WRITTEN EXAMINATIONS FOR THE DEGREES
MASTER OF PUBLIC HEALTH AND MASTER OF SCIENCE IN PUBLIC HEALTH

in the

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

assembled and edited by

DANA QUADE

Institute of Statistics Mimeo Series #1329
Third Edition: July 1988
WRITTEN EXAMINATIONS

for the degrees

MASTER OF PUBLIC HEALTH

and

MASTER OF SCIENCE IN PUBLIC HEALTH

in the

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

assembled and edited by

DANA QUADE

First Edition: February 1981
Second Edition: September 1984
Third Edition: July 1988
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Date</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1966</td>
<td>4</td>
</tr>
<tr>
<td>1967</td>
<td>7</td>
</tr>
<tr>
<td>27 April</td>
<td>11</td>
</tr>
<tr>
<td>3 May</td>
<td>13</td>
</tr>
<tr>
<td>1969</td>
<td>14</td>
</tr>
<tr>
<td>4 April</td>
<td>15</td>
</tr>
<tr>
<td>24 April</td>
<td>19</td>
</tr>
<tr>
<td>2 August</td>
<td>27</td>
</tr>
<tr>
<td>8 April</td>
<td>31</td>
</tr>
<tr>
<td>12 May</td>
<td>37</td>
</tr>
<tr>
<td>31 March</td>
<td>40</td>
</tr>
<tr>
<td>13 April</td>
<td>47</td>
</tr>
<tr>
<td>12 April</td>
<td>53</td>
</tr>
<tr>
<td>3 April</td>
<td>58</td>
</tr>
<tr>
<td>2 April</td>
<td>62</td>
</tr>
<tr>
<td>8 April</td>
<td>65</td>
</tr>
<tr>
<td>7 April</td>
<td>69</td>
</tr>
<tr>
<td>12 April</td>
<td>76</td>
</tr>
<tr>
<td>11 April</td>
<td>84</td>
</tr>
<tr>
<td>3 April</td>
<td>93</td>
</tr>
<tr>
<td>29 May</td>
<td>103</td>
</tr>
<tr>
<td>10 April</td>
<td>104</td>
</tr>
<tr>
<td>14 April</td>
<td>110</td>
</tr>
<tr>
<td>13 April</td>
<td>120</td>
</tr>
<tr>
<td>12 April</td>
<td>126</td>
</tr>
<tr>
<td>7 May</td>
<td>130</td>
</tr>
<tr>
<td>5 May</td>
<td>138</td>
</tr>
</tbody>
</table>
INTRODUCTION

This publication contains the written examinations which the Department of Biostatistics has set for candidates for the degrees of Master of Public Health (MPH) and Master of Science in Public Health (MSPH). A Department-wide written examination for these degrees was first instituted in 1966, in order to satisfy the Graduate School requirement that every Master's candidate pass a comprehensive examination covering all course work done for the degree. This examination has been scheduled each Spring starting from 1966 to the present; there have also been several special reexaminations.

The rules under which the examination was conducted for the first four years are not entirely clear. In 1966 it consisted of 3 questions, while in 1967, 1968, and 1969 (both regular and special) there were only 2 questions. Starting from 1970, however, the rules have been explicit and formal. From 1971 through 1976 the examination was closed-book, with 3 hours allowed in which to answer 3 of 6 questions (exceptions: 2 of 3 in the 1971 special exam, and 3 of 4 in the 1972 special exam). In 1977 the examination was changed to open-book, with 4 hours allowed in which to answer 3 of 5 questions. In 1980 the rules changed again, so that the same examination could be taken not only by MPH and MSPH candidates but also by candidates for the degree Master of Science (MS). (Up till this time the written examination for MS students had been the same as the Basic Written Examination for doctoral students.) The examination was then given in two Parts, each with two Sections, as follows:

Part One: Theory (Closed Book)
- MPH: 2 hours; any 2 questions
- MSPH and MS: 3 hours; 3 questions, at least 1 from each Section

Part Two: Applications (Open Book)
- MPH and MSPH: 2 hours; any 2 questions
- MS: 3 hours; 3 questions, at least 1 from each Section
(MSPH candidates not desiring to go on to the DrPH were permitted to answer 2 questions in Part One and 3 in Part Two.)

