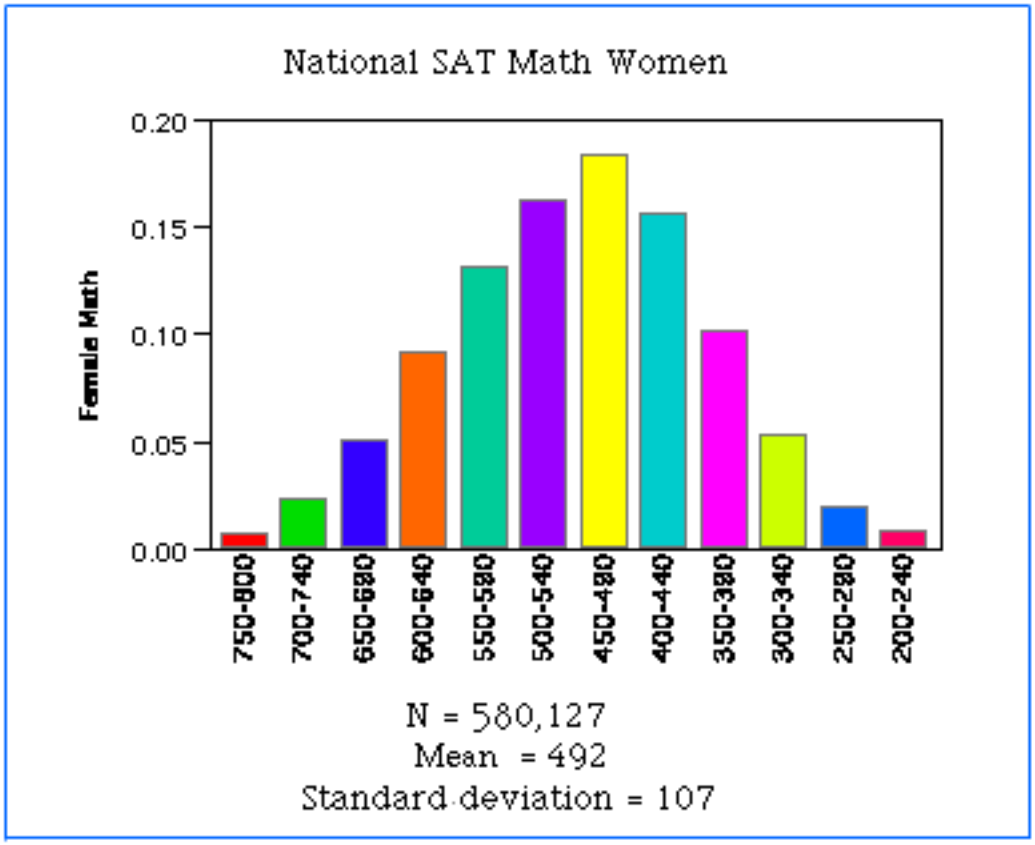
INTRODUCING THE USE OF A STATISTICAL PACKAGE

Written by the Chance Team



INTRODUCING THE USE OF A STATISTICAL PACKAGE

To understand articles on statistical experiments, it is useful for the students to have some experience exploring datasets For this purpose we have the students use a simple statistical package in the Chance course. In this unit we will describe how we introduce the students to the use of a statistical package.

Three packages that we have used are Minitab, Data Desk, and JMP. Of course, which package you use will be depend on your own preferences and what is available at your school. Our presentation here is based on Data Desk; however, it will be easy to modify what we do for another statistical package. Obviously, what follows is not intended to be a complete guide to Data Desk, but rather an illustration of strategies we use to introduce computing early in the Chance course.

On the first day of class we ask the students to fill out a survey to get some information about their fellow students. In the 1996 Chance class we used the following survey.

Survey given to students in a Chance class at Dartmouth in the Fall of 1996.

