Department of Mathematics
University of Toledo

Master of Science Degree
Comprehensive Examination
Applied Statistics

April 18, 1998

*Instructions:*
Do all four problems.

Show all of your computations.
Books, notes, and calculators *may be used.*
This is a three hour test.

The examination committee tries to proofread the exams as carefully as possible. Nevertheless, the exam may contain misprints. If you are convinced that a problem has been stated incorrectly, mention this to the proctor and indicate your interpretation in your solution. In such cases, do not interpret the problem in such a way that it becomes trivial.

1. (40 points) For this problem, 15 pairs of the variables x and y were simulated in such a way that they are related. The problem asks you to explore the relationships between these variables in several different ways. A substantial amount of MINITAB output is provided to assist you in this journey.

a. We start by analyzing whether or not the distributions are centered in a similar place. The data-generation model described indicates one logical parametric methodology for testing whether or not the mean of x equals the mean of y. Use the statistics provided to perform this test at level of significance equal to .05. Don't forget to consider the important assumptions for the validity of this test.

b. Find a pair of (approximate) 90% simultaneous confidence intervals for the population means of x and y.

c. Do the most appropriate nonparametric test for answering the same basic question. State your model and your hypotheses. Defend your choice of test. Compare your answer to that obtained in part a. If they are substantially different, explain why that may be the case.

d. Use any nonparametric method you wish to find a (roughly) 90% confidence interval for the difference between the median of x and the median of y.

e. Use parametric methods to test whether the correlation of x and y is zero.

f. Test the same basic question as in part e using nonparametric methods. Compare your answer to that obtained in part e. If they are substantially different, explain why that may be the case.

g. Relate your answer in part c to the linear regression shown. Explain how you can tell if the two give the same or different answers.

h. Comment of the quality of the linear regression. Among other things, discuss the validity of using this model for describing the relationship between y and x and of using this model to predict y from x. Also discuss any further steps you would consider taking in modeling y as a function of x. Finally, comment on the meaning of all four plots. What do they say about the validity of the regression? Why is it that I really did not need to plot residuals versus x for this problem?

2. (15 points) A researcher wants to estimate the proportion of residents in a city who would be interested in joining a new and unique kind of health club. She estimates to begin with that only about 2% of the population would likely be interested and wants to end up with an estimate with a (roughly 95%) margin of error of .5% or less.

a. If the city has 100,000 potential joiners, how many should she sample in order to achieve the desired accuracy?

b. Say in addition, she is able to identify a region of the town where the fraction of joiners might be as high as 8%. If this region houses about 1/5 of the total target population, what is her preliminary guess for the fraction of likely joiners in the rest of the town?

c. Use your answer in part b to find the smallest sample sizes from the two regions which will allow for estimation of the whole-population proportion with the stated accuracy. How many fewer subjects are required than for a completely random sample?

**3.** [20 points] According to genetic theory, the seeds collected from a field of pink pea should produce plants with white, pink, and red flowers in the proportion 1:2:1. Of 400 plants grown from such seeds, 93 were white, 211 were pink, and 96 were red. Does this result contradict genetic theory?

**4.** [25 points] Let $k$ be a positive integer. For the $3 \times 4$ contingency table

$$\begin{array}{cccc} k & k & k & k \\ k & k & 0 & k \\ k & k & k & k \end{array}$$

find $k$ such that the hypothesis about independence of two classifications will be rejected at the 0.05 significance level.