Beginning in 1983 an examination for the MSPH was no longer needed, that degree having been discontinued; and the examination for the MS was separated out and scheduled for an earlier date each year. The rules then adopted for the MPH written examination provide that there will be two Parts (Theory and Applications), both open-book, with 2 hours allowed in which to answer 2 of 3 questions.

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 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 between their scores. Their joint report may 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.
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 Examinations 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. A candidate whose performance is not of the standard required may be reexamined at the next regularly scheduled examination, or at an earlier date set by the Examinations Committee. One reexamination is permitted automatically.

Candidates whose native language is not English are not to be 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.
1. A medical research worker wanted to know whether, for a certain specified type of patient, a new drug X would be more effective in relieving pain (i.e., produce higher "pain relief scores") than the standard drug C. He divided a sample of 30 patients into two equal groups at random, gave one drug to each group, and obtained the following scores:

Drug X: 0,0,1,1,2,2,4,5,6,6,6,6,6,6,6,6

\[ \sum x_i = 57 \quad \sum x_i^2 = 303 \]

Drug C: 0,1,1,2,2,3,3,3,3,4,4,4,4,4,6

\[ \sum y_i = 42 \quad \sum y_i^2 = 150 \]

Give several essentially different statistical descriptions of these data, and base a test of significance on one of them. Justify your choice of test and state the assumptions it involves. What conclusion do you draw with respect to the research worker's original question? Write a brief report including your descriptions, analysis, and conclusion.
2. The following data give the number of pounds of weight gained in a random sample of 20 children in the first and second years of life. It has been suggested that the weight gain in the second year might be predicted by simply subtracting a constant from the weight gain in the first year. If so, what constant should be subtracted? Can you find a significantly better prediction method?

<table>
<thead>
<tr>
<th>Child Number</th>
<th>First Year (x)</th>
<th>Second Year (y)</th>
<th>Child Number</th>
<th>First Year (x)</th>
<th>Second Year (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>20.2</td>
<td>17.2</td>
<td>11.</td>
<td>29.6</td>
<td>23.3</td>
</tr>
<tr>
<td>2.</td>
<td>20.4</td>
<td>22.3</td>
<td>12.</td>
<td>30.3</td>
<td>26.4</td>
</tr>
<tr>
<td>3.</td>
<td>22.5</td>
<td>19.5</td>
<td>13.</td>
<td>30.4</td>
<td>30.3</td>
</tr>
<tr>
<td>4.</td>
<td>25.7</td>
<td>24.2</td>
<td>14.</td>
<td>32.3</td>
<td>29.4</td>
</tr>
<tr>
<td>5.</td>
<td>25.8</td>
<td>21.0</td>
<td>15.</td>
<td>32.7</td>
<td>28.7</td>
</tr>
<tr>
<td>6.</td>
<td>25.9</td>
<td>23.9</td>
<td>16.</td>
<td>35.0</td>
<td>33.2</td>
</tr>
<tr>
<td>7.</td>
<td>27.7</td>
<td>24.5</td>
<td>17.</td>
<td>37.1</td>
<td>31.7</td>
</tr>
<tr>
<td>8.</td>
<td>28.1</td>
<td>27.5</td>
<td>18.</td>
<td>37.4</td>
<td>32.1</td>
</tr>
<tr>
<td>9.</td>
<td>28.3</td>
<td>24.4</td>
<td>19.</td>
<td>38.1</td>
<td>38.1</td>
</tr>
<tr>
<td>10.</td>
<td>29.4</td>
<td>28.7</td>
<td>20.</td>
<td>39.1</td>
<td>32.9</td>
</tr>
</tbody>
</table>

