The written Master’s Examination

Option Probability and Statistics, Option I

FALL 2004

Full points may be obtained for correct answers to 8 questions. Each numbered question (which may have several parts) is worth the same number of points. All answers will be graded, but the score for the examination will be the sum of the scores of your best 8 solutions.

Use separate answer sheets for each question. DO NOT PUT YOUR NAME ON YOUR ANSWER SHEETS. When you have finished, insert all your answer sheets into the envelope provided, then seal and print your name on it.

Any student whose answers need clarification may be required to submit to an oral examination.
1. (Math 310)
Let A be a real m×n matrix with rank m. Show that the matrix AAᵀ is positive definite. What can we say about AᵀA?

2. (Math 413)
Suppose \{a_n : n ≥ 1\} and \{b_n : n ≥ 1\} be sequences of non-negative real numbers such that
lim_{n→∞} b_n = b and limsup_{n→∞} a_n = a, where both a and b are positive real numbers. Show that
limsup_{n→∞} a_n b_n = ab.

3. (Stat 401)
Suppose the p.d.f. of a random variable X is given by

\[ f_X(x) = \begin{cases} \frac{1}{2} x^2 e^{-x^2}, & \text{if } x > 0 \text{ and } f_X(x) = 0, & \text{if } x ≤ 0. \end{cases} \]

Suppose that given X=x (x>0), the random variables Y₁, Y₂ are independent, identically distributed and the conditional p.d.f. of each of them is as follows:

\[ g(y \mid x) = \begin{cases} \frac{1}{x}, & 0 < y < x, \text{ and } g(y \mid x) = 0, \text{otherwise}. \end{cases} \]

Find the conditional expectation E(X|Y₁=y₁, Y₂=y₂).

4. (Stat 411)
Let \(X₁, X₂, \ldots, Xₙ\) be a random sample of size n from a distribution with p.d.f.:

\[ f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0, \text{ zero elsewhere}. \]

Find the minimum variance unbiased estimator of \(e^{-3/\theta}\).

5. (Stat 461)
Consider a branching process whose offspring follow the geometric distribution

\[ P_k = (1-c)c^k, \quad k = 0,1,\ldots, \text{ where } 0 < c < 1. \]

Determine the probability of eventual extinction when: (i) \(X₀ = 1\); (ii) \(X₀ = 2\).

6. (Stat 411)
Let \(X₁, X₂, \ldots, Xₙ\) be a random sample from \(N(\mu, \theta)\), where \(\mu\) is the expectation and \(\theta\) is the variance, \(\mu, \theta\) both unknown. Derive the likelihood ratio test for \(H₀: \mu = 0\) versus \(H₁: \mu ≠ 0\) and show that it can be expressed as a two-tailed test based on a t-statistic. What is the degree of freedom of the t?
7. (Stat 416)
Four different fabrics are compared on a Martindale wear tester that can compare four materials in a single run. The weight loss (in milligrams) from five runs is measured and the following results are obtained.

<table>
<thead>
<tr>
<th>Run</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fabric A</td>
<td>36</td>
<td>17</td>
<td>30</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>Fabric B</td>
<td>38</td>
<td>18</td>
<td>39</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>Fabric C</td>
<td>36</td>
<td>26</td>
<td>41</td>
<td>38</td>
<td>28</td>
</tr>
<tr>
<td>Fabric D</td>
<td>30</td>
<td>17</td>
<td>34</td>
<td>33</td>
<td>21</td>
</tr>
</tbody>
</table>

Our main interest is in comparing the four fabrics.

(a) Describe the basic model of the Friedman test, give its test statistic S, and compute S.

(b) Explain, why the Friedman test is here more appropriate than the Kruskal-Wallis test.

8. (Stat 431)
1-Define the margin of error in survey sampling and indicate what will be the margin of error if we carry a survey under SRS (1000, 49) for a binary survey variable with unknown percentage \( \pi(+) \) and \( \pi(-) \) in case our sample contains 20 pluses and 29 minuses. Indicate all the assumptions that you made in your calculations.

