

Applied Statistics Qualifier Examination  
January, 2010

**Instructions:**

- (1) The examination contains 4 Questions. You are to **answer 3 out of 4** of them.
- (2) You may use any books and class notes that you might find helpful in solving these problems.
- (3) You are on the *double honor system*. **Do not consult with any individual** about this examination until after the examination is over. With a double honor system, you are required to only consult books, journals and class notes for help in solving these problems *and* you are required to report any violations of this policy that you observe in any other student who is taking this test.
- (4) **Print your name and the problem number on every page you turn in. Number every page and staple pages for each solution together. Then use a paper clip to attach all your solutions to this cover sheet. Place the entire file in an envelope.**
- (5) Please **fill in the information below** and submit your solutions in person to Dr. Nancy Mendell in Math Tower Room 1-111 by FRIDAY, January 29 at 2 PM or 24 hours after you receive the exam (should you receive it early or late for some special reason).

**Should you have any questions while doing this exam** please feel free to email or phone Professor Mendell ([nancy.mendell@stonybrook.edu](mailto:nancy.mendell@stonybrook.edu) ; 631 6328373, 631 6899059 or 5162972970 ).

**Please be sure to fill in the appropriate information below:**

I am submitting solutions to QUESTIONS \_\_\_\_\_, \_\_\_\_\_, and \_\_\_\_\_ of the applied statistics  
qualifier examination. There are \_\_\_\_\_ pages of written solutions.

**Please read the following statement and sign below:**

This is to certify that I have taken the applied statistics qualifier and have used no other person as a resource nor have I seen any other student violating this rule.

\_\_\_\_\_  
(Signature)

\_\_\_\_\_  
Print your name here.

### QUESTION 1

A randomized block design is used to compare four treatments in six blocks.

Block	Treatment			
	1	2	3	4
1	89	81	84	86
2	91	86	87	88
3	85	84	83	81
4	93	83	84	82
5	86	78	83	84
6	89	92	88	90

- (a) Use the Friedman test to detect differences in location among the four treatment distributions with confidence level 0.05. Give the value of the Friedman statistic.
- (b) Find the approximate p-value for the test in part (a). What is your conclusion?
- (c) Perform an analysis of variance and give the ANOVA table for the analysis.
- (d) Give the value of the F-statistic for testing the equality of the four treatment means. Find the p-value for the F-statistic.
- (e) Compare the two p-values, and explain the implications of your comparison.

## QUESTION 2

Suppose we have two independent random samples from two normal populations, that is,  $X_1, X_2, \dots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ , and  $Y_1, Y_2, \dots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ .

Furthermore, it is known that  $\sigma_2^2 = 2\sigma_1^2$  although the true values of these variances are unknown.

- (a) Please derive the exact  $100(1-\alpha)\%$  confidence interval for  $(\mu_1 - \mu_2)$ . Please include the entire derivation for full credit.
- (b) Suppose the total sample size  $n = n_1 + n_2$  is fixed due to budget constraints. Let  $n_1 = np$ ,  $n_2 = n(1-p)$ , where  $0 \leq p \leq 1$ . Please derive the optimal value of  $p$  that will yield the shortest  $100(1-\alpha)\%$  confidence interval for  $(\mu_1 - \mu_2)$ . Hint: For simplification, you may approximate the sample variances using the corresponding population variances in the final stages of your computation.

### QUESTION 3.

The attached data is simulated. A patient who is at risk of developing a disease has blood and tissue drawn. The amount of a protein in the patient's blood is an important precursor of the disease and is recorded as the variable  $Y$ , the rightmost variable in the data file. The leftmost column is a number indicating the patient's identification and, in principle, should not be related to  $Y$ . The patient ID is between 1 and 800. The concentration of three potentially toxic chemicals in the patient's tissue is measured and recorded as variables  $X_1$ ,  $X_2$ , and  $X_3$ . These variables are recorded in the second through fourth columns from the left. There are eleven genes that might be associated with  $Y$ . If the patient has a genotype on gene  $i$  that puts the patient at risk, the variable  $G_i$  is scored as 1; if the patient is not at risk,  $G_i$  is scored as zero. The values of these indicators are given in the next eleven columns. The genes are sorted by the prevalence of the at risk genotype.

The dependent variable may be associated with any gene, any chemical concentration, any interaction of chemical concentration and gene, and any interaction of gene and gene. Analyze the data. In a short report, summarize each step in your data analysis and briefly discuss your findings on how the explanatory variables affect the dependent variable. Make sure that you clearly label and fully discuss your final model.

#### QUESTION 4

Below we give the results of a study which evaluates the chromosomal status of the fetus “C”, normal or trisomy 21, with respect to 2 factors (1) “S” Maternal smoking; mother currently smoking vs. mother not smoking and (2) “A” Maternal age; Age of mother in years.

Maternal Smoking (S)	Maternal Age (A)	Chromosome Trisomy 21	Status of fetus (C) Normal
Smoker	15-19	5	43
Nonsmoker	15-19	4	27
Smoker	20-24	6	81
Nonsmoker	20-24	20	76
Smoker	25-29	10	64
Nonsmoker	25-29	32	99
Smoker	30-34	13	40
Nonsmoker	30-34	24	80
Smoker	35-39	10	13
Nonsmoker	35-39	23	46
Smoker	40-44	5	7
Nonsmoker	40-44	7	11

- (a) Make some graphs of the  $\ln(\text{odds})$  of trisomy vs. age for each smoking status group. From these graphs conjecture whether you believe that there is an age and smoking interaction and whether you think that the assumption of linearity with age applies for one or both smoking categories.
- (b) Use logistic regression to fit a model for predicting the odds of trisomy 21 includes only main effects of age and smoking. In this first attempt consider age as a categorical variable. Test for the presence of a two way interaction between smoking and age on trisomy. If a two way interaction is significant then consider the smoking groups separately.
- (c) If there is no significant two way interaction test for main effects of age and smoking: otherwise test for a main effect of age for each smoking group.
- (d) Consider age as a quantitative variable and test for goodness of fit to a logistic regression where age is a linear predictor of trisomy 21. If smoking is a significant main effect then include it in the model. If you are considering the two smoking groups separately then test the fit to a linear predictor.
- (e) Repeat (b) and (c) using loglinear models. In this case we have a third variable chromosome 21 status : trisomy or normal. We also first test for a 3 way association followed by testing for two way associations in the case of no significant 3 way association. Indicate what added information is obtained using loglinear models about the relationship between maternal age and smoking.

- (f) Write a short report on what you see as the important factors for trisomy 21. Where you observe that a statistically significant factor (and/or and interaction between factors) as a predictor of trisomy 21 interpret the relevant odds ratio(s).