DEPARTMENT-WIDE DOCTORAL WRITTEN EXAMINATIONS

of the

DEPARTMENT OF BIOSTATISTICS
School of Public Health
University of North Carolina at Chapel Hill

VOLUME I: Closed-Book Parts

edited by Dana Quade

DEPARTMENT-WIDE DOCTORAL WRITTEN EXAMINATIONS
of the
DEPARTMENT OF BIOSTATISTICS
School of Public Health
University of North Carolina at Chapel Hill

VOLUME I
CLOSED-BOOK PARTS

assembled and edited by
DANA QUADE *

First Edition: June 1981
Second Edition: November 1983
Third Edition: May 1988

* with MICHAEL J SYMONS (First and Second Editions)
**TABLE OF CONTENTS**

<table>
<thead>
<tr>
<th>Date</th>
<th>Part</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>16 September 1967:</td>
<td>Part I</td>
<td>5</td>
</tr>
<tr>
<td>16 September 1967:</td>
<td>Part II</td>
<td>7</td>
</tr>
<tr>
<td>21 September 1968:</td>
<td>Part I</td>
<td>12</td>
</tr>
<tr>
<td>21 September 1968:</td>
<td>Part II</td>
<td>14</td>
</tr>
<tr>
<td>21 February 1970:</td>
<td>Part I</td>
<td>17</td>
</tr>
<tr>
<td>7 March 1970:</td>
<td>Part III</td>
<td>20</td>
</tr>
<tr>
<td>26 June 1970:</td>
<td>Part I</td>
<td>26</td>
</tr>
<tr>
<td>10 July 1970:</td>
<td>Part III</td>
<td>28</td>
</tr>
<tr>
<td>20 February 1971:</td>
<td>Part I</td>
<td>32</td>
</tr>
<tr>
<td>25 June 1971:</td>
<td>Part I</td>
<td>37</td>
</tr>
<tr>
<td>25 June 1971: Extra Questions (for one student taking a special re-examination)</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>29 January 1972:</td>
<td>Part I</td>
<td>41</td>
</tr>
<tr>
<td>27 January 1973:</td>
<td>Part I</td>
<td>44</td>
</tr>
<tr>
<td>25 May 1973:</td>
<td>Part I</td>
<td>48</td>
</tr>
<tr>
<td>26 January 1974:</td>
<td>Part I</td>
<td>52</td>
</tr>
<tr>
<td>18 May 1974:</td>
<td>Part I</td>
<td>55</td>
</tr>
<tr>
<td>25 January 1975:</td>
<td>Part I</td>
<td>59</td>
</tr>
<tr>
<td>24 January 1976:</td>
<td>Part I</td>
<td>64</td>
</tr>
<tr>
<td>22 January 1977:</td>
<td>Part I</td>
<td>68</td>
</tr>
<tr>
<td>3 August 1977:</td>
<td>Part I</td>
<td>72</td>
</tr>
<tr>
<td>21 January 1978:</td>
<td>Part I</td>
<td>77</td>
</tr>
<tr>
<td>26 August 1978:</td>
<td>Part I</td>
<td>81</td>
</tr>
<tr>
<td>11 August 1979:</td>
<td>Part I</td>
<td>85</td>
</tr>
<tr>
<td>9 August 1980:</td>
<td>Part I</td>
<td>90</td>
</tr>
<tr>
<td>15 August 1981:</td>
<td>Part I</td>
<td>95</td>
</tr>
<tr>
<td>16 August 1982:</td>
<td>Part I</td>
<td>99</td>
</tr>
<tr>
<td>8 August 1983:</td>
<td>Part I</td>
<td>103</td>
</tr>
<tr>
<td>30 July 1984:</td>
<td>Part I</td>
<td>109</td>
</tr>
<tr>
<td>23 July 1985:</td>
<td>Part I</td>
<td>115</td>
</tr>
<tr>
<td>29 July 1986:</td>
<td>Part I</td>
<td>121</td>
</tr>
<tr>
<td>3 August 1987:</td>
<td>Part I</td>
<td>127</td>
</tr>
</tbody>
</table>
INTRODUCTION

This publication (Mimeo Series #1343) and its companion (Mimeo Series #1344) contain the general written examinations which the Department of Biostatistics has set for doctoral degrees: the closed-book parts here and the take-home parts in #1344.

The first Department-wide doctoral-level written examinations were given in September 1967, at the initiation of the new PhD program in Biostatistics. (Prior to that time the Department had offered the PhD only in Public Health, for which it had not required any Department-wide examination.) The original format was: (A) Basic Written Examination, with two Parts, (I) Theory and (II) Analysis, each including 5 closed-book questions, of which candidates were to complete 4 in 3 hours; and (B) Advanced Written Examination, again with two Parts, (I) General and (II) Specialty Areas, each one week take-home, where Part I had 5 questions and Part II 10, and candidates were to complete 4 from each Part.

In the 1969-1970 academic year, it was decided to drop the Advanced Written Examination and expand the Basic Written Examination to three Parts: (I) Theory, with 5 closed-book questions of which 4 were to be completed in 3 hours; (II) Applications, with 5 take-home questions of which 4 were to be completed in one week; and (III) Specialty Areas, with 6 closed-book questions of which 3 were to be completed in 3 hours. This format was used twice, in February/March and again in June of 1970.

On 3 March 1971, the Department voted to abolish Part III of the Basic Written Examination, and to let the Literature Review for the Dissertation satisfy the Graduate School requirement for a
Preliminary Written Examination. This action produced a format for "the Basics" which has continued essentially without change until the present time. In 1972 the time allowed for Part I (closed-book) was expanded from 3 hours to 4. Also, the following instruction was added to the rubrics for both Parts:

Most questions should be answered in the equivalent of less than 7 typewritten pages (300 words per page) and under no circumstances will more than the first 15 typewritten pages or the equivalent be read by the grader.

This rubric was dropped from Part I in 1974. It continues for Part II, and indeed changed from 7 and 15 pages to 5 and 10 starting in 1978; but it has never been strictly enforced.

Originally the Basics served not only as a screening examination for the PhD, but also as the Master's Written Examination for the MS. Thus, each candidate was evaluated using one of three grades: pass at doctoral level, fail at doctoral level but pass at master's level, or fail. This practice continued through 1979, since when the MS examination has been given separately. In addition, although the Basics were originally intended for the PhD only, beginning in 1979 candidates for the DrPH have also been required to pass Part II (not Part I).

The Basics are given annually; originally in Winter, but starting 1979 in Summer. Special examinations have been given on several occasions when there were sufficient students to warrant it; the current regulations require this only if there are 5 or more students to retake a Part.

The Examinations Committee prepares and conducts all Department-wide written examinations, and handles arrangements for
their grading. A team of two graders is appointed for each question. Where possible, all graders are members of the Department of Biostatistics and of the Graduate Faculty, and no individual serves on more than one team for the same examination (the two Parts counting separately in this context). The members of each grading team are asked to prepare for their question an "official answer" covering at least the key points. They agree beforehand on the maximum score possible for each component, the total for any question being 25.

The papers are coded so that the graders are unaware of the candidates' identities, and each candidate's answer is marked independently by each of the two graders. The score awarded reflects the effective proportion correctly answered of the question. The two graders then meet together and attempt to clear up any major discrepancies (more than 5 points, say) between their scores. Their joint report is to include comments on serious shortcomings in any candidate's answer.

On the basis of a candidate's total score on a paper, the Examinations Committee recommends to the faculty whether the candidate is to be passed, failed, or passed conditionally. In the last case, the condition is specified together with a time limit. All final decisions are by vote of the faculty. The faculty has never specified in advance what the passing score will be, out of the possible 200 total points per Part, but has set cut-points varying from about 100 to 140, usually a little higher for Part II than for Part I. Examination papers are not identified as to candidate until after the verdicts of PASS and FAIL have been rendered.
Once the decisions have been made, advisors are free to tell their students unofficially; the official notification, however, is by letter from the Chairman of the Examination Committee. Actual scores are never released, but the "official answers" are made public, and candidates who are not passed unconditionally are permitted to see the graders' comments on their papers. When performance is not of the standard required, candidates are reexamined at a later date set by the Examinations Committee.

Candidates whose native language is not English are not allowed extra time on Department-wide (not individual course) examinations. This condition may be waived for individual candidates at the discretion of the Department Chairman upon petition by the candidate at least one week prior to the examination.

NOTE. Most of what follows reproduces the examinations exactly as they were originally set; however, minor editorial changes and corrections have been made, particularly in order to save space. Thus, journal articles which were included with the original examinations are not reproduced here, but complete references to them are provided in all cases. Also, in recent years the larger datasets have been provided in mainframe files and/or on diskettes. Some of these datasets are still available to the interested reader; inquiries may be addressed to the Editor.
1. Let $X$ be the number of successes in $n$ independent Bernoulli trials with probability $p$ of success on each trial. Under appropriate conditions, the distribution of $X$ is approximately Poisson. State and prove the limit theorem pertaining to this.

2. (a) Derive the density function of the ratio of two independent standard normal random variables:

$$U = \frac{X}{Y}, \quad X \sim N(0,1), \; Y \sim N(0,1); \; X \text{ and } Y \text{ independent.}$$

(b) What is the name of this distribution?

(c) What is $P(U < 0)$?

3. (a) State and prove the Neyman-Pearson Lemma.

(b) To what extent is this Lemma applicable in significance tests involving a composite hypothesis?

(c) Name and define a test criterion applicable to cases where the Lemma cannot be used.
4. Statistician A decided to observe a certain random process for one hour; on doing so, he recorded that 50 events occurred during that hour. Independently, Statistician B decided to observe the process until 50 events occurred; on doing so, he recorded that one hour was required for that to happen.

(a) Find the maximum likelihood estimate of the mean rate of occurrence of events, using A's data; using B's data.

(b) Are these estimates unbiased? What are their standard errors? If they are not unbiased, show how to correct them for bias.

(c) An onlooker remarked that

(i) "A and B have really observed the same thing, i.e., 50 events in one hour" and

(ii) "When 2 statisticians have observed the same thing, they should make the same estimates."

Discuss these points carefully.

5. \(X_1\) and \(X_2\) are independent and identically distributed random variables with density function

\[
f(x) = \begin{cases} 
\frac{e^{-(x-\theta)}}{\theta} & \text{if } x > \theta, \\
0 & \text{otherwise}, \quad -\infty < \theta < \infty.
\end{cases}
\]

Show that \(\min(X_1, X_2)\) is a sufficient statistic for \(\theta\), and find its density function.
1. In order to determine whether the addition of an anticariogenic agent to a dentifrice makes an appreciable difference in flavor, it was decided to run a duo-trio test with ten panelists.

The duo-trio test consists of presenting each judge with a known standard (the present dentifrice) and two unknowns (the new dentifrice and the present dentifrice). Each judge is then asked to state which of the unknown dentifrices differs from the standard.

(a) What is the exact probability that 8 or more panelists will be right when all are actually guessing?

(b) What is the exact probability that 9 or more panelists are right when all are actually guessing?

(c) What is the probability that all 10 are right when all are actually guessing?

(d) If we want $\alpha \leq .05$, how many panelists must be right to reject the null hypothesis that all panelists are guessing?

What is the alternate hypothesis? Is this a two-tail test? Explain.

(e) If it is desired to assure 90% probability of detecting a true 2:1 chance of correct identification, while keeping $\alpha \leq .05$, how many judges must be used?
A test has been conducted to compare the effectiveness of a new deodorant product (A) with the effectiveness of a competitive product (C). Effectiveness is measured in terms of underarm odor scores at a fixed time after use of the product. The objective is to reach one of the following conclusions:

1. A is good enough to justify a formal claim that A is better than C.
2. A is not good enough to justify such a claim.

Scores are defined so that they increase as effectiveness increases. The true mean scores for products A and C are $\mu_A$ and $\mu_C$ respectively. In the standard test $\mu_A$ is estimated by $\bar{Y}_A$ and $\mu_C$ is estimated by $\bar{Y}_C$. Conclusion (1) is reached if $(\bar{Y}_A - \bar{Y}_C)$ is at least 1.0 unit.

In Figure 1 the performance of a single test is presented in graphical form. The plotted curve represents the probability of reaching conclusion (1) as a function of the true difference $\mu_A - \mu_C$.

![Figure 1. Probability of Reaching Conclusion (1)](image)

(a) What is the null hypothesis in this test?

(b) What is the $\alpha$-risk? (Give numerical value.)

(c) If a difference of 1.8 units is considered important, what is the $\beta$-risk? (Give numerical value.)

(continued)
For the remaining questions we assume that $\mu_A - \mu_C = .5$. Define two events as follows:

- $G$ is the event that $\tilde{Y}_A - \tilde{Y}_C \geq 1$,
- $H$ is the event that $\tilde{Y}_A - \tilde{Y}_C \geq 0$.

From Figure 1 the probability of occurrence of $G$ is $P(G) = .2$. The probability of occurrence of $H$ is known to be $P(H) = .8$.

(d) What is the conditional probability that we will reach conclusion (1) given that $\tilde{Y}_A$ is at least as large as $\tilde{Y}_C$?

(e) What is the expected value of the random variable $(\tilde{Y}_A - \tilde{Y}_C)$?

(f) Suppose two independent tests are run.

(i) What is the conditional probability that $\tilde{Y}_A - \tilde{Y}_C \geq 1$ in the second test given that we reached conclusion (1) on the basis of the first test?

(ii) What is the joint probability of reaching conclusion (1) on the basis of the first test and then finding $\tilde{Y}_A - \tilde{Y}_C < 1$ in the second test?

3. Suppose the following experiment was performed.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Production Unit 1</th>
<th>Production Unit 2</th>
<th>Production Unit 3</th>
<th>Production Unit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AB</td>
<td>AB</td>
<td>aB</td>
<td>ab</td>
</tr>
<tr>
<td>2</td>
<td>Ab</td>
<td>aB</td>
<td>ab</td>
<td>AB</td>
</tr>
<tr>
<td>3</td>
<td>aB</td>
<td>ab</td>
<td>AB</td>
<td>Ab</td>
</tr>
<tr>
<td>4</td>
<td>ab</td>
<td>AB</td>
<td>Ab</td>
<td>aB</td>
</tr>
</tbody>
</table>

Note that the treatments form a $2^2$ factorial, where

- $a$ represents the low level of 1st treatment,
- $A$ represents the high level of 1st treatment,
- $b$ represents the low level of 2nd treatment,
- $B$ represents the high level of 2nd treatment.

(a) Exhibit the complete breakdown of degrees of freedom in the ANOVA for this experiment.

(b) Comment on the interactions which may be investigated in this experiment.
4. Given below are real data from a recent study to compare certain dental treatments. The children at a North Carolina orphanage were divided at random into 2 groups, and those in one group were given one treatment while those in the other group were given a different treatment. For each child a determination was made of the DMF rate (number of decayed, missing, or filled teeth) at the beginning of the experiment and two years later.

(a) Make a comparison of the treatments, including a test of significance, using a method of analysis so simple that you can complete it now.

(b) Describe briefly how you might attack a problem like this if you had plenty of time and computational assistance (and if there were more data, so that an elaborate analysis would be worthwhile).

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>Treatment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child #</td>
<td>Sex</td>
</tr>
<tr>
<td>1</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>F</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
</tr>
<tr>
<td>15</td>
<td>M</td>
</tr>
<tr>
<td>16</td>
<td>M</td>
</tr>
<tr>
<td>17</td>
<td>M</td>
</tr>
<tr>
<td>18</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
</tr>
<tr>
<td>20</td>
<td>M</td>
</tr>
<tr>
<td>21</td>
<td>M</td>
</tr>
<tr>
<td>22</td>
<td>M</td>
</tr>
</tbody>
</table>

*Not available; child left orphanage before end of 2-year period.
It has been suggested that a linear relationship exists between the amount of mail handled in a post office and the man-hours required to handle it. In order to test out this hypothesis the following data were collected:

<table>
<thead>
<tr>
<th>Four-week period, fiscal year 1962</th>
<th>Pieces of mail handled (X) (in millions)</th>
<th>Man-hours used (Y) (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>157</td>
<td>572</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>570</td>
</tr>
<tr>
<td>3</td>
<td>169</td>
<td>599</td>
</tr>
<tr>
<td>4</td>
<td>186</td>
<td>645</td>
</tr>
<tr>
<td>5</td>
<td>183</td>
<td>645</td>
</tr>
<tr>
<td>6</td>
<td>184</td>
<td>671</td>
</tr>
<tr>
<td>7</td>
<td>268</td>
<td>1053</td>
</tr>
<tr>
<td>8</td>
<td>184</td>
<td>655</td>
</tr>
<tr>
<td>9</td>
<td>180</td>
<td>637</td>
</tr>
<tr>
<td>10</td>
<td>188</td>
<td>667</td>
</tr>
<tr>
<td>11</td>
<td>184</td>
<td>656</td>
</tr>
<tr>
<td>12</td>
<td>182</td>
<td>640</td>
</tr>
<tr>
<td>13</td>
<td>179</td>
<td>609</td>
</tr>
</tbody>
</table>

Assume the following model:

\[ Y = \beta_0 + \beta_1 X + \epsilon, \]

where \( X \) is a fixed (independent) variate, and \( \epsilon \) is \( N(0, \sigma^2) \)

Calculations

\[
\begin{align*}
\Sigma X &= 2405 \\
\Sigma Y &= 8619 \\
\Sigma XY &= 1633249 \\
\Sigma X^2 &= 453517 \\
\Sigma Y^2 &= 5892485 \\
\Sigma X^2/N &= 185 \\
\Sigma Y/N &= 663 \\
\Sigma XY - (\Sigma X)(\Sigma Y)/N &= 38734, \\
\Sigma Y^2 - (\Sigma Y)^2/N &= 178088
\end{align*}
\]

(a) Calculate the ANOVA table for the assumed model using the above figures.

(b) Assuming all the above calculations are correct, test the hypothesis that \( \beta_1 = 0 \) using an \( \alpha \)-risk of 5%.

(c) Comment on the validity of the error term in the Analysis.

(d) How do you explain that the best estimate of the intercept is negative?

(e) Is there any evidence here that the linear model is inadequate? Justify your answer.
1. Define the term "sufficient statistic". Explain the importance of sufficient statistics in the theory of estimation: prove your statements as far as you can.

Prove that the most powerful test of a simple null hypothesis against a simple alternative is based on a sufficient statistic, if such a statistic exists.

2. Find the probability that, in a random sample of size 3 from an exponential distribution, the mean will exceed the median.

3. Assume that a curve of the form \( y = b_0 + b_1x + b_2x^2 \) has been fitted and that the usual assumptions of normality hold. Derive a confidence interval for that \( x \) which produces the maximum or minimum value of \( y \).
4. (a) Assume $X$ and $Y$ are random variables with means $\mu_X$ and $\mu_Y$ ($\mu_Y \neq 0$), variances $\sigma_X^2/n$ and $\sigma_Y^2/n$, and covariance $\sigma_{XY}/n$. Derive the asymptotic variance of $X/Y$.

(b) Assume that $X$ is a $(1 \times p)$ random vector with mean vector $\mu$ and variance matrix $\Sigma$: define $q < p$ functions of the elements of $X$, viz. $f_i(X)$, $i = 1, 2, \ldots, q$, which have derivatives up to second order. Derive the asymptotic covariance matrix of the vector $\mathbf{F} = [f_1(X) \ldots f_q(X)]$.

For what kinds of functions is this expression exact?

5. Let $X$ be a Poisson random variable with expected value $\theta$.

Let $Y_1, Y_2, \ldots$, be independent variables each having the same expected value $\mu$ and variance $\sigma^2$.

Let

$$Z = \begin{cases} \frac{X}{\sum_{j=1}^{\infty} Y_j} & X > 0 \\ 0 & X \leq 0 \end{cases}$$

Show that the correlation coefficient between $X$ and $Z$ does not depend on $\theta$. 


1. Suppose a population consists of four strata, with the following weights:

\[ w_1 = 10\%, \quad w_2 = 50\%, \quad w_3 = 25\%, \quad w_4 = 15\%. \]

Suppose the variances within the strata are approximately:

\[ s^2_1 = 16, \quad s^2_2 = 4, \quad s^2_3 = 9, \quad s^2_4 = 12.25. \]

A sample of size \( n = 100 \) has been drawn from the population, with Neyman allocation used to determine the number of units selected from each stratum. What are these sample sizes? Suppose the observed means within the strata are \( \bar{y}_1 = 10, \bar{y}_2 = 12, \bar{y}_3 = 14, \bar{y}_4 = 15; \) and the observed within-strata variances are

\[ s^2_1 = 15, \quad s^2_2 = 5, \quad s^2_3 = 10, \quad s^2_4 = 12. \]

Give the appropriate estimate for the population mean; also give an estimate of the sampling variance of this estimate.

NOTE: Problem #2 is on the next page.

3. An investigator has approached you with the following data concerning the response of white rats to a certain drug. He tells you that he has examined 10 animals at each drug level, and the responses given are means. He has also calculated variances at each drug level.