2- How would you compute the variance of the HTE \( \pi(+) \) and its estimator for the sampling plan in 1- above. Also put a upper bound on your variance and variance estimator.

9. (Stat 481)
Ninety graduating engineers were classified according to two categories: GPA (low, average, high) and starting salary (low, high). The data are given in the table below:

<table>
<thead>
<tr>
<th>Salary</th>
<th>GPA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Average</td>
</tr>
<tr>
<td>Low</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>High</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>
At the 5% level of significance, are the two ways of classification independent? Use $\chi^2(0.05;2) = 5.991$.

10. (Stat 481)

For the simple linear regression model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, we compute, from $n = 16$ observations, the following summary statistics:

$$\hat{\beta}_1 = 0.35, \text{MSE} = 2.3, \sum (x_i - \bar{x})^2 = 100.$$ 

(i) Based on the ‘standard’ assumptions, construct a 95% confidence interval for $\beta_1$ 

[Given: $t(0.05;14)=1.761$]

(ii) What are the ‘standard’ assumptions?

(iii) Is the slope of the regression line significantly different from zero (at the 5% level of significance)? Justify your answer.
Problem: Linear Algebra: Let $A = (a_{ij})_{m \times n}$ with rank $m$.

The matrix $AA^T$ is an $n \times n$ matrix: $m \times n \times n \times m = m \times m$.

Let $AA^T$. Consider $(x^T AA^T x) = (x^T A A^T x) = (A^T x)^2 \geq 0$.

Thus $AA^T$ is non-negative definite. Suppose $(x^T AA^T x) = 0$.

$\Rightarrow (A^T x)^2 = 0 \Rightarrow A^T x = 0$. Now, # linearly indep.

Solutions $x$ to $A^T x = 0 = \text{rank}(A^T) - m - \text{rank}(A^T) = m - m = 0$.

Thus $A^T x = 0 \Rightarrow x = 0$ & so $AA^T$ is positive definite.

However, $A^T A$ is an $n \times n \times n \times n = n \times n$ matrix.

Again, $(x^T A^T A x) = (A x)^2 \geq 0$. Thus $A^T A$ is non-negative definite. Suppose $Ax = 0$. # indep. solutions $= n - m$.

If $n > m$ then we can find such $x \neq 0$ and $A^T A$ is only non-negative definite. If $n = m$, it is positive definite.
Since \( a e' + b e' + e^2 \downarrow 0 \) as \( e \downarrow 0 \), given \( \varepsilon > 0 \) we can find \( \varepsilon' > 0 \) such that \( a e' + b e' + e^2 < \varepsilon \).

Now for this \( \varepsilon' > 0 \) find \( N_b \) and \( N_a \)
for all \( n > N_b \), \( b_n < b + \varepsilon' \).

That \( n > N_a \), \( a_n < a + \varepsilon' \). Consequently for

\[ n > \max(N_a, N_b), \quad a_n b_n = (a + \varepsilon')(b + \varepsilon') = ab + a \varepsilon' b + b \varepsilon' + \varepsilon^2 < ab + \varepsilon \]

\[ \therefore \limsup_{n \to \infty} a_n b_n < ab \]

Next since \( a e' + b e' \downarrow 0 \) as \( e' \downarrow 0 \) find \( \varepsilon' > 0 \) such that \( a e' + b e' < \varepsilon' \), \( e' < b \), \( e' < a \)

For this \( \varepsilon' \) find \( N_b \) such that

for all \( n > N_b \), \( b_n > b - \varepsilon' \),

Now given \( N \geq 1 \) find \( N_0 > \max(N, N_b) \)

such that \( a_n > a - \varepsilon' \)