1. How many CD's do you own?_____

2. Which year are you? (97, 98, 99, or 00)_____

3. Are you male or female? (M = male, F = female)_____

4. What is your height in inches? _____

5. What is your shoe size (length, not width)?_____

6. Record your pulse rate (the number of beats in one minute) in class or in a similar setting._____

7. Are you left or right handed? (0 = left, 1 = right, 2 = ambidextrous)___

8. Do you smoke? (0 = no, 1 = occasionally, 2 = regularly)_____

9. What is your birth order? (1 = oldest or only child, 2 = second oldest, etc.) _____

10. How many siblings (i.e., brothers and sisters) do you have?_____

11. What was your SAT verbal score?_____
    What was your SAT math score? _____

12. What is your current gpa (grade point average)?_____

13. On average, how many hours of television or videos do you watch per week? _____

14. How much exercise do you get per week (hours)?_____

15. How much did you spend on your last haircut (including tip) $_____

16. What is the average class size for the classes you are taking this semester? _____

17. How old do you think your professors are? Shunhui Zhu_____, John Finn _____

18. How many hours do you sleep per night on average?_____

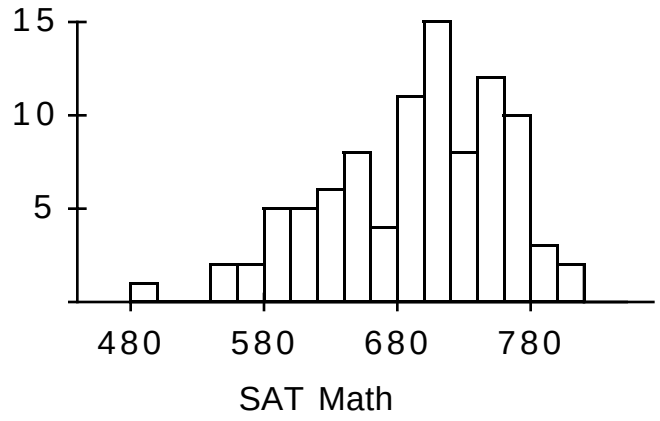
19. How much change do you have with you right now?_____

20. We may discuss the following topics in class, rank them in order of preference (1 = most interesting, etc.)

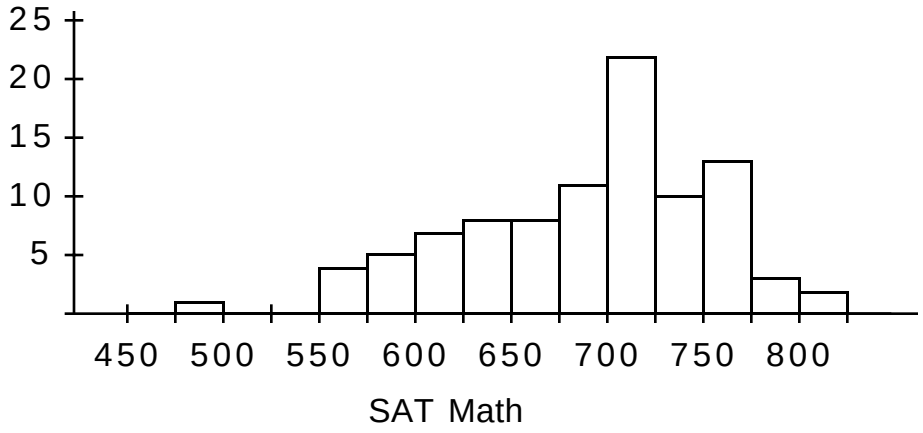
    _____health, ____ weather, ___ polling, ___ sports, _____gambling

We collect the data on the first class and make it available to the students by the second class. In the second or third class, or in a special laboratory session, we show the students how to use Data Desk, with the survey data as an example. We then ask them to further explore this data set on their own. If it can be arranged, a laboratory session is more successful since the students can carry out the operations as we illustrate them. Because we will be getting data from a variety of sources, we start by showing the students how to import data in Data Desk when it is available as a text file in standard spreadsheet form.

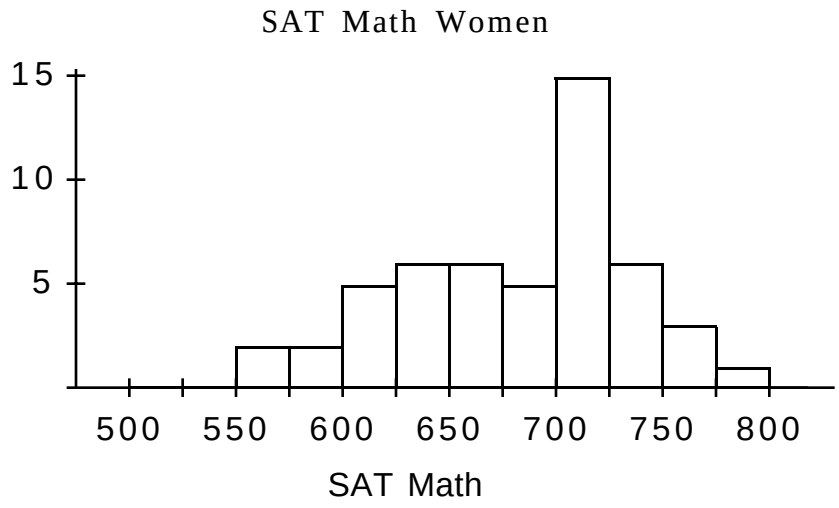
To get started, we show how to look at a few questions we are interested in, leaving other questions that the students might be interested in to them. For example, suppose we want to look at SAT math scores for the class. We might begin with a simple graphical display, say a histogram. To do this, highlight the icon for the SAT math variable, and then from the menu bar choose Plot > Histograms. This gives



