Proceedings of the 1972 Conference on Computers in Undergraduate Curricula

A COURSE ON COMPUTING AND STATISTICS FOR SOCIAL SCIENCE STUDENTS

Herbert L. Dershem
Hope College
Holland, Michigan 49423
Telephone: (616) 392-5111

Origin of the Course

Before the introduction of the course to be described in this paper, most economics and social science students at Hope College enrolled in two mathematics courses. These were introductory statistics and computer programming. In the academic year 1969-70 two developments occurred which pointed to the advisability of combining these courses.

First, Dr. Jay Folkert conducted experimental sections of the introductory statistics course at Hope College in which he used the computer as a tool to obtain illustrative information. He did this by allowing students to use previously written programs on prepared data decks. This was done in conjunction with the project to study the use of the computer in statistical instruction sponsored by the National Science Foundation and the University of North Carolina, in which Dr. Folkert was a participant. At the close of the experiment it was Dr. Folkert's feeling that the computer was an asset in such a course but that something was lost because the students were not able to participate in the preparation of the programs.

At the same time, a course on computing for social science students was introduced into the Hope curriculum. This course is basically a FORTRAN programming course. Those who were involved with teaching this course found that some knowledge of statistics would be valuable to those students enrolled. With some statistics background, the students could be assigned projects pertinent to their fields of interest.

It was therefore the consensus of the Hope College mathematics department that it would be more valuable to combine statistics and computer programming into one course for social science students rather than to continue with two separate courses. A proposal was made in the Fall of 1970 to the National Science Foundation for the development of such a course, in conjunction with the development of a laboratory for the mathematical probability and statistics course, and this project was funded.

Value of the Computer

There are three major reasons that the computer is an asset to the introductory statistics course. First, exposure to the computer and computing is a necessary experience for any social science student. He should be aware of the economic, sociological and psychological impacts of the computer as well as the application of the computer to the solution of problems in his discipline.

Second, the computer can serve as an aid in teaching statistics theory. The student has a much greater mastery of a concept after he has explained it to someone else. When a student writes a program he is doing just this, explaining to the computer how to solve the problem. For example, in the assignment shown in Figure 3, the student is asked to explain to the computer how to test hypotheses. Also the computer can be used to provide the student with illustrations which further enhance his understanding. An example of this is found in the assignment given in Figure 2 in which the student is asked to illustrate the normal approximation to the binomial.

Third, the computer allows the student to apply the statistical procedures he is learning to useful sets of data, thus giving him valuable experience in interpreting results and an interesting incentive for learning the statistics.

Description of the Course

The course being developed is entitled "Applied Statistics and Computer Programming." The only prerequisite is high school algebra. It is a two semester course for three hours credit each semester. An outline of the topics presented in the course is found in Figure 1.

The students are introduced to a simplified form of input and output so that they can begin programming without being exposed to FORMAT statements. This is done by subroutines written for this course because Hope College has an IBM 1130 which has no simplified input/output included in the system. The author has prepared notes to serve as a text for the class for the FORTRAN portion of the course because existing texts present the language in a sequence different from that determined to be optimal for this course. For example, we present subscripted variables very early in the course because they are needed to program examples and procedures in descriptive statistics.

First Semester

1. Introduction to computers - Computers, algorithms, lang.
2. Elements of FORTRAN - Constants, variables, assignment statements, transfer statements, simplified input/output.
3. Probability - Rules of probability, counting techniquues
4. Subscripted variables
5. Descriptive statistics - Frequency distributions, mean, standard deviation, other measures.
6. Probability distributions - Discrete, continuous, Chebyshev's inequality, Binomial, Normal.
7. Random sampling - Random number generators, sample means.
8. Subprograms - Functions, subroutines.
9. Other computer languages and systems - COBOL, list processing languages and their applications, time-sharing, man-machine interaction.

Second Semester

1. Input/Output - FORMAT statements.
2. Estimation - Confidence intervals, t-distribution.
3. Testing Hypotheses - Type I and II errors, testing means, testing proportions.
4. Correlation and regression - Multiple linear regression, non-linear regression, interpretation of results.
5. Multiple subscripts
6. Chi-square - Contingency tables, goodness of fit.
7. Analysis of variance - use of library programs.
8. Nonparametric tests - Sign test, rank-sum test.
9. Miscellaneous - Implications of computers, experimental design, simulation.

Figure 1. Outline of the course.

<u>Purpose</u>: The purpose of this assignment is to illustrate how the normal distribution can be used as an approximation to the binomial, when the approximation is good, and how to use the normal table program XNORM.

<u>Description</u>: Write a program which reads n and p and using subprogram BINOM computes the binomial probability that x = k for k = 0,1,...,n, and the normal probability that $k - \frac{1}{2} \leq x \leq k + \frac{1}{2}$ for the normal distribution with mean np and variance np(1-p), and the same values of k. Use XNORM described in Appendix E to find normal probability.