<table>
<thead>
<tr>
<th>( \bar{y} ) (mean response)</th>
<th>( x ) (drug level)</th>
<th>( s^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>1</td>
<td>1.40</td>
</tr>
<tr>
<td>3.9</td>
<td>2</td>
<td>.80</td>
</tr>
<tr>
<td>3.2</td>
<td>3</td>
<td>.65</td>
</tr>
<tr>
<td>8.4</td>
<td>4</td>
<td>3.30</td>
</tr>
<tr>
<td>7.5</td>
<td>5</td>
<td>3.00</td>
</tr>
<tr>
<td>6.8</td>
<td>6</td>
<td>2.70</td>
</tr>
</tbody>
</table>

Indicate what steps are needed to complete the analysis. How would you explain your results to this investigator?
2. An amateur weather forecaster made "rain" or "shine" predictions for 100 days. During this period he was successful on 8 out of his 20 "rain" forecasts and 63 out of his 80 "shine" forecasts.

This information might be organized for $\chi^2$ computation in any of the following three ways. In each case the numbers in parentheses are the "expected frequencies" based on a null hypothesis which seems natural from the way the table is set up.

(a) Successes 71 (50)
Failures 29 (50) $\chi^2 = 17.7$
Total 100

(b) Prediction

<table>
<thead>
<tr>
<th></th>
<th>Rain</th>
<th>Shine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>8(14.2)</td>
<td>63(56.8)</td>
</tr>
<tr>
<td>Failure</td>
<td>12(5.8)</td>
<td>17(23.2)</td>
</tr>
</tbody>
</table>

$\chi^2 = 11.7$

(c) Predicted

<table>
<thead>
<tr>
<th></th>
<th>Rain</th>
<th>Shine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>8(5)</td>
<td>17(20)</td>
</tr>
<tr>
<td></td>
<td>12(15)</td>
<td>63(60)</td>
</tr>
</tbody>
</table>

$\chi^2 = 3.0$

For each case state the null hypothesis on which the expected frequencies are computed. Which of these three analyses leads to the most meaningful answer to the question: "Has the forecaster demonstrated real ability?" Justify your answer.
4. Two judges each ranked eleven contestants in a beauty contest, with the following results:

<table>
<thead>
<tr>
<th>Contestant</th>
<th>Judge 1</th>
<th>Judge 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>I</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>J</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>K</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

Using three essentially different methods: obtain a measure of the agreement between the judges; interpret this measure; and test whether the judges agree significantly (i.e., calculate the test statistic and indicate exactly where and how you would look up the critical value).

5. A common practice in medicine is to use each patient as his own control when comparing two treatments A and B. That is: \( n/2 \) patients receive treatment A followed by B, and \( n/2 \) receive B followed by A. Assume that neither drug produces any appreciable residual effect and that the effect of treatment can be measured by a normally distributed random variable. Under what conditions is thus using each patient as his own control better, equal, or worse as a design than administering each of drugs A and B to a separate group of \( n \) patients?
2. Suppose a certain population contains two strata of sizes $N_1$ and $N_2$, respectively. Suppose samples of sizes $n_1$ and $n_2$ are taken from the respective strata independently according to simple random sampling. Suppose that the selected elements are classified according to two binomial attributes, with one of them being the variable of principal interest such as occurrence of automobile accidents, and the other being a related covariable, such as age classified as under or over 35. The data appear as follows:

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th></th>
<th>Stratum 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Accident</td>
<td>No accidents</td>
<td>Total</td>
</tr>
<tr>
<td>&lt; 35</td>
<td>$n_{111}$</td>
<td>$n_{121}$</td>
<td>$n_{1..}$</td>
</tr>
<tr>
<td>&gt; 36</td>
<td>$n_{112}$</td>
<td>$n_{122}$</td>
<td>$n_{1..}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{1.1}$</td>
<td>$n_{1.2}$</td>
<td>$n_{1..}$</td>
</tr>
<tr>
<td>Age</td>
<td>Accident</td>
<td>No accidents</td>
<td>Total</td>
</tr>
<tr>
<td>&lt; 35</td>
<td>$n_{211}$</td>
<td>$n_{221}$</td>
<td>$n_{2..}$</td>
</tr>
<tr>
<td>&gt; 36</td>
<td>$n_{212}$</td>
<td>$n_{222}$</td>
<td>$n_{2..}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{2.1}$</td>
<td>$n_{2.2}$</td>
<td>$n_{2..}$</td>
</tr>
</tbody>
</table>

a. Indicate what the appropriate estimates of the incidence rate of automobile accidents are for each stratum and the overall population when no information other than that given above is available. Also indicate an estimate of the standard error of these statistics.

b. How is the analysis in (a) changed if it is known that the population is subdivided as follows?

<table>
<thead>
<tr>
<th>Stratum</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age ≤ 35</td>
<td>$N_{1.1}$</td>
<td>$N_{1.2}$</td>
<td>$N_{1....}$</td>
</tr>
<tr>
<td>Age &gt; 35</td>
<td>$N_{2.1}$</td>
<td>$N_{2.2}$</td>
<td>$N_{2....}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_{..1}$</td>
<td>$N_{..2}$</td>
<td>$N_{....}$</td>
</tr>
</tbody>
</table>
1. A and B make independent measurements on a certain material having magnitude $a$. Both A's measurement and B's measurement are subject to random error, the former being uniformly distributed over $a \pm 2$ and the latter uniformly distributed over $a \pm 1$. What is the probability that A's and B's measurements will be within $u$ of each other? What is the expected value of the magnitude of the discrepancy between the two measurements? (Hint: Take $U = |X - Y|$ where $X$ is uniform on $[-2, 2]$ and $Y$ is uniform on $[-1, 1]$.)

NOTE: Problem #2 is on the previous page.

3. Let $F$ be a continuous distribution function with density function $f$ and unique median $\eta$; let $X_1, X_2, \ldots, X_n$ be a random sample of $n$ observations from $F$, and let $Y_n$ be the sample median of the $X$'s. (Assume $n$ is odd.)

a. Find the density function of $Y_n$, say $g$.

b. Show that if $f$ is symmetric about $\eta$ then so is $g$.

c. Show that if $E[X^k]$ exists for some $k > 0$ then so does $E[Y_n^k]$.

d. Show that $\{Y_n, n=1, 3, 5, \ldots\}$ is a consistent estimator sequence for $\eta$.

4. Suppose we wish to estimate a set of Poisson probabilities. Using the maximum likelihood estimate of $\lambda$, we find the maximum likelihood estimates are

$$\hat{p}_i = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i=0, 1, 2, \ldots$$

a. Show that the mean of $\hat{p}_i$ is

$$e^{-n\lambda(1-e^{-\frac{1}{n}})} \frac{E[k^i]}{i!}$$

where $k$ is a Poisson variable with parameter

$$\mu = n\lambda e^{-1/n}.$$ 

b. Show that

$$\lim_{n \to \infty} E(\hat{p}_i) = e^{-\lambda}.$$ 

c. What properties are associated with the estimates $\hat{p}_i$?
5. A traffic accident researcher found an average of 0.5 accidents per driver (for a certain time period). In his report he called attention to the fact that 40% of the accidents were "caused by" (i.e., were on the records of) only 9% of the drivers. Assuming an adequate sample was taken, would you accept this as evidence of "accident proneness" of a relatively small percentage of the drivers? Justify your answer.

Poisson distribution with $\lambda = 0.5$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6065</td>
</tr>
<tr>
<td>1</td>
<td>0.3033</td>
</tr>
<tr>
<td>2</td>
<td>0.0758</td>
</tr>
<tr>
<td>3</td>
<td>0.0126</td>
</tr>
<tr>
<td>4</td>
<td>0.0016</td>
</tr>
<tr>
<td>5</td>
<td>0.0002</td>
</tr>
<tr>
<td>6</td>
<td>0.0000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

1.0000
Instructions

a) This is a closed-book examination.

b) The time limit is three hours.

c) Answer any three of the six questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Return the examination with a signed statement of the honor pledge.

1. A recently-reported Taiwan study used a "matching" study in an attempt to "estimate the real demographic impact of the IUD". The introduction and methods (design) sections of the paper are reproduced and attached*. The pre-IUD and post-IUD fertility data from the household register were used to obtain three year pre-IUD and post-IUD to December 31, 1967 fertility rates, which were compared to estimate the demographic impact.

a) Discuss the advantages of this design as compared with a study of acceptors alone.

b) Discuss the potential biases when using the two registers to identify IUD acceptors and non-acceptors in the household register system.

c) Assume that acceptors and non-acceptors can be accurately and completely identified before "matching". Discuss how other factors including migration might bias the results of the present study.

d) How might one analyze the available data to get some idea of whether biases exist?

2. In an experiment on the response of three variables to two drugs, the following data were obtained:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Drug 1</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Drug 2</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>8</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>9</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

\[
\bar{x}_1 = (5,11,8), \quad \bar{x}_2 = (9,12,8), \quad S = \frac{1}{8} \begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & 3 \\ 0 & 3 & 8 \end{pmatrix}, \quad S^{-1} = \frac{1}{8} \begin{pmatrix} 1/6 & 0 & 0 \\ 0 & 8/39 & -1/13 \\ 0 & -1/13 & 2/13 \end{pmatrix}
\]

a) Is the given inverse correct?

b) Assuming multivariate normality: which, if any, variables are independent? If multivariate normality is not assumed, what statements can be made about independence?

c) Test for equality of means in the two groups. (i.e., calculate the test statistic and indicate how to look up its significance).

d) Discuss the effect of sample size, and number of variables, on the test you have just performed.

3. Assume that a health department serving a metropolitan area with a population of approximately one million persons has initiated a cervical cancer screening program aimed at reaching approximately 5000 women. You have been asked to serve as statistical consultant for the program, with specific responsibility for preparing a brief statistical report suggesting:

a. Population groups the program might seek to reach, and the best location of program services for this purpose.

b. Statistical data and analyses needed in program operation and evaluation.

Briefly highlight the type of report you might prepare. Include, for example, ways you might use mortality and any available morbidity data in determining target groups, types of statistics on program services that might be maintained, and means by which you would seek to measure program success.
4. The following experiments were undertaken to measure the effect of an observable stationary patrol car on the speed of vehicles. The method of measurement used consisted in placing hidden cars with Vascar speed measuring equipment at the extreme points of a section of highway to determine the actual speed of vehicles at these points. In the control experiment I, no observable patrol vehicle is used; in experiment II, an observable patrol car is midway between the two hidden points; and in experiment III, the observable patrol car is nearer the terminal hidden point. These may be diagrammed as follows:

![Diagram of experiments]

where A is the initial point and C is the terminal point. The data were as follows:

<table>
<thead>
<tr>
<th>Experiment</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed at C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60</td>
<td>19</td>
<td>139</td>
<td>46</td>
</tr>
<tr>
<td>&gt;60</td>
<td>8</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>154</td>
<td>50</td>
</tr>
<tr>
<td>Speed at A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60</td>
<td>1</td>
<td>139</td>
<td>46</td>
</tr>
<tr>
<td>&gt;60</td>
<td>4</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>154</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>228</td>
<td>65</td>
</tr>
</tbody>
</table>

Discuss the analysis of these data in terms of a general categorical data model. Indicate the nature of the effects of the observable patrol car and differences among the experiments. Finally, give approximate values for test statistics directed at the various hypotheses of interest.
5. A test for the presence or absence of gout is based on the serum uric acid level $x$ in the blood. Assume $x$ is a normally distributed random variable. In healthy individuals $x$ has a mean of 5.0 mg/100ml and a standard deviation of 1.0 mg/100ml and in diseased persons it has a mean of 8.5 mg/100 ml and a standard deviation of 1.0 mg/100ml.

A test $T$ for the presence of gout is to classify those persons with a serum uric acid level of at least 7.0 mg/100ml as diseased.

(a) In epidemiologic terminology, the "sensitivity" of a test for a disease is defined as the probability of detecting the disease when it is present. Express the sensitivity of $T$ as a test for gout in terms of standard tabulated functions.

(b) The "specificity" of a test is defined as the probability of the test not detecting the disease when it is absent. Express the specificity of $T$ similarly.

Alternatively, this can be regarded as a statistical test of the null hypothesis

$$H_0: \mu = 5.0 \text{mg/100ml}$$

against the alternative hypothesis

$$H_a: \mu = 8.5 \text{mg/100ml}$$

The data is a single observation on the serum uric acid $x$ of the person to be classified.

What is the rejection region of the test $T$?

What is the size of the test (expressed in terms of tabulated functions)?

Relate the epidemiological concepts of sensitivity and specificity to $\alpha$, $\beta$, and $1-\beta$, the Type I and Type II error probabilities and the power of $T$ respectively.
6. In his continuing research into ways of improving our environment, a scientist has developed a fly ointment which, when applied to the legs of young flies, may or may not inhibit (slow down) the growth of their legs. Since short-legged flies are known to have more trouble with takeoffs and landings (fall on their little faces...), if the ointment works the scientist feels it will be a boon to future generations of mankind.

In order to prove that his ointment does indeed inhibit growth of flies' legs the scientist performed the following experiment. Three concentrations of ointment were prepared, containing 0(control), 0.2 mg/\text{fly} and 0.4 mg/\text{fly} of the active ingredient. Thirteen young flies (6 females and 7 males) were captured and assigned to treatments at random. After being assigned to treatments, the flies in each group were numbered 1,2,...,n; as shown in the following table:

<table>
<thead>
<tr>
<th>Concentration</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cell #</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>fly #</td>
<td>1,2,3,4</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cell #</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>fly #</td>
<td>1,2</td>
<td>1,2,3</td>
<td>1</td>
</tr>
</tbody>
</table>

For example, there were n_3 = 2 male flies who received the 0.4 concentration ointment.

Before the ointment was applied, the lengths of each fly's fore and hind legs were measured and recorded. The appropriate ointment was then applied to the right fore and hind legs of each fly, and the flies were allowed to mature under controlled conditions, after which their legs were re-measured.

The experimenters chose to analyze the data with the following model:

$$Y_{ij} = \mu_i + \alpha X_{ij} + e_{ij}, \quad j = 1,2,...,n_i$$

$$i = 1,2,...,6$$

(1)

where:

- $Y_{ij}$ is the difference between growth of the right foreleg (which received the ointment) and the growth of the left foreleg of the j-th fly in the i-th cell.

- $X_{ij}$ is the average length of the two forelegs of the j-th fly in the i-th cell before the ointment was applied.

continued
Questions:

a. In the model (1) above, what is the interpretation of the $\mu_i$'s?

b. In the model above, what is the interpretation of the $e_{ij}$'s?

c. What assumptions about the $e_{ij}$'s are necessary if we are to use the Gauss-Markov theorem to justify using least squares estimators of $\alpha$ and the $\mu_i$'s?

d. What further assumptions about the $e_{ij}$'s are necessary if we want to perform the usual tests of hypothesis?

Let $\beta = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \alpha)'$.

We can test hypotheses of the form $C\beta = 0$. Write down the C-matrices for testing the usual hypotheses of an analysis of covariance. That is, write down the C-matrices to test:

e. For a "Sex" main effect;

f. For a "Concentration" main effect;

g. For a Sex x Concentration interaction.

h. Whether $\alpha = 0$.

i. Write a C-matrix to test whether there is a significant linear trend in response, $y$, as concentration increases.

j. Write a C-matrix to test whether there is a significant quadratic trend in the response, $y$, as concentration changes. Write this C-matrix so that the resulting test statistic is not correlated with the test statistic generated by the test in part i.
Instructions

a) This is a closed-book examination.
b) The time limit is three hours.
c) Answer any four of the five questions which follow.
d) Put the answers to different questions on separate sets of papers.
e) Return the examination with a signed statement of the honor pledge.

1. In a certain population the number of children per family has a distribution with mean 3 and variance 6; and every child, independently of all others, has probability 1/2 of being male. Let a male child be chosen at random from this population, and let the random variable \( X \) be the number of brothers he has. Find \( E[X] \).

2. For each of the following distributions: write down the probability function; specify a model, including all necessary assumptions, which would produce data having the distribution; and give an illustrative example from the biological or health sciences.
   a) Negative binomial
   b) Multinomial
   c) Hypergeometric

3. Suppose \( X_1 \) is exponential with mean \( \lambda_1 \) and \( X_2 \) is exponential with mean \( \lambda_2 \). What is the distribution of \( X_1/X_2 \)? What is the mean of this distribution? Find an expression for the \( p^{th} \) percentile.
4. Given independent random samples, of size \( n \) each, from three normal populations having a common known variance \( \sigma^2 \) but unknown means:

\[
\begin{align*}
\{X_{1i}, i=1,2,\ldots,n\} & \text{ from } N(\mu_1,\sigma^2), \\
\{X_{2i}, i=1,2,\ldots,n\} & \text{ from } N(\mu_2,\sigma^2), \\
\{X_{3i}, i=1,2,\ldots,n\} & \text{ from } N(\mu_3,\sigma^2),
\end{align*}
\]

construct a most powerful test of size \( \alpha \) for testing the null hypothesis

\[
H_0: \mu_1 = \mu_2 = \mu_3 = 0
\]

against the alternative hypothesis

\[
H_A: \mu_1 = \alpha_1, \mu_2 = \alpha_2, \mu_3 = \alpha_3,
\]

where \( \alpha_1, \alpha_2, \alpha_3 \) are given constants. Be precise as to details of the critical region.

For the specific case \( \sigma = 2, \alpha_1 = .3, \alpha_2 = .4, \alpha_3 = 1.2, \alpha = 0.025 \), find the smallest sample size required to ensure that the power of the test against \( H_1 \) will be not less than 0.975.

5. In a certain region a 15\% random sample of hotels is taken and classified according to whether or not every bedroom has a telephone. The results are:

<table>
<thead>
<tr>
<th>Total Number of bedrooms</th>
<th>Hotels in Region</th>
<th>Hotels in Sample</th>
<th>Sample hotels with full number of telephones</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>400</td>
<td>80</td>
<td>2</td>
</tr>
<tr>
<td>6-20</td>
<td>1500</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>21-50</td>
<td>600</td>
<td>100</td>
<td>45</td>
</tr>
<tr>
<td>51-100</td>
<td>300</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>101-</td>
<td>200</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>3000</td>
<td>450</td>
<td>150</td>
</tr>
</tbody>
</table>

Compute the appropriate unbiased estimates of the total number of hotels with telephones in every room and the corresponding variance, assuming the data arose from (a) a simple random sample; (b) a stratified random sample as indicated in the table. Find the relative efficiency of the two methods of selection.
BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART III

(9-12 a.m., Friday, July 10, 1970)

INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is three hours.

c) Answer any three of the six questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Return the examination with a signed statement of the honor pledge.

1. Briefly specify

a. Some ways in which vital statistics might be used in development and evaluation of programs for detection and care of chronic illness, such as cancer and heart disease — including specific reference to some analytic techniques that might be employed.

b. Several ways in which other types of health statistics, e.g. on morbidity and on health services provided, might be used in development and evaluation of the chronic disease programs.

c. Some special needs for improved analytic techniques that might be used in applying vital or other health statistics in development and evaluation of chronic disease programs.
2. The following data were obtained from a sample of motorcycle operators on campus in order to investigate the dependence of average daily mileage on the operator's class status and marital status.

<table>
<thead>
<tr>
<th>Mileage</th>
<th>Very Low</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate Single</td>
<td>39</td>
<td>39</td>
<td>42</td>
<td>49</td>
<td>169</td>
</tr>
<tr>
<td>Graduate Single</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>Undergraduate Married</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Graduate Married</td>
<td>15</td>
<td>13</td>
<td>9</td>
<td>2</td>
<td>39</td>
</tr>
</tbody>
</table>

Discuss the statistical analysis of these data in terms of a model which accounts for the relative importance of the marital status effects, class status effects, and marital status x class status interaction on mileage.

3. In this problem, $N_p(\mu, V)$ refers to the $p$-variate normal distribution with mean vector $\mu$ (px1) and variance-covariance matrix $V$.

(a) What is the joint frequency function (probability density function) of $N_p(\mu, V)$?

(b) What is the moment generating function of $N_p(\mu, V)$?

(c) If $A$ is an $m \times p$ matrix ($m \leq p$) of rank $m$, and $\mathbf{\gamma}$ has the $N_p(\mu, V)$ distribution, what is the distribution of $\mathbf{z} = A\mathbf{\gamma}$? Prove your answer.

(d) Prove: If $\mathbf{\gamma}$ has the $N_p(\mu, V)$ distribution then the components $y_i$ are jointly independent if and only if $\text{cov}(y_i, y_j) = 0$ for all $i \neq j$, i.e., iff $V$ is diagonal.
4. Consider n randomized blocks of $k(>2)$ plots each where $k$ different treatments are applied (once each). The response of the plot in the $i^{th}$ block receiving the $j^{th}$ treatment is $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})'$, where

$$X_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}; \varepsilon_{ij} \sim N_2(0, \Sigma),$$

and $\Sigma$ is positive definite.

(a) Show how to test the null hypothesis $H_0^{(1)}: \tau_1 = \ldots = \tau_k$ vs $\tau_j \neq \tau_j'$, for at least one $(j, j')$.

(b) Treating $X^{(2)}$ as the concomitant variate and assuming that $\beta_i^{(2)} = 0, \forall i$, and $\tau_j^{(2)} = 0, \forall j$, show how to carry out the analysis of covariance test for the null hypothesis $H_0^{(2)}: \tau_1^{(1)} = \ldots = \tau_k^{(1)}$ vs $\tau_j^{(1)} \neq \tau_j^{(1)}$, for at least one $(j, j')$.