Data Desk allows great flexibility in what kind of analysis can be done at any time. To provide some guidance, most Data Desk windows have a submenu indicated by a triangle located in the upper left corner of the title bar. This so-called HyperView menu suggests general actions related to the analysis or display in the window. In the present case, we might consider adjusting the axis scale. This is done from the Plot Scale option in the HyperView menu. Since many people informally summarize SATs in multiples of 50, we have started the first bar at 450 and used a bar width of 25. (More generally, students can use this feature to investigate effects of changing bin widths on a histogram).

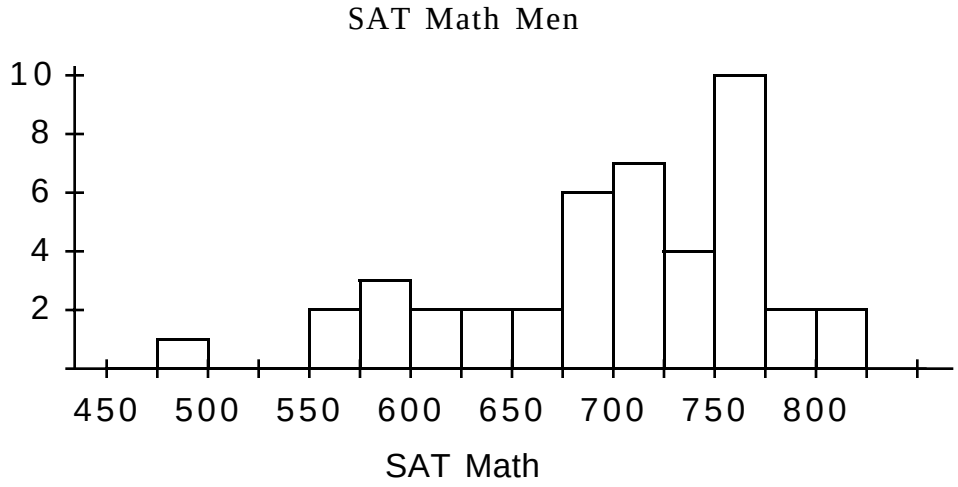


News articles frequently remark that men and women perform differently on the SATs. It is easy to have Data Desk perform computations by groups. To do this, highlight the variable that gives the group names (Gender in our example), and then from the menu bar choose Special > Group > Assign. This will create a button at the bottom of the screen that toggles on and off. While it is on, separate analyses will be done for each category represented in the group. Now with the toggle on, Plot > Histograms produces the following.



N = 51
Mean = 677.6
Standard Deviation = 55.16

We next do the same thing for the Men's SAT scores and obtain:



N = 43
Mean = 697.21
Standard Deviation = 75.03

Notice that we have added numerical summaries to the figures above. These were obtained from the menu bar by choosing Calc > Summaries > Reports. We see that the men have an average SAT math score about 20 points higher than the women, but also that the men have a standard deviation about 20 points higher than the women. Finally, we point out that there are differences in the shapes of the distributions, and that the men's in partcular does not look "bell shaped."

Picking up on the variability issue, the fact that men have larger standard deviations than women in IQ type tests is well known. Here is an article from Chance News 4.10 relating to this phenomenon.

Men at extreme ranges of IQ tests, study says.
Sacramento Bee, 7 July, 1995 A1
Byline: Bee News Services

The article reports that a new study, by Larry Hedges and Amy Nowell of the University of Chicago (reported in "Science", 7 July 1995), has found that the average man and average woman share about the same level of intelligence, but men account for a higher proportion of both geniuses and the mentally deficient. The report analyzed six large national surveys of American male and female teenagers' performance on tests of mental ability, conducted over the past thirty years.

Seven of every eight people in the top 1% of IQ tests are men, but men also represent an almost equally large percentage of the mentally disadvantaged. Neuroscientist Richard Haier of the University of California, Irvine, says that the findings of a higher percent of men in the top IQ levels is nothing new, but what is new is that "there were more males in the low end."

