Appendix: Sixty Projects

1. Go to a local grocery store and collect these data for at least 75 breakfast cereals: cereal name; grams of sugar per serving; and the shelf location (bottom, middle, or top). Group the data by shelf location and use three boxplots to compare the sugar content by shelf location. [Observational data; using boxplots to summarize data, can also be used for ANOVA test; high-sugar cereals are often at child-eye height.]
2. Use computer software to simulate 1,000 flips of a fair coin. Record the fraction of the flips that were heads after 10, 100, and 1,000 flips. Repeat this experiment 100 times and then use three histograms to summarize your results. [Simulation data; using histograms to summarize data; demonstrates central limit theorem and effect of sample size on standard deviation.]
3. Estimate the average number of hours that students at this school sleep each day, including both nighttime sleep and daytime naps. Also estimate the percentage who have been up all night without sleeping at least once during the current semester. [Survey data; confidence intervals for quantitative and qualitative data; students sleep less than 8 hours and many have all-nighters; if done at the beginning and end of the term, the differences are as expected.]
4. Estimate and compare the average words per sentence in *People*, *Time*, and *New Republic*. [Observational data; confidence interval with quantitative data; the order given is from fewest words to most; *New Republic* has some outlier sentences with close to 100 words.]
5. Estimate the percentage of the seniors at this college who regularly read a daily newspaper, the percentage who can name the two U.S. senators from their home state, the percentage who are registered to vote, and the percentage who would almost certainly vote if a presidential election were held today. [Survey data; confidence intervals for qualitative data; far more students are registered and will vote than read a newspaper or can name their senators.]
6. Conduct a taste test of either Coke versus Pepsi or Diet Coke versus Diet Pepsi. Survey at least 50 randomly selected students who identify themselves beforehand as cola drinkers with a definite preference for one of the brands you are testing. Give each subject a cup of each cola that has been coded in a way known only to you. Calculate the fraction of your sample whose choice in the taste test matches the brand identified beforehand as their favorite. (Do not tell your subjects that this is a test of their ability to identify their favorite brand; tell them it is a test of which tastes better.) Determine the two-sided p-value for a test of the null hypothesis that there is a 0.5 probability that a cola drinker will choose his or her favorite brand. [Experimental data; hypothesis test using binomial model; most students prefer Coke, but neither group is very successful at identifying its favorite.]
7. Find five avid basketball players and ask each of them to shoot 100 free throws. Do not tell them the purpose of this exercise, which is to determine if a missed free throw is equally likely to bounce to the same or opposite side as their shooting hand. Use your data for each of these players to calculate the two-sided p-value for testing the null hypothesis that a missed free throw by this player is equally likely to bounce to either side. [Experimental data; hypothesis test using binomial model; coaches often say that the ball will bounce to shooting-hand side, but the data are unpersuasive.]
8. Ask 50 female students these four questions: Among female students at this college, is your height above average or below average? Is your weight above average or below average? Is your intelligence above average or below average? Is your physical attractiveness above average or below average? Ask 50 male students these same questions (in comparison to male students at this college). Try to design a survey procedure that will ensure candid answers. For each gender and each question, test the null hypothesis that $p = 0.5$. [Survey data; hypothesis test using binomial model; most males think that they are above average.]
9. Young children who play sports are often separated by age. In 1991, for example, children born in 1984 might have been placed in a 7-year-old league while children born in 1983 were placed in an 8-year-old league. Someone born in January 1984 is eleven months older than someone born in December 1984. Because coaches give more attention and playing time to better players, children with early birth dates may have an advantage when they are young that cumulates over the years. To test this theory, look at a professional sport and see how many players have birth dates during the first six months of the year. [Observational data; hypothesis test using binomial model; seems to be true.]

