

MATHEMATICS 150

(2001-2002)

PROBLEM SET 3

Solutions are **due on Monday, November 19**. Solutions may be submitted in classe or may be delivered to the instructor's office **by 4:00 pm**. *Please remember to print your name on the upper right-hand corner of the front page.*

1. It is difficult and expensive to classify accurately subjects who should be excluded from a particular experimental project. For any subject, a classification procedure could produce a clear 'suitable' classification (correctly or not), an unclear result, or a clear 'unsuitable' classification (correctly or not). Four relatively inexpensive preliminary classification procedures are to be compared for accuracy on the basis of sample testing of 400 subjects *known to be unsuitable*. Procedure A was used with 150 of the subjects; procedures B, C and D were used with 100, 75 and 75 of the subjects, respectively. The procedures were assigned randomly to subjects until the numbers above were obtained. The resulting classifications were stored in a data file **classify.dat** which is available in the usual way from

<http://www.trentu.ca/math/courses/stat/files/classify.dat>

Each line of the file contains the procedure used — A, B, C or D — in column 2 and the resulting classification suitable, unclear or unsuitable, left justified in columns 4 through 13. (You must use an appropriate FORMAT to enter the data in MINITAB.)

Use MINITAB or other software of your choice to display these data with a cross-tabulation display with appropriate detail. Arrange the display so that the classifications are in rows and the procedures are in columns.

2. In the following data, x represents a measure of environment size and y represents a measure of species success.

x	10	14	11	7	24	28	9	17	21	20	4	15
y	39	43	42	29	52	56	36	48	51	49	17	45

Initially, the data are to be analyzed to fit a linear prediction model $y = b_0 + b_1x$.

- a) Plot a scatter diagram of y vs x .
 - b) Does the relationship appear to be linear for the range sampled?
 - c) Calculate the correlation coefficient for x and y . Does it indicate that a linear prediction model is reasonable?
 - d) Determine the least squares regression line for y vs x .
 - e) Predict success values for size values of $x = 5$ and 25.
 - f) Plot the regression line on a scatter diagram of the data.
 - g) Comment on the appropriateness (in terms of accuracy and reality) of predictions obtained by extrapolating beyond the sampling range.
3. Suppose that, for the study in Problem 2, it is decided to use the least squares procedure with transformed data to produce a fitted equation for the curve $y = a_0 + a_1 \ln(x)$
 - a) What are x_{new} and y_{new} ?
 - b) Using the sample data, determine the linear regression line for y_{new} vs x_{new} and plot this fitted line on a scatter diagram of the *transformed data*.
 - c) What is the correlation between y_{new} and x_{new} ?
 - d) Determine the fitted *curve* for the *original data* that results from part b) and plot it on a scatter diagram of the *original data* with the regression line in Problem 2.
 - e) Predict success values for size values of $x = 5$ and 25.
 - f) What extrapolation problem(s) would this model have, if any?
 4. Suppose that, for the study in Problem 2, it is decided to use the least squares procedure with transformed data to produce a fitted equation for the curve $y = a_1 \ln(x)$
 - a) Using the sample data, determine the linear regression line without constant term for y_{new} vs x_{new} and plot this fitted line on a scatter diagram of the *transformed data*.
 - b) Determine the fitted *curve* for the *original data* that results from part a) and plot it on a scatter diagram of the *original data*.
 - c) Predict success values for size values of $x = 5$ and 25.
 - d) What extrapolation problem(s) would this model have, if any?