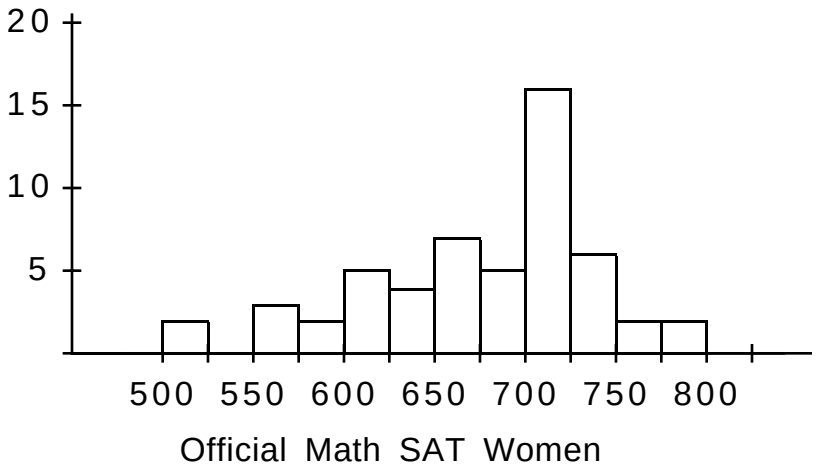
The article mentions that, while men and women differ in brain sizes and male and female brains function differently, such physiological differences do not account for the differences in the abilities of the sexes. The study sheds little light on the origin of sex difference in aptitude.

In the "Science" article the authors stress that it is important to analyze representative samples instead of samples selected from talent searches etc. While they analyze a number of studies, their main conclusions are based on their analysis of the National Assessment of Educational Progress program which periodically tested large samples (70,000 to 100,000 students) in the areas of reading, mathematics, science, and writing. They found that in all four areas the men had higher variances than the women, typically of the order 3 to 15%. Men had higher average scores in mathematics and science and the women in reading and writing. They suggest that both the small number of women in the top 10% of math and science and the high number of men in the bottom 10% in writing and reading have policy implications. Hedges and Nowell suggest that intensive recruiting will be necessary to achieve a fair representation of women in science and that some men will have difficulty finding employment in an increasingly information driven economy.

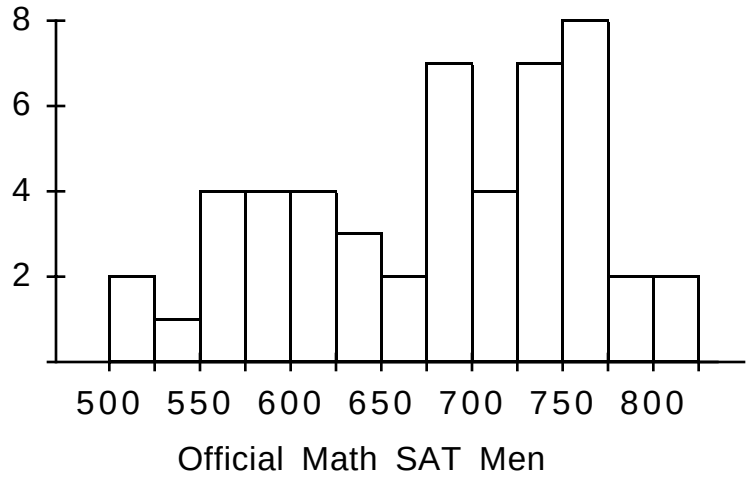
Since survey results occur frequently in the news, it is worth having the students start thinking about potential problems with survey design. For example, in our survey the students are self reporting their SAT scores and this can bias the results. Some students may not know if they should give their best score or their last score. While we emphasize that the survey is anonymous, there may be students who don't believe us and fear that reporting a low score will penalize them. There were seven students

who did not give their SAT scores and these could have been mostly students with low scores.