10. College students are said to experience the Frosh 15 -- an average weight gain of 15 pounds during their first year at college. Test this folklore by asking at least 100 randomly selected students how much weight they gained or lost during their first year at college. Determine the two-sided p-value for testing the null hypothesis that the population mean is a 15-pound gain, and also determine a 95 percent confidence interval for the population mean. [Survey data; hypothesis test using t distribution; strongly rejected (is it a myth or do students misreport?).]
11. What percentage of the seniors at your college expect to be married within five years of graduation? What percentage expect to have children within five years of graduation? How many biological children do the seniors at your college expect to have during their lives? Do males and females differ in their answer to these questions? [Survey data; two-sample test; few expect to be married or have children within five years of graduation; males plan to have slightly more children; if possible, a comparison with alumni records is interesting.]
12. Ask a random sample of at least 50 students the following question: "During the school year, how many hours a week do you spend, on average, on school-related work -- for example, reading books, attending class, doing homework, and writing papers?" Ask a random sample of at least 25 professors this question: "During the school year, how many hours a week do you spend, on average, on school-related work -- for example, preparing lectures, teaching, grading, advising, serving on committees, and doing research?" Determine the p-value for a test at the 5 percent level of the null hypothesis that the two population means are equal. [Survey data; two-sample test; professors work twice as many hours as students.]
13. Ask at least 100 randomly selected college students to write down their grade point average (GPA) and to indicate where they typically sit in large classrooms: in the very front, towards the front, in the middle, towards the back, or in the very back. If feasible, restrict your sample to students who are taking the same class or similar classes. Use an ANOVA F-test to see if the differences in the GPAs among these five categories are statistically significant at the 5 percent level. [Survey data; ANOVA; no statistically persuasive patterns.]
14. Ask randomly selected college students if they have had a serious romantic relationship in the past two years and, if so, to identify the month in which the most recent relationship began. When you have found 120 students who answer yes and can identify the month, make a chi-square test of the null hypothesis that each month is equally likely for the beginning of a romantic relationship. [Survey data; chi-square test; the start of each term is a popular time for romance.]
15. Ask 50 randomly selected students this question and then compare the male and female responses: "You have a coach ticket for a nonstop flight from Los Angeles to New York. Because the flight is overbooked, randomly selected passengers will be allowed to sit in open first-class seats. You are the first person selected. Would you rather sit next to: (a) the U.S. president; (b) the president's wife; or (c) Michael Jordan? [Survey data; chi-square test; females choose the president's wife, males the president.]
16. For each of the 50 states, calculate Bill Clinton's percentage of the total votes cast for the Democratic and Republican presidential candidates in 1992; do not include votes for other candidates. Do the same for the 1996 election. Is there a statistical relationship between these two sets of data? Are there any apparent outliers or anomalies? [Observational data; simple regression; extremely strong correlation with a few anomalies.]
17. Select an automobile model and year (at least three years old) that is of interest to you -- for example, a 1993 Saab 900S convertible. Now find at least 30 of these cars that for sale (either from dealers or private owners) and record the odometer mileage (x) and asking price (y). As best you can, try to keep the cars as similar as possible. For example, ignore the car color, but do not mix together 4-cylinder and 6-cylinder cars or manual and automatic transmissions. Estimate the equation $y = a + bx + e$ and summarize your results. [Observational data; simple linear regression; good fit with reasonable coefficients and interesting outliers.]
18. Pick a date and approximate time of day (for example, 10:00 in the morning on April 1) for scheduling nonstop flights from an airport near you to at least a dozen large U.S. cities. Determine the cost of a coach seat on each of these flights and the distance covered by each flight. Use your data to estimate a simple linear regression model with ticket cost the dependent variable and distance the explanatory variable. Are there any outliers? [Observational data; simple linear regression; good fit with reasonable coefficients and interesting outliers.]