In the second problem, what is a simultaneous test for the paired differences $\tau_j - \tau_j', 1 \leq j < j' \leq k$?

5. A "gross nuptiality table" is a "life table" in which a population of single individuals suffers decrementation through marriage in the absence of mortality.

(a) Describe such a table including the assumptions involved in constructing it.

(b) Derive an expression for the average age at marriage in the stationary life table population.

(c) Show how mortality can be incorporated as an additional source of decrementation to arrive at what is called a "net nuptiality table".
6. The following data on the incidence of pneumococcal pneumonia have been collected in Chapel Hill during the past year.

<table>
<thead>
<tr>
<th>Age at last birthday years</th>
<th>Boys</th>
<th>Attack rate (per 1000)</th>
<th>Girls</th>
<th>Attack rate (per 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Pop.</td>
<td>Cases</td>
<td>Pop.</td>
</tr>
<tr>
<td>Under 1</td>
<td>15</td>
<td>200</td>
<td>13</td>
<td>195</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>225</td>
<td>26</td>
<td>305</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>210</td>
<td>16</td>
<td>215</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>202</td>
<td>12</td>
<td>205</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>195</td>
<td>7</td>
<td>180</td>
</tr>
</tbody>
</table>

Suppose you are interested in learning whether there is a sex difference in susceptibility or exposure to the disease among children under five years of age.

a. List at least six different tests that might be performed to test this hypothesis.

b. Describe in general terms the power of each test or reasons why one would want to use it, or not use it, in the present instance.

c. Perform at least two of these tests, and indicate what inferences you can draw about a sex differential from each test.
1. Consider a finite population of size N with population mean \( \bar{X} \) and population variance \( S^2 \). A simple random sample of size 3 is drawn with replacement from this population. Let \( \bar{x}' \) denote the unweighted mean over the distinct units in the sample.

   a) Show that the probabilities that the sample contains 1, 2, or 3 distinct units are

   \[
   P_1 = \frac{1}{N^2}, \quad P_2 = \frac{3(N-1)}{N^2}, \quad P_3 = \frac{(N-1)(N-2)}{N^2}.
   \]

   b) Show that, for estimating \( \bar{X} \), \( \bar{x}' \) is conditionally unbiased (i.e., conditional upon a fixed number of distinct units). Show also that \( \bar{x}' \) is unconditionally unbiased for \( \bar{X} \).

   c) Find the unconditional variance of the estimator \( \bar{x}' \).
2. A bivariate random variable $\mathbf{x} = (x_1, x_2)$ has a bivariate lognormal distribution if $\mathbf{y} = (y_1, y_2)$, where $y_1 = \log e x_1$ and $y_2 = \log e x_2$, has a bivariate normal distribution. That is

$$f(\mathbf{y}; \mu, \Sigma) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y} - \mu)^\top \Sigma^{-1} (\mathbf{y} - \mu)\right],$$

where $\mu = (\mu_1, \mu_2)$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$

a) Give explicitly the distribution of $\mathbf{x} = (x_1, x_2)$.

b) What is the distribution of $z = x_1 / x_2$?

c) What is $E(z)$?
3. In a two-organ system (e.g., kidneys, lungs) assume that the life-times of the two organs are independent exponentials, each with density function

\[ f(x) = \lambda_0 e^{-\lambda_0 x}, \quad \lambda_0 > 0, x \geq 0. \]

Assume also that as soon as one organ fails, at time \( x_o \), say, the surviving organ has density function

\[ g(y) = \lambda_1 e^{-\lambda_1(y-x_o)}, \quad y \geq x_o \]

with \( \lambda_1 > 2\lambda_0 \)

a) Show that at time \( t \) both organs are still functioning with probability

\[ P_0(t) = e^{-2\lambda_0 t}, \quad t \geq 0. \]

b) Show that at time \( t \) the system is still functioning on one organ with probability

\[ P_1(t) = \frac{2\lambda_0}{\lambda_1 - 2\lambda_0} (e^{-2\lambda_0 t} - e^{-\lambda_1 t}), \quad t \geq 0. \]

c) Show also that the density function of system life is

\[ h(t) = \frac{2\lambda_0 \lambda_1}{\lambda_1 - 2\lambda_0} (e^{-2\lambda_0 t} - e^{-\lambda_1 t}), \quad t \geq 0. \]
4. The following table records 292 litters of mice classified according to the size of the litter and the number of females in the litter.

<table>
<thead>
<tr>
<th>SIZE OF LITTER</th>
<th>NUMBER OF FEMALES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

a) With the assumption that for any specified size of litter the number of females is binomially distributed, show that there is no significant difference at the 5% level among litter sizes with respect to the proportion of females.

b) Test the significance of deviations from the assumption of the binomial distribution (the numbers have been chosen so that the arithmetic is easy), and then comment (no calculations are required) on how your procedure for this test would require modification if the first test had shown a variation in the proportion of females with litter size.
5.

a) Define the following terms:

i) Unbiased estimator
ii) Consistent estimator
iii) Efficient estimator
iv) Sufficient estimator

b) Give an example of an estimator for a parameter of a normal distribution which is unbiased but not consistent and prove your result.

c) Give an example of an estimator for a parameter of a normal distribution which is consistent but not unbiased and prove your result.

d) Show that there is no sufficient estimator for the location parameter \( \theta \) of the Cauchy distribution

\[ f(x) = \frac{1}{\pi[1+(x-\theta)^2]} \]
1. It is desired to select a sample from a certain population in order to estimate the mean of a particular variate, within ±2% of its true value with probability at least 95%. It is known from previous experience with this population that this variate has approximately the following distribution:

<table>
<thead>
<tr>
<th>Variate Value</th>
<th>Approximate Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Interval</td>
<td></td>
</tr>
<tr>
<td>0 &lt; X ≤ 10</td>
<td>10%</td>
</tr>
<tr>
<td>10 &lt; X ≤ 20</td>
<td>20%</td>
</tr>
<tr>
<td>20 &lt; X ≤ 30</td>
<td>40%</td>
</tr>
<tr>
<td>30 &lt; X ≤ 40</td>
<td>15%</td>
</tr>
<tr>
<td>40 &lt; X ≤ 50</td>
<td>10%</td>
</tr>
<tr>
<td>50 &lt; X ≤ 60</td>
<td>5%</td>
</tr>
</tbody>
</table>

Using the above information, what sample size will be necessary in order to satisfy the objectives of the survey (at minimum cost where total cost is presumed to be a linear function of the number of observations)?

2. A 2 x 5 factorial experiment is to be replicated four times using 40 individuals chosen from a population of 10,000 individuals. Use this situation to discuss
   a) a completely randomized design
   b) a randomized block design (What might be used for blocking factors? What is randomized? Describe the differences in the sampling procedures from a completely randomized design.)
   c) a components of variance vs. a fixed effects model.

Give the mathematical model in each case with the appropriate formulas for the partitioned sum of squares and give the corresponding degrees of freedom.
3. Let $X_1$, $X_2$, $X_3$ be independent Poisson variables with parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, respectively. (a) Obtain the UMP Test for

$$H_0: \lambda = \lambda_1 + \lambda_2 + \lambda_3 = \lambda_0 \text{ vs. } \lambda = \lambda^* > \lambda_0$$

(b) Suggest a test for

$$H_0: \lambda_1 = \lambda_2 = \lambda_3 \text{ vs. } \lambda_i \neq \lambda_j \text{ for at least one } (i,j): 1 \leq i < j \leq 3.$$ 

4. Suppose $X_1, X_2, \ldots, X_n$ (n$\geq$2) are mutually independent and normally distributed random variables with zero mean and unit variance. Let

$$U = \frac{X_1}{(X_1^2 + X_2^2 + \ldots + X_n^2)^{1/2}} \quad \text{and} \quad V = \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{X_1^2 + X_2^2 + \ldots + X_n^2}.$$ 

Show that the random variables $U$ and $V$ are mutually independent and find the distribution of $U$.

5. a) Find the distribution function of the largest observation in a random sample of size $n$ from a distribution with c.d.f. $F(x)$.

b) Given a sample of size $n$ from the uniform distribution on $[0, \theta]$, obtain a confidence interval for $\theta$ of confidence coefficient $1-\alpha$ based on the maximum likelihood estimator of $\theta$. 
RE-EXAMINATION FOR

BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9 A.M. - 5 P.M., Friday, June 25, 1971)

Instructions

a) This is a closed-book examination.

b) The time limit is eight hours.

c) Answer four of the five questions on Part I of the Basic Written Examination and the two attached questions.

d) Put the answers to different questions on separate sets of papers.

e) Put your code letter not your name on each question.

f) Return the examination with a signed statement of the honor pledge on a page separate from your answers.

Special Question for

Consider a model of the form

$$E(y) = \beta_1 x + \beta_2 x$$

(a) State an appropriate error model and derive the least squares equations for estimating the parameters $\alpha_1$, $\beta_1$, $\alpha_2$, $\beta_2$.

(b) What problems arise in the least squares estimation of the parameters $\alpha_1$, $\alpha_2$, $\beta_1$, $\beta_2$?

(c) Suggest an approximate method of solving these equations.

(d) Suppose it is known that $\beta_1=\beta_2=\beta$. Consider an alternative error model and find an explicit formula for an estimator of $\beta$. You may assume $\alpha_1+\alpha_2=\alpha$ is known in this case.
We have a random sample of size \( n \) from a distribution whose probability density function is assumed to be

\[
f_\theta(x) = \frac{1}{2}(1 + \theta x), \quad -1 \leq x \leq 1,
\]

\( f_\theta(x) = 0 \) otherwise, where the parameter \( \theta (-1 \leq \theta \leq 1) \) is unknown.

a) Find that unbiased estimator, \( T_0 \), of \( \theta \) which is a linear function of the sample mean \( \bar{X} \).

b) Show that if the true value of \( \theta \) is 0, then no unbiased estimator has a smaller variance than \( T_0 \).

c) Show that \( T_0 \) minimizes the maximum (with respect to \( \theta \)) of the variance of any unbiased estimator of \( \theta \).

d) Do you notice an undesirable property of \( T_0 \)?
1. Given a $2 \times 2$ (two way) analysis of variance design,

(10 points) a) Reformulate as a one way design by showing the relationships between the parameters of the two models.

(8 points) b) Show how the main effects and interactions may be tested by way of orthogonal contrasts applied to the one way model.

(7 points) c) Compare the Scheffe and Tukey methods of multiple comparisons discussing the relative advantages and disadvantages of each.

2. The concept of a hazard function is used in reliability. The hazard function $h(t)$ is defined such that $h(t)dt$ is the probability that the failure of a system will occur in the time interval $t$ to $(t+dt)$ given that no failure occurred prior to time $t$.

(8 points) a) If $f(t)$ and $F(t)$ are the density function and cumulative distribution function of the random variable $T$, the time to failure, show that

$$h(t) = f(t)/[1-F(t)].$$

(10 points) b) Find $f(t)$ if $h(t) = \alpha \beta t^{\alpha-1}$ ($\alpha > 0$, $\beta > 0$, $t > 0$).

(5 points) c) Find the mean and variance of $f(t)$ found in (b).

(2 points) d) Show that the exponential distribution is a special case of this distribution.
3. Let $x_i (i=1,2,\ldots,n+1)$ be a random sample of size $n+1$ from a normal $N(\mu,\sigma^2)$ population.

a) Prove that

$$\left[ \frac{\sum (x_i - \bar{x})^2}{b}, \frac{\sum (x_i - \bar{x})^2}{a} \right]$$

is a confidence interval for $\sigma^2$ which has confidence coefficient $1-\alpha$ provided

$$\int_a^b f_n(y) dy = 1-\alpha \ldots \text{(i)}$$

where

$$f_n(y) = \frac{y^{m-1} e^{-\frac{y}{2}}}{2^m \Gamma(m)}, \quad y>0.$$  

b) If the interval is to be the shortest, show that $a$ and $b$ must satisfy (i) and

$$f_{n+4}(a) = f_{n+4}(b).$$

c) Show that the likelihood-ratio test, at significance level $\alpha$, of $H_0: \sigma^2 = \sigma_0^2$ against $H: \sigma^2 \neq \sigma_0^2$ is to accept $H_0$ if

$$\frac{\sum (x_i - \bar{x})^2 / \sigma_0^2}{\sigma_n^2}$$

lies in the interval $(a,b)$, where $a,b$ now satisfy (i) and

$$f_{n+3}(a) = f_{n+3}(b).$$

[You may assume that $Y = \sum (X_i - \bar{X})^2 / \sigma_0^2$ has density function $f_n(y)$.]
4. Suppose two independent samples of sizes $n_1$ and $n_2$ are taken from $N(\mu_i, \sigma^2)$ for $i = 1, 2$. Denote then by $X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2}$ where the first $n_1$ come from population 1 and the remainder from population 2. For each $X_i$, an indicator variable $Z_i$ is associated, where $Z_i = 1$ if $X_i$ is in population 1 and $Z_i = 0$ if $X_i$ is in population 2. Let $r$ be the sample correlation between the $X$'s and the $Z$'s. Find the relation between $r$ and the usual unpaired two sample t-test.

5. Discuss the relative merits of using personal interviews, telephone interviews, and mailed questionnaires as methods of data collection for each of the following situations:

a) A television executive wants to estimate the proportion of viewers in the country who are watching his network at a certain hour.

b) A newspaper editor wants to survey the attitudes of the public toward the type of news coverage offered by his paper.

c) A city commissioner is interested in determining how homeowners feel about a proposed zoning change.

d) A county health department wants to estimate the proportion of dogs that have had rabies shots within the last year.
1. In the analysis of family records on a certain inherited disease, the distribution of affected children is often given in the form

\[ Pr(X=r) = \binom{s}{r} \frac{\theta^r(1-\theta)^{s-r}[1-(1-\pi)^r]}{1-(1-\pi\theta)^s} \]  \hspace{1cm} (1)

for \( r=1,2, \ldots, s \), where

- \( s \) = the number of children in a family (family size);
- \( \theta \) = segregation probability;
- \( \pi \) = ascertainment probability.

(a) Show that (1) is a probability mass function.

(b) Evaluate the expected number of affected children in a family of size \( s \), using distribution defined in (1).

(c) For small \( z \), \( (1-z)^k \approx (1-kz) \). Using this fact, derive an approximation to (1) for small \( \pi \). Show that the approximation is a probability mass function, and find its expectation.
2. Let $X$ be a random variable with a gamma distribution, having density

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty,$$

$$\alpha, \beta > 0.$$

Let $x_1, \ldots, x_n$ be a random sample from $f$.

(a) If $\beta$ is known, find a sufficient statistic for $\alpha$.

(b) If $\beta$ is known, what is the best linear unbiased (BLUE) estimator for $\alpha$? What is the variance of this estimator?

(c) If $\alpha$ is known, find a sufficient statistic for $\beta$.

(d) If $\alpha$ is known, $\beta$ unknown, find the Cramer-Rao lower bound on the variance of an unbiased estimator of $\beta$. Find an estimator which achieves this lower bound. You may assume that a family of gamma laws with fixed $\alpha$ and unknown $\beta$ is complete.

(e) Find the estimator for the variance of $f$ ($\alpha$ known, $\beta$ unknown) which is optimal in the sense of part d).

3. Given a random sample $x_1, \ldots, x_n$ from a normal distribution $N(\mu, \sigma^2)$:

(a) Construct the likelihood ratio test criterion for the hypothesis

$$H_0: \mu = \sigma^2$$

against alternatives

$$H_1: \mu \neq \sigma^2.$$

It is not necessary to reduce the test criterion to lowest terms.

(b) Transform the test criterion to one with known asymptotic distribution, and show exactly how you construct a critical region for a test of level $\alpha$.

(c) Find the asymptotic variance of the maximum likelihood estimator of $\mu$ under $H_0$. 

4. The Duchy of Grand Fenwick experiences a period of unprecedented economic growth from year 1 to year \( n_1 \) resulting from a massive American aid program. The end of such aid following year \( n_1 \) combined with a new policy of open immigration extended to citizens of underdeveloped countries, initiates a decreasing trend in per capita income from years \( n_1 \) to \( n \), in stark contrast to the increasing trend experienced during the period of American aid. It is suggested that the change in per capita income, years 1 to \( n \), may be appropriately described by two straight lines, necessarily intersecting at year \( n_1 \).

(a) Let \( y_i \) = per capita income in year \( i \).

\[ y = \text{vector of } y_i\text{'s, } i=1, \ldots, n \]

Define a parameter vector \( \beta \) and an appropriate \( X \) matrix for the model \( E y = X \beta \) described above.

(b) Give algebraic expressions for the elements of \( (X'X) \).

(c) Give a matrix expression for \( \hat{\beta} \), the least-squares estimates of \( \beta \), (in terms of \( X \) and \( y \)).

(d) Find \( \hat{\beta} \) for the following data, where \( n_1 = 3 \)

<table>
<thead>
<tr>
<th>year</th>
<th>Per capita income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

4 points

(e) Suggest two things to look at in evaluating the fit of such a model.

(f) Discuss briefly the effect of sampling considerations (how per capita income was measured) and possible distributional assumptions on the interpretation and use of this model.
5.

(a) Show that for a non-negative random variable \(X\) for which \(E|X|^k\) exists for some \(k>0\),

\[
E X^k = k \int_0^\infty x^{k-1} [1-F(x)] \,dx,
\]

where \(F(x) = P(X \leq x),\ 0 \leq x < \infty\).

Hence, or otherwise, show that

\[
EX = \int_0^\infty [1-F(x)] \,dx.
\]

(b) Let \(X\) be a non-negative integer valued random variable with

\[
p_i = P(X=i),\ i=0,1,2,\ldots,\infty.
\]

Also, let

\[
Q_i = P(X>i) = p_{i+1} + p_{i+2} + \ldots,\ i \geq 0;
\]

\[
p[q] = p(p-1)\ldots(p-q+1).
\]

Let then

\[
\mu^i[k] = E(X[k]) = \sum_{i=0}^\infty i[k] p_i,\ \text{k an integer (\geq 1)}.
\]

Show that

\[
\mu^i[k] = k \sum_{i=0}^\infty i[k-1] Q_i.
\]

(c) Let \(X\) have the Poisson distribution viz.,

\[
P(X=i) = e^{-m} \frac{m^i}{i!},\ i=0,1,\ldots,\infty;\ m>0,
\]

and let

\[
\mu_r = E(X-m)^r,\ \text{for } r=0,1,2\ldots.
\]

Show that

\[
\mu_{r+1} = mr\mu_{r-1} + \frac{d}{dm}(\mu_r),\ r \geq 1.
\]

Hence, or, otherwise, evaluate the first four central moments of \(X\).
1. Simple random samples drawn from each of three strata give data as follows:

<table>
<thead>
<tr>
<th>Stratum I</th>
<th>Stratum II</th>
<th>Stratum III</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>12</td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Estimate the population mean, and the variance of your estimate, under the assumption that the sample shown was drawn under proportional allocation.

b) Estimate the population mean and the variance of your estimate, if the true stratum weights are equal.

c) Assume the combined sample above was drawn by simple random sampling without stratification. Estimate the population mean and its variance.

d) Now assume, under the conditions of (a), that data is available on a concomitant variable x, such that for each of the y values observed above, the corresponding x value is (y-5). Suppose also that the actual totals for the x variable in the population are known to be 8, 12, and 20 million in stratum I, II, III, respectively. Calculate the combined and separate ratio estimates of the population total Y. Which would you prefer in this case, and why?
2. Let the probability $p_n$ that a family has exactly $n$ children be $\alpha^np$ when $n \geq 1$, and $p_0 = 1 - \alpha(1 + p + p^2 + \ldots)$. Suppose that the probability of giving birth to a boy is $\frac{1}{2}$ and that all births are independent events.

a) Show that for $k \geq 1$ the probability that a family has exactly $k$ boys is $2\alpha p^k/(2-p)^{k+1}$.

b) Given that a family includes at least one boy, what is the probability that there are two or more?

NOTE: Problem #3 is on the next page.

