STATISTICS DEPARTMENT
M.S. EXAMINATION

PART II
OPEN BOOK

Tuesday, May 18, 2004

9:00 a.m. - 1:00 p.m.

College of Science Telecom Lab, SC S138

Instructions: **Complete *only four of the five* problems. Each problem counts 25 points. Unless otherwise noted, points are allocated approximately equally to lettered parts of a problem. Spend your time accordingly.**

The web site address for data and program files for this exam is:

http://www.sci.csuhayward.edu/~esuess/msexam/

Begin each problem on a new page. Write the problem number and the page number in the specified locations at the top of each page. Also write your chosen ID code number on every page. Please write only within the black borderlines, leaving at least 1" margins on both sides, top and bottom of each page. Write on one side of the page only.

At the end of this part of the exam you will turn in your answer sheets, but you will keep the question sheets and your scratch paper. Please put your answer sheets in the proper order, and do not submit answers to more than four problems.

You may use a computer to work any of the problems, but your answers must be handwritten on standard paper provided for the examination. Printers may *not* be used during the exam, and pages printed out by computer may *not* be submitted. As indicated, some problems have data files available on disk.

1. We wish to establish that women like the movie Titanic more than men do (based on population means). The variable of interest is 'titanr', which is the rating of the movie on a scale from 0-4. We also keep track of 'gender', which takes two values 0 = male and 1 = female. Suppose we take a random samples of n_1 women and n_2 men. For large n_1 and n_2, assume that the sample standard deviation s_1 of the women is approximately .5 and that the sample standard deviation s_2 of the men is approximately .8.

(a) Explain why, for large n_1 and n_2, the random variable

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has approximately a standard normal distribution. (Here the $\bar{X}$'s refer to sample means for women and men, and μ's correspond to the population means for women and men.)

(b) Using the appropriate hypothesis test and supposing that $n_1 = 52$ and $n_2 = 39$, explain in terms of $\bar{X}_1 - \bar{X}_2$ when it can be proven that women like the movie more than men do. Use the Type 1 error probability $\alpha = .025$.

(c) Still assuming that $n_1 = 52$ and $n_2 = 39$ and also assuming that $\Delta = \mu_1 - \mu_2 = .36$, obtain the probability that we can establish that women like Titanic more than men. That is, find the **power** at $\Delta = \mu_1 - \mu_2 = .36$.

(d) Assume that $n_1 = n_2 = n$. How large must n be so that the power at $\Delta = \mu_1 - \mu_2 = .36$ is .975?

(e) Refer to the file 'rating' (saved in 'mtw', 'xls' and 'dat' formats). In this file $n_1 = 52$, and $n_2 = 39$, with 3 missing observations (denoted by '*'). That is, there are 94 observations altogether. At $\alpha = .025$, can one conclude that women like Titanic more than men?

2. A pharmaceutical company has three possible Processes (A, B, and C) for making a drug. One question is whether the three processes yield the same potency. Also, a key ingredient for making the drug (by any process) is delivered to the company in Batches. Another question is whether Batches vary in a way that affects potency. To try to answer these questions, an experiment is performed in which each Process is run with each of ten randomly chosen Batches. Potencies are determined for two samples from each of the resulting 30 runs.

The data are shown below and are available in the files DRUG.TXT (plain text) and DRUG.MTP (Minitab portable). In each file there are six rows of ten observations each, corresponding to the format of the table below.

	Batch									
Process	1	2	3	4	5	6	7	8	9	10
A	570	556	577	615	644	618	581	670	632	613
	570	555	575	614	641	616	582	669	633	612
B	564	545	559	579	608	597	571	641	598	602
	566	542	560	578	607	597	571	641	598	601
C	525	511	512	535	569	547	522	600	586	560
	524	509	512	535	570	547	522	600	587	559

(a) Write the most complete possible ANOVA model for this experiment. What kind of ANOVA is this? Say whether each factor is fixed or random and whether there is any nesting of factors. Define your symbols. Give ranges of subscripts, restrictions on parameters, and distributions of random variables.

(b) Give the ANOVA table corresponding to your model in part (a). Say which effects are significant.

(c) Two questions were mentioned in the statement of the problem. Try to answer them in terms meaningful to a manager at the company. Does any one of the three Processes yield significantly higher potency than the other two? Compared with random error, is variability among batches a significant source of variability in potency? Discuss interaction.

(d) Perform appropriate diagnostics to test model assumptions. Why is the distribution of residuals perfectly symmetrical? Why do the residuals take so few values?

3. Let X be a random variable with the probability function p, where $p(0) = P[X = 0] = 0.2$, $p(1) = P[X = 1] = 0.1$, and $p(2) = P[X = 2] = 0.7$. In this problem assume that $X_{n,k}$'s are independent random variables with the same probability function as X.