19. Go to a large bookstore that has a prominent display of best-selling fiction and nonfiction hardcover books. For each of these two categories, record the price and number of pages for at least ten books. Use these data to estimate a multiple regression model with price the dependent variable and three explanatory variables: a dummy variable that equals 0 if nonfiction and 1 if fiction, the number of pages, and the dummy variable multiplied by the number of pages. Are there any apparent outliers in your data? [Observational data; multiple regression; good fit with reasonable coefficients and interesting outliers.]
20. Ask 100 randomly selected students to estimate their height and the heights of both of their biological parents. Also note the gender of each student in your sample. Now estimate a multiple regression model with the student's height as the dependent variable and the student's gender, mother's height, and father's height as the explanatory variables. [Survey data; multiple regression; good fit with reasonable coefficients and evidence of regression toward the mean.]

A particularly interesting regression project from the Fall 1995 version of the OPRE 404 course examined a proprietary chemical application, with nine independent variables, several dependent variables of interest (in particular, current) and a subset of 937 observations from a 12,000 observation database. The conclusions were somewhat surprising to the investigator, and led to a renewed focus on experimental design within the firm, and a series of follow-up reports by the student over the course of the term.

28 Another regression project, done by a student in Accounting, studied the nature of the relationships among a firm's property, plant, and equipment expenditures over a two-year cycle, expecting that a firm's profitability, age of assets, and ability to make capital expenditures would all be tied in to the fixed asset expenditures made in the following year.

29 In an interesting experimental design project, a sample of 150 toys from the J. C. Penney 1995 Christmas Catalog was studied to see if the price of a toy depended on the age group or gender (0 = either, 1 = female, 2 = male) for which the toy was intended. Data were also collected on the section of the catalog where the toy appeared. A two-way ANOVA studying price as a function of age and gender revealed no clear interaction, and indicated that age alone was a significant factor.

30 A group of students studied a very interesting dataset that was published in the August 20, 1995, issue of the *Cleveland Plain-Dealer* on Northeast Ohio's public school districts. They looked primarily at the effect of the wealth of the community (taken at two levels -- median family income below or above 25,000 dollars) and the percentage of parents who attended college (at three levels -- <40%, 40-60%,> 60%) on the quality points given the school system by the newspaper.

31 Another group studied a wide array of packaged foods (in total, 59) in three categories: cookies, crackers, and salad dressings. In particular, they investigated the idea that the fat and sodium levels of these items would be related to their prices and dietary sugar and carbohydrate levels. The foods were divided into non-fat (0 g of fat), low-fat (according to package labels), or full-fat, and also into lower and higher (more than 10% of U.S. RDA) sodium levels.

32 Some of the students adapted statistical tools to problems in operations management. One student studied total throughput in a manufacturing system in light of the shift time (three 8-hour shifts), as well as the number of packing machines, the number of people working the shift, and the number of supervisors available. The shift factor did seem to have an effect on total throughput, which was primarily explained by the larger numbers of people working on the first shift as opposed to the other two.

33 A more theoretically minded student developed a series of simulations to test the impact of changing arrival time distributions in a first come, first serve queue on the resulting steady state service times in the queue (on average). Exponential arrivals appeared to result in longer service lengths than similarly centered uniform and gamma arrivals.

34 For time series analysis projects, several students looked at data from local firms. One analyzed data from a local high-end sunglasses manufacturer on monthly sales figures (in thousands of dollars) for a three-year period, both to measure the success of the corporation and to assess staffing needs for production of the sunglasses. Another student analyzed the monthly sales from January 1993 through October 1995 of an automotive after-market rear-view mirror adhesive, which turned out to be a highly seasonal product.

35 Two students with interests in finance studied the daily prices of Reebok stock, in an attempt to assess the effect of a series of independent variables. These included economic measures, stock market activity figures, and company earnings. The main results were blasted into space by a series of enormous and especially relentless collinearity problems, and the class suggested a series of exotic transformations of the data which, unfortunately, failed to solve the problem.

36 A student working at the Federal Reserve Bank (FRB) made several attempts to predict the demand M2, also known as velocity (broadly, the money supply), to determine the effects of money demand on the interest rate, controlled by the FRB. The best choice, according to several schemes of model selection, used board opportunity cost, nominal Gross Domestic Product, personal consumption expenditure, effective return on M2, and level of thrift deposits.