4. Let $X_1, \ldots, X_n$ be $n$ independent random variables distributed according to the $(0, \theta)$, $\theta > 0$, rectangular distribution, and let

$$U = \max_{1 \leq i \leq n} X_i \quad \text{and} \quad V = \min_{1 \leq i \leq n} X_i.$$ Show that

a) $U_n$ and $V_n$ are stochastically independent,

b) $U_n$ is the maximum likelihood estimator of $\theta$ and is also a sufficient statistic for $\theta$.

c) Is $U_n$ an unbiased estimator of $\theta$?

d) Obtain a 95% confidence interval for $\theta$ based on $U_n$. 
3. Suppose that a random variable $Y$ is related to a non-stochastic factor $X$ over the interval $-1 \leq X \leq 1$ by the relationship

$$\eta = E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2.$$ 

Let $N$ pairs of observations $(X_i, Y_i)$, $i = 1, 2, \ldots, N$, be available for which $\text{Var}(Y_i) = \sigma^2$, $\text{Cov}(Y_i, Y_{i'}) = 0$ if $i \neq i'$; further, let $\mu_k = \frac{1}{N} \sum_{i=1}^{N} X_{i}^{k}$, $k = 1, 2, 3$, and assume that $\mu_1 = \mu_3 = 0$. It is desired to approximate $\eta$ over $[-1, 1]$ using

$$\hat{Y} = b_0 + b_1 X,$$

where $b_0 = \frac{1}{N} \sum_{i=1}^{N} Y_i$ and $b_1 = \frac{\sum_{i=1}^{N} X_i Y_i}{\sum_{i=1}^{N} X_i^2}$.

a) Find $E(b_0)$ and $E(b_1)$.

b) Show that, in general,

$$J = \int_{-1}^{1} E[(\hat{Y} - \eta)^2]dX$$

can be written as $J = V + B$ where

$$V = \int_{-1}^{1} \text{Var}(Y)dX$$

and

$$B = \int_{-1}^{1} [E(\hat{Y}) - \eta]^2dX.$$ 

c) For the particular situation described earlier, show that $V$ is minimized when $\mu_2 = 1$, that $B$ is minimized when $\mu_2 = 1/3$, and that the value of $\mu_2$ which minimizes $J$ is the solution of a cubic equation.

NOTE: Problem #4 is on the preceding page.
5. Assume that an individual is subject to two types of failure and the two failure times are represented by random variables $X$ and $Y$. The joint distribution of $X$ and $Y$ can be described in terms of the probability function

$$P[X>s \text{ and } Y>t] = e^{-\lambda_1 s - \lambda_2 t - \lambda_{12} \max(s,t)}$$

for $s>0$, $t>0$

$$= 1 \text{ for } s<0, \quad t<0$$

where

$$\max(s,t) = s \text{ if } s \geq t$$
$$= t \text{ if } s < t$$

and $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\lambda_{12} > 0$.

a) Find the marginal distribution functions of $X$ and $Y$.

b) Find $P[X \leq x, \text{ and } Y \leq y]$

c) Prove $P[X>x+u, \text{ and } Y>y+u \mid X>x \text{ and } Y>y]$

$$= P[X>u \text{ and } Y>u]$$
1. Consider a population composed of the following

\[ y_1=1, y_2=2, y_3=3, \ldots, y_{29}=29, y_{30}=30 \]

a. What is the variance of the mean for a random sample of size \( n=6 \) selected without replacement?

b. What is the variance of the mean for the systematic sample which selects every \( 5^{th} \) sampling unit from a random start?

c. If the population were divided into 6 strata as follows:

\[ (y_1, y_2, \ldots, y_5); (y_6, y_7, \ldots, y_{10}); \ldots; (y_{26}, y_{27}, \ldots, y_{30}), \]

what is the variance of the estimate of the population mean for a stratified random sample of one observation per stratum?

d. What is the efficiency of the three sampling schemes relative to simple random sampling?

2. The probability distribution of a random variable \( Y \) is given by

\[ P[Y=y] = p \cdot q^{y-1}, \quad y=1,2,\ldots, \]

and \( q = 1-p \).

a. Show that the moment generating function of \( Y \) is given by

\[ M(s) = (\lambda e^{-s} - \lambda + 1)^{-1}, \text{ where } \lambda = \frac{1}{p}. \]

Let \( Y_1 \) and \( Y_2 \) be two independent observations from the above distribution, and let \( Z_1 = \max(Y_1, Y_2) \) and \( Z_2 = \min(Y_1, Y_2) \).

b. Find the probability distribution of \( Z_1 \).

c. Find the conditional distribution of \( Z_1 \) given \( Z_2 \).
3. Let
\[ Y_i = \alpha + \beta t_i + \varepsilon_i, \quad i=1,2,\ldots,n, \]
where \( \varepsilon_i \sim N(0, \sigma_i^2) \), \( t_i \) is a known positive constant for every \( i \), and \( \alpha \) and \( \beta \) are parameters to be estimated. Obtain the maximum likelihood estimate of \( \beta \) when

- a. \( \alpha \) is unknown and \( \sigma_i^2 = \sigma^2 \), \( i=1,2,\ldots,n \).
- b. \( \alpha = 0 \) and \( \sigma_i^2 = \sigma^2 \), \( i=1,2,\ldots,n \).
- c. \( \alpha = 0 \) and \( \sigma_i^2 = t_i \sigma^2 \), \( i=1,2,\ldots,n \).
- d. \( \alpha = 0 \) and \( \sigma_i^2 = t_i^2 \sigma^2 \), \( i=1,2,\ldots,n \).

4. During a holiday season, a merchant will make a net profit of \( \alpha \) dollars for each unit amount of some commodity sold and will lose \( \beta \) dollars for each unit amount left unsold at the end of the season. Suppose that the merchant can purchase this commodity only before the season starts, and that the total amount his customers will want to buy during the season can be viewed as a continuous random variable with a given probability density function \( f(x) \) for \( x > 0 \).

- a. If \( k \) is the number of units which the merchant should purchase in order to maximize the expected value of his profit \( G \) on the commodity, show that \( k \) is given by

\[ F(k) = \frac{\alpha}{\alpha + \beta}. \]

- b. Compute \( k \) when \( \alpha = 1, \beta = 2, \) and

\[ f(x) = .0002xe^{-0.001x^2}, \quad x > 0. \]
Clinical studies in which several clinics participate in evaluation of therapies using a standard protocol have become relatively common. Assume an experimental design in which patients who meet protocol requirements are randomly assigned to treatments within clinics, and that the clinics that participate in the study are a random sample from the population of clinics that might use the therapies. An appropriate linear model might be

\[ y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \epsilon_{ijk}, \]

where \( E(y_{ijk}) = \mu + \alpha_i \); \( \mu \) is the general mean, \( \alpha_i \) is the differential effect of the \( i \)-th treatment, \( \beta_j \) is a random clinic effect, \( (\alpha \beta)_{ij} \) is a random effect due to interaction between the \( j \)-th clinic and the \( i \)-th treatment, \( \epsilon_{ijk} \) is a random effect due to the \( k \)-th patient in the \( i \)-th treatment and \( j \)-th clinic; \( \beta_j \), \( (\alpha \beta)_{ij} \), and \( \epsilon_{ijk} \) are independently distributed, each with mean 0 and respective variances \( \sigma_B^2 \), \( \sigma_{\alpha \beta}^2 \), \( \sigma_c^2 \) for \( i=1, \ldots, t \), \( j=1, \ldots, c \); \( k=1, 2, \ldots, n_{ij} \).

a. What is the variance of \( y_{ijk} \)?

b. What is the covariance between two observations in the same clinic and treatment?

c. What is the covariance between two observations in the same clinic, but different treatments?

d. What is the covariance matrix for a vector of treatment means from a specified clinic?

e. Find the covariance between \( \sum_j c_{1j} \bar{y}_{ij} \) and \( \sum_j c_{2j} \bar{y}_{ij} \), where \( \sum_j c_{1j} = 0, \sum_j c_{2j} = 0 \), \( \bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \), and \( i \) is fixed.
1. Suppose that $X$ and $Y$ are random variables with joint density

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} (x^2-2\rho xy+y^2)\right\},$$

$-\infty < x < \infty, -\infty < y < \infty.$

a) Obtain the moment generating function of the joint distribution of $X$ and $Y$.

b) Show that the correlation between $X^2$ and $Y^2$ is given by $\rho^2$.

c) Find the distribution of

$$Q = (X^2-2\rho XY+Y^2)/(1-\rho^2)$$

d) Show that $U = X+Y$ and $V = X-Y$ are independently distributed.
2. Let \( Y_1 < \ldots < Y_n \) be the ordered random variables of a sample of size \( n \) from the rectangular \((0, \theta)\) distribution, where \( 0 < \theta < \infty \). By a careless mistake, the observations \( Y_{(k+1)}, \ldots, Y_{(n)} \) were recorded incorrectly, so they were discarded. (Here, \( k \) is a positive integer less than \( n \).)

(a) Show that the conditional distribution of \( Y_1, \ldots, Y_{(k-1)} \), given \( Y_{(k)} = y \), is independent of \( \theta \).

(b) Hence, or otherwise, obtain the maximum likelihood estimator of \( \theta \) and show that it is a function of \( Y_{(k)} \) alone.

(c) If \( k/n \to p \) as \( n \to \infty \), for some \( 0 < p < 1 \), what can you say about the asymptotic distribution of the maximum likelihood estimator of \( \theta \)?

3. A bus line has length \( L \). The probability density for the point \( X \) at which a passenger gets on the bus is proportional to \( X(L-X)^2 \), and the probability density for the point \( Y \) at which a passenger who entered at point \( X \) gets off the bus is proportional to \( (Y-X) \). Find the probability that

(a) A passenger will get on the bus before point \( z \),

(b) A passenger who got on the bus at point \( x \) will get off after point \( z \),

(c) A passenger who gets off at point \( y \) got on before point \( z \).
4. Suppose that $N_1$ represents the number of traffic accidents occurring at a certain highway location during the year prior to the installation of a certain improvement at that location, and that $N_2$ represents the number occurring during the year following the improvement. Suppose, moreover, that $N_1$ and $N_2$ are assumed to be independent Poisson random variables with means $\mu_1$ and $\mu_2$, respectively.

a) What is the probability that $n_1 + n_2$ accidents occur during the two-year period?

b) What is the conditional probability that $n_2$ accidents occur during the second year given that $n_1 + n_2$ occur during the two-year period?

c) Show that the conditional probability of part (b) is the binomial probability $B(n_1; n_1 + n_2, \theta)$, where

$$\theta = \frac{1}{1+p}$$

and

$$p = \frac{\mu_2}{\mu_1}.$$

d) Suggest a test of $H_0: \rho = 1$ versus $H_1: \rho < 1$ with approximate significance level $\alpha$, and comment on any properties of your suggested test which you think are particularly good or bad.
5. Consider the following set of data:

<table>
<thead>
<tr>
<th>GROUP</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

a) Perform a simple linear regression of Y on X for the twelve bivariate observations given.

b) For each observation, calculate the residual from the fitted regression line.

c) Perform a one-way analysis of variance on the residuals, to see if the mean residuals differ significantly from group to group.

d) Place the ANOVA test you have just performed in the context of general multiple regression analysis. (It is not necessary to get very technical.) Do parts a)-c) relate to a classical procedure familiar to you? Give a crucial restriction on the use of this technique.
1. Let $\psi(x)$ be a monotonic increasing function of $x$ such that $\psi(x) \to 0$ as $x \to x_0$ and $\psi(x) \to \infty$ as $x \to \infty$. Let $W(x)$ be the cumulative distribution function (CDF) of a random variable $X$. Let $O$ be the CDF of a random variable $X$.

a) Prove that $Y = 2\psi(X)$ has a $\chi^2$ distribution with 2 degrees of freedom.

b) Let $X_1, X_2, \ldots, X_k$ be $k$ ($k>2$) independent random variables, $X_j$ having the CDF

$$F_{X_j}(x_j) = 1 - \exp\left[-\lambda u_j(x_j)\right], \lambda > 0, j = 1, 2, \ldots, k,$$

where $u_j(x_j)$ satisfies the conditions of $\psi(x)$ in part (a) and $\lambda (>0)$ is a constant.

Using part (a), show that

$$Y = 2\lambda \sum_{j=1}^{k} u_j(x_j)$$

has a $\chi^2$ distribution with $2k$ degrees of freedom.

c) Suppose that $\lambda$ is an unknown parameter and is estimated from the formula

$$\hat{\lambda} = (k-1) / \sum_{j=1}^{k} u_j(x_j)$$

Show that $\hat{\lambda}$ is an unbiased estimator of $\lambda$.

**HINT:** Note that $\hat{\lambda} = 2\lambda(k-1)/\chi^2_{2k}$.

d) Find the variance of $\hat{\lambda}$. 
2. Consider a simplistic model of morbidity (illness) for some ailment, say migraine headache. For any given calendar week, an individual belongs to one of two states:

- **Well (W)** if he experiences no symptoms or signs of migraine headache at any time during the week;
- **Morbid (M)** if he experiences symptoms or signs of migraine headache at some time during the week.

For a population P under study, assume that the probability that a randomly chosen individual will be well in week (k+1) depends only on his health statuses (W or M) in weeks k and (k-1). Denote the four probabilities for the status "well in week (k+1)" as below and assume that all probabilities are strictly between 0 and 1.

<table>
<thead>
<tr>
<th>HEALTH STATUS WEEK (k-1)</th>
<th>WEEK k</th>
<th>PROBABILITY OF BEING WELL IN WEEK (k+1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>W</td>
<td>( p_{ww} )</td>
</tr>
<tr>
<td>W</td>
<td>M</td>
<td>( p_{wm} )</td>
</tr>
<tr>
<td>M</td>
<td>W</td>
<td>( p_{mw} )</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>( p_{mm} )</td>
</tr>
</tbody>
</table>

a) Let the states of a Markov chain be defined by pairs \( z=(z_{k-1}, z_{k}) \), where \( z_j \) is the health status of an individual in week \( j \).

Rename the states as: \( 1=(W,W) \); \( 2=(W,M) \); \( 3=(M,W) \); \( 4=(M,M) \). Let \( X_n \) be the state random variable. Verify that \( \{X_n : n \geq 0\} \) is a homogeneous Markov chain over the state space \( \{1,2,3,4\} \).

b) Write out the one- and two-step transition matrices.

c) Write down the equilibrium equations for this chain.

d) Suppose that \( p_{ww} = 0.8 \), \( p_{wm} = 0.2 \), \( p_{mw} = 0.4 \), and \( p_{mm} = 0.6 \). Assume no emigration, immigration, births or deaths relative to the population P. After a large number of weeks, what is the probability that an individual selected randomly from P will be well in the calendar week of his selection? In the week following?
3. Consider a sequence of independent random variables $X_1, X_2, X_3, \ldots$, each having the same expected value $\mu$ and standard deviation $\sigma$.

The distribution of
$$Y_i = \frac{X_i - \mu}{\sigma}$$
does not depend on $\mu$ or $\sigma$, and
$$E[(Y_i - Y_j)^{-2}] = \lambda (\ll \infty) \text{ for } i \neq j.$$

Firstly, $X_1$ and $X_2$ are observed. Then, a further set of $N$ observations $X_3, X_4, \ldots, X_{N+2}$ is taken, where $N$ is the least integer not less than $M(X_1 - X_2)^2$, $M$ being a fixed positive number. The quantity
$$\bar{X} = \frac{1}{N} \sum_{k=1}^{N} X_{k+2}$$
is then calculated.

a) Show that
$$E(N) \geq 2M\sigma^2.$$

b) Show that
$$\Pr\left[ | \bar{X} - \mu | < \varepsilon \right] \geq 1 - \frac{\lambda}{M\varepsilon^2}$$

for any positive number $\varepsilon$. 
4. It is required to take samples of sizes \( n_1 \) and \( n_2 \) from two normal populations having known variances \( \sigma_1^2 \) and \( \sigma_2^2 \), the purpose being to test the null hypothesis of equality of the population means \( \mu_1 \) and \( \mu_2 \) against the alternative \( \mu_1 > \mu_2 \).

a) Setting 
\[
V = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
\]
and using the notation which defines \( z_k \) by \( \Pr(Z > z_k) = k \), find the expression for \( V \) which satisfies the conditions that the level of significance be \( \alpha \) and the power of the test against the alternative \( (\mu_1 - \mu_2) = \delta \) be \( (1-\beta) \).

b) For the \( V \) given in any such situation, find formulas for \( n_1 \) and \( n_2 \) (in terms of \( V, \sigma_1, \sigma_2 \)) which will minimize sampling cost if the cost of making an observation in Population 1 is four times that in Population 2. What are the specific sample size values if \( \sigma_1 = 5 \), \( \sigma_2 = 4 \), \( \alpha = .05 \), and the power is to be .90 when \( \delta = 3 \)?
5. Data is available of the form \((X_i, Y_i, Z_i), i=1,2,\ldots,n\).

a) Consider the statistical model

\[ Y_i = \alpha + \beta X_i + \varepsilon_i, \]

where \(\alpha\) and \(\beta\) are fixed (unknown) parameters and the \(\varepsilon_i\) are independent random variables with zero mean and common variance \(\sigma^2\). Give formulae for \(\hat{\alpha}\) and \(\hat{\beta}\), the least squares estimators of \(\alpha\) and \(\beta\). Give formulae for the variances of \(\hat{\alpha}\) and \(\hat{\beta}\). Derivations are not necessary.

b) Consider the statistical model

\[ Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i, \]

where \(\alpha, \beta\) and \(\gamma\) are fixed (unknown) parameters and the \(\varepsilon_i\) are independent random variables with zero mean and common variance \(\sigma^2\). An analysis of variance for the data, under this model, is given below:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>1</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Z/X</td>
<td>1</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Residual</td>
<td>23</td>
<td>65.0</td>
<td>2.8</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>78.0</td>
<td></td>
</tr>
</tbody>
</table>

The correlation (Pearson) between \(Y\) and \(Z\) is \(r_{YZ} = .4\). Calculate:

i) \(r_{XY.Z}\), the partial correlation between \(Y\) and \(X\), adjusting for \(Z\);

ii) \(r_{Y.XZ}\), the multiple correlation of \(Y\) with \(X\) and \(Z\);

iii) \(r_{XZ}\), the Pearson correlation of \(X\) with \(Z\).

c) Consider the statistical model

\[ Y_i = \alpha^* + (\beta^* + Z_i^2) X_i + \varepsilon_i^*, \]

where \(\alpha^*\) and \(\beta^*\) are fixed (unknown) parameters and the \(\varepsilon_i^*\) are independent random variables with zero mean and common variance \(\sigma^2\). Derive formulae for \(\hat{\alpha}^*\) and \(\hat{\beta}^*\), the least squares estimators of \(\alpha^*\) and \(\beta^*\). What are the variances of these estimators?
1. Let \( \{X_i\} \) be a sequence of independent and identically distributed random variables and let \( S_0 = 0, S_k = X_1 + \ldots + X_k, \ k \geq 1. \) Also, let \( N \) be a non-negative integer-valued random variable, independent of the \( X_i. \)

(a) Show that whenever the moments exist,

\[
E(S_N) = E(N)E(X_1)
\]

\[
V(S_N) = E(N)V(X_1) + V(N)[E(X_1)]^2
\]

(b) Suppose that

\[
X_i = \begin{cases} 
0, & \text{with probability } 1-p \\
1, & \text{with probability } p
\end{cases}, \quad 0 < p < 1,
\]

and \( N \) has the Poisson distribution with parameter \( \lambda(>0). \)

Find \( E(S_N) \) and \( V(S_N). \)

(c) What is the actual distribution of \( S_N \) in part (b)?
2. Consider a bivariate normal distribution with mean vector \( \mu = (\mu_1, \mu_2) \) and dispersion matrix \( \begin{pmatrix} 2.25 & 1.68 \\ 1.68 & 4.00 \end{pmatrix} \). Also, let \( (\overline{X}_n, \overline{Y}_n) \) be the mean vector of a sample of size \( n \) from this distribution.

(i) Show that for testing

\[ H_0: \theta = \mu_1 - \mu_2 = 0 \ vs \ H_1: \theta > 0 \]

the UMP test based on the set of \( n \) differences \( (X_i - Y_i, 1 \leq i \leq n) \) consists in rejecting \( H_0 \) if \( \overline{X}_n - \overline{Y}_n > c \) where \( P(\overline{X}_n - \overline{Y}_n \geq c | H_0) = \alpha \), the level of significance.

(ii) Find an expression for the power function of the UMP test for \( \theta = \theta_1 (>0) \).

(iii) Determine \( n \) such that for \( \theta_1 = 0.5 \) and \( \alpha = 0.05 \), the power is equal to 0.95.

(iv) For the same \( n \) and \( (\alpha, \theta_1) \), determine the power of the classical sign-test based on the signs of the difference of the two variates for the \( n \) observations (vectors).
5. Suppose that a set of three measurements \((Y, X_1, X_2)\) is available on each experimental unit in a study.

(a) Determine the values of \((\beta_1, \beta_2)\) which minimize the variance of

\[ Z = Y - \alpha - \beta_1 X_1 - \beta_2 X_2, \]

where you may assume that the dispersion matrix of \((Y, X_1, X_2)\) (say, equal to \(\Sigma\)) is known. Does the solution depend on \(\alpha\)?

(b) Given a sample \((Y_i, X_{1i}, X_{2i}), i = 1, \ldots, n\) from a trivariate normal distribution, find out the maximum likelihood estimator of the conditional mean of \(Y\) given \(X_1 = x_1\) and \(X_2 = x_2\).

(c) Suppose that in (b), it is given that the mean vector of the trivariate normal distribution is equal to \(\mathbf{0}\). Does the solution remain the same?

4. (a) Let \(X\) be a random variable with a continuous distribution function \(F(x), -\infty < x < \infty\). Find the distribution function of the random variable \(Y = F(X)\). What is this result usually called?

(b) Illustrate the use of this result in simulating a random sample from a simple exponential distribution.

(c) Let \(X_1, \ldots, X_n\) be \(n\) independent random variables with continuous distribution functions \(F_1(x), \ldots, F_n(x)\), all defined on \((-\infty, \infty)\).