In this case, we can see how accurate our sample data are in this case by asking the registrar for a list of the SAT scores for students in the class. When we did this we obtained the following results:



N = 54
Mean 672.40
Standard Deviation 63.90



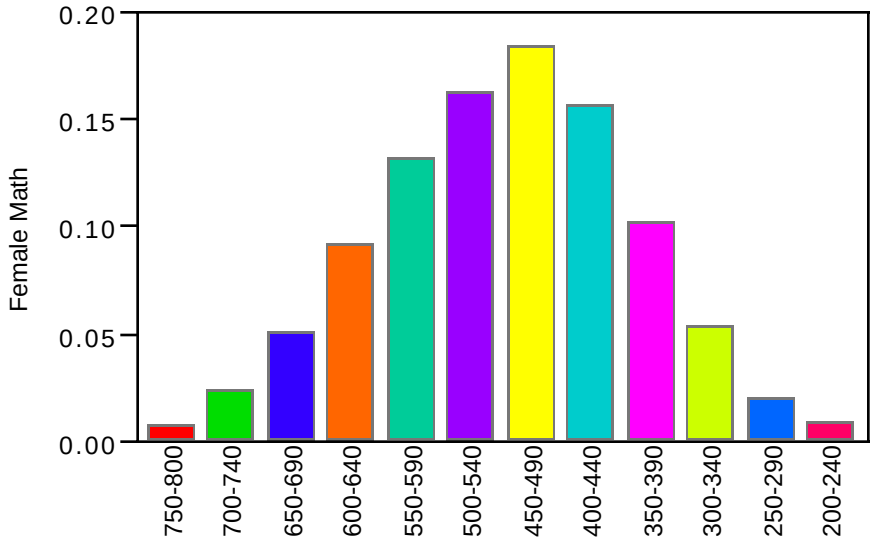
N= 50
Mean = 677.20
Standard Deviation = 81.57

We see that there are some differences between the estimate from the class survey and the official records of the registrar. The largest difference occurs in the estimate of the average scores for men which are 20 points higher than those in the official records. As we have mentioned, there are number of possible explanations for this.

Of course, Dartmouth SAT scores are not typical of SAT scores nationally. We can see how different they are by looking at data from the College Board's 1996-7 report "College-Bound Senior: A Profile of SAT Program Test Takers."

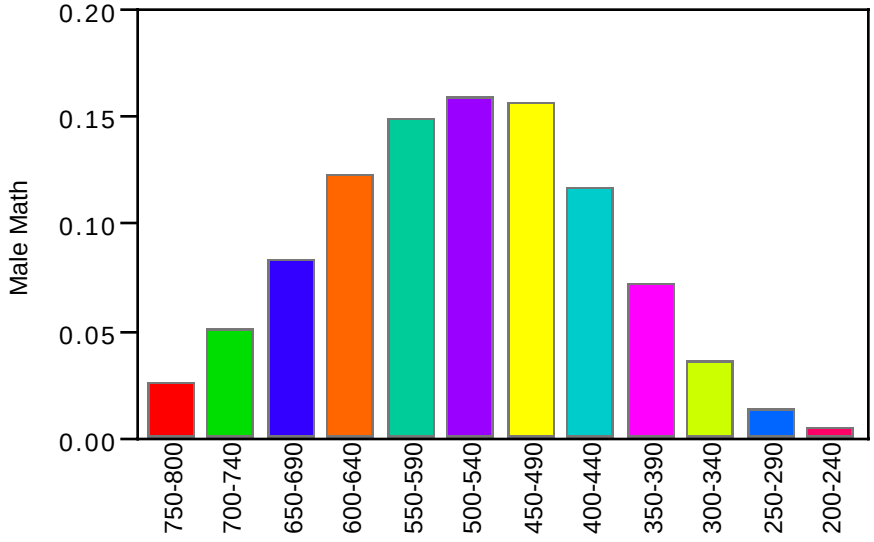
Reports like this tend to provide summary data rather than raw data. It is not always easy to deal with summary data with a simple statistical package. In particular, we could not see an easy way using Data Desk to provide the following Bar Graph for the summary data provided in this report. (This happens to be easy in JMP).

National SAT Math Women



N = 580,127
Mean = 492
Standard deviation = 107

National SAT Math Men



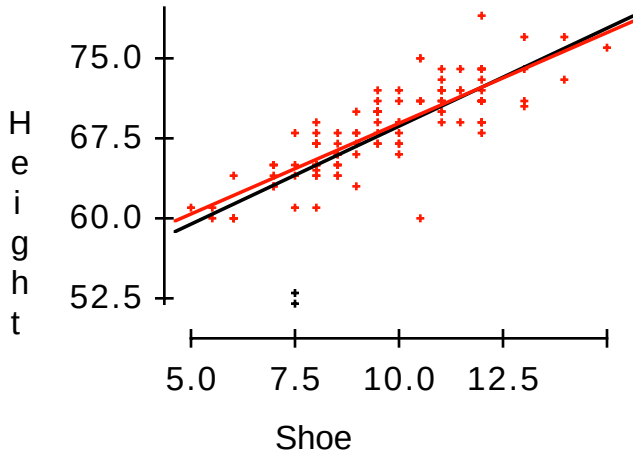
N = 504,598
Mean = 527
Standard deviation = 115

We see from these graphs that the men averaged about 35 points higher than the women on the Math SAT, and had a slightly higher standard deviation.