\[
\Sigma x = 596.0 \\
\Sigma y = 539.3 \\
\Sigma x^2 = 18361.12 \\
\Sigma y^2 = 15064.33 \\
\Sigma (x-x) = 600.3200 \\
\Sigma (y-y) = 522.1055 \\
\bar{x} = 29.800 \\
\bar{y} = 26.965 \\
\Sigma xy = 16585.00 \\
\Sigma (x-\bar{x})(y-\bar{y}) = 513.8600
\]
3. The Congress of the United States is currently discussing the question whether or not to continue the food supplementation program in the public schools. This consists of providing milk and other foods for the school lunch program so that every child can participate either at no cost or at a very nominal price.

Suppose the Director of Public Health Nutrition in X State Health Department consults you as the staff biostatistician to help him to evaluate the effectiveness of this program in X state. He feels that if sound data can be presented to the Department of Agriculture regarding the effectiveness of this program, that Department will have a stronger case in arguing the merits of continuing the school lunch program.

Describe what you would recommend to the Director of Public Health Nutrition and what questions you might ask him about the school lunch program. Assuming a given set of answers to those questions, how would you proceed to design a study to answer the question of effectiveness of the school lunch program. If your recommendation includes the collection of data in certain schools, indicate how you would select the schools, how many schools, how many children, and other details necessary in the planning of such a study.
The aim of this examination is to find out something about the student's general ability and ingenuity in tackling a problem of the kind which might be brought to him as a statistician in a health agency. You will thus see that the two questions which follow ask you to describe approaches to solution; they do not call for detailed solutions. Be as brief and specific as you can in giving your answers, with the purpose of showing clearly what you have in mind and why it is appropriate.

1. It is believed that the position of the lower incisor (tooth) plays an important role in the aesthetics of facial features. It is commonly assumed that the angle called IMPA (Lower Incisor - Mandibular Plane Angle), which we will denote by \( y \), should have a certain value related to general facial features; this value varies between about 80° and 110° for various combinations of facial features.

In view of the above, an orthodontic treatment is to place (push or pull) the mandible so that the desired angular position can be approximately obtained.

But it has been observed that, after some period, the teeth of some individuals return back to the initial position, those of others move partially back, and the teeth of some stay on the position to which they were placed by the treatment.

We will, then, distinguish three positions of the lower incisor:

(1) Position 1, that is, the position at the initial stage, before the treatment was applied; the measurements of IMPA here are denoted by \( y_1 \);
(ii) Position 2, that is, the position of the lower incisor at the conclusion of the treatment, at the time where the patient was dismissed; measurements of IMPA here are denoted by $y_2$.

(iii) Position 3, that is, the position which the lower incisor reaches at some stated time after the treatment; the measurements of IMPA here are denoted by $y_3$.

In a certain study, there were three time periods after which $y_3$ was measured in three independent groups of patients:

- Group I $y_3$ measured after 1-2 years (15 patients)
- Group II $y_3$ measured after 2-3 years (14 patients)
- Group III $y_3$ measured after 3 years (7 patients).

The general problem which arises here is: Why does the mandible, after treatment, try to return back to the initial position in some individuals, but stay in the position attained by the treatment in other individuals?

An important question in the matter is: Does the "amount of relapse" depend on the amount of treatment (magnitude of change in $y$ from $y_1$ to $y_2$), or is it related to the initial position, or to both: amount of treatment and initial position?

1. Describe and comment on statistical methods which you would like to apply to answer this question. Use the enclosed data to make some representative calculations. (A complete analysis is not expected.)