Using the result in (a) show that each of \(U_n = -2\sum_{i=1}^{n} \log F_i(X_i)\) and \(V_n = -2\sum_{i=1}^{n} \log(1-F_i(X_i))\) is distributed as chi-square with \(2n\) degrees of freedom.
5. One way of dealing with the problem of non-response in a survey is to make efforts to collect information from a sub-sample of the units not responding in the first attempt. In a mail enquiry, \( n \) units were selected from a population of \( N \) using simple random sampling without replacement (SRSWOR). Of these, \( n_1 \) units responded. From the \( n_2 (= n-n_1) \) non-responding units, \( r_2 \) were selected with SRSWOR and the required information was obtained by personal interview.

(a) Let \( n_2 = vr_2 \), \( v (>1) \) fixed in advance, \( N_1 \) be the number of units in the population that would have responded in the first attempt, \( N_2 = N - N_1 \) and \( w_j = N_j / N, j = 1, 2 \). Show that

\[
E(n_1/N_1) = vE(r_2/N_2).
\]

(b) Let \( y \) be the variable under study, \( \bar{y}_1 \) and \( \bar{y}_2 \) be the sample means based on the \( n_1 \) units responding in the 1st attempt and on the sub-sample of size \( r_2 \) units respectively. Show that

\[
\bar{y}' = (n_1\bar{y}_1 + n_2\bar{y}_2)/n
\]

is an unbiased estimator of the population mean \( \bar{y} \).

(c) Show that

\[
V(\bar{y}') = \frac{N-n}{N} \frac{S^2}{n} + \frac{(v-1)}{n} \frac{w_2 S^2_2}{n}
\]

where \( S^2 \) is the variance of the whole population and \( S^2_2 \) is the variance within the non-response stratum.
1. Suppose that after suitable observation, it is decided that the number (N) of auto accidents at a given intersection this year has a Poisson distribution with parameter \( \lambda \). A typical accident involves one or more cars. Let \( X \) be a random variable denoting the number of autos involved in an accident. We assume that the numbers for different accidents are independent and identically distributed with some distribution.

(a) Derive the distribution of \( T \), the total number of autos involved in \( N \) accidents at the intersection this year. Name the resulting distribution.

(b) Find the generating function of \( T \).

(c) If we are given the rule of thumb that, in practice, only one, two or three autos are involved in a given accident with corresponding frequencies in the ratio 5:3:2, what is the mean of \( T \)?

2. Let \( X_1, X_2, \ldots, X_n \) be a random sample from a Poisson distribution with parameter \( \theta \), i.e.,

\[
f(x; \theta) = \frac{(x!)^{-1} \theta^x e^{-\theta}}{x!}
\]

with \( \theta > 0 \) and \( x = 0, 1, 2, \ldots \).

Let \( \Pr(X \leq 1) = (1+\theta)e^{-\theta} = \lambda \).

(a) Show that \( Y = \sum_{i=1}^{n} X_i \) is a complete sufficient statistic for \( \theta \).

(b) Let \( u(X_1) = 1 \) for \( X_1 \leq 1 \) and \( u(X_1) = 0 \) for \( X_1 > 1 \). Show that \( u(X_1) \) is unbiased for \( \lambda \).

(c) Obtain the minimum variance unbiased estimate for \( \lambda \).

(d) Show that the statistic obtained in (c) is indeed unbiased for \( \lambda \).
3. In a certain region a 15% random sample of hotels is taken and classified according to whether or not every bedroom has a telephone. The results are:

<table>
<thead>
<tr>
<th>Total Number of Bedrooms</th>
<th>Hotels in Region</th>
<th>Hotels in Sample</th>
<th>Sample Hotels With Full Number of Telephones</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>400</td>
<td>80</td>
<td>2</td>
</tr>
<tr>
<td>6-20</td>
<td>1500</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>21-50</td>
<td>600</td>
<td>100</td>
<td>45</td>
</tr>
<tr>
<td>51-100</td>
<td>300</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>101+</td>
<td>200</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>3000</td>
<td>450</td>
<td>150</td>
</tr>
</tbody>
</table>

Compute the appropriate unbiased estimates of the total number of hotels with telephones in every room and the corresponding variance, assuming the data arose from

(a) a simple random sample without the knowledge of the strata,
(b) a stratified random sample as indicated in the table.
(c) Find the relative efficiency of the two methods of selection.
4. Let $X_1, X_2, \ldots, X_n$ be a random sample from the following probability density function:

$$f(x; \theta_1, \theta_2) = \theta_2^{-1} e^{-(x-\theta_1)/\theta_2}$$

where $\theta_1 < x < \infty$, $-\infty < \theta_1 < \infty$, and $0 < \theta_2 < \infty$.

(a) Determine the most powerful test of size $\alpha$ for testing

$$H_0: \theta_2 = \theta_2', \quad \text{with } \theta_1 \text{ known}$$

against

$$H_1: \theta_2 = \theta_2'' < \theta_2', \quad \text{with } \theta_1 \text{ known}.$$ 

(b) Is the test obtained in (a) the uniformly most power test of size $\alpha$ for testing

$$H_0: \theta_2 = \theta_2', \quad \text{with } \theta_1 \text{ known}$$

against

$$H_1: \theta_2 < \theta_2', \quad \text{with } \theta_1 \text{ known}.$$ 

(c) Obtain the likelihood ratio test for testing

$$H_0: \theta_1 = \theta_1', \quad \text{with } \theta_2 \text{ unknown}$$

against

$$H_1: \theta_1 > \theta_1', \quad \text{with } \theta_2 \text{ unknown}.$$ 

(d) Show that the test in (c) may be based on a statistic which has an $F$ distribution under the null hypothesis.
5. A wrist-watch manufacturing concern intended to study the effect of humidity on the rusting of a particular metal they used for the watches. Two countries A (a tropical one) and B (a temperate one) were chosen and within each country five different places with varying degrees of humidity were selected. For each place, ten specimens of the metal were exposed to the normal environment for a year. For the i-th country, j-th place and k-th specimen, let \( Y_{ijk} \) denote the amount of rusting, and let \( t_{ij} \) be the average amount of humidity (during the experimental year) in the j-th place of the i-th country. Note that \( i = A, B \); \( j = 1, \ldots, 5 \); \( k = 1, \ldots, 10 \).

It is conceived that for each country, a simple linear regression of \( Y \) on \( t \) describes the model adequately. The concern wants to know:

(a) Whether the rate of growth of rusting with humidity is the same for both the countries, and

(b) Whether the typical rusting at 0 amount of humidity is the same for both the countries.

(i) State your assumptions carefully,

(ii) Provide suitable statistical tests for both (a) and (b),

(iii) Comment on their merits and demerits.
BASIC WRITTEN EXAMINATION IN BIOSTATISTICS

SPECIAL OFFERING OF PART I

(9:00 a.m. - 1:00 p.m., Wednesday, August 3, 1977)

1. Sampling

Simple random samples drawn from each of three strata give data as follows:

<table>
<thead>
<tr>
<th>Stratum I</th>
<th>Stratum II</th>
<th>Stratum III</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>12</td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>38</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. **Hypothesis Testing**

Define

(a) Type I and Type II errors;
(b) Power of a test;
(c) Uniformly most power test;
(d) Unbiased test.

A box contains \( N \) items numbered from 1 to \( N \). One item is selected at random and its number \( X \) observed. We wish to test the hypothesis \( N = 5 \) at \( \alpha = 0.2 \). The following 5 tests are thus available at the desired level:

- \( T_1 \) rejects \( H_0 \) if \( X = 1 \) or \( X > 5 \);
- \( T_2 \) rejects \( H_0 \) if \( X = 2 \) or \( X > 5 \);
- etc.

(e) Calculate and sketch the power curve of each of the tests \( T_1, T_2, \ldots, T_5 \) against the alternatives \( N = 1, 2, 3, 4, 5, 6, 7, \ldots \). Are these tests unbiased?

(f) Is there a uniformly most powerful test among \( T_1, T_2, \ldots, T_5 \)? Justify your answer.
3. **Estimation**

The random variable $X$ is said to have a **Pareto distribution** if $F(x) = P\{X \leq x\}$ is given by

$$F(x) = \begin{cases} 
0, & x \leq x_0 \\
1 - (x_0/x)^a, & x > x_0 
\end{cases}$$

where both $x_0$ and $a$ are positive numbers.

(a) Derive the form of the probability density function $f(x)$ corresponding to $F(x)$.

(b) Find $E(X^k)$ for $k = 1, 2$ and $V(X)$.

(c) Under what condition on $a$ does $E(X^k)$ exist for $k = 1, 2$.

(d) If $a$ is known and if $X_1, \ldots, X_n$ are independent random variables having the same distribution $F(x)$, show that there exists a sufficient statistics for $x_0$.

(e) For a specified $a$, find a 95% confidence interval for $x_0$ based on the sufficient statistic.
4. **Stochastic Processes**

The special case of the general birth and death process in which the birth rate is $\lambda_n = \lambda n + \alpha$, $\lambda$ and $\alpha > 0$, and the death rate is $\mu_n$, $\mu > 0$, is called a linear growth process with immigration.

(a) Write down the transition probabilities corresponding to the above birth and death rates. (We assume time homogeneity.)

(b) Is this a Markov process? Show why or why not.

(c) Determine differential-difference equations satisfied by the transition probabilities in (a).

(d) Use (c) to determine the (time-dependent) mean.

(e) Give conditions for which limits of the expressions in (a) and (d) exist and write down the limits.
5. **Experimental Design**

The problem often arises in experimental design over the large area: how many sites and how many replications for each experiment should be used to obtain some 'optimal' solution under certain conditions. We will consider the problem as follows:

Let \( y_{ij} \) denote the yield of some crop observed in the \( i \)-th experiment (site) \((i = 1, 2, \ldots, m)\) on the \( j \)-th plot, \((j = 1, 2, \ldots, r)\). (We assume that the number of sites is very large, and for practical purpose infinite.)

Let

\[
y_{ij} = u_i + u_{ij} + \epsilon_{ij},
\]

with

\[
\epsilon_{ij} \sim N(0, \sigma_1^2), \quad u_i \sim N(0, \sigma_2^2),
\]

\[
\text{cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0, \quad \text{cov}(\epsilon_{ij}, u_i) = 0.
\]

(a) Let

\[
\bar{y} = \frac{1}{rm} \sum_{i=1}^{m} \sum_{j=1}^{r} y_{ij}
\]

denote the observed overall average yield of this crop. Find \( \text{var}(\bar{y}) \).

(b) As costs of an experiment, let \( c_1 \) be the cost of obtaining each plot and \( c_2 \) be the cost of preparing and conducting a trial within a plot. Give an expression for the total cost \( C \) of an experiment with \( m \) plots and \( r \) trials within a plot.

Find the optimal allocation with respect to \( r \) and \( m \), when it is desired:

(c) to minimize \( \text{var}(\bar{y}) \) for given total cost, \( C = C_0 \);

(d) to minimize total cost \( C \) for given variance \( \text{var}(\bar{y}) = V_0 \).
1. Consider a continuous process, denoting the interarrival time between the \((i-1)\)th and the \(i\)th events by \(T_i\) for \(i \geq 1\). Assume that the \(T_i\) are independent and identically distributed as an exponential, specifically

\[
f(t; \theta) = \theta e^{-\theta t}
\]

for \(t \geq 0\) and \(\theta > 0\).

a) Verify that the no memory waiting time assumption is satisfied, i.e.,

\[
\Pr(T > t + u | T > u) = \Pr(T > t) \quad \text{for every } t \geq 0, u \geq 0.
\]

b) Let \(T_1, T_2, \ldots, T_n\) be the interarrival times for the \(n\) events observed between time zero and time \(t\). Find the distribution of the total time \(W_n\) required to observe \(n\) events.

c) Let \(N_t\) be the number of events occurring in an interval of length \(t\). Find the distribution of \(N_t\).

d) Suppose that

\[
f(t; \theta) = \theta^2 t e^{-\theta t} \quad \text{for } t \geq 0 \text{ and } \theta > 0.
\]

Verify whether the no memory waiting time assumption is satisfied or not.
2. Let a random variable $Y = \log X$ be normally distributed with expected value $\xi$ and standard deviation $\sigma$.
   a) Obtain the expected value of $X$, $E(X) = \mu$, say, in terms of $\xi$ and $\sigma$.
   b) Let $X_1, X_2, \ldots, X_n$ be a random sample from this distribution. Find the maximum likelihood estimator of $\mu$ ($\hat{\mu}$, say).
   c) Find the expected value and variance of $\hat{\mu}$ (in terms of $\xi$ and $\sigma$).

3. For the linear model

   $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

   suppose that $(Y_i, x_{1i}, x_{2i}), i = 1, \ldots, n$ are given.

   a) What are the interpretations of the parameters $(\beta_0, \beta_1, \beta_2)$?
   b) Suppose that you want to test

   $$H_0: \beta_0 = 0 \text{ vs } H_1: \beta_0 \neq 0$$

   where $\beta_1, \beta_2$ are treated as nuisance parameters. Is $H_0$ a linear hypothesis?
   c) State clearly the standard assumptions usually made in this context.
   d) Construct the standard analysis of variance table for this testing problem.
   e) What is the distribution of the test statistic when (i) $H_0$ holds and (ii) $H_0$ does not hold?
Define each of the following: i) Simple random sampling, ii) Stratified random sampling, iii) Systematic sampling, iv) Cluster sampling, and v) Two stage sampling.

b) Consider the ratio method of estimation, and let us denote by

$$y_i \quad \text{and} \quad x_i$$

the values of the primary and an ancillary characteristic under study for the $i^{th}$ unit of the population ($i = 1, 2, \ldots, N$)

$$R = \frac{\overline{y}}{\overline{x}}$$

the ratio of the population mean of $y$ to the population mean of $x$ and

$$\hat{R} = \frac{\overline{y}}{\overline{x}}$$

the corresponding ratio for a sample of size $n$.

Show that $E(\hat{R})$ is approximately equal to

$$R \left(1 + \frac{N-N}{Nn} \left( C_x^2 - \rho C_x C_y \right) \right),$$

where $C_x$, $C_y$ denote the coefficient of variation of $x$ and $y$ respectively, and $\rho$ denotes the correlation coefficient between $x$ and $y$.

c) Suppose that a population is divided into two mutually exclusive classes with $N_1$ and $N_2$ members respectively, so that

$$N_1 = Np, \quad N_2 = Nq \quad \text{and} \quad N_1 + N_2 = n.$$

Assume that a simple random sample of size $n$ is drawn without replacement from the $N$ units and $n_1$ of the observations in the sample are in class 1 and $n_2$ in class 2. Define

$$\hat{R} = \frac{n_1}{n_2} \quad \text{and} \quad R = \frac{N_1}{N_2}.$$

Show that for large values of $N$, the relative bias in $\hat{R}$ is given by $\frac{1}{nq}$.
5. 

a) For testing a simple null hypothesis vs a simple alternative, what is a most powerful test? Does it exist always?

b) What is a uniformly most powerful test?

c) Let $X_1, \ldots, X_n$ be $n$ independent random variables each having the normal distribution with mean $\theta$ and variance $\sigma$, where $\theta > 0$.

Let

$$H_0: \theta = 1 \quad H_1: \theta = 1.5.$$ 

(i) What is the most powerful test for this problem?

(ii) Verify whether the test derived in (i) is uniformly most powerful when

(I) $H_0: \theta = 1$ vs $H_1: \theta > 1$,

(II) $H_0: \theta = 1$ vs $H_1: \theta \neq 1$. 

QUESTION 1

a) State the components of variance model for the one-way layout consisting of \( k \) samples of \( n \) observations each. Use the notations \( \sigma_B^2 \) and \( \sigma_W^2 \) for the between-sample and within-sample components of variance, respectively.

b) Using the notations \( SSB \) and \( SSW \) for the between-sample and within-sample sums of squares, respectively, set up and fill out an analysis of variance table for this one-way layout, with rows "between" and "within" and columns "DF", "SS", "MS", "EMS", and "VR".

c) Let \( \theta = \sigma_B^2 / \sigma_W^2 \). Give formulas for the point estimator of \( \theta \) and for a two-sided \( 100(1 - \alpha)\% \) confidence interval on \( \theta \). Note that \( \theta \) cannot be negative. How do your formulas take this fact into account?

d) State how to test the hypothesis that \( \theta = \theta_0 \) against the one-sided alternative \( \theta > \theta_0 \). What is the power of the test against the specific alternative \( \theta = \theta_1 > \theta_0 \)?

e) Suppose \( k = 3, \ n = 6 \); the 3 sample means are \( \bar{x}_1 = 5, \ \bar{x}_2 = 6, \ \bar{x}_3 = 7 \); and \( SSW = 1800 \), \( \alpha = .10, \ \theta_0 = 1, \ \theta_1 = 2 \). Write down the analysis of variance table for these data; calculate the point of estimate and the confidence interval of (c); perform the test of (d) and evaluate its power.

NOTE: If \( X \) is distributed as \( F \) with \( 2 \) and \( m \) degrees of freedom, then

\[
P(X > c) = \left( \frac{m/2}{c + m/2} \right)^{m/2}.
\]
QUESTION 2

Suppose that $X$ is normally distributed with mean $\xi$ and variance $\sigma^2$.

a) Let $-\infty < L < U < \infty$. Derive the moment generating function of the conditional distribution of $X$ given $L < X < U$.

b) Let $Y = (X - \xi)/\sigma$ and $\phi(y)$ be the probability density function of $Y$. Show that

$$(d/dy)\phi(y) = -y\phi(y)$$

$$(d^2/dy^2)\phi(y) = (y^2 - 1)\phi(y), \quad \forall -\infty < y < \infty.$$  

c) Using (a) or (b), derive

$$E(X|L < X < U) \quad \text{and} \quad \operatorname{Var}(X|L < X < U)$$

in terms of $\phi$ and the corresponding cdf $\Phi$.

d) Verify that as $L \rightarrow -\infty$ and $U \rightarrow +\infty$, the expressions in (c) converge to $\xi$ and $\sigma^2$, respectively.

QUESTION 3

a) What is a likelihood function? Define a maximum likelihood estimator (MLE) of a parameter. State the usual properties of a MLE of a parameter $\theta$.

b) Suppose that $X$, the length of life of an electric tube, has the simple exponential density function

$$f_\theta(x) = \theta e^{-\theta x}, \quad 0 < x < \infty \quad (\theta > 0).$$

Based on a sample of size $n$, $n_t$ have been observed to have lifetime less than $t(> 0)$ while $n - n_t$ have lifetime $\geq t$.

(i) Find the MLE $\hat{\theta}_n$ of $\theta$.

(ii) Verify whether $\hat{\theta}_n$ is unbiased for $\theta$ or not.

(iii) What is the asymptotic variance of $\hat{\theta}_n$?
QUESTION 4

a) Define
   (i) Markov Process
   (ii) Markov Chain

b) For $n = 1, 2, \ldots$, let $X_n = U_1 + U_2 + \ldots + U_n$ where $\{U_n: n \geq 1\}$ is a sequence of independent random variables. We consider two cases
   (i) each $U_n$ normally distributed,
   (ii) each $U_n$ takes the values 0 and 1 with probability $p$ and $q$ respectively ($p + q = 1; 0 < p < 1$).

State in each case whether or not the stochastic process $\{X_n: n \geq 1\}$ is a Markov Process or Markov Chain. Explain your reasoning.

c) Let $\{X_1, X_2, \ldots\}$ be a sequence of independent and identically distributed random (non-negative) variables with p.d.f. $f(x)$ and let $R$ be an integer valued random variable with $\text{Probability}\{R = r\} = p_r$, the $X$'s and $R$ being independent. A finite renewal process is one in which events occur at $X_1, X_2 + X_2, \ldots, X_1 + \ldots + X_R$.

Calculate
   (i) the mean and variance of the distribution of the time to $X_1 + \ldots + X_R$
   (ii) the expected number of events in $(0, t)$.
QUESTION 5

a) Explain each of the following

(i) Simple random sampling with or without replacement;
(ii) Stratified random sampling;
(iii) Cluster sampling.

b) In a population of size $N$, let $p$ be the proportion of individuals with a certain characteristic. Let

$$Y_i = \begin{cases} 
1, & \text{if the } i\text{th individual has the characteristic;} \\
0, & \text{otherwise,}
\end{cases}$$

for $i = 1, \ldots, N$.

Suppose that a simple random sample of size $n$ is drawn without replacement. Denote by $\bar{y}_n$ the sample arithmetic mean of the $Y$-values. Show that $\text{E} \bar{y}_n = p$. Find $\text{Var}(\bar{y}_n)$.

c) Assume that in the population described in (b) the proportion of individuals possessing the characteristic varies by age and sex. Let $N_{hk}$ and $n_{hk}$ denote respectively the number of people in $h$-th age and $k$-th sex group and the size of a simple random sample (drawn without replacement) from the corresponding group ($h = 1, 2, \ldots, H$, $k = 1, 2$). Use the following scoring system. Denote

$$Y_{ihk} = \begin{cases} 
1, & \text{if } i\text{-th individual in the } h\text{-th age and } k\text{-th sex group possesses the characteristic} \\
0, & \text{otherwise}
\end{cases}$$

for $i = 1, 2, \ldots, n_{hk}$, $h = 1, 2, \ldots, H$, $k = 1, 2$.

