Applied Statistics Qualifier Examination
SPRING 2005

Instructions:

(1) The examination contains 4 problems. You are to answer 3 out of 4 of these questions.

(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 to 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 solutions 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 Wednesday, January 26 at 2 PM or 24 hours after you receive the exam (should you receive it early or late for some special reason).

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.

(Signature) ______________________________________________________

Print your name here.
Question 1.
The following experiment was designed to find out to what extent a particular type of fabric gave homogeneous results over its surface for a standard wear test. In a single run the test machine could accommodate four samples of fabric, at positions 1, 2, 3, and 4. On a large sheet of the fabric four areas A, B, C, and D were marked out at random at different places over the surface. From each area 4 samples were taken, and the 16 samples thus obtained were compared in the machine with the following results, given in milligrams of wear:

<table>
<thead>
<tr>
<th>Position</th>
<th>Run I</th>
<th>Run II</th>
<th>Run III</th>
<th>Run IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A 26.1</td>
<td>D 26.0</td>
<td>C 22.4</td>
<td>B 26.7</td>
</tr>
<tr>
<td>2</td>
<td>B 25.3</td>
<td>A 26.8</td>
<td>D 21.5</td>
<td>C 35.0</td>
</tr>
<tr>
<td>3</td>
<td>C 29.1</td>
<td>B 22.4</td>
<td>A 26.3</td>
<td>D 22.8</td>
</tr>
<tr>
<td>4</td>
<td>D 30.6</td>
<td>C 28.8</td>
<td>B 25.2</td>
<td>A 26.4</td>
</tr>
</tbody>
</table>

What type of design is this? Analyze the data, stating carefully your assumptions.
Question 2.  
In a random sample of 498 motorists on a major toll road, each driver was classified according to the following categorical variables:  

1. Transportation status \( (T) \) = number of passengers including the driver.  
   Values = 1, 2, 3, 4+  
2. Destination area of the city for the driver \( (D) \). Values = NE, SE, NW, SW.  
3. Public transportation use anticipated \( (P) \). Values = Yes, No.  

\[
\begin{array}{cccccccc}
T & Y & N & Y & N & Y & N & Y & N \\
--- & --- & --- & --- & --- & --- & --- & --- & --- \\
1  & 7 & 10 & 15 & 22 & 31 & 27 & 15 & 7 \\
2  & 4 & 8 & 13 & 14 & 18 & 16 & 9 & 4 \\
3+ & 13 & 10 & 21 & 29 & 12 & 13 & 4 & 3 \\
4+ & 28 & 17 & 32 & 35 & 15 & 18 & 16 & 12 \\
\end{array}
\]

(a) Make a graph which shows the IR (odds) of using public transportation with respect to destination and transportation status. Interpret your graph in terms of effects of \( T \) and \( D \) on \( P \) and also interactions between \( T \) and \( D \). 

(b) Use logistic regression to evaluate whether transportation status, destination or the interaction of these two is related to use of public transportation. 

(c) Interpret your findings in terms of (i) the relationship between the number of passengers and intention to use public transportation, (ii) destination and use of public transportation and (iii) non-uniform effects of the number of passengers and intention to use public transportation. 

(d) Use log-linear models to test for partial association between destination and the number of passengers. What do you conclude? 

(e) Pretend you have been hired to report to the transportation commissioner on the relationships between these three variables. Write a brief report using odds ratios where they are relevant. Compare \( D=SE, NW, \) and \( SW \) to \( D=NE \) and \( T=1, 2, 3 \) to \( T=4+ \).
QUESTION 3.

An experiment was conducted to compare three methods of grass seed preparation: mechanical scarification (Ms), acid dip (Ad) and hot water dip (Hw). One hundred seeds were put in each of 150 petri dishes. Among the 150 dishes 50 were assigned to Ms, 50 to Ad and 50 to Hw. After a period of two weeks the germination rates in each dish was recorded. The summary statistics are given below:

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean rate</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms</td>
<td>65.3</td>
<td>7.2</td>
<td>50</td>
</tr>
<tr>
<td>Ad</td>
<td>82.1</td>
<td>5.4</td>
<td>50</td>
</tr>
<tr>
<td>Hw</td>
<td>73.1</td>
<td>6.5</td>
<td>50</td>
</tr>
</tbody>
</table>

(a) Test a hypothesis of equal within group variance for the three treatments.
(b) Use the information here to obtain the mean germination rate for all 150 petri dishes.
(c) Use the information here to compute the pooled estimate of variance or MSW.
(d) Use (b) and (c) to obtain analysis of variance table.
(e) Suppose the goal of the study was to compare Ad and Hw to Ms. Do the appropriate multiple comparisons procedure and indicate your conclusions.
(f) Write a summary of your findings. Include 95% confidence interval estimates of the difference between mean germination rates for (i) Ms and Ad (ii) Ms and Hw.
Question 4
A subject 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 should not be related to Y. The amount of three potentially toxic chemicals in the patient’s tissue is measured and recorded as variables X1, X2, and X3. These variables are recorded in the second through fourth columns from the left. There are five genes that are hypothesized to affect Y. If the subject has a gerotype on gene i that puts the subject at risk, the variable Gi is scored as 1. If the subject is not at risk, Gi is scored as 0.

Analyze the data. In a report less than 4 pages, summarize each step in your data analysis and briefly discuss your findings on how the explanatory variables affect the dependent variable.

Here is a partial listing of the data sent to you via email as "Q4data_s2005.xls":

```
   Y  X1  X2  X3  G1  G2  G3  G4  G5
----------  ----------  ----------  ----------  ----------  ----------  ----------  ----------  ----------
  2.00  1.50  2.00  1.75  0.00  0.00  0.00  0.00  0.00
  3.00  2.00  2.50  2.25  0.00  0.00  0.00  0.00  0.00
  4.00  2.50  3.00  2.75  1.00  1.00  1.00  1.00  1.00
  5.00  3.00  3.50  3.25  0.00  0.00  0.00  0.00  0.00
  6.00  3.50  4.00  4.25  0.00  0.00  0.00  0.00  0.00
  7.00  4.00  4.50  5.25  0.00  0.00  0.00  0.00  0.00
  8.00  4.50  5.00  6.25  0.00  0.00  0.00  0.00  0.00
  9.00  5.00  5.50  7.25  0.00  0.00  0.00  0.00  0.00
```

---

5