Name

Final Exam
Thursday the 16th of December, 1998

This exam is open book. Write your answers on the exam sheets. If you need more space, continue your answer on the back of the page. Make sure you have all 13 pages!

The exam is 180 minutes long. There are four questions, worth a total of 100 points. They are not equally weighted, nor are they of equal difficulty. The number of points each question is worth is printed with the question. Read the questions carefully. If you are unsure of the interpretation, come ask. **You must show your work** to obtain full credit. If you use a result from class, state what result you are using. If you can’t complete a problem for any reason, explain what concepts are at issue, and how you would attack the problem. It is, in any case, a good idea to explain what your reasoning is in English. If I can’t tell that you understood what you were doing, I can’t give you credit, particularly if you get the wrong numerical answer. GOOD LUCK!

**Question 1 (25 Points, 45 minutes)**

We are interested in the average sales ($\mu_y$) of firms in a certain industry. (All numbers are in millions). We take a random sample of 100 firms and ask them for their last years sales. This gives $\bar{Y} = 27$. However we also ask them for last years profit ($X$). This gives $\bar{X} = 15$. It is known that the average profit in the industry last year was $\mu_x = 10$. Use the following information to answer the questions.

$$
\begin{align*}
\mu_x &= 10 & \sigma_x &= 50 \\
\mu_y &= ? & \sigma_y &= 100 \\
\rho &= 2 & \rho &= .95 \\
\end{align*}
$$

**Part a**

Give the simple random sample ($\bar{Y}_{SRS}$) and ratio ($\bar{Y}_R$) estimates for $\mu_y$.

**Part b**

Give the exact bias of $\bar{Y}_{SRS}$ and the approximate bias of $\bar{Y}_R$ (ignore the finite population correction).

**Part c**

Give the exact variance of $\bar{Y}_{SRS}$ and the approximate variance of $\bar{Y}_R$ (ignore the finite population correction).

**Part d**

Give the mean squared errors of $\bar{Y}_{SRS}$ and $\bar{Y}_R$ (ignore the finite population correction).
**Question 2 (30 Points, 54 minutes)**

Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables and that

$$X_i \sim \text{Poisson}(a_i \lambda)$$

where $a_1, a_2, \ldots, a_n$ are known constants.

**Part a**

Show that the MLE for $\lambda$ is

$$\hat{\lambda} = \frac{\sum X_i}{\sum a_i}$$

**Part b**

Calculate $E[\hat{\lambda}]$ and $Var(\hat{\lambda})$.

**Part c**

Calculate $I_{X_i}(\lambda)$ the information for one observation $X_i$.

**Part d**

Is $\hat{\lambda}$ the UMVUE for $\lambda$? Note, since the $X_i$ are not iid, the Cramer-Rao inequality implies that for any unbiased estimator $\hat{\theta}$

$$Var(\hat{\theta}) \geq \frac{1}{\sum_i I_{X_i}(\lambda)}$$

**Part e**

Suppose you wish to test $H_0 : \lambda = \lambda_0$ vs the alternative $H_A : \lambda > \lambda_0$. Use Neyman-Pearson to find the form of the optimal hypothesis test. (Note you do not need to specify how to calculate the constant $C_*$)

**Part f**

We are performing a wildlife study and are interested in whether animals in a certain species are randomly distributed about a region or whether they tend to clump together. We divide the region into 5 different sub-regions and count the number of animals in each sub-region. Unfortunately because of the geography of our region we can not divide it into 5 equal sized sub-regions. Let $\lambda$ be the average number of animals per square mile and let $X_i$ be the number of animals observed in the $i$th sub-region. If the animals are randomly distributed then

$$X_i \sim \text{Poisson}(a_i \lambda)$$

where $a_i$ is the area of the region. Use the following data to perform a goodness of fit test to see whether the animals are randomly distributed or not.
<table>
<thead>
<tr>
<th>Sub-region</th>
<th>$X_i$</th>
<th>Area (sq miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>59</td>
<td>20</td>
</tr>
</tbody>
</table>

**Question 3 (15 Points, 27 minutes)**

Suppose you are performing an analysis of CEO compensation across 4 different industries. For each of the 4 industries you randomly choose 10 companies and record the compensation for their CEOs. The mean compensation for each industry was as follows. (All numbers are in thousands of dollars.)

$$ \bar{Y}_1 = 410, \quad \bar{Y}_2 = 511, \quad \bar{Y}_3 = 399, \quad \bar{Y}_4 = 488 $$

This gave a grand mean of $\bar{Y}_0 = 452$

(a) Use this information to complete the partial One Way Anova table below. Is there significant evidence that there is a difference in compensation across the four industries?

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups (Industries)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>123406</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Suppose that Industries 1 and 3 are heavily technology oriented while Industries 2 and 4 are not. As part of your study you are interested in comparing technology with non technology industries. Produce a 95% confidence interval for the difference in the average level of compensation for technology industries with that of non technology industries. (Assume that the data was not used to make this decision).

**Question 4 (30 Points, 54 minutes)**

You are performing a study of the "hits" on a certain website. Let $N$ denote the number of people that attempt to connect to the website in a one day period. Suppose that $N \sim \text{Poisson}(\lambda)$ with $\lambda$ unknown. As with all websites a certain percentage of the time people are unable to connect. These people are unobserved. Let $p$ denote the probability a random person manages to connect. Assume the probability of connecting is independent for all people and that $p$ is known. Let $X$ denote the actual number of hits i.e. the number of people that connect. $X$ is observed but $N$ is the quantity we are interested in.

**Part a**

What is the conditional distribution of $X$ given $N = n$?
Part b

What is the joint probability function of $X$ and $N$ i.e. $P(X = x, N = n)$?

Part c

Show that the marginal distribution of $X$ is $Poisson(p \lambda)$. Hint recall

$$
\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda
$$

Part d

Assuming that $p$ is known give both the maximum likelihood and method of moments estimate for $\lambda$.

Part e

What is the conditional distribution of $N$ given $X = x$?

Part f

Show that the expected value of $N$ given $X = x$ is $x + (1 - p) \lambda$.

Part g

Suppose $p = .95$ and on a given day $X = 10,000$. Give reasonable estimates for $\lambda$, $N$, and the number of lost customers i.e. the number of people that failed to connect.