Use the scores to obtain an unbiased estimate of $p$, the proportion of individuals possessing the characteristic in the population. Find the variance of your estimate.
Question 1

Conditional on θ fixed, X and Y are independent random variables each having a Poisson distribution with expected value θ. Also, suppose that θ follows a distribution with the probability density function

\[ g(\theta) = \frac{1}{\Gamma(\alpha)} e^{-\theta} \theta^{\alpha-1}, \quad 0 < \theta < \infty, \]

where α is a positive integer.

(a) Find the marginal distribution of X (and Y).

(b) Find the joint distribution of X and Y.

(c) Find the value of the (linear) correlation coefficient between X and Y.

(d) Find the regression function of Y on X, namely, \( E(Y|X=x) \).
Question 2

Let \((X_{11}, X_{12}, \ldots, X_{1n_1})\) and \((X_{21}, X_{22}, \ldots, X_{2n_2})\) be two independent random samples from Poisson distribution with parameters \(\mu_1\) and \(\mu_2\), respectively, \((\mu_1 > 0, \mu_2 > 0)\).

(a) Construct the likelihood and derive the maximum likelihood estimator for \(\mu_1\).

(b) Suppose that \(\mu_1 = \mu_2 = \mu\). Obtain the maximum likelihood estimator of \(\mu\), using data from the two independent samples.

(c) Construct the (extended) likelihood ratio criterion for testing \(H_0: \mu_1 = \mu_2 = \mu\) against the alternative \(H_1: \mu_1 \neq \mu_2\), using a suitable distributional approximation for the likelihood ratio criterion.

(d) Suppose that the following data are available:

\[
\begin{align*}
  n_1 &= 50, \quad \bar{x}_1 = 2.70; \quad n_2 = 60, \quad \bar{x}_2 = 3.05,
\end{align*}
\]

where

\[
\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad i = 1, 2.
\]

Apply the test you construct in (c) to these data, using significance level \(\alpha = 0.05\). What is your decision in regard to \(H_0\)?
Question 3

(a) Write down the estimator for a population proportion in the case of a stratified random sample.

(b) Derive a computational formula that would enable you to compute the variance of your estimator.

(c) Suppose that a simple random sample of size $n_h$ was selected from each stratum of size $N_h$, $h = 1, \ldots, H$, and $m_h$ respondents were obtained from each stratum. Then, for each stratum $h = 1, \ldots, H$, a simple random sub-sample of size $t_h$ was selected for intensive follow-up from the $(n_h - m_h)$ non-respondents. Suppose further that the sub-samples of size $t_h$ all responded, $h = 1, \ldots, H$. Consider the following estimator $\hat{p}_i$ for a population proportion, $p_i$, in category $i$,

$$
\hat{p}_i = \frac{1}{N} \sum_{h=1}^{H} \frac{m_{hi}}{n_h} \left[ \frac{m_{hi} + (n_h - m_h) t_{hi}}{n_h} \right],
$$

where $N = \sum_{h=1}^{H} n_h$, $m_{hi}$ is the number of respondents in category $i$ out of the $m_h$ initial respondents, and $t_{hi}$ is the number of respondents in category $i$ out of the sub-sample of size $t_h$, for $h = 1, \ldots, H$.

Show that the variance of $\hat{p}_i$ may be estimated using the following computational formula

$$
\text{var}(\hat{p}_i) = \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h (n_h - m_h)(m_h - m_{hi})m_{hi}}{n_h^2 (m_h - 1)} + \sum_{h=1}^{H} \frac{N_h^2 (n_h - m_h - t_h)(t_h - t_{hi})(m_h - m_{hi} t_{hi})}{n_h^2 t_h^2 (t_h - 1)}.
$$

(d) Suppose now that in the actual survey, some of the sub-sample of non-respondents did not respond even with intensive follow-up causing a few of the $t_h$ to be zero. How might you modify the computational formula in part (c) to deal with this situation?
Question 4

Consider a one-way analysis of variance (ANOVA) model with fixed effects. There are \( k \geq 2 \) treatments, indexed \( i = 1, \ldots, k \) and \( n \geq 2 \) replicates for each treatment, indexed \( j = 1, \ldots, n \). The response for the \( j \)th replicate of the \( i \)th treatment is denoted by \( Y_{ij}, 1 \leq j \leq n, 1 \leq i \leq k \).

(a) Write down the usual model for this situation and any assumptions about the model and its parameters.

(b) In terms of the model parameters, write down the null and alternative hypotheses for testing whether there are any treatment effects.

(c) Provide a complete breakdown of the usual ANOVA table, including degrees of freedom, sum of squares, mean squares.

(d) State the test statistic used for the test in part (b) and its distribution under the null hypothesis of no treatment effects.

(e) Derive the expected values of the mean square due to error and the mean square due to treatment when the null hypothesis may not be true.

(f) Write the formula for the noncentrality parameter of the distribution of the test statistic in part (d). Of what use is this parameter?
Question 5

(a) Give definitions of the following terms (define any notation you use):

(i) persistent positive (or positive recurrent) state
(ii) persistent null (or null recurrent) state
(iii) transient state
(iv) stationary distribution

(b) Suppose you are given a finite Markov chain with transition matrix

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & .1 & 0 & 0 & .9 \\
.3 & .1 & .1 & .3 & .2 \\
.2 & .2 & .2 & .2 & .2 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

(i) Find the conditional probability of eventual absorption into state 0, given \(X_0 = i\), for \(i = 0, 1, 2, 3, 4\).

(ii) Find the expected waiting time for absorption into state 0 or state 4, given \(X_0 = i\), for \(i = 0, 1, 2, 3, 4\).
INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is four hours.

c) Answer any four (but only four) of the five questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Put your code letter, not your name, on each page.

f) Return the examination with a signed statement of honor pledge on a page separate from your answers.

g) You are required to answer only what is asked in the questions and not all you know about the topics.
Question 1

a) Clarify the confusion of a person who makes the following statement:

"Stratified sampling and cluster sampling are conceptually identical. Both are types of designs in which a probability sample is selected from a population which is divided into groups of elements."

b) Briefly describe stratified (simple) random sampling.

c) Let

\[ Y_{hi} = \begin{cases} 1 & \text{if the i-th element in the h-th stratum possesses some attribute} \\ 0 & \text{if otherwise} \end{cases} \]

and suppose we wish to estimate the proportion (P) of elements in a population of size \( N \) which possess the attribute. The sampling design we use is proportionate stratified (simple random) sampling and our estimator of \( P \) takes the general form

\[ P = \frac{\sum_{h} W_{h} p_{h}}{\sum_{h} W_{h}} \]

where \( W_{h} = N_{h}/N \), \( N_{h} \) is the total number of elements in the h-th stratum, and \( p_{h} \) is the proportion of \( n_{h} \) sample elements possessing the attribute in the h-th stratum. Show that the true variance of \( P \) is

\[ \text{Var}(p) = \frac{\sum_{h} W_{h} p_{h} q_{h}}{n} - \frac{\sum_{h} W_{h} p_{h} q_{h}}{N} \]

where \( q_{h} = 1 - p_{h} \) is the proportion of all elements in the h-th stratum which possess the attribute.

d) Assuming that \( N_{h} = N_{h} - 1 \) and \( N = N - 1 \), show that the true variance of the estimator \( (p_{o}) \) of \( P \) from a simple random sample of size \( n \) taken from the same population can be expressed as

\[ \text{Var}(p_{o}) = \text{Var}(p) + \frac{1}{n} \sum_{h} W_{h} (p_{h} - P)^2 \]

where \( f = n/N \).
Question 2

Suppose that a system has two components whose life times (X and Y, say) are independent and each has the same exponential distribution with mean \( \theta > 0 \). The system fails as soon as at least one of its components does so. Let \( Z \) be the life-time of the system.

(a) What is the probability density function of \( Z \)?

(b) For \( n \geq 1 \) systems of the same type, let \( Z_1, \ldots, Z_n \) be the respective life times. Obtain the maximum likelihood estimator of \( \theta \) (say, \( \hat{\theta}_n \)) based on \( Z_1, \ldots, Z_n \).

(c) Obtain \( E\hat{\theta}_n \) and \( \text{Var}(\hat{\theta}_n) \).

(d) Compare \( \text{Var}(\hat{\theta}_n) \) with the Cramér-Rao bound and comment on the efficiency and sufficiency of \( \hat{\theta}_n \).

Question 3

a) Suppose that the probability of an insect laying \( r \geq 0 \) eggs is \( e^{-m r/r!} \) and that the probability of an egg developing is \( p \) \( (0 < p < 1) \). Assuming mutual independence of the eggs, show that the probability of a total \( k \geq 0 \) survivors is \( e^{-mp}(mp)^k/k! \).

b) Suppose that two insects of the same species are placed in two different incubators with different temperatures, and you suspect that the temperature may have some effect on the probability of an egg developing, though the distribution of the number of eggs layed should be the same. Let \( p_1 \) and \( p_2 \) be these probabilities for the two insects and let \( k_1 \) and \( k_2 \) be the actual number of survivors in the two incubators.

(i) Find the joint distribution of \( (k_1, k_2) \).

(ii) Find the conditional distribution of \( k_1 \), given \( k=k_1+k_2 \).

(iii) We wish to test the hypothesis \( H_0: p_1 = p_2 \) against \( H_1: p_1 > p_2 \). It would be easier to construct the test if we use the distribution defined in (ii). How in this case \( H_0 \) is expressed? Construct a test of size \( \alpha \) \( (0 < \alpha < 1) \), given \( k_1, k_2 \).
Question 4

Suppose that a random variable \( Y \) is related to a non-stochastic factor \( x \) over the interval \(-1 \leq x \leq 1\) by the relationship

\[
\eta_x = E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2.
\]

Let \( N \) pairs of observations \((x_i, Y_i), i = 1, \ldots, N\), be available, for which \( \text{Var}(Y_i) = \sigma^2(< \infty) \), \( \text{Cov}(Y_i, Y_j) = 0 \), \( i \neq j = 1, \ldots, N \), and let

\[
\mu_k = N^{-1} \sum_{i=1}^{N} x_i^k, \quad k = 1, 2, 3.
\]

Assume further that \( \mu_1 = \mu_2 = 0 \). It is desired to approximate the function \( \eta_x \) over the interval \(-1 \leq x \leq 1\) by a linear function

\[
\hat{\eta}_x = b_0 + b_1 x, \quad -1 \leq x \leq 1,
\]

where

\[
b_0 = N^{-1} \sum_{i=1}^{N} Y_i \quad \text{and} \quad b_1 = \frac{\sum_{i=1}^{N} x_i Y_i}{\sum_{i=1}^{N} x_i^2}.
\]

(a) Express \( E(b_0) \) and \( E(b_1) \) in terms of \( \beta \)'s and \( \mu \)'s.

(b) Express \( V(b_0) \) and \( V(b_1) \) in terms of \( \sigma^2, N, \beta \)'s and \( \mu \)'s.

(c) Show that

\[
\text{Cov}(b_0, b_1) = 0 \quad \text{(since} \quad \mu_1 = 0)\]

(d) Show in general that the integrated mean square error

\[
J = \int_{-1}^{1} E[(\hat{\eta}_x - \eta_x)^2]dx
\]

can be written as the sum of the integrated variance \( V \) and the integrated squared bias \( B \),

where \( V = \int_{-1}^{1} \text{Var}(\hat{\eta}_x)dx \) and \( B = \int_{-1}^{1} [E(\hat{\eta}_x) - \eta_x]^2dx \).

(e) Show that, for the particular situation described earlier, \( V \) is minimized when \( \mu_2 = 1 \).
Question 5

Let $X_1, \ldots, X_n$ be $n$ independent and identically distributed random variables with the probability density function (pdf) $f(x)$ and distribution function $F(x)$. Let $Z_1 \leq \ldots \leq Z_n$ be the order statistics corresponding to $X_1, \ldots, X_n$.

(a) For an arbitrary $k (1 \leq k \leq n - 1)$, write down the joint pdf of $Z_1, \ldots, Z_{k+1}$.

(b) Hence or otherwise, obtain the conditional pdf of $Z_{k+1}$ given $Z_1, \ldots, Z_k$.

(c) Show that (b) depends on $Z_1, \ldots, Z_k$ only through $Z_k$.

(d) Define a Markov process.

(e) Define a Markov chain.

(f) Is the process $(Z_1, \ldots, Z_n)$ a Markov process? Is it a Markov chain? Does it have independent increments? Does it have homogeneous (stationary) increments.
Q.1. Estimates of differences are a common part of analysis from survey data. Suppose we wish to do an initial sample survey in a population with N elements at a point in time, T₁, and then a second survey in the same population of N elements at a later point in time, T₂. Population means for T₁ and T₂ are \( \bar{Y}_1 \) and \( \bar{Y}_2 \), respectively. Our objective is to estimate the difference, 

\[ \Delta = \bar{Y}_2 - \bar{Y}_1. \]

Suppose furthermore that the samples for the two surveys are both of size n are selected by the following process:

1) An SRS sample of size n is selected at T₁.

ii) The SRS sample at T₂ consists of m "old" elements (0 ≤ m ≤ 1) from the sample at T₁ plus (1 - m)n "new" elements that are independently selected by SRS from the T₂ population elements which did not fall in the T₁ sample.

We assume that n is small relative to N so that finite population corrections can be ignored.

(a) Show that the variance of our estimator of Δ,

\[ d = \bar{y}_2 - \bar{y}_1, \]

is

\[ \text{Var}(d) = \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_2) - 2m \text{ Cov}(\bar{y}_1, \bar{y}_2) \]

where \( y_{1j} \) is the j-th sample observation at T₁ and

\[ \bar{y}_1 = \frac{1}{n} \sum_{j=1}^{n} y_{1j}; \quad \bar{y}_2 = \frac{1}{n} \sum_{j=1}^{n} y_{2j}. \]
(b) If element variances at $T_1$ and $T_2$ are equal, i.e.,

$$s_1^2 = s_2^2 = s^2,$$

then show that

$$\text{Var}(d) = 2s^2(1-\pi\rho_{12})$$

where

$$\rho_{12} = \sqrt{\frac{\sum_{i=1}^{N} (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{\sum_{i=1}^{N} (Y_{1i} - \bar{Y}_1)^2 \sum_{j=1}^{N} (Y_{2j} - \bar{Y}_2)^2}}.$$

(c) Discuss the relative statistical precision of $d$ in the following situations:

i) Situation 1: $-1 < \rho_{12} < 0$ and $0 < \pi < 1$

ii) Situation 2: $\rho_{12} = 0$ and $\pi = 0$

iii) Situation 3: $0 < \rho_{12} < 1$ and $0 < \pi < 1$

iv) Situation 4: $0 < \rho_{12} < 1$ and $\pi = 1$

Q.2. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables from the Poisson distribution with parameter $\lambda$. Let $Y_n = X_1 + \ldots + X_n$.

(a) Show that the probability distribution of $Y_n$ is given by

$$P(Y_n = k) = (n\lambda)^k e^{-n\lambda} / k!$$

(b) Calculate the moment generating function of $Y_n$: $M_n(t) = E(e^{tY_n})$, for real $t$.

(c) Show that $Y_n$ is asymptotically normally distributed with mean $n\lambda$ and variance $n\lambda$.

(d) Show that for $m$, not necessarily an integer,

$$\lim_{n \to \infty} \sum_{k=0}^{[m]} \frac{e^{-m} m^k}{k!} = \frac{1}{2}$$

where $[m]$ stands for the largest integer contained in $m$.

[Hint: You may use (c) to prove (d)].
Q.3. (a) Define sufficient statistics, minimal sufficient statistics and completeness in the context of estimation theory.

(b) Let \( X_1, \ldots, X_n \) be \( n \) independent and identically distributed random variables with the probability density function

\[
 f(x, \theta) = \begin{cases} 
 (\theta_2 - \theta_1)^{-1} & \text{if } \theta_1 \leq x \leq \theta_2, \\
 0, & \text{otherwise}, \\
\end{cases} 
\]

where \( \theta = (\theta_1, \theta_2) \) is an unknown parameter (vector). Also, let \( X_{(1)} \leq \ldots \leq X_{(n)} \) be the ordered variables corresponding to \( X_1, \ldots, X_n \).

(i) Show that \( X_{(1)} \) and \( X_{(n)} \) are jointly sufficient for \( \theta_1 \) and \( \theta_2 \).

(ii) Assuming completeness and using (i), obtain a minimum variance unbiased (MVU) estimator of \( \delta = \theta_2 - \theta_1 \) and derive the formula for its mean square error.

(iii) Suppose that it is given that \( \theta_1 = 0 \). Then, what is the MVU estimator of \( \delta = \theta_2 \)?

Q.4. Let \( Y_1, \ldots, Y_n \) be \( n \) independent random variables with a common probability density function (p.d.f.)

\[
 f_Y(y) = \theta \exp(-\theta y), \quad 0 \leq y < \infty, \quad \theta > 0,
\]

where \( \theta \) is an unknown parameter. Further, let \( Z_1, \ldots, Z_n \) be an independent set of \( n \) independent random variables with a common p.d.f.

\[
 g_Z(z) = c \exp(-cz), \quad 0 \leq z < \infty, \quad c > 0
\]

where \( c \) is a specified constant. Let then

\[
 X_i = \min(Y_i, Z_i), \quad i = 1, \ldots, n.
\]

(a) Write down the likelihood function of \( X_1, \ldots, X_n \).

(b) Based on \( X_1, \ldots, X_n \), obtain the maximum likelihood estimator \( \hat{\theta}_n \) of \( \theta \).

(c) Obtain the expressions for the bias and variance of \( \hat{\theta}_n \). What are the minimum values of \( n \) for which these are finite?

(d) Show that \( n^{-1} (\hat{\theta}_n - \theta) \) has a bias converging to 0 as \( n \to \infty \).

(e) Obtain the Cramér-Rao bound for the variance of an estimator and hence, comment on the asymptotic efficiency of \( \hat{\theta}_n \).
Q.5. In a mathematical modeling approach to representation of biological phenomena, we often start with a very simple model so as to achieve mathematical tractability and then build increasingly sophisticated models to reflect more realism and to obtain a better fit to the empirical process. In particular, if we consider births in a population, we may start with the simple Poisson process and then, by generalizing the parameters of the resulting model, attempt to model more complicated probabilistic phenomena. For the stochastic process $X$, we use the distribution

$$p_n(t) = P[X(t) = n], \quad n=0,1,...$$

to represent the number of births in a population in the interval $(0, t)$.

(i) State three assumptions about the above probability distribution which are necessary and sufficient to insure that the distribution is Poisson. State these assumptions very carefully being sure to specify appropriate parameters where needed.

(ii) State the Chapman-Kolmogorov equations for continuous-time Markov chains (discrete state space) and, using the above assumptions, show that the aforementioned distribution satisfies these assumptions and that they may be used to imply the following set of differential equations (or the differential-difference equation) where parameter $\lambda$ is associated with your aforementioned statement of the Poisson property:

$$p'_n(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \quad n \geq 1$$

$$p'_0(t) = -\lambda p_0(t)$$

for all $t \geq 0$, where $\lambda > 0$.

(iii) Define the probability generating function for the stochastic process $\{X(t): t \geq 0\}$. Use this generating function and the above equation to prove that the solution to the system is the Poisson distribution.

(iv) Suppose that we generalize this model by allowing the birth rate to become a function of $t$. Recompute the above generating function and derive the new expression for $p_n(t)$. What is the expected value of this new distribution?
Q.1. \(X_1, \ldots, X_n\) are independent random variables, each having the uniform distribution with probability density function

\[
f_X(x) = \begin{cases} (\theta_2 - \theta_1)^{-1}, & \theta_1 \leq x \leq \theta_2, \\ 0, & \text{otherwise,} \end{cases}
\]

where \(\theta_1, \theta_2\) are unknown parameters.

(i) Find a minimal sufficient statistic for \((\theta_1, \theta_2)\) [In this context, define clearly sufficient statistics and minimal sufficient statistics.]

(ii) Given that \(\theta_2 > 0\) and that \(\theta_1 = k\theta_2, k\) known, find a minimal sufficient statistic for \(\theta_2\) when

(a) \(0 < k < 1\);

(b) \(k = 0\);

(c) \(k < 0\).
Q.2. (a) Let $X_1, \ldots, X_n$ be $n$ independent random variables with a continuous distribution function $F$, and let $X_{n:1}, \ldots, X_{n:n}$ be the order statistics corresponding to $X_1, \ldots, X_n$. If $\xi_p$ is the population $p$-quantile of $F$, so that $F(\xi_p) = p$ for $0 < p < 1$, then show that

$$P\{X_{n:r} < \xi_p < X_{n:s}\} = \sum_{k=r}^{s-1} \binom{n}{k} p^k (1-p)^{n-k}$$

for every $(r,s): 1 < r < s < n$. Also, show that if

$$T_n^+ = \sum_{i=1}^{n} R_i \psi_i,$$

where $R_i$ is the rank of $|X_i|$ among $|X_1|, \ldots, |X_n|$, for $i=1, \ldots, n$, and $\psi_i$ is equal to 1 or 0 according as $X_i$ is $> 0$ or $< 0$, then

$$P\{1 \leq T_n^+ \leq n(n-1)/2 \mid H_0\} = 1 - 2^{-n+1},$$

where $H_0$ refers to the null hypothesis that $F$ is symmetric about 0.