2. How could you improve the design of the experiment?
### Stability of the Mandibular Incisor

<table>
<thead>
<tr>
<th>Patient No.</th>
<th>Group I (1-2 years)</th>
<th>Group II (2-3 years)</th>
<th>Group III (over 3 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$y_1$</td>
<td>$y_2$</td>
<td>$y_3$</td>
</tr>
<tr>
<td>1</td>
<td>85</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>95</td>
<td>97</td>
<td>96</td>
</tr>
<tr>
<td>3</td>
<td>94</td>
<td>88</td>
<td>84</td>
</tr>
<tr>
<td>4</td>
<td>99</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>86</td>
<td>85</td>
<td>89</td>
</tr>
<tr>
<td>6</td>
<td>93</td>
<td>87</td>
<td>88</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>90</td>
<td>88</td>
</tr>
<tr>
<td>8</td>
<td>92</td>
<td>98</td>
<td>103</td>
</tr>
<tr>
<td>9</td>
<td>99</td>
<td>96</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>91</td>
<td>91</td>
<td>94</td>
</tr>
<tr>
<td>11</td>
<td>83</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>12</td>
<td>77</td>
<td>87</td>
<td>87</td>
</tr>
<tr>
<td>13</td>
<td>98</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>14</td>
<td>97</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>91</td>
<td>89</td>
<td>89</td>
</tr>
</tbody>
</table>

2. Assume that a program has just been initiated in a region of the size of North Carolina to provide for more widespread use of the latest medical advances in diagnosis and treatment of heart disease, cancer, and stroke. The program will be administered by a medical school and place primary emphasis on dissemination of new knowledge about care for these diseases to medical personnel throughout the region for use in patient care. Numerous special short training courses will be given to assist in providing up to date information to medical personnel.
The region to be served by the program has a population of approximately 5,000,000 persons about equally divided between urban and rural communities. About one fourth of the population is over 45. Over half of the deaths have been due to heart disease, cancer, or stroke in recent years. The region has three medical schools, about a dozen major hospitals affiliated with the medical schools, and scattered smaller general hospitals and physicians in private practice.

You have been asked to serve as a consultant in planning and subsequent evaluation of training programs for the region. Briefly indicate analyses you would suggest on:

1. The nature and magnitude of the problem of morbidity and mortality due to heart disease, cancer, and stroke among different population groups in the region.

2. Availability of and needs for trained physicians, nurses, and related personnel for care of patients with the specified diseases.

3. Training provided under the program, such as number of persons of various medical specialties trained, kinds of training, and costs.

4. Effectiveness of the training programs in bringing better patient care and reduced morbidity and mortality due to heart disease, cancer, and stroke.

5. Possible uses of computers in preparing and maintaining desired information for use in the program.

Give particular attention to the fourth item in your discussion.
Experiment:

It has been postulated that the volume of a cake is influenced very substantially by batter temperature. Using a specific cake recipe, a large amount of batter was made, and then divided into 20 equal portions (aliquots). It was decided to see if the cake volume was different at 4 different batter temperatures, 60°, 70°, 80°, and 90°. Each aliquot was allotted to a temperature at random until there were 5 aliquots on each temperature. The cakes were baked one at a time and the data were recorded as follows:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60°</td>
<td>6</td>
<td>11</td>
<td>80°</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>60°</td>
<td>8</td>
<td>12</td>
<td>80°</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>60°</td>
<td>8</td>
<td>13</td>
<td>80°</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>60°</td>
<td>4</td>
<td>14</td>
<td>80°</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>60°</td>
<td>4</td>
<td>15</td>
<td>80°</td>
<td>+2</td>
</tr>
<tr>
<td>6</td>
<td>70°</td>
<td>6</td>
<td>16</td>
<td>90°</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>70°</td>
<td>9</td>
<td>17</td>
<td>90°</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>70°</td>
<td>10</td>
<td>18</td>
<td>90°</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>70°</td>
<td>8</td>
<td>19</td>
<td>90°</td>
<td>-2</td>
</tr>
<tr>
<td>10</td>
<td>70°</td>
<td>7</td>
<td>20</td>
<td>90°</td>
<td>-2</td>
</tr>
</tbody>
</table>

The Analysis of Variance was computed for this example and is recorded as follows:

