METHODS QUALIFYING EXAM

JANUARY 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.

QUESTION 1

A crop scientist has asked your help in analyzing her data on plant growth for three different varieties of cotton and four different row spacings ($4", 8", 12"$, and $16"$). Each row spacing was randomly assigned to two plots. Each plot was divided into three subplots, with a variety randomly assigned to each subplot.

(A) One of the response variables is the total yield from each subplot.

i. Give an appropriate model for analyzing these data. For each term, indicate if the term is a fixed effect or random effect.

ii. Give the sources and degrees of freedom for the appropriate ANOVA table. Where appropriate, identify the denominator mean square for the F-test.

iii. Are there any additional test procedures you would recommend? If yes, identify the procedures.

(B) Suppose instead of measuring the yield for the entire subplot, she measured the yield for randomly selected plants within each subplot.

i. If she measured six plants within each subplot, what changes would you make to the analyses in part (A)?

ii. Suppose some subplots have measurements for six plants, others for five plants, others for four plants and others for only two plants. Does this change your recommendations? Does this create any problems for the analyzes? If yes, what are the problems?

(C) Suppose she has six plants for each subplot. In addition to the yield for each plant, she has recorded the plant height. She believes the yield for an individual plant may be linearly related to the plant height.

i. How would you incorporate plant height into your analysis?

ii. How would you determine if it is reasonable to assume the linear relationship between plant height and plant yield is the same, except for perhaps the intercept, for all varieties?

QUESTION 2

(A) Let θ $(0 \leq \theta \leq 1)$ denote the probability of pain relief at dose x of a certain drug. One model for θ is to take $\theta = \int_{-\infty}^{x} f(u)du$ where $f(u)$ is a probability density function. If $f(u)$ represents the extreme value distribution

$$f(u) = \beta \exp[(\alpha + \beta u) - \exp(\alpha + \beta u)], \; -\infty < u < \infty$$

then find $\log\{-\log(1-\theta)\}$.

(B) Let m_i, $i = 1, 2, \cdots, n$ denote the number of patients exposed at dose level x_i, and y_i denote the number of patients relieved from pain at that dose level. Suppose you model the probability of pain relief θ_i at the dose level x_i by a logistic model with dose level (x) as the only explanatory variable with intercept parameter α and the slope parameter β.

In the logistic model we use logistic link function so that $\theta = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$.

i. Write the model with a graphical presentation.

ii. Find the maximum likelihood equations for estimating α and β.

QUESTION 3

A company designs a study to evaluate two methods (M_1, M_2) for converting recycled automobile tires into surfaces for tennis courts. The company wants to compare the average surface traction of the material produced by the two processes. Method M_1 is the conventional method of conversion and Method M_2 is a new method which is more expensive in its conversion of the tires. The company wants to determine if M_2 produces a surface having a higher average traction rating than M_1. Since M_2 is more expensive, the mean traction of M_2 must be at least 5 units larger than the mean for M_1 in order for it to be considered economically feasible. The company decides to take a random sample of material on 50 consecutive days of production from each of the two methods. The data consists of the surface traction measurements of the 50 samples from

$M_1 : X_1, \ldots, X_{50}$ with process mean μ_1 and process standard deviation σ_1

and the surface traction measurements from 50 specimens from

M_2: $Y_1, \ldots, Y_{50}$ with process mean μ_2 and process standard deviation σ_2.

(A) The company is interested in the research hypothesis $\mathrm{H}_1 : \mu_1 + 5 < \mu_2$.

i. Write down a general formula for the t test statistic for this hypothesis test (It is presumed that M_2 will produce a product having a more consistent surface traction than M_1.).

ii. Write down the decision rule for this hypothesis test. Use $\alpha = 0.05$.

iii. State the necessary conditions needed for your procedure to be valid and how you would verify whether the conditions are satisfied in this experimental setting.

(B) In the context of the hypothesis test presented in (A), give clear, explicit definitions of the following terms, **Make Sure to Frame Your Definitions in Terms of This Specific Problem**

i. Type I error

ii. Type II error

iii. Power of the test.

(C) For parts (C) and (D) of this question, you may assume that $\sigma_1 = 3$ and $\sigma_2 = 1$ and that the sample sizes are large enough to invoke the central limit theorem if necessary.

i. Calculate the power of your test for the following six values of the parameter:

$$\mu_2 - \mu_1 = \quad 4.5, \quad 5.0, \quad 5.5, \quad 6.0, \quad 6.5, \quad 7$$

ii. Use your results from (C.i) to sketch a power curve for your test. Be sure to label your axes clearly.

(D) The company's engineer examines your results from (A) through (C) states, "The power of the test when $\mu_2 - \mu_1 = 5.5$ is not large enough. Determine the minimum sample size necessary to achieve a power of at least 0.90 when $\mu_2 - \mu_1 \geq 5.5$.