(b) Suppose two independent observations on the random variable $X$ yield the values 3 and 26. Consider using $[3, 26]$ as a confidence interval for the population median ($\xi_{0.5}$). What confidence coefficient can be assigned to this interval, assuming $X$ is normally distributed? Without the normality assumption (based on the sign test, and also based on the Wilcoxon signed-rank test)?

(c) A third independent observation on the same random variable yields the value 4. With what confidence coefficient can it now be asserted that the population median is in the interval $[3, 26]$, based on the assumption that $X$ is normally distributed? Without the normality assumption (again based on the sign test, and also on the Wilcoxon signed-rank test)?

EDITORIAL NOTE: A table of the cumulative distribution function of Student's $t$ for 1(1)5 degrees of freedom was provided.
Q.3. A sample consists of $n$ independent observations $X_1, X_2, \ldots, X_n$.

It is known that $X_1$ has been chosen randomly from a parent population with probability density function

$$f_{X_1}(x; \alpha_1, \beta) = \frac{\alpha_1^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha_1 x}, \quad 0 < x < \infty; \quad \alpha_1 > 0, \quad \beta > 0.$$  

(Note that $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx(t > 0)$)

(a) Find $E(X_1^r)$, $r > 0$.

(b) Use the result in part (a) to find explicit expressions for the mean and variance of the geometric mean $\bar{X} = (\prod_{i=1}^n X_i)^{1/n}$.

(c) Assuming that $\beta$ is known, show that the likelihood ratio test criterion for testing the null hypothesis that $\alpha_1$ has a common, but unspecified, value (say, $\alpha$) in all $n$ distributions involves comparing $\bar{X}$ with the arithmetic mean $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Q.4. Suppose that in a model for onset of a disease we distinguish two states: $S_1$ - a vulnerability state and $S_2$ - onset of a disease after reaching state $S_1$.

Suppose that a healthy individual (at state $S_0$) was exposed to an agent which makes him liable to enter state $S_1$. Let $X_1 > 0$ denote the time to entry from $S_0$ to $S_1$, and $f_1(x_1)$ for $0 < x_1 < a$ and $f_1(x_1) = 0$ otherwise be the probability density function (PDF) of $X_1$. Let $X_2 > 0$ be the time from incidence of vulnerability to onset of a disease, and $f_2|_1(x_2|x_1)$ be the conditional PDF of $X_2$ given $X_1$.

Let $Z = X_1 + X_2$ denote the total time to onset.

(a) Obtain the expression for the probability $Pr(X_1 + X_2 \leq z)$.

Consider a special case, where $X_1$ and $X_2$ are independent with

$f_1(x_1) = \frac{1}{T} \quad 0 < x_1 < T$ (uniform) and $f_2(x_2) = \lambda e^{-\lambda x_2}, \lambda > 0, \quad x_2 > 0$ (exponential).
(b) Obtain $\Pr[X_1 + X_2 < z]$.  
(c) Find the PDF $f_z(z)$.  
(d) Hence, or otherwise, find the mean and variance of $Z$.  

Q.5. Suppose we have a two-stage sampling design of equal-sized clusters in which we select SRS (a of A) clusters in the first stage and SRS (b of B) elements within each selected cluster. If the object is to estimate the population mean per element ($\bar{Y}_0$), the true variance of the estimator ($\bar{y}_{2cs}$) is known to be  

$$ \text{Var}(\bar{y}_{2cs}) = (1-f_a) \frac{s^2_a}{a} + (1-f_b) \frac{s^2_b}{ab} $$  

where $f_a = a/A$, $f_b = b/B$, $S^2 = S^2_a + S^2_b$, and $S^2$ is the element variance in the population.  

(a) Ignoring all finite population corrections and allowing that, for all intents and purposes, $S^2 = S^2_a + S^2_b$ and the intraclass correlation coefficient can be expressed as $\rho = 1 - S^2_b/S^2$, show that  

$$ \text{Var}(\bar{y}_{2cs}) = \frac{s^2}{ab} [1 + \rho(b-1)] \quad (1) $$  

(b) Explain what is meant by "the sampling distribution of $\bar{y}_{2cs}$" and how does it relate to the result of Equation (1)?  
(c) Define a "design effect." What part of Equation (1) represents the design effect for $\text{Var}(\bar{y}_{2cs})$?  
(d) Derive the formula you would use to determine the sample size, $n = ab$, required to yield some prespecified coefficient of variation of $\bar{y}_{2cs}$, given that reasonably good prior guesses about the sizes of $S^2$, $Y_0$, and the design effect are available.
INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is four hours.

c) Answer any four (but only four) of the five questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Put your code letter, not your name, on each page.

f) Return the examination with a signed statement of honor pledge on a page separate from your answers.

g) You are required to answer only what is asked in the questions and not all you know about the topics.
Q.1 Suppose you are designing a two-stage cluster sample of households for a health survey being planned in a large city. You propose to obtain personal interviews from 1000 households selected from a sample of 50 primary sampling units, which for this survey would be area segments consisting of a small number of neighboring blocks. The principal goal of the survey is to estimate the proportion (P) of households in the city with health insurance. Your best guess is that P is likely to be about 0.20 in the population and that the design effect (Deff) for your proposed sample will be about 1.50.

a) Describe in words how you would select the area segments and households within segments to produce an epsem sample of 1000 households.

b) An appropriate estimator for P in the resulting sample would be the proportion (p) of sample households with health insurance. Ignoring the finite population correction, how large would you expect the coefficient of variation to be for p as obtained in the proposed survey?

Hint: Recall that the true variance for an estimated mean from a simple random sample of size n is

\[ \text{Var}(\bar{y}_0) = \frac{1-f}{n} S^2 \]

where f is the sampling rate and S^2 is the variance among elements in the population.

c) Assuming that the anticipated design effect of 1.50 is correct, what would be your estimate of intraclass correlation for the estimate of P in your design?
Q.2

a) Find a probability distribution \( \{ r_n : n = 0, 1, \ldots \} \) which satisfies the difference equation:

\[
  r_{n+1} + p r_n = r_n
\]

for \( 0 < p < 1 \). What is this distribution?

b) Show that the pgf of \( \{ r_n : n = 0, 1, \ldots \} \), say \( R(s) \), satisfies the equation:

\[
p R'(s) = q [R(s)]^2
\]

where \( q = 1 - p \).

c) Use part (b) to derive the mean and variance of \( \{ r_n : n = 0, 1, \ldots \} \).

d) What is the probability of occurrence at time \( n \) given no previous occurrence?

e) Give an empirical example for which one might use this probability distribution and discuss the implications of the assumptions inherent in its definition; e.g., look at the ratio \( r_{n+1}/r_n \). What is the mode of the distribution; i.e., at what point is the distribution at its maximum?

Q.3 Suppose that the number of new cases \( X_i \) of lung cancer developing in region \( i \) (\( i = 1, 2 \)) over a certain specified period of time is to modeled by the Poisson distribution with parameter \( M_i \lambda_i \), where \( M_i \) is the total number of person-years of residence for all individuals living in the \( i \)-th region of various times during the specified time period, and where \( \lambda_i \) is the rate of lung cancer development per person-year of residence during that time period. Theoretically, \( M_i \) represents the sum over all individuals residing in region \( i \) (during the specified time period) of the number of years of residence for each individual.
a) Treating $M_1$ and $M_2$ as fixed constants, show that the likelihood ratio test statistic for testing $H_0: \lambda_1 = \lambda_2$ versus $H_1: \lambda_1 \neq \lambda_2$ can be written in the form

$$\frac{M_1 + M_2}{[\frac{M_1}{M_1 + M_2} Y_1]} X_1 - \frac{M_1 + M_2}{[\frac{M_1}{M_1 + M_2} (1 - Y_1)]} X_2$$

where $Y_1 = X_1/(X_1 + X_2)$. You can assume that $X_1$ and $X_2$ are independent random variables.

b) Under $H_0: \lambda_1 = \lambda_2$, and conditional on $(X_1 + X_2) = k$, a fixed positive integer, show that the rejection region for the likelihood ratio test statistic developed in part (a) can be based on the binomial distribution with "sample size" $k$ and parameter $p = M_1/(M_1 + M_2)$ [HINT: Find the distribution of $X_1$ given $(X_1 + X_2) = k$ under $H_0: \lambda_1 = \lambda_2 (= \lambda, \text{say})$.].

c) Suppose that $M_i$ has a gamma distribution with parameters $\alpha$ and $\beta$. Assuming that the conditional distribution of $X_i$ given $M_i = m_i$ is Poisson with parameter $m_i \lambda$, find the unconditional distribution of $X_i$. When $\beta$ is a positive integer, show that this distribution is negative binomial.

Q.4 Let $(X_1, Y_1, \ldots, X_n, Y_n)$ be $n$ independent (bivariate) random vectors with a continuous distribution $F(x, y)$, and let $F_1(x) = F(x, \infty)$ and $F_2(y) = F(\infty, y)$, $(x, y) \in \mathbb{R}^2$. Let $R_i (Q_i)$ be the rank of $X_i$ ($Y_i$) among $X_1, \ldots, X_n$ ($Y_1, \ldots, Y_n$), for $i = 1, \ldots, n$. Consider a statistic of the form

$$L_n = \sum_{i=1}^{n} a_n(r_i) b_n(Q_i)$$

where $a_n(1) \leq \ldots \leq a_n(n)$ and $b_n(1) \leq \ldots \leq b_n(n)$ are given scores. Consider the null hypothesis $H_0$ that $F(x, y) = F_1(x)F_2(y)$ i.e., $X$ and $Y$ are stochastically independent. Further let
\[ L^*_n = \sum_{\alpha=1}^{n} a_n(\alpha)b_n(S_{\alpha}), \]

where \( (S_1, \ldots, S_n) \) takes on each permutation of \((1, \ldots, n)\) with the common probability \((n!)^{-1}\).

a) Show that under \( H_0 \), \( L_n \) and \( L^*_n \) both have the same distribution (which does not depend on \( F_1 \) and \( F_2 \)).

b) Using (a) (or otherwise), prove that

\[
\begin{align*}
E[L_n | H_0] &= \frac{na}{n} \\
V[L_n | H_0] &= (n-1)A_n^2B_n^2
\end{align*}
\]

where

\[
\hat{\alpha}_n = n^{-1}L_i=1 a_n(k), \quad \hat{A}_n^2 = \frac{1}{n-1} \sum_i=1^n [a_n(i) - \hat{\alpha}_n]^2
\]

and \( \hat{B}_n^2 \) are defined similarly.

c) Show that the distribution of \( L_n \) (under \( H_0 \)) is symmetric if either \( a_n(i) + a_n(n-i+1) = \) constant, for all \( i (\leq n) \) or \( b_n(i) + b_n(n-i+1) = \) constant, for all \( i (\leq n) \).

d) Consider the Spearman rank correlation and comment on (b) and (c) for this statistic.

Q.5

a) State the components of variance model for the one-way layout consisting of \( k \) samples of \( n \) observations each. Use the notations \( \sigma_B^2 \) and \( \sigma_W^2 \) for the between-sample and within-sample components of variance, respectively.

b) Using the notations \( SSB \) and \( SSW \) for the between-sample and within-sample sums of squares, respectively, set up and fill out an analysis of variance table for this one-way layout, with rows "between" and "within" and columns "DF", "SS", "MS", "EMS", and "VR".
c) Let \( \theta = \sigma_B^2 / \sigma_W^2 \). Give formulas for the point estimator of \( \theta \) and for a two-sided 100(1-\( \alpha \))% confidence interval on \( \theta \). Note that \( \theta \) cannot be negative. How do your formulas take this fact into account?

d) State how to test the hypothesis that \( \theta = \theta_0 \) against the one-sided alternative \( \theta > \theta_0 \). What is the power of the test against the specific alternative \( \theta = \theta_1 > \theta_0 \)?

e) Suppose \( k = 3 \), \( n = 6 \); the 3 sample means are \( x_1 = 5 \), \( x_2 = 6 \), \( x_3 = 7 \); and \( SSW = 1800 \), \( \alpha = .10 \), \( \theta_0 = 1 \), \( \theta_1 = 2 \). Write down the analysis of variance table for these data; calculate the point of estimate and the confidence interval of (c); perform the test of (d) and evaluate its power.

NOTE: If \( X \) is distributed as \( F \) with 2 and \( m \) degrees of freedom, then

\[
P(X > c) = \left( \frac{m/2}{c + m/2} \right)^{m/2}.
\]
INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is four hours.

c) Answer any four (but only four) of the five questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Put your code letter, not your name, on each page.

f) Return the examination with a signed statement of honor pledge on a page separate from your answers.

g) You are required to answer only what is asked in the questions and not all you know about the topics.
Q1. Suppose that \( n \) integers are drawn at random and with replacement from the integers 1, 2, \ldots, \( N \). That is, each sampled integer has probability \( \frac{1}{N} \) of taking on any of the values 1, 2, \ldots, \( N \), and the sampled values are independent observations. Further, assume that \( N \) is an unknown parameter to be estimated.

a) If \( Y_1, Y_2, \ldots, Y_n \) represent the random sample of \( n \) observations, show that the estimator of \( N \) obtained by the method of moments is

\[
\hat{N}_1 = (2\bar{Y} - 1)
\]

where \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \).

b) Show that

\[
\text{var}(\hat{N}_1) = \frac{1}{3n} (N^2 - 1)
\]

c) Show that the maximum likelihood estimator of \( N \) is

\[
\hat{N}_2 = \max\{Y_1, Y_2, \ldots, Y_n\}
\]

d) Find an exact expression for \( \Pr(\hat{N}_2 = k) \).

e) Assuming that

\[
\text{var}(\hat{N}_2) \approx \frac{nN^2}{(n+1)^2(n+2)} \quad \text{for large } N,
\]

and that \( (\frac{n+1}{n})\hat{N}_2 \) is approximately an unbiased estimator of \( N \), show that

\[
\text{var}[ (\frac{n+1}{n})\hat{N}_2 ] < \text{var}(\hat{N}_1)
\]

for large \( N \) when \( n > 1 \).
Q2. Suppose that a random sample of length-of-life measurements \( Y_1, Y_2, \ldots, Y_n \) is to be taken on components whose length of life has an exponential distribution with mean \( \theta \). It is of interest to estimate unbiasedly the parameter

\[
\overline{F}_T(t) = 1 - F_T(t) = \Pr(T > t) = e^{-t/\theta},
\]

the reliability at time \( t \) of a component of this type.

a) Find the maximum likelihood estimator of \( e^{-t/\theta} \). Note that this estimator is not an unbiased estimator of \( e^{-t/\theta} \).

b) Use the Rao-Blackwell Theorem to show that

\[
\overline{F}_T^*(t) = (1 - \frac{t}{U})^{n-1},
\]

where \( U = \sum_{i=1}^{n} Y_i \) is an unbiased estimator of \( \overline{F}_T(t) \).

[HINT: Let \( V = \begin{cases} 1, & Y_1 > t \\ 0, & \text{otherwise} \end{cases} \); then, find \( \mathbb{E}(V|U=u) \).]

c) Show that \( \overline{F}_T^*(t) \) found in part (b) is the minimum variance unbiased estimator of \( e^{-t/\theta} \).
Q3. Let \( X_1, \ldots, X_n \) be \( n \) independent random variables with continuous distributions \( F_1, \ldots, F_n \), all defined on \( (-\infty, \infty) \).

Let \( R_i \) be the rank of \( X_i \) among \( X_1, \ldots, X_n \), for \( i = 1, \ldots, n \) and for each \( n \), let \( a_n(1), \ldots, a_n(n) \) be a set of real numbers (scores). Consider the statistic

\[
L_n = \sum_{i=1}^{n}(c_i - \overline{c}_n)a_n(R_i),
\]

where \( c_1, \ldots, c_n \) are known constants, not all equal, and \( \overline{c}_n = n^{-1}(c_1 + \ldots + c_n) \).

a) Show that under \( H_0: F_1 = \ldots = F_n = F \) (unknown),

\[
E(L_n) = 0 \quad \text{and} \quad V(L_n) = c_n^2 A_n^2
\]

where

\[
c_n^2 = \sum_{i=1}^{n}(c_i - \overline{c}_n)^2 \quad \text{and} \quad A_n^2 = \frac{1}{n-1} \sum_{i=1}^{n}(a_n(i) - \overline{a}_n)^2,
\]

\[
\overline{a}_n = n^{-1}(a_n(1) + \ldots + a_n(n)).
\]

b) Show that the Wilcoxon-Mann-Whitney two-sample rank sum statistic is a particular case of \( L_n \).

c) Show that the Brown-Mood two-sample median test statistic is a particular case of \( L_n \).

d) For the case (c) derive the exact probability law of \( L_n \), under \( H_0 \), when \( n = n_1 + n_2 \) and \( n_1, n_2 \) are the individual sample sizes.

e) What can you say about the distribution of \( L_n \) (under \( H_0 \)) when \( n \) is small and what large sample approximation is available?
Q4. Suppose that stratified simple random sampling is the design of choice for a survey you're helping to design and that the survey's unit costs will be the same among all strata (i.e., $J_h = J^*$, $h = 1, 2, \ldots, H$).

a) Assuming the conditions stated above, show that the optimum stratum sample size will be

$$n_{h}^{\text{opt}} = n \left( \frac{W_{h} S_{h}}{\sum_{h} W_{h} S_{h}} \right)$$

where $n$ is the total sample size, $W_{h}$ is the proportion of the population in the $h$-th stratum, $S_{h}$ is the stratum element standard deviation, and "optimum" here refers to minimizing the true variance of an estimated population mean given prespecified total survey cost following the model,

$$\text{Total Cost} = J - J_0 + nJ^*,$$

where $J_0, J^*$ are fixed costs not dependent on the sample size.

b) Assuming the element coefficient of variation for each stratum is the same, show that $n_{h}^{\text{opt}} = K Y_{h}$, where $Y_{h}$ is the stratum total and $K$ is a constant of proportionality.

c) What is the name given to the design which results when stratum unit costs and element variances are equal among strata (i.e., when $J_h = J^*$ and $S^2_h = S^*{^2}$, $h = 1, 2, \ldots, H$). Is this design epsem? Explain.
Q5. Suppose that \( \{X_1(t), t\geq 0\} \) and \( \{N_2(t), t\geq 0\} \) are independent Poisson processes with rates \( \lambda_1 \) and \( \lambda_2 \):

\[
P(X_1(t) = n) = \frac{(-\lambda_1 t)^n}{n!} e^{\lambda_1 t}
\]

\[
P(N_2(t) = n) = \frac{(-\lambda_2 t)^n}{n!} e^{\lambda_2 t}
\]

a) Show that \( \{N_1(t) + N_2(t), t\geq 0\} \) is Poisson with rate \( \lambda_1 + \lambda_2 \).

b) Show that the probability that the first event of the combined process comes from \( \{X_1(t), t\geq 0\} \) is \( \lambda_1 / (\lambda_1 + \lambda_2) \) independently of the time of the event.

c) Fix \( 0 = t_0 < t_1 < \ldots < t_n = 1 \). Find the conditional distribution of the number of arrivals for \( X_1(t) + N_2(t) \) in the disjoint time intervals \( (t_0, t_1], (t_1, t_2], \ldots, (t_{n-1}, t_n] \) given a total of \( X \) arrivals in \( [t_0, t_n] \).
BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(10:00 a.m.- 2:00 p.m., Tuesday, July 23, 1985)

INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is four hours.

c) Answer any four (but only four) of the five questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Put your code letter, not your name, on each page.

f) Return the examination with a signed statement of honor pledge on a page separate from your answers.

g) You are required to answer only what is asked in the questions and not all you know about the topics.
Q1. The normally distributed random variables $X_1, X_2, \ldots, X_n$ are said to be serially correlated or to follow an autoregressive model if we can write

$$X_i = \theta X_{i-1} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $X_0 \equiv 0$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent $N(0, \sigma^2)$ random variables.

a) Find $\text{corr}(X_1, X_2)$.

b) Find $f_{X_2}(x_2 | X_1 = x_1)$, the conditional distribution of $X_2$ given $X_1 = x_1$.

c) Find $f_{X_1, X_2}(x_1, x_2)$, the joint distribution of $X_1$ and $X_2$.

d) Prove that the joint distribution of $X_1, X_2, \ldots, X_n$ is

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta x_{i-1})^2\right\},$$

$-\infty < x_i < +\infty$, $i = 1, 2, \ldots, n$, $X_0 \equiv 0$.

e) Using one observation $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from the joint density given in part (d), show that the likelihood ratio test of $H_0: \theta = 0$ versus $H_A: \theta \neq 0$ can be expressed directly as a function of the statistic

$$\left[\frac{\sum_{i=1}^{n-1} x_i x_{i-1}}{\sum_{i=1}^{n} x_i^2}\right]^{n-1} \left[\frac{\sum_{i=1}^{n-1} x_i^2}{\sum_{i=1}^{n} x_i^2}\right].$$
Q2. Suppose you are asked to design the sample for a survey of patients seen during 1934 in a large hospital. The main purpose of the survey will be to estimate \( D = P_M - P_F \), where \( P_M \) and \( P_F \) are, respectively, the proportions of all male and female patients without health insurance. You propose to select a stratified simple random sample (selected without replacement) of size \( n \) from among the 5000 patients in the survey population. Stratification will be by sex, with each stratum containing half of the population.