ANOVA

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>d.f.</th>
<th>s.s.</th>
<th>m.s.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>19</td>
<td>211</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bet. Temperatures</td>
<td>3</td>
<td>145</td>
<td>48.33</td>
<td>11.71*</td>
</tr>
<tr>
<td>Residual</td>
<td>16</td>
<td>66</td>
<td>4.125</td>
<td></td>
</tr>
</tbody>
</table>

Requirements:

1. State the assumptions underlying the analysis of variance done in the above table.

2. Indicate where each assumption is needed in the analysis of the above data.

3. Devise a method for checking at least two of the assumptions.

4. Complete the analysis of the experiment indicating your conclusions (use $\alpha = 5\%$).
II.

As statistician in the State Health Department, you have been given the task of designing a study to determine the relationship (if any) between maternal rubella (German measles) and congenital malformations in infants. The choice between a retrospective and prospective study is yours. Ignoring cost, prepare and briefly discuss your design for such a study for the State of North Carolina. Emphasize statistical problems which might be encountered at all stages (including analysis) and your solution to them.
I. The data below constitute independent random samples from two Poisson distributions with parameters $\mu_1$ and $\mu_2$.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>18</td>
<td>19</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Sample 2</td>
<td>9</td>
<td>16</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>50</td>
</tr>
</tbody>
</table>

a) Describe two distinct ways of testing the hypothesis $H_0: \mu_1 = \mu_2$.

b) Perform one of these tests, with $\alpha = .05$. State why you chose this test. Write up your conclusions.

II. Suppose you have been called into a county as a statistical consultant to help set up a Comprehensive Health Planning Agency. Primarily you are asked to construct the health information system necessary for community health planning on a broad scale.

The permanent population of this county is about 200,000, but it increases seasonally with an influx of tourists and migrant laborers. The principal sources of income are agriculture, tourism, and light industry. It is approximately 60% urban.

You will have a generous budget, with computing facilities available and personnel to staff and operate any system that you design. The Planning Agency has liaison with all county and city government agencies, hospitals, the local chapter of the Medical Society, and other voluntary agencies. You can secure information from any of these sources.

The Health Information System will gather, tabulate, analyze, and interpret data as well as monitor community health, evaluate programs, and define unmet needs.

a) Describe the type of system and data needed to accomplish the tasks, using examples to illustrate the general considerations.

b) What statistical techniques or studies will be used to gather and analyze the various types of needed information? Again, examples may be employed to suggest specific illustrations.

c) Discuss the type of quality checks necessary to test adequacy of the data. Also outline the phases that the system will go through, distinguishing between the information needed for initial planning and what will be gathered as the planning agency becomes more sophisticated in its needs and operation.
(1969 RE-EXAMINATION)

MASTER'S EXAMINATION

1. According to binomial theory, if we take two fair coins and toss them 100 times each, we should observe 0 heads 25 times, 1 head 50 times, and 2 heads 25 times. That is, the expected outcomes should be:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 heads</td>
<td>T,T</td>
</tr>
<tr>
<td>1 head</td>
<td>T,H or H,T</td>
</tr>
<tr>
<td>2 heads</td>
<td>H,H</td>
</tr>
</tbody>
</table>

In a trial, the following results were obtained:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 heads</td>
<td>20</td>
</tr>
<tr>
<td>1 head</td>
<td>40</td>
</tr>
<tr>
<td>2 heads</td>
<td>40</td>
</tr>
</tbody>
</table>

a) Test to see if these results agree with binomial theory. Use $\alpha = 0.05$.

b) Is the test you use a test of independence, homogeneity or goodness-of-fit?

2. Race-specific perinatal mortality rates* for North Carolina are computed annually for each county in the state. The Maternal and Child Health Director wishes to set up some scheme whereby he can determine whether the annual race-specific rates for each county are "normal", "too high", or "too low". His ultimate objective is to identify counties with rates "too low" and then attempt to find out why they are so low. Similarly, he wants to identify counties with rates "too high", find out why, and take remedial action to lower the rates.