We also can use Data Desk to explore relations among the answers to our student questionnaire. For example is shoe size related to height? To investigate this, we click on the Height variable, labeling it as the dependent variable Y . Then, holding down the shift key, we click on Shoe labeling it as the independent variable X . We then choose from the File Menu Plot > Scatterplot to obtain a scatterplot of Shoe and Height. From the HyperView menu we choose Add Regression Line. The result is:

Not surprisingly, we see that there is a definite relation between shoe size and height and it can be described quite well by a simple straight line. From the HyperView menu we can also choose Correlation of Height vs. Shoe Size and find that the correlation is $r = .793$ indicating a strong relation between Shoe size and Height. Of course this invites us to mention the difference between correlation and causation.

This is also a good place to discuss outliers. We see two points that fall distinctly below the regression line, namely the points correspond to heights 52 and 53 (question: what do you think of the rightmost point with height 60.0?). Is it possible that these were supposed to be 5'2" and 5'3", making them 62 and 63 inches? We can change these to 62 and 63 and see what effect that has, or we can remove these points and redo the analysis. We make the latter choice. We will first investigate the change graphically. Using the knife tool, we highlight all but the suspect points, and change their color on the scatterplot. Then from the HyperView menu, we can Add Color Regression Lines, which adds a new regression line drawn using only the selected points, as shown below.



We can see that the new regression line has a slightly smaller slope, but that the suspect points did not have dramatic influence. If we want to update the numerical results, we need to define a selector variable for the colored points. A selector variable is Data Desk's mechanism for restricting analysis to a subset of the data. Assuming your colored points are still highlighted (otherwise highlight them again now), go to the menu bar and choose Modify > Selection > Record. This creates a new 0-1 indicator variable in your data with value 1 for the selected points, and 0 for the suspect points. If you now drag this icon into the correlation analysis window and drop it on the "selection line" (which currently reads No Selector), the computation will be updated to include only these selected points. We see that the correlation increases from .793 to .829 when the suspect points are removed. If you click on the selection line in the window, the resulting pop-up menu includes a Remove Selector command to restore the full analysis.

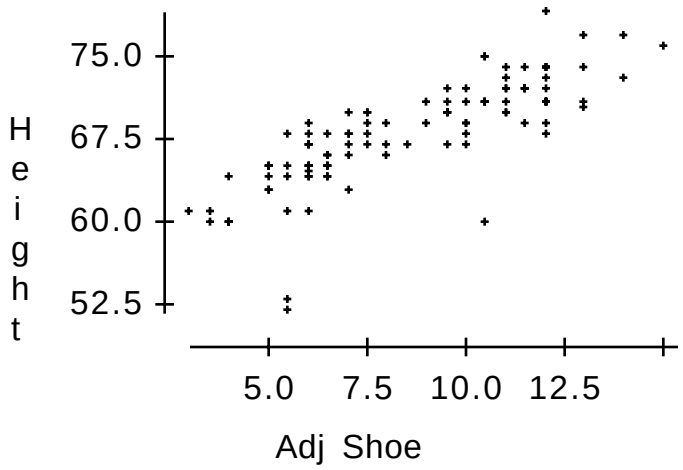
We note briefly another possible point for discussion, which we would probably not mention in a first introduction, but could be used later in the course. It can be argued that we have effectively used shoe size as a proxy for foot length. But women's shoe sizes run on a different scale than men's: a women's size 7 corresponds to a men's size 5. So we might consider standardizing things by subtracting 2 from the women's sizes. First we create a 0-1 derived variable called Female (0 = M, 1 = F) by defining

IF TextOF(Gender) = "F" THEN 1 ELSE 0

Such indicator variables are of course useful in their own right. Here we can create an adjusted size variable by defining