Consequently

\[ a_n b_n > (a - \varepsilon')(b - \varepsilon') = ab - a \varepsilon' b - b \varepsilon' + \varepsilon^2 \]

\[ > ab - (b \varepsilon' + \varepsilon') > ab - \varepsilon \]

\[ \therefore \limsup_{n \to \infty} a_n b_n > ab \]
$$f_{\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{X}}(y_1, y_2, x) = f_{\mathbf{Y}_1, \mathbf{Y}_2 | \mathbf{X}}(y_1, y_2 | x) \cdot f_X(x)$$

$$= \frac{f(y_1, y_2 | x)}{y_1, y_2 | x} \cdot \frac{f(y_2 | x)}{y_2 | x} \cdot f_X(x)$$

$$= \left\{ \begin{array}{ll}
\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot x^2 e^{-x} & 0 < y_1 < x, 0 < y_2 < x \\
0 & \text{otherwise}
\end{array} \right.$$
$\textcircled{3}$

\[ e^{(y_1 y_2)} = (e^{y_1}) (e^{y_2}) + 1 = 1 + (y_1 y_2) \]

\[ \therefore e^{x (y_1 + y_2)} = 1 + y_1 y_2 = 1 + \text{max} \{y_1, y_2\} \]
Solutions for STAT 411 Option I

Fall 04, prob. # 4:

\[ Y = \sum_{i=1}^{n} X_i \] is a complete sufficient statistic for \( \theta \). Let \( U = I[X_i > 3] \). Clearly \( U \) is an unbiased estimator for \( e^{-\frac{3}{\theta}} \). So \( E[UY] \) is the minimum variance unbiased estimator for \( e^{-\frac{3}{\theta}} \).

Now \( E[UY; y] = P[X_i > 3 | Y = y] = \begin{cases} 0 & \text{if } y \leq 3 \\ \int_{y}^{\infty} f(x_i | y) \, dx_i & \text{otherwise} \end{cases} \)

where \( f(x_i | y) = \begin{cases} \frac{f(x_i) f(y-x_i)}{f_Y(y)} & \text{if } 3 \leq x_i \leq y \\ 0 & \text{otherwise} \end{cases} \)

where \( Y - X_i = \sum_{i=2}^{n} X_i \) is Gamma \( \alpha = n-1, \beta = \theta \).

So \( P[X_i > 3 | Y = y] = \int_{y}^{\infty} \frac{(n-1)(y-x_i)^{n-2}}{y^{n-1}} \, dx_i = \frac{1}{y^{n-1}} (y-3)^{n-1} = 0 \) for \( y \leq 3 \), else \( \frac{1}{y^{n-1}} (y-3)^{n-1} \).

So the m.v.u.e. for \( e^{-\frac{3}{\theta}} \) is \( \left(1 - \frac{3}{y}\right)^{n-1} I[Y > 3] \).
Fall 04:

1) \( \mu = E(\xi) = \sum_{k=0}^{\infty} k (1-c)^k \frac{c^k}{1-c} \). Now if \( \frac{c}{1-c} \leq 1 \), i.e. \( c \leq \frac{1}{2} \), then prob. of eventual distinction, \( \mu_\infty \), is 1. But if \( c > \frac{1}{2} \), then \( \mu > 1 \), and the prob. of eventual distinction is the smallest positive solution for the equation \( \phi(s) = s \), where \( \phi \) is the prob. gen. func. of \( \xi \), given by:

\[
\phi(s) = \sum_{k=0}^{\infty} s^k (1-c)^k \frac{c^k}{1-s} = \frac{1-s}{1-sc}
\]

Now consider the equation \( \frac{1-c}{1-sc} = s \iff c s^2 - s + 1-c = 0 \)

\[
\therefore s = \frac{1 \pm \sqrt{1-4c(1-c)}}{2c} = \frac{1 \pm (2c-1)}{2c}
\]

\[
\text{prob. of eventual extinction} = \frac{1-(2c-1)}{2c} = \frac{2(1-c)}{2c} = \frac{1-c}{c}
\]

\[\begin{align*}
(\text{i}) \quad \mu_{\infty} &= (1)^2 = 1 \quad \text{if} \quad \mu \leq 1 \quad [c \leq \frac{1}{2}] \\
&= \left(\frac{1-c}{c}\right)^2 \quad \text{if} \quad \mu > 1 \quad [c > \frac{1}{2}]
\end{align*}\]
6. (Stat 411)