What scheme would you recommend to him for making the determinations of "normal", "too low", and "too high" perinatal rates? If there are limitations in your scheme which the MCH Director should know about, specify what they are. Also, specify any other precautions (if any) which he should be aware of in using your scheme for his intended purpose.

In 1967 there were 1841 white and 1477 nonwhite perinatal deaths in N. C. The rates were: Total State 35.2; white 28.3; nonwhite 50.6.

*Perinatal mortality rate = $\frac{\text{Fetal deaths 20 wks. gestation or over plus deaths to live born infants occurring under 28 days}}{\text{all live births}} \times 1000$
EDITORIAL NOTE: A table was included which gave the natural logarithms of the integers from 2 to 22.

1. a) Massachusetts and Mississippi each claim a lower tuberculosis death rate than the other in 1950. Using the following data, show how both claims can be substantiated, and explain the paradox.

<table>
<thead>
<tr>
<th>State</th>
<th>Total</th>
<th>White</th>
<th>Non-white</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massachusetts</td>
<td>4,690,514</td>
<td>4,611,503</td>
<td>79,011</td>
</tr>
<tr>
<td>TB deaths *</td>
<td>1,005</td>
<td>942</td>
<td>63</td>
</tr>
<tr>
<td>Mississippi</td>
<td>2,178,914</td>
<td>1,188,632</td>
<td>990,282</td>
</tr>
<tr>
<td>TB deaths *</td>
<td>573</td>
<td>191</td>
<td>382</td>
</tr>
</tbody>
</table>

* Deaths from all forms of tuberculosis

b) Given that the United States had a population of 151,325,798 in 1950, of which 89.31% was White, calculate suitably standardized tuberculosis mortality rates for Massachusetts and Mississippi.

c) Discuss how it might be decided whether the rates for these two states are significantly different.

(But do not actually carry out the calculations for any statistical test.)
2.

a) List the major functional components of a computer.

b) List two sources or packages of statistical programs.

c) If you were about to store a large data set, what features of the data set would determine whether to store it on cards, disk, or tape?

d) What is the essential characteristic of a sequential data set? Give an example of (i) a sequential data set; (ii) a data set with some organization other than sequential (machine readable or not).

d) What is the purpose of having physical records which contain more than one logical record? Elaborate.

f) In the following example:

(i) Which field is major?
(ii) Which field is minor?
(iii) Is field 1 ascending or descending?
(iv) Is field 2 ascending or descending?
(v) Is field 3 ascending or descending?

<table>
<thead>
<tr>
<th>Field:</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>(zero)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
An opinion poll showed that, of 3600 people interviewed, 2304 were in favor of allowing birth control information to be made available to single women.

Note: \( \frac{2304}{3600} = .64 \)

(a) Find an approximate two-sided 95% confidence interval for \( p \) (the proportion in the population who are in favor).

(b) Test the hypothesis

\[
H_0 : p = .7 \quad \text{against} \quad H_A : p \neq .7 \quad \text{with} \quad \alpha = .05
\]

(c) What is the power of this test if \( p = .65 \)?

4. One of the so-called Pascal distributions has the probability function

\[
f(x) = P(X = x) = \frac{n^x}{(1+n)^{x+1}}, \quad n \geq 0, \quad x = 0, 1, 2, \ldots
\]

(a) Find the maximum likelihood estimate of \( \theta \) parameter \( \theta \), given a random sample of size \( n \) from a population having the above distribution.

(b) Given a random sample as in (a), derive the form of the critical region for the likelihood ratio test of the hypothesis \( H_0 : \theta = \theta_0 \) against the alternative \( H_A : \theta \neq \theta_0 \).

(c) Using the fact that, for large \( n \), the distribution of \(-2(n)\), where \( \lambda \) is the likelihood ratio statistic, is approximately the chi-square distribution with 1 degree of freedom, apply the test (b) to testing \( H : \theta = 2 \) versus \( H_A : \theta \neq 2 \), at the 5% level of significance, given a random sample of 30 observations yielding sample mean 1.