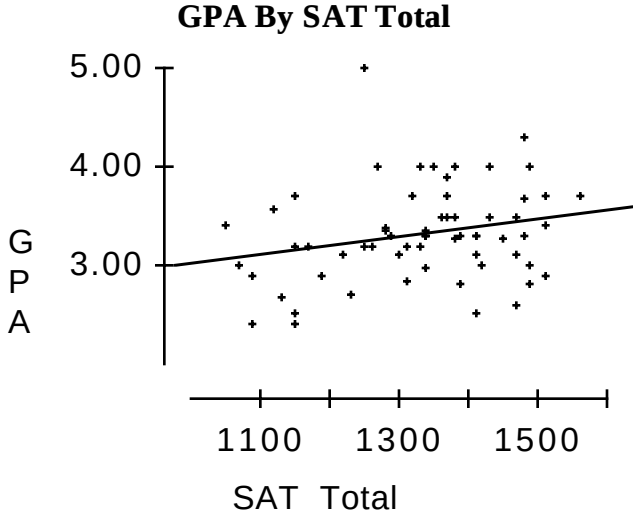
Adj Shoe = Shoe – (2 * Female)

Now a scatterplot of Height vs. Adj. Shoe looks like



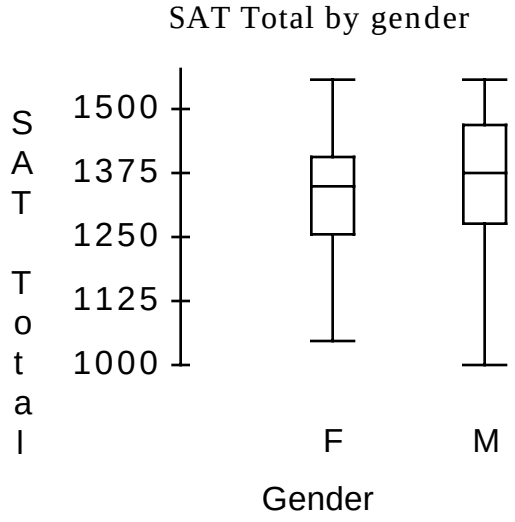
and has a correlation coefficient of .803, which is somewhat higher than the .793 for the original. Also a potential third outlier (which we tentatively flagged earlier, but did not remove) seems more prominent.

We consider next the relation between SAT scores and the students' current grade point average. We will use the total SAT score. To do this wee make a new derived variable SAT Total that is simply the sum of SAT Verbal and SAT Math. Then we make a scatterplot and add the regression line.

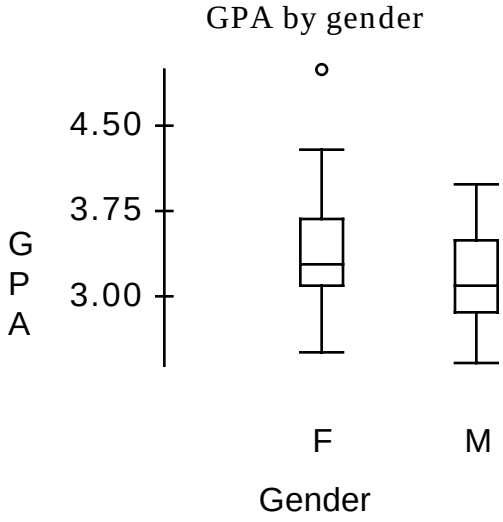


We see that our straight line does not give a close fit, but the slight positive indicates that there is some tendency for those with higher Total SAT scores to have a higher grade point averages. However, we find a correlation of only .247. The relation would have been stronger if we were considering the relation between SAT scores and freshman grade point average. This is what the College Board uses to describe the validity of their SAT scores.

When the independent variable is categorical and the dependent variable is quantitative, we can use Plot > Boxplots y by x to compare the dependent variable for various values of the independent variable. For example, we obtain the following by selecting SAT Total as Y and Gender as X and using Boxplots y by x.



We see again that the men in the class did better then the women on total SAT score but with more variation. Let's now compare GPA by gender.



Here we see that the women in the class are doing better on GPA than the men. This is consistent with larger studies of college performance and has caused some controversy over how the SAT scores are used in the awarding of National Merit Scholarships. Here is a recent article on this subject from Chance News 5.11.

College board revises test to improve chances for girls.
New York Times, 2 October, 1996, B8
Karen W. Arenson

In an agreement with the Office of Civil Rights, the College Board will add a multiple-choice test on writing to its Preliminary Scholastic Assessment Test (PSAT). This test will include questions involving the structure of language and standard written English. The PSAT test is taken by high school juniors and is the main determinant in awarding the National Merit Scholarships.