a) Suppose (1) that \( D = 0 \), and (2) that total survey cost can be expressed as

\[
C = C_M n_M + C_F n_F ,
\]

where the average per-unit costs for males \( (C_M) \) and females \( (C_F) \) are equal, and \( n_M \) and \( n_F \) are, respectively, the male and female sample sizes. Show that the optimum allocation of the \( n \) patients among strata is proportionate allocation.

b) Suppose that, in fact, \( P_M = 0.40 \) and \( P_F = 0.20 \); how large a sample from your proportionate stratified sample would be needed to yield a 10 percent coefficient of variation for the unbiased estimator, \( d = p_M - p_F \), of \( D \), where \( p_M \) and \( p_F \) are, respectively, the sample proportions of male and female patients without insurance?

c) If an unstratified simple random sample has been proposed instead of the stratified simple random sample, show that the expected size of \( n_F \) (the number of females in the sample) would have been \( 2500(n/N) \).
Q3. a) Define the following:
   i) Markov chain;
   ii) homogeneous Markov chain;
   iii) persistent state, transient state.

b) Let \( X_1, X_2, \ldots \) be independent integer-valued random variables with density \( f \). Let \( Y_0 \) be an integer-valued random variable that is independent of the \( X_i \)'s and set
   \[
   Y_n = Y_0 + X_1 + X_2 + \ldots + X_n, \quad n \geq 1.
   \]
   Show that the sequence \( \{Y_n, n \geq 0\} \) is a Markov chain whose state space is the set of all integers and whose transition function \( P(x,y) \) is given by
   \[
   P(x,y) = f(y - x).
   \]

c) Let \( \{Y_n, n \geq 0\} \) be a Markov chain. Given the \( n \)-step transition probabilities
   \[
   p_{ij}(n) = P(Y_{m+n} = j | Y_m = i),
   \]
   write down the Chapman-Kolmogorov equations.
Q4. Let $X_1, X_2, \ldots, X_n$ be random variables such that $E(X_i) = \mu$, $V(X_i) = \sigma^2$, $\text{corr}(X_i, X_j) = \rho$ for every $i \neq j$, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$. Define

$$L_1 = \sum_{i=1}^{n} a_i X_i \text{ and } L_2 = \sum_{i=1}^{n} b_i X_i,$$

where the $a_i$'s and $b_i$'s are constants.

a) Find a general expression for $\text{corr}(L_1, L_2)$ as a function of the $a_i$'s, $b_i$'s and $\rho$.

b) Using the result derived in part (a), or otherwise, find $\text{corr}(L_1, L_2)$ when $L_1 = (X_1 - X_3)$ and $L_2 = (X_1 - 2X_2 + X_3)$.

c) When $L_1 = X_1$ and $L_2 = (X_2 - X_1)$, and when $E(X_2|X_1 = x_1) = \alpha + \beta x_1$, express $\text{corr}(L_1, L_2)$ as a function of $\beta$. What is $\text{corr}(L_1, L_2)$ when $\beta = 0$?
Q5. Consider the density function

\[
f_X(x; \theta) = c^{1/\theta} \theta^{-1} x^{-\left(1 + \frac{1}{\theta}\right)}, \quad x > c, \quad \theta > 0, \quad c > 0.
\]

Here, \(\theta\) is an unknown parameter and \(c\) is a known constant.

a) For \(r > 0\), find \(E(X^r)\). For what set of values of \(r\) does \(E(X^r)\) exist?

b) For \(\theta < 1\), if \(X_1, X_2, \ldots, X_n\) constitute a random sample of size \(n\) from \(f_X(x; \theta)\), show that

\[
\hat{\xi} = \left(\frac{\bar{X} - c}{c}\right), \quad \text{where} \quad \bar{X} = n^{-1} \sum_{i=1}^{n} X_i,
\]

is an unbiased estimator of \(\xi = \theta/(1 - \theta)\).

c) Find \(V(\hat{\xi})\). Under what conditions on \(\theta\) does \(V(\hat{\xi})\) exist?

d) Find an explicit expression (as a function of \(\xi\)) for the Cramér-Rao lower bound for the variance of any unbiased estimator of \(\xi\). Is \(\hat{\xi}\) the minimum variance bound unbiased estimator of \(\xi\)?
BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

PART I

(9:00 a.m. - 1:00 p.m., Tuesday, July 29, 1986)

INSTRUCTIONS

a) This is a closed-book examination.

b) The time limit is four hours.

c) Answer any four (but only four) of the five questions which follow.

d) Put the answers to different questions on separate sets of papers.

e) Put your code letter, not your name, on each page.

f) Return the examination with a signed statement of honor pledge on a page separate from your answers.

g) You are required to answer only what is asked in the questions and not all you know about the topics.
Q1. Suppose that we have available \( n \) sets of observations of the form 
\[(Y_i, x_{1i}, x_{2i}), \quad i = 1, 2, \ldots, n,\]
again under the following conditions:

i) the \( x \) values, namely \( \{x_{1i}\}_{i=1}^n \) and \( \{x_{2i}\}_{i=1}^n \), are fixed, known constants;

ii) for each \( i \), \( Y_i \) is a normally distributed random variable with mean
\[E(Y_i | x_{1i}, x_{2i}) = \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2),\]
where \( \bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{1i} \) and \( \bar{x}_2 = \frac{1}{n} \sum_{i=1}^{n} x_{2i} \),
and with variance
\[V(Y_i | x_{1i}, x_{2i}) = \sigma^2;\]

iii) \( Y_1, Y_2, \ldots, Y_n \) constitute a set of mutually independent random variables.

Define
\[S^2_x = (n-1)^{-1} \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2,\]
\[S^2_{x_2} = (n-1)^{-1} \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2,\]

and
\[\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = (n-1) \Sigma_{x_1 x_2}.\]

a) Assuming that the regression model in (ii) above is to be fit by the method of least squares, show that the ordinary least squares estimator \( \hat{\beta}_2 \) of \( \beta_2 \) has variance
\[V(\hat{\beta}_2) = \frac{\sigma^2}{(n-1)S^2_x (1-r^2)}.\]

b) Now, suppose that we have available an extraneous unbiased estimator \( b_1 \) of \( \beta_1 \) (i.e., an independent estimator of \( \beta_1 \) not based on the \( n \) - 1 observations of each \( i \)).
observations currently under consideration). Let \( V(b_1) = \sigma^2 \). To estimate \( \beta_2 \), consider the regression of \( \{Y_i - b_1 (x_{i1} - \bar{x}_1)\} \) on \( (x_{i2} - \bar{x}_2) \). In particular, show that the estimator
\[
\hat{b}_2 = \frac{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)(Y_i - b_1 (x_{i1} - \bar{x}_1))}{(n-1) S_{x_2}^2}
\]
is an unbiased estimator of \( \beta_2 \) with variance
\[
V(\hat{b}_2) = \frac{\sigma^2 + (n-1) S_{x_1}^2 \sigma^2 \rho^2}{(n-1) S_{x_2}^2}.
\]

(c) Using the results in parts (a) and (b), show that the use of the extraneous information about \( \beta_1 \) (namely, \( b_1 \)) increases precision over just the use of \( \hat{b}_2 \) when estimating \( \beta_2 \) if and only if
\[
(n-1)(1-\rho^2) S_{x_1}^2 < \frac{\sigma^2}{\sigma_{b_1}^2}.
\]

Q2. Let \( X_1, \ldots, X_n \) be a random sample from a Poisson distribution with parameter \( \lambda \), i.e.,
\[
Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \ k = 0, 1, \ldots.
\]
Define \( \pi = Pr(X \leq 1) = (1+\lambda)e^{-\lambda}, \ \lambda > 0. \)

(a) Show that \( \prod_{i=1}^{n} x_i \) (=T, say) is a complete sufficient statistic for \( \pi \).

(b) Show that \( I_{[X_1 \leq 1]} \) is an unbiased estimator of \( \pi \). (\( I_A \) stands for the indicator function of the set \( A \)).

(c) Obtain the minimum variance unbiased estimator of \( \pi \).

(d) Show that the statistic obtained in (c) is a function of \( T \), and is indeed an unbiased estimator of \( \pi \).
Q3. Consider the situation where our goal is to estimate the proportion \( P \) with some attribute in a population of \( N \) elements that has been divided into \( H \) strata, with the proportion \( (P_h) \) of the \( N_h \) elements in the \( h \)-th stratum possessing the same attribute.

8 pts.

a) Suppose that a (without replacement) simple random sample of \( n_h \) elements is chosen independently in each stratum, and that the estimator for \( P \) is \( \hat{p} = \sum_{h}^{H} W_h P_h \), where \( P_h \) is the proportion of the \( n_h \) elements with the attribute and \( W_h = N_h / N \). Derive the true variance of \( \hat{p} \).

[Hint: Recall that the true variance of the stratum sample mean \( \bar{y}_h \) in this setting would be
\[
\text{Var}(\bar{y}_h) = \frac{(1-f_h)}{n_h} \sum_{i}^{n_h} \left( \bar{y}_h - \bar{y} \right)^2
\]
where \( y_{hi} \) is the \( i \)-th observation in the \( h \)-th stratum, \( \bar{y}_h \) is the overall stratum mean, and \( f_h = n_h / N_h \).]

5 pts.

b) If proportionate stratified simple random sampling is used in Part (a), show that
\[
\text{Var}(p) = \frac{(1-f)}{n} \sum_{h}^{H} W_h \frac{N_h}{(N_h-1)} P_h (1-P_h),
\]
where \( n = \sum_{h}^{H} n_h \) and \( f = n/N \).

8 pts.

c) If \( \text{Var}(p) \) represents the true variance of the proportion \( p_0 \) of a simple random sample of size \( n \) who possess the attribute, show that
\[
\text{Var}(p_0) - \text{Var}(p) \approx \frac{(1-f)}{n} \sum_{h}^{H} W_h \frac{N_h}{(N_h-1)} (P_h - P)^2.
\]

[Hint: Partition \( S^2 = \sum_{hi}^{HN} (y_{hi} - \bar{y})^2/(N-1) \) into between and within stratum components, and note that \( N(N_h-1)/N_h(N-1) \geq 1 \).]
d) Show how the result from part (c) supports the general notion that stratified sampling (as opposed to no stratification in sampling) is most effective when internally homogeneous strata can be formed prior to selection.

Q4. Suppose that $X$ is a random variable with density function $f(x)$, and with cumulative distribution function (cdf) $F(z) = \Pr(X \leq z)$.

a) Show that $E(X) = \int_{0}^{\infty} x f(x) \, dx$.

b) Use part (a) to show that $E(X) = \int_{0}^{\infty} [1 - F(z)] \, dz$.

Suppose that $X$ is the failure time of a certain piece of equipment. A replacement policy might be to replace the piece either upon failure or upon reaching age $c$, whichever event happens first. Then, $X_c$, the time of replacement, is expressible as

$$X_c = \min(c, X) = \begin{cases} X & \text{if } X \leq c, \\ c & \text{if } X > c \end{cases}.$$

c) Express $\Pr(X_c > z)$ in terms of the cdf of $X_c$.

d) Use parts (b) and (c) to express $E(X_c)$ in terms of the cdf of $X$.

e) Evaluate $E(X_c)$ for the special case when $X$ is exponentially distributed, i.e., when $f(x) = \theta e^{-\theta x}$. 
Q5. Let \( \{X_{ij}\}, i=1, 2, 3, \text{ and } j=1, 2, \ldots, n, \) be a set of independent normal random variables with

\[
E(X_{ij}) = \mu_i \text{ and } Var(X_{ij}) = \sigma^2.
\]

Confidence intervals of the general form

\[
\bar{x}_i \pm \lambda s_i
\]

are calculated for \( i=1, 2, 3, \) where

\[
\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2;
\]

\( \lambda \) is chosen (using appropriate percentiles of a Student's t distribution) so that each confidence interval has exact confidence coefficient \((1-\alpha)\).

For notational simplicity, let \( I_i \) denote the set of points in the \( i \)-th interval \((\bar{x}_i \pm \lambda s_i)\), \( i=1, 2, 3; \) further, let the event \( E_{ii'} \), \( i \neq i' \), denote the event

\[
I_i \cap I_{i'} = \emptyset;
\]

i.e., \( E_{ii'} \) is the event that the confidence intervals \((\bar{x}_i \pm \lambda s_i)\) and \((\bar{x}_{i'} \pm \lambda s_{i'})\) do not overlap.

a) Show that

\[
pr(E_{ii'}) = pr(\left|\bar{x}_i - \bar{x}_{i'}\right| > \lambda (s_i + s_{i'})).
\]

b) Under the null hypothesis \( H_0: \mu_1 = \mu_2 = \mu_3 \), show, for \( i \neq i' \), that

\[
pr(E_{ii'}|H_0) \leq pr(\left|t_{2(n-1)}\right| > t_{(n-1), 1-\frac{\alpha}{2}}|H_0),
\]

where \( t_v \) is a Student's t random variable with \( v \) degrees of freedom and \( pr(t_v > t_{v, 1-\frac{\alpha}{2}}) = \frac{\alpha}{2}. \)

c) It is proposed to test the null hypothesis \( H_0: \mu_1 = \mu_2 = \mu_3 \) versus the alternative hypothesis \( H_A: \mu \)'s are not all equal by not rejecting \( H_0 \) if and only if the three confidence intervals all have at least one point in common (i.e., there is at least one point which is simultaneously in all three confidence intervals). Using earlier results, show that the significance level for this proposed statistical testing procedure is at most

\[
3\alpha(1 - \alpha + \frac{\alpha^2}{3}).
\]
Q.1. Specimens of a new high-impact plastic are tested by repeatedly striking them with a hammer until they fracture. If each specimen has a constant probability $\theta$ of surviving a blow, independently of the number of previous blows received, the number of blows $X$ required to fracture a specimen will have the geometric distribution

$$p_X(x; \theta) = (1-\theta)^{x-1} \theta, \quad x = 1, 2, \ldots, \infty.$$ 

a) Suppose that the results of tests on $n = 200$ specimens were as follows:

<table>
<thead>
<tr>
<th>Number of blows $x$ required:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>$\geq 4$</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of specimens fracturing after $x$ blows:</td>
<td>112</td>
<td>36</td>
<td>22</td>
<td>30</td>
<td>200</td>
</tr>
</tbody>
</table>

Find the maximum likelihood estimator of $\theta$, and develop an appropriate 95% confidence interval for $\theta$.

b) Based on an inspection of the data in part (a), a statistician suggests that, while the geometric distribution applies to most specimens, a fraction $(1 - \lambda)$ of them have structural flaws and therefore always fracture on the first blow. Given this additional consideration, show that the proportions of specimens fracturing after one, two, three, and four or more blows are, respectively,

$$(1-\lambda \theta), \lambda \theta(1-\theta), \lambda^2 \theta(1-\theta), \text{ and } \lambda^3 \theta^3.$$
c) If \( x_i \) specimens are observed in the \( i \)-th category \((i = 1, 2, 3, 4)\), where \( \sum_{i=1}^{n} x_i = n \), use the results in part (b) to find explicit expressions for the maximum likelihood estimators of \( \lambda \) and \( \theta \).

0.2. Consider systematic sampling from a serially numbered list of \( N \) population elements in which a sample is chosen by drawing a random integer between 1 and \( k \), say \( g \), and selecting the element with the corresponding serial number and every \( k \)-th element thereafter on the list. Instead of the simple case where \( N \) is an integral multiple of the sample size \( (n) \) for all possible random starts, suppose that \( N = nk + r \), where \( r \) is an integer such that \( 1 \leq r < k \), i.e., \( N \) divided by the actual sample size is not an integer.

a. Explicitly describe how this design is equivalent to a one-stage sample of a single cluster from a set of unequal sized clusters.

b. Let \( y_{gj} \) denote the measurement of interest for the element bearing the serial number \( g + (j-1)k \) in the population; \( g = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, n_g \), where

\[
\frac{n_g}{n} = \begin{cases} 
\frac{n+1}{n} & \text{if } 1 \leq g \leq r \\
1 & \text{if } r < g \leq k
\end{cases}
\]

Also let \( \bar{y}_g = \frac{1}{n_g} \sum_{j=1}^{n_g} y_{gj} \) denote the sample mean when the sample's random start is \( g \). Show that the expected value of \( \bar{y}_g \) among all possible samples is

\[
E(\bar{y}_g) = \frac{1}{k} \sum_{g=1}^{k} \bar{y}_g/k
\]

and that \( \bar{y}_g \) will be an unbiased estimator of the population mean,

\[
\bar{Y} = \frac{1}{N} \sum_{g=1}^{k} \sum_{j=1}^{n_g} y_{gj}/N,
\]

if \( N \) is an integral multiple of the sample size for all possible random starts.
c. Show that \( \bar{y}^* = kN / \sum v_j \) is an unbiased estimator of \( \bar{y} \), where
\[
\bar{y}_g = \frac{1}{n} \sum_{j=1}^{n} v_{gj}
\]
d. Derive the true variance of \( \bar{y}^* \) among all possible samples.

Q.3

a) Suppose that \( Y_1, Y_2, \ldots, Y_n \) are independent Poisson variables with respective means \( \mu_1, \mu_2, \ldots, \mu_n \). Let \( k_1, k_2, \ldots, k_n \) be known positive constants. Consider the null hypothesis
\[
H_0: \mu_i = k_i \theta, \quad i = 1, 2, \ldots, n,
\]
where \( \theta (>0) \) is an unknown parameter. Show that the likelihood ratio statistic \( -2 \ln \lambda \) for testing \( H_0 \) versus the alternative hypothesis \( H_A \) that there are no structural restrictions on the \( \mu_i \)'s takes the form
\[
-2 \ln \lambda = 2 \sum_{i=1}^{n} Y_i \ln(Y_i/\bar{\mu}_i),
\]
where
\[
\bar{\mu}_i = k_i \theta \quad \text{and} \quad \theta = \frac{\sum Y_i}{\sum k_i}.
\]

b) Suppose that \( k_1 = k_2 = k_3 = 1,000, \ Y_1 = 10, \ Y_2 = 11, \) and \( Y_3 = 12 \). Test \( H_0 \) versus \( H_A \) using part (a). What is the P-value of your test? Do you reject \( H_0 \) or not?

c) Use the data in part (b) to construct an appropriate 95% confidence interval of \( \theta \).
Q.4. Data have been gathered from hospital records to study the effect of treatment of a certain disease. A Markov model with 4 states is to be used for the study:

0 - state of being under treatment
1 - death after treatment and due to the disease
2 - recovery
3 - state of being lost after recovery or death from other causes.

DATA:

<table>
<thead>
<tr>
<th>Previous State</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60</td>
<td>12</td>
<td>83</td>
<td>0</td>
<td>155</td>
</tr>
<tr>
<td>2</td>
<td>72</td>
<td>0</td>
<td>77</td>
<td>27</td>
<td>176</td>
</tr>
</tbody>
</table>

Assume only one state transition is possible between observations.

a) What is your estimate of the matrix of transition probabilities?

b) How would you clarify each of the states? What is the period of the chain?

c) What is the probability that a patient who is under treatment will survive for 4 observation periods or longer?

d) What is the probability that a patient in recovery will be 'lost' to observation or die of other causes?

e) What is the probability that a patient in treatment will eventually die of the disease?
Q.5. Let $X_1, \ldots, X_n$ denote a random sample from the density $f_\theta$ defined by:

$$f_\theta(x) = \begin{cases} \frac{\theta}{(\theta+x)^2} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0; \end{cases}$$

where $\theta \in \Theta = \{\theta: \theta > 0\}$

a) Compute the Fisher information $I_n(\theta)$ for this set of $n$ observations. It $t_n$ is the maximum likelihood estimator (you don't have to solve the likelihood equation), find the asymptotic distribution of $\sqrt{n}(t_n - \theta)$ as $n \to \infty$.

b) For $0 < p < 1$, show that if $Y_{n,p}$ is the sample $p$-tile, then $\sqrt{n}(Y_{n,p} - \frac{p \theta}{1-p})$ is asymptotically normal $(0; \frac{p \theta^2}{(1-p)^3})$.

[Hint: You may use without proof that if $\xi_p$ is the $p$-tile of a distribution $F(\cdot)$ with continuous density $f(\cdot)$, then the sample $p$-tile $Y_{n,p}$ is asymptotically normal $(\xi_p, \frac{p(1-p)}{f^2(\xi_p)n})$.]

c) Find a constant $c_p$ such that $\{c_p Y_{n,p}\}$ is a consistent sequence of estimators of $\theta$, and determine the asymptotic law if $\sqrt{n}(c_p Y_{n,p} - \theta)$.

d) Determine the asymptotic relative efficiency of $\{c_p Y_{n,p}\}$ to $\{t_n\}$. What value of $p$ yields the best efficiency?

e) Since $t_n$ is difficult to compute, give a computable sequence of estimators $\hat{t}_n$ that are asymptotically as efficient as $t_n$. 