(E) The 50 observations from each process represent the surface traction obtained from a single batch of production. Thus, there may be a strong *positive* correlation within the 50 surface traction measurements from a given process. However, the measurements between the two processes remain independent. Given this additional information, answer the following questions.

i. If the correlations between all pairs of daily measurements from method M_1 are equal to $\rho_1 > 0$ and the correlations between all pairs of daily measurements from method M_2 are equal to $\rho_2 > 0$, how does this positive correlation between the daily measurements affect the estimated standard error of $\hat{\mu}_1 - \hat{\mu}_2$? Justify your answer mathematically.

ii. Suppose you **did not** adjust for the positive correlation between the daily measurements and proceeded to use the test you proposed in part (A). Will the positive correlation in the data increase or decrease the numerical values of power you calculated for the test statistic in part (C)? Explain.

iii. Suppose you **did not** adjust for the positive correlation between the daily measurements and proceeded to obtain a 95% C.I. for μ_2 using procedures for independent random samples. What is the effect of the positive correlation in the data on the level of confidence of your C.I.? What is the effect of the positive correlation in the data on the width of your C.I.?

iv. Suppose that it is known that $\rho_1 = \rho_2 = .9$ and you use this information to adjust your test statistic to account for the positive correlation.

Taking into account the effect of the positive correlation on the standard error of $\hat{\mu}_1 - \hat{\mu}_2$ and assuming that $\sigma_1 = 3$, $\sigma_2 = 1$, $\rho_1 = \rho_2 = .9$ and that the sample sizes are large enough to invoke the central limit theorem:

i. Calculate the power of an $\alpha = 0.05$ test for the following six values of the parameter:

$$\mu_2 - \mu_1 = \quad 4.5, \quad 5.0, \quad 5.5, \quad 6.0, \quad 6.5, \quad 7$$

ii. Use your results to sketch a power curve for the adjusted test. Compare this curve to the curve from (C.ii).

QUESTION 4

(A) A data set contained 42 observations. Each observation consisted of a value of a response varible, Y, and four predictor variables: X1,X2,X3, and X4. The table shown below gives the Error Sum of Squares (SSE) for each of the possible models of from 1 to 4 of the predictor variables. (An X indicates that the predictor variable is in the model. All models include an intercept term.)

Model	X1	X2	X3	X4	SSE
1	X				1062.92
2		X			2431.72
3			X		1941.64
4				X	3530.88
5	X	X			1055.03
6	X		X		968.99
7	X			X	1020.51
8		X	X		1923.71
9		X		X	2322.45
10			X	X	1931.20
11	X	X	X		896.57
12	X	X		X	1009.58
13	X		X	X	958.85
14		X	X	X	1853.74
15	X	X	X	X	884.51

i. Use the information in the above table to calculate the numerator and denominator sum of squares for the F-statistics to test the null hypotheses listed below:

$H_o : \beta_3 = 0 | \beta_o, \beta_1, \beta_2, \beta_4 \neq 0$

$H_o : \beta_1 = 0, \beta_4 = 0 | \beta_o, \beta_2, \beta_3 \neq 0$

$H_o : \beta_2 = 0 | \beta_o, \beta_1, \beta_4 \neq 0, X_3$ not in the model

ii. Suppose you want to test $H_o : \beta_3 = 3, \beta_2 = -2\beta_4 | \beta_o, \beta_1 \neq 0$. Indicate the appropriate numerator and denominator mean square for the F-statistic. If the appropriate SSE is not in the table, explain how you would find the appropriate SSE.

(B) The ANOVA table and information about the coefficients for a data set with seven predictor variables are shown below. Use this information to answer the questions following the tables.

Model	Sum Squares	df	Mean Square	F	Significance
Regression	31210146	7	4458592.252	20.187	.000
Residual	7509549	34	220869.100		
Total	38719695	41			

	Unstandarized Coefficients		Stand.Coefficient			Collinearity Sta	
Term	$\hat{\beta}_i$	Std. Error	$\hat{\beta}_i'$	t	Significance	Tolerance	
Constant	-151.869	348.256		-.436	.666		
X1	-4.24E-05	.000	-.059	-.314	.755	.160	
X2	6.321	.972	.678	6.501	.000	.524	
X3	.622	.545	.267	1.141	.262	.104	
X4	-.884	14.911	-.015	-.059	.953	.084	
X5	-1.848	2.563	-.090	-.721	.476	.365	
X6	1.268E-03	.007	.036	.195	.847	.166	
X7	17.875	10.720	.179	1.668	.105	.497	

i. From the information in the above table, can you determine that both X4 and X6 may be deleted from the model, using a significance level of 0.10? If not, why not?

ii. If BACKWARD variable selection is to be used, with significance level 0.10, which variable would be first deleted? Why?

iii. Is there an indication of multicollinearity? If so, which variables do you suspect? Support your answer with results shown in the tables.