37 Three students looked at data from student course evaluation forms for courses in five departments at the Weatherhead School during the Fall 1994 semester. They studied which questions on the form best predicted the overall course and instructor ratings, differences between departments and courses, and the effect of type of instructor (lecturer, Ph.D. candidate, assistant professor, associate professor, or professor) on the ratings.

38 Thanks to the success of the Indians and the disappearance of the artists formerly known as the Cleveland Browns, baseball dominates the conversation of Cleveland sports fans these days, and two students spent a good part of the semester with three different datasets involving baseball players, teams, and their salaries for the 1987 season. This project sparked a good deal of conversation in class, and led the instructor to bring in papers by Lackritz (1990) and, eventually, Hoaglin and Velleman (1995) on the analysis of baseball salaries. Eventually, the class collected more current data on several teams in an attempt to provide a financial justification for the Cleveland Indians' recent successes. The students found the Hoaglin and Velleman paper extremely interesting because it discussed several different analyses of a similar dataset from the 1988 American Statistical Association Data Analysis Exposition, some of which were closely related to the approaches we had developed in class. The students had reached many of the same conclusions on their own, and were very pleased to see their ideas and approaches for tackling the data justified in this way. This project led the students to the World Wide Web and other data sources, and encouraged them to get more involved in studying the underlying issues, particularly in model selection and diagnostics.

Some Projects Done by the Students

Surveys:

1. Favorite color jelly beans; a survey was taken of a random sample of the student body as to their favorite color jelly bean and then separated by gender.

2. Candidate preferences; a survey was taken of a random sample of the student body as to which presidential candidate they had favored before the November 1992 election and which candidate they thought would be doing the best job in the spring of 1993. Relationships between initial candidate preference, likelihood of change and party affiliation were examined.

3. Weight room facilities; a survey was taken of a random sample of users of the new campus weight room as to their satisfaction level. Users were categorized by the extent of their use of the facilities, by gender and by whether they were varsity or non-varsity athletes.

4. Changes in physical condition, eating habits and exercise habits; a random sample of freshmen were surveyed to see if their physical condition or eating and exercising habits had changed from high school to college. Results were tabulated for all freshmen and by gender.

Experiments and Observational Studies:

1. Economy vs. name brand laundry detergent; two different laundry detergents were compared as to their effectiveness in removing stains. An unbiased judge ranked the cleanliness of socks, stained with four different types of stains and then washed in one of the two types of detergents.

2. Mailing times; the length of time letters took to arrive at several different destinations, with and without zipcodes, was measured (this was a suggestion from the textbook). Letters were sent to six different towns in different regions of the country, two with and two without zipcodes.

3. Car manufacturers; the number of foreign and domestic cars passing a given location during a specified time period were recorded. Several similar intersections in different neighborhoods near Philadelphia were observed to see if the geographic location affected the ratio.

Studies based on available data:

1. Church attendance; the attendance records at a particular church were examined for Sundays when Holy Communion was offered and for Sundays when Holy Communion was not offered to see if this had any effect.

2. Fraternity GPAs; the average grade point averages of all the students in a fraternity were compiled by semester and then grouped depending on the semester during which the student pledged. The GPAs were then compared with the average GPA of the general student population.

3. Temperature and homicides; the average high temperature by month for Philadelphia (compiled from newspapers) and the number of homicides per month (obtained from the District Attorney's office) for a three year period were examined to see if there was a correlation between temperature and the number of homicides.

4. Motor vehicle accidents, fatalities and DWI arrests; data from the _Statistical Abstract of the United States_ were used to examine the numbers of registered drivers and the rate of motor vehicle accidents, fatalities, and driving while under the influence arrests in the United States. The data were given for all drivers and then broken down by gender and by different age groups. The data were examined for several different years to see if there were any trends.