<u>Output</u>: The output is to consist of one line containing n, one containing p, followed by n+1 lines each containing a value of k, the corresponding binomial probability and the normal probability.

<u>Questions</u>: 1. Try out your program for a variety of values of n and p. Do the cases where np and n(1-p) are both greater than 5 show good accuracy? How is the accuracy when the above rule is violated?

2. Does the accuracy of the approximation tend to vary with k for fixed n and p? If so, how?

3. Do you notice that one probability is always larger than the other? Can you explain this?

<u>Extra things to Try</u>: Write a program which is the same as the one described above but which computes the probability that $x \leq k$ instead of x = k. Answer questions 1-3 for this program.

Figure 2. Sample Assignment

Normal Approximation to the Binomial

Purpose: The purpose of this assignment is to introduce the student to testing a hypothesis about a sample mean and illustrate type I and type II errors.

Description: Write a function subprogram which has the following arguments: XMU0, the hypothesized mean, XMU, the actual mean, SIG the actual standard deviation, and N, the sample size. The subprogram is to generate 100 samples of size N from a normal distribution with mean XMU and standard deviation SIG. For each sample, a 95% confidence interval is constructed assuming sigma known, and a test is made as to whether XMU0 is in the confidence interval, i.e., whether μ = XMU0 is accepted. A count is made of the number of times the hypothesis is accepted This is the value to be returned for the function.

Write a calling program which calls this function four times, each time with XMU0=20, SIG=5, N=10, and for XMU=20,22,25,30.

Output: Your output should consist of XMU0, XMU, SIG, N and the value of the function for each call of the function.

Questions: 1. Relate the results of each call to function to either type I or type II errors. Specify which.

2. Punch a card summarizing your results.

3. What would be the effect on your answers if SIG were 10 instead of 5? What if N were 20 instead of 10? What if we used a 99% instead of 95% confidence interval?

4. Compute the theoretical probability of making an error in each of the four performed tests of hypotheses. Indicate for each whether it is a type I or a type II error. Compare these with your results.

5. The above program tests the hypothesis $\mu = 20$ against the two-sided alternative $\mu \neq 20$. How would your answers be changed if a one-sided alternative $\mu > 20$ were used? What if $\mu < 20$ were the alternative?

Extra things to try: Add enough generality to your program that you can try some of the things suggested in questions 3 and 5 above.

Figure 3. Sample Assignment Testing of Hypotheses

The statistics text chosen is *Elementary Statistics* by Paul G.Hoel, and the topics covered follow the presentation of the text with exceptions noted below. Descriptive statistics and probability are reversed in order that the students may gain some familiarity with FORTRAN and subscripted variables before they are needed for descriptive statistics. Some discussion is added concerning random number generators along with experience in their use when random sampling is treated. Also, the students are given practice in using canned subroutines and interpreting their results for regression and analysis of variance.

Five data sets, each consisting of several variables, are stored on a disk. The students learn early in the course how to access this data. These sets are data actually used for research purposes in the areas of sociology, psychology, education and economics, and have been contributed by faculty on the Hope campus from their research and from other books and articles. Already three additional sets have been contributed for next year. Each student is assigned one data set and one variable from that set which he uses throughout the year. This use ranges from finding the mean and standard deviation to taking random samples to obtain confidence intervals to correlation and regression with other variables in the same data set.

The students work fewer textbook problems than in the standard statistics course. Instead, they write computer programs for solving these problems and then apply these programs to their data sets. The students are given a total of 25 assignments involving the computer throughout the year. This averages out to slightly less than one assignment per week. The assignments typically involve the writing of a program, the application of that program, and the answering of some questions intended to bring out the important points. In many cases students are asked to punch cards summarizing their results so that a mean result may be obtained for the entire class. Extra credit problems are given along with each assignment to challenge the better students. Two sample assignments are found in Figures 2 and 3.

Description of the Project

The project for developing this course is of two years duration. In the summer of 1971 this course was organized with the assistance of four senior mathematics majors who wrote the necessary subroutines and programming examples as well as testing the computer assignments. The course is being taught for the first time in the academic year 1971-72 with a starting enrollment of 26 students. The summer of 1972 will be devoted to revising the course according to the experience gained during the preceding academic year. The course will then be taught in revised form during the academic year 1972-73 and the following summer a final report will be made along with the final preparation of materials such as course outline, lecture notes and assignments. These materials will be distributed to allow other staff members to teach the course.

Results

At this writing it is the middle of the first year of the project and hence too early for any firm discussion of results. Thus far the reaction of the students has been most favorable. Some have indicated that they feel the statistics is made easier to understand by the use of the computer. I suspect, however, that there are others for whom the computer simply clouds the issue. More students have been completing the extra credit portion of the assignments than was expected.

The author feels that the morale and interest of the students are much greater in this course than in either of the two parent courses, and for this reason, the course is a pleasure to teach. There has also been a favorable response to this course from our social science faculty who feel it is a most valuable course and are encouraging their students to take it as well as assisting us in its development.