Qualifying Exam, Part I
Thursday, January 3, 1991

You have 3 hours. Do as many problems as you can. No one is expected to answer all the problems correctly. Partial credit will be given.

Please put each problem on a separate sheet of paper.

You are not given any tables. Some of the problems request that you carry out hypothesis tests or find confidence intervals. In these problems, clearly indicate what tabled value you need (specify the distribution, tail probability, degrees of freedom, etc.) and how it is to be used.

Problem 1: We are analyzing a standard two-way ANOVA with 5 treatments, 7 blocks, and 3 replications per cell. We suppose that

$$E(Y_{ijk}) = \mu + \alpha_i + \beta_j$$

for $1 \leq i \leq 5$, $1 \leq j \leq 7$ and $1 \leq k \leq 3$ where $\sum \alpha_i = \sum \beta_j = 0$. According to the computer printout $\hat{\alpha}_1 = 4.2$ with a standard error of 1.1, and $\hat{\alpha}_2 = 2.7$ with a standard error of 1.1 , etc.

(a) What is the standard error of $\hat{\alpha}_2 - \hat{\alpha}_1$?
(Give a numerical answer.)

(b) Give a 90% confidence interval for $\alpha_2 - \alpha_1$.
(If you could not find the standard error requested in part (a), just denote this by **SE** in your solution to (b).)

Problem 2: Assume that $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$. Suppose you are given **only** the values $\overline{X} = \frac{1}{m}\sum_i X_i$ and $\overline{Y} = \frac{1}{n}\sum_j Y_j$. How would you construct a 90% confidence interval for μ? Justify your answer.
(If you cannot do this, for less credit you may answer the problem setting $m = n$.)

Problem 3: Answer the following questions.
(The parts can be answered independently of each other. In particular, you do not need to know the answer to (a) in order to do (b), (c) and (d).)

(a) A researcher has collected data on a response variable Y and 15 potential predictor variables $X_1, X_2, \ldots, X_{15}$. It is thought that Y can be well explained by using a subset of the predictor variables. Name two different procedures for identifying a subset of the predictor variables that may do as well or better in predicting Y than the whole group of 15 predictors. Give an advantage and a disadvantage of each method.

(b) Suppose that variables X_3, X_7, and X_{15} are selected. A new sample of 200 observations is obtained and the model

$$Y = \beta_0 + \beta_3 X_3 + \beta_7 X_7 + \beta_{15} X_{15} + \epsilon$$

is fitted. What residual analyses would you perform in order to assure yourself that the usual hypothesis tests for the regression coefficients are valid?

(c) Suppose that these tests indicate no problem with the model assumptions. Now your client is perplexed because the overall F-statistic testing $H_0 : \beta_3 = \beta_7 = \beta_{15} = 0$ is significant at the .0001 level while none of the 3 partial F-statistics testing the individual hypotheses $H_0 : \beta_i = 0$, $i = 3, 7, 15$ was significant at the .05 level. How can this happen?

(d) Relative to predictions made using a regression model and to the estimation of the regression coefficients and error variance σ^2, what are the consequences of ...

(i) including a variable in the fitted model whose true regression coefficient is zero?

(ii) omitting a variable from the fitted model whose true regression coefficient is non-zero?

Problem 4: Let X and Y be independent binomial random variables, each with parameters n and p.

(a) What is the moment generating function of $X + Y$?

(b) Determine the conditional probability distribution of X , given that $X + Y = m$.

Problem 5: Let $X_1, X_2, \ldots, X_n$ be a random sample from the discrete uniform distribution on $\{1, 2, \ldots, p\}$. You have been asked to provide an estimate of p. However, records of the original data are damaged and all that is available is a list of the indices j $(1 \leq j \leq n - 1)$ for which $X_j = X_{j+1}$. Find a consistent estimator of p based on these data and determine its asymptotic distribution as $n \to \infty$. (Hint: first find the natural unbiased estimate of $1/p$.)

Problem 6: If Y_1 and Y_2 are independent standard exponential random variables (with density e^{-y} for $y > 0$), prove that $Y_1/(Y_1 + Y_2)$ has a uniform distribution. (Do not do this by quoting a more general result.)

Problem 7: There are 70 equal-sized plots at Tall Timbers Research Station. On each plot it is recorded whether plant species A is present or not. 37 of the plots have not been burned in 13 years, 22 have been burned a few times, and 11 plots have been burned many times in 13 years. Species A was present on 15 of the 37 plots, on 11 of the 22 plots, and 8 of the 11 plots. Is there evidence that Species A prefers one type of plot over another? (Carry out an appropriate test.)