Let \( X_1, X_2, \ldots, X_n \) be a random sample from \( N(\mu, \theta) \), where \( \mu \) is the expectation and \( \theta \) is the variance, \( \mu, \theta \) both unknown. Derive the likelihood ratio test for \( H_0: \mu = 0 \) versus \( H_1: \mu \neq 0 \) and show that it can be expressed as a two-tailed test based on a \( t \) statistic. What is the degree of freedom of the \( t \)?

\[
L(\mu, \theta) = \left( \frac{1}{2\pi \theta} \right)^{n/2} e^{-\frac{1}{2\theta^2} \sum_{i=1}^{n} (x_i - \mu)^2}
\]

Mles

\[
\hat{\mu} = \bar{X}, \quad \hat{\theta} = S^2 = \frac{1}{n} \sum (x_i - \bar{X})^2
\]

\[
L(\Omega) = \max_{\Omega} L(\mu, \theta) = \left( \frac{1}{2\pi} \right)^{n/2} \left( \frac{1}{S^2} \right)^{n/2} e^{-\frac{n}{2}}
\]

Mles under \( H_0: \mu = 0 : \)

\[
\mu = 0, \quad \hat{\theta} = S_0^2 = \frac{1}{n} \sum x_i^2
\]

\[
L(\hat{\omega}) = \max_{H_0} L(\mu, \theta) = \left( \frac{1}{2\pi} \right)^{n/2} \left( \frac{1}{S_0^2} \right)^{n/2} \ e^{-\frac{n}{2}}
\]

Likelihood ratio

\[
\lambda = \frac{L(\hat{\omega})}{L(\hat{\omega})} = \left( \frac{S^2}{S_0^2} \right)^{n/2}
\]

Since \( \sum x_i^2 = \sum (x_i - \bar{X})^2 + n \bar{X}^2 \)

\[
\lambda \leq c \Leftrightarrow \frac{n \bar{X}^2}{\sum (x_i - \bar{X})^2} \geq c_1 \Rightarrow \frac{\sqrt{n} |\bar{X}|}{\sqrt{\sum (x_i - \bar{X})^2 / (n-1)}} \geq c_2
\]

Since, under \( H_0, \) \( \frac{\sqrt{n} |\bar{X}|}{\sqrt{\sum (x_i - \bar{X})^2 / (n-1)}} \sim t \) with \( (n-1) \) degrees of freedom.

This is a 2-tail \( t \) test.
7. (Stat 416) SOLUTION

(a) The basic model assumptions of the Friedman test are given in Hollander and Wolfe (1999) on pages 271-272.

The relevant statistics for computing the test statistic \( S \) are:

<table>
<thead>
<tr>
<th>Run</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fabric A</td>
<td>2.5</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>2.5</td>
<td>8.5</td>
<td>1.7</td>
</tr>
<tr>
<td>Fabric B</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2.5</td>
<td>16.5</td>
<td>3.3</td>
</tr>
<tr>
<td>Fabric C</td>
<td>2.5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>17.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Fabric D</td>
<td>1</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>7.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

The Friedman test statistic is

\[
S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left( R_j - \frac{k+1}{2} \right)^2
\]

where \( k = 4 \) and \( n = 5 \).

\[
S = 3 \left( (1.7 - 2.5)^2 + (3.3 - 2.5)^2 + (3.5 - 2.5)^2 + (1.5 - 2.5)^2 \right)
\]

\[
= 3(3.28)
\]

\[
= 9.84
\]

(b) The Friedman test is here more appropriate than the Kruskal-Wallis test since each run may be set on a different level and has to be treated as a block.