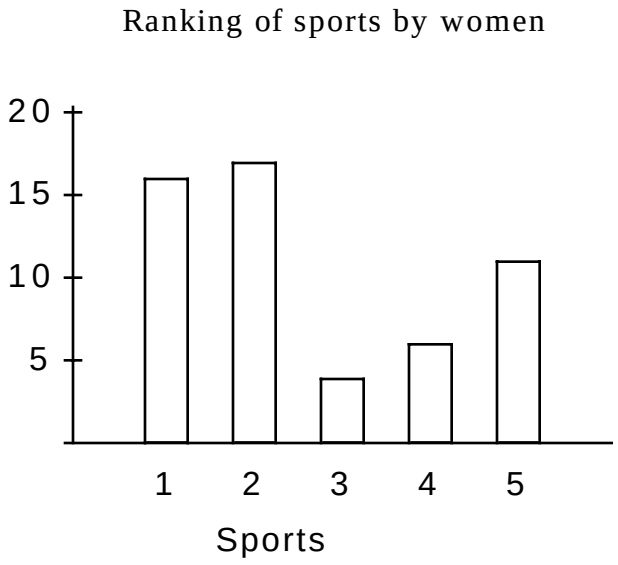
This followed a complaint filed with the Education Department in 1994 by civil rights activists who said that girls tend to score lower on the PSAT than boys, even though their high school and college grades were better.

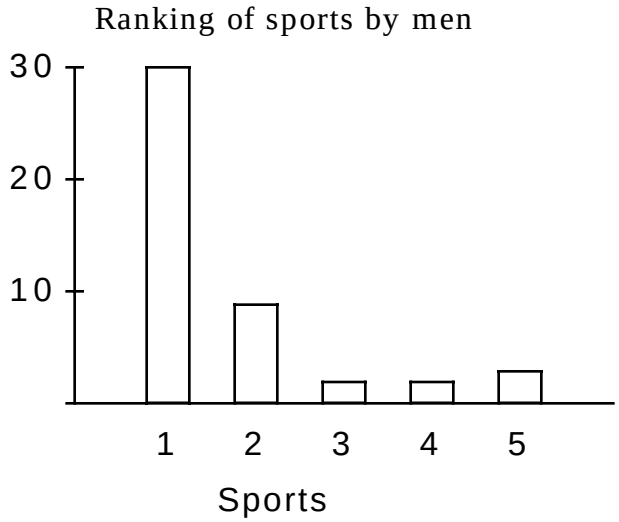
Each year more than a million high school juniors take the PSAT tests. About 55% of these are girls. Those in the top 15,000 scores, of which about 60% are boys, automatically become National Merit semifinalists. The students then submit grades, extracurricular activities, recommendations and essays; and about 14,000 are chosen as finalists.

In 1989, a Federal District Court in Manhattan ruled that the New York Regents Scholarships discriminated against girls because they were based on SAT scores. When New York State relied on standardized test, girls won 43 percent of the scholarships. A year after the court ruling, when grades were also taken into consideration, the girls won 51% of the scholarships.

As a last example, we consider a situation in which both of the variables are categorical. Recall that we asked the students to rank the topics health, weather, polling, sports, and gambling, with 1 being most favored and 5 least favored. We have had arguments among ourselves about whether women were turned off by the use of too many sports examples. We can use this example to shed some light on this question.

In our data set we have a variable Sports that indicates where the students ranked sports. Using our group variable gender, we can compare the students responses by plotting bar charts for men and for women.





In each of these graphs, we could get bar counts from the HyperView menus. But a more efficient way to proceed is to do a two-way table. This is obtained by highlighting the Gender and Sports variables, and choosing from the menu bar Calc > Contingency Table. The HyperView menu in the resulting table gives Table Options. In addition to the counts, we chose to report row margins and percent of row total.

Rows are levels of **Gender**
Columns are levels of **Sports**
No Selector
101 total cases of which 1 is missing

	1	2	3	4	5	total
F	16	17	4	6	11	54
	29.6	31.5	7.41	11.1	20.4	100
M	30	9	2	2	3	46
	65.2	19.6	4.35	4.35	6.52	100

table contents:
Count
Percent of Row Total

From this we see that about 65% of the men ranked sports as their favorite topic while only about 30% of the women ranked sports as their favorite topic. We also see that more than 20% of the women ranked sports as their least favorite subject.

We have found that the students enjoy working with a statistical package to explore this data set. They enjoy looking for relations between sex and number of CD's, cost of haircuts, amount of change carried etc. We make available, on our web site, surveys from previous classes and from different schools and so the students can explore the differences between the answers by different groups of students both at Dartmouth and at other schools.

Finally, we note that there are certainly alternative (and fancier) ways in Data Desk of doing some of the analyses we have described here. An advantage of the package is that it can be learned gradually by point-and-click exploration, and students readily incorporate more advanced features into their work as the course progresses.