Let $Z_1 = X_{0,1}$.

If $Z_1 = 0$, then let $Z_2 = 0$.

If $Z_1 > 0$, then let $Z_2 = X_{1,1} + X_{1,2} + \cdots + X_{1,Z_1}$.

If $Z_2 = 0$, then let $Z_3 = 0$.

If $Z_2 > 0$, then let $Z_3 = X_{2,1} + X_{2,2} + \cdots + X_{2,Z_2}$.

Continuing, for $n = 1, 2, \ldots$,

If $Z_n = 0$, then let $Z_{n+1} = 0$.

If $Z_n > 0$, then let $Z_{n+1} = X_{n,1} + X_{n,2} + \cdots + X_{n,Z_n}$.

(a) Find $P[Z_1 = 0]$.

(b) Find $P[Z_2 = 0]$.

In what follows you may take the following equation for granted for positive integers n:

$$P[Z_{n+1} = 0] = 0.2 + .1P[Z_n = 0] + 0.7P[Z_n = 0]^2.$$

(c) Find $P[Z_3 = 0]$.

(d) Find $P[Z_5 = 0]$. You may use a computer.

(e) Find $\lim_{n\to\infty} P[Z_n = 0]$ either numerically or using a theorem.

4. Budgeting difficulties at the University are a substantial problem for all. The table below contains information taken from the University websites along with some assumptions about data that is at this moment not available.

Spendable budget[1]	Academic year ending	Student Fees	Academic affairs budget	Student enrollments (FTES)	Faculty (FTEF)
111,278,162[2]	1998	34,775,921	76,543,885[3]	9,813.56	517.06
126,556,396[4]	1999	39,933,460	81,978,123	9,677.51	528.99
110,086,347	2000	29,819,693	67,247,637	9,671.24	545.89
117,102,515	2001	32,550,290	71,614,421	10,244.12	557.69
116,640,684	2002	36,262,684	71,446,807	10,611.42	557.69
126,144,561	2003	43,616,461	74,276,309	10,779.37	546.923
124,594,749	2004	47,993,049	74,823,779	10,779.37	526.466

a. Using SAS® create a multiple regression model for spendable budget based on the information available.

b. Comment on any surprises that you see. Evaluate the fit of the model or suggest a way to assess the fit at a later time. Be sure to discuss multicollinearity.

c. Comment on the appropriateness or lack thereof of the model assumptions, particularly independence, and suggest ways to handle any assumption failures.

d. The value for the 2004 spendable budget is derived from the Governor's budget put out in January. Use your model to predict the spendable budget in 2005-06, including 95% confidence limits for the prediction. Assume that all values in 2004-05 are held at the 2004 level.

---

[1] From Budget report except where noted.
[2] Estimated from Financial Report 1999.
[3] From Instruction and Instructional Support in Financial Report.
[4] Estimated from Financial Report 2000.

5. Suppose there is a medical diagnostic test for a type of cancer. The *sensitivity*, η, of the test is 0.90. This means that if a person has cancer, the probability that the test returns a positive result is 0.90. The *specificity*, θ, of the test is 0.95. This means that if a person does not have the disease, the probability that the test returns a negative result is 0.95, or that the *false positive*, $1-\theta$, rate of the test is 0.05. In the population, 1% of the people are infected with this type of cancer.

   (a) What is the probability that a person tested has cancer, given that the result of the test is positive?

   (b) Does this show that screening is effective in detecting this cancer?

   (c) Beyond what rate of infection in the population would a majority of those testing positive be infected? That is, for this test what value of $P(D)$ would give $P(D|+) \geq 0.50$? Find a general solution to the problem in terms of the parameters η, θ, $P(D)$, and $P(\overline{D}|+)$.

   (d) Using the following S-Plus code, verify your answers to (a) and (c).

```
# test parameters
eta <- 0.90
theta <- 0.95

# prior probability of disease in the population
p.prior <- 0.01

# Bayes' Rule
p.post <- (p.prior*eta)/(p.prior*eta+(1-p.prior)*(1-theta))

# answer
p.post

# plot p.post versus p.prior
p.prior <- seq(0,1,0.001)
p.post <- (p.prior*eta)/(p.prior*eta+(1-p.prior)*(1-theta))
plot(p.prior,p.post)

# Find the p.prior value for a posterior of 0.5
p.prior[round(p.post,digits=2)==0.50]
```

   (e) What values of sensitivity and specificity would be needed to detect a majority of those testing positive to be infected if 1% of the population were infected? (Hint: Change the sensitivity and specificity in the program above and find values that give a posterior $P(D|+) \geq 0.50$.)