METHODS QUALIFYING EXAM

AUGUST 2002

INSTRUCTIONS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER LEFT HAND CORNER of EACH PAGE of your exam.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. Be sure to attempt all parts of every question. It may be possible to answer a later part of a question without having solved the earlier parts.
5. Be sure to hand in all of your exam. No additional material will be accepted once the exam has ended and you have left the exam room.

PROBLEM #1

You have been asked to help a researcher design her study of the effect of temperatures on the germination of tomato seeds. The experiment is a two-factor factorial, with combinations of four temperatures (Factor T) and two different types of seeds (Factor S) i.e., eight Atreatment combinations@. A fixed number of seeds will be placed in a growth chamber, with the temperature of the growth chamber set at one of the four temperatures. There is only one growth chamber available for the study. After one hour (I know this is too short a time, but humor me), the percentage of seeds that have germinated is determined. Warning: The descriptions below may include more factors than necessary for the appropriate ANOVA.

Various study conditions are described below. For each description (i) state briefly how you would assign the different levels of factors T and S, (ii) provide the appropriate ANOVA table (sources and degrees of freedom), and (iii) indicate the denominators of the F-tests for testing main effects and interactions for the factors T and S.

(A) For each of six days (Factor D), there are eight time periods (Factor P) available. Seeds of only one type can be in the growth chamber at any one time. The temperature of the chamber may be changed each time period, if necessary. (Total of 48 observations)

(B) Suppose it is inconvenient to change the temperature, so it is preferred that both types of seeds be used at a temperature setting before changing to a different setting. As before, only one type of seed can be in the chamber at any one time. There are still six days (Factor D) with eight time periods (Factor P) per day. (Total of 48 observations)

(C) Another change. There may only be one temperature per day for the chamber, and there are only two periods (Factor P) per day, but now the study will be conducted over 24 days (Factor D). Only one type of seed may be in the chamber at any one time. (Total of 48 observations)

PROBLEM #2

The American Red Cross analyzes blood samples from $n = 100$ randomly selected donors in Texas. The samples were then analyzed and produced a measurement,

X_i = Serum lead level (in micrograms per deciliter) for donor i.

From past experience, the distribution of X_i is a lognormal distribution. Thus, a statistician defined the transformed variable $Y_i = \ln(X_i)$. For purposes of our analysis, we will assume that

$$Y_i = \mu + e_i \tag{1}$$

where μ is a fixed mean parameter for blood donors in Texas and $e_i,\ i = 1, \ldots, 100$ are independent and identically distributed normal random variables with mean zero and variance σ^2. For our sample of 100 donors, some summary statistics for the Y_i values are the sample mean $\bar{y} = 1.36$ and the sample variance $s_y^2 = 1.44$.

(A) Use model (1) and the information above to compute a 95% confidence interval for μ.

(B) Now consider a new observation Y_{n+1} that also satisfies model (1). Compute a 95% *prediction interval* for Y_{n+1}, based on the old data $Y_i,\ i = 1, \ldots, 100$.

(C) Give a customary interpretation of your prediction interval in (B), paying special attention to: (i) what is fixed; (ii) what is random; and (iii) to what probability does the term "95%" refer?

(D) Explain how the interpretation of your prediction interval from (C) differs from the customary interpretation of the confidence interval in (A).

(E) Recall that our transformed observations $Y_i = \ln(X_i)$ were recorded on a log-transformed scale. Use the results from the preceding steps to compute a 95% prediction interval for a new observation X_{n+1} on the *original* (untransformed) scale.

(F) Explain why the procedure you used in (E) to obtain the 95% prediction interval for X_{n+1} would not yield an **exact** 95% confidence interval for $E(X_i)$.

(G) Give a justification for using the procedure in (E) to obtain an **approximate** 95% confidence interval for $E(X_i)$.

(H) The prediction interval in part (B) is based on the assumption that the Y_i values are normally distributed. List *two* specific methods that you could use to check this assumption based on the transformed observations Y_i. For *each* of these two methods, give explicit rules (omitting numerical critical values) you would follow to decide whether the observations are consistent with the normal-distribution assumption.

PROBLEM #3

A study was conducted to understand features of growth curves for boys and girls between age $7-12$. n_1 boys and n_2 girls are randomly selected into the study. The following variables are considered:

Y: height in inches.

X: age.

Z: a binary indicator; 0 for boys and 1 for girls.

It is widely accepted that the best parametric model for growth curves is a quadratic polynomial model. Answer the following questions.

(A) Let β's denote the regression parameters in the mean function. Please write down one general regression model that could be used to describe all boys and girls growth data.

(B) For a regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,$$

we let RSS(X_1, X_2) denote the sum of squared residuals, $\sum_{i=1}^{n}(Y_i - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i})^2$, where $(\widehat{\beta}_1, \widehat{\beta}_1)$ are the least square estimate of the β's. For the following tests, please provide, (i) a model which could be a sub (reduced) model of the model in (A); (ii) H_0 and H_A; (iii) Test statistic; (iv) the proper test procedure which includes the correct specification of the degree of freedoms (if needed).

(a) There are 2 different growth curves for boys and girls.

(b) It is widely believed that the growth of boys and girls have the same starting point. Based on this believe, please test there are 2 different growth curves.

(C) Several assumptions are needed for the test(s) in (B). Please state them. Please also make comments on would large n_1 and n_2 allow each of the assumptions stated be eliminated.

PROBLEM #4

You have been asked to help with the analysis of data from a study using a 3x4 factorial arrangement of four different growth stimulators at three different dose levels on the growth of a particular type of animal. There were three different dosage rates (say R1, R2, and R3) are equally spaced. Each dosage rate by stimulator type combination was feed to six randomly selected animals. The response variable is a measure of growth for a fixed period of time. (Total of 72 observations)

(A) What analysis would you recommend in addition to the usual ANOVA?

(B) Whoops, you learn that R1 = 0, i.e., an absence of any stimulator. Does this change your analysis? If so, how?