5. In the United States, the vital registration system and the decennial Census provide data essential in many areas. Evaluate this "dual system" from the point of view of providing demographic statistics needed to adequately measure population change. Suggest remedies for any deficiencies you find.
6. **A school mental health consultation program is to be evaluated at a public elementary school. The purpose of the consultation program is to increase the level of mental health of each school child and/or decrease the level of mental malfunction. This is to be accomplished by offering consultation to the classroom teacher so that she or he can handle mental health problems which occur in the classroom as well as promote positive mental health in the classroom. Assume for this design that the teacher can, with consultation, handle all children in the classroom and does not need to refer any children for more extensive, individualized aid.**

Four "equivalent" classrooms in the school are chosen. Two of the classrooms are given a pre-test which measures the level of mental health of every child in the classroom. One of these classrooms gets the consultation program; the other classroom does not. The other two classrooms do not receive the pre-test. One of these classrooms receives the consultation program, and the other classroom does not. All four classrooms receive the post-test at the end of the school year. (The post-test is the same test as the pre-test.) These classrooms are equivalent in the sense that if all four classrooms were pre-tested, the average level of mental health would be the same for all four classrooms.

This design can be represented as

<table>
<thead>
<tr>
<th>Classroom</th>
<th>Pre-Test</th>
<th>Consultation Program</th>
<th>Post-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom 1</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Classroom 2</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Classroom 3</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Classroom 4</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

The first three questions below are statistical "design" questions. In answering these questions, simply state which classrooms you would use and in what manner (e.g. pre-test scores on Classroom X compared to difference scores between pre-test and post-test on Classroom Y). Do not specify statistical analyses such as analysis of variance, contingency tables, etc.

a) **How would you test for the effectiveness of the consultation program?**

b) **How would you test for the possibility that the pre-test affects the post-test in the absence of the consultation program?**

c) **How would you test for the possibility that the pre-test affects the post-test in the presence of the consultation program, i.e. that the pre-test and the consultation program interact?**

d) **In this design classrooms are confounded with teachers since the classrooms in the school are self-contained. Assuming that you can find four classrooms that are equivalent, what variables would you like to include to assess whether or not the four teachers are equivalent for the purposes of this program evaluation?**

e) **Comment on the use of the same test for pre-test and post-test measures.**
1. (15 points) a) The following data are from "Increases in Divorces - United States - 1967," Series 21, Number 20 of Vital and Health Statistics, published by National Center for Health Statistics, December 1970. The table below gives the age specific death rates for married and single males in the U. S. in the period 1959-1961. (Data for divorced and widowed males are available in the publication but not included here.)

Death Rates per 1000 population, by Age and Marital Status: U. S. Males, 1959-61

<table>
<thead>
<tr>
<th>Age in Years</th>
<th>Married</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total, 15+</td>
<td>12.8</td>
<td>7.5</td>
</tr>
<tr>
<td>15-19</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>20-24</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td>25-29</td>
<td>1.3</td>
<td>2.8</td>
</tr>
<tr>
<td>30-34</td>
<td>1.6</td>
<td>4.1</td>
</tr>
<tr>
<td>35-39</td>
<td>2.3</td>
<td>5.9</td>
</tr>
<tr>
<td>40-44</td>
<td>3.8</td>
<td>8.8</td>
</tr>
<tr>
<td>45-49</td>
<td>6.3</td>
<td>12.5</td>
</tr>
<tr>
<td>50-54</td>
<td>10.6</td>
<td>18.1</td>
</tr>
<tr>
<td>55-59</td>
<td>15.9</td>
<td>23.5</td>
</tr>
<tr>
<td>60-64</td>
<td>25.0</td>
<td>37.4</td>
</tr>
<tr>
<td>65-69</td>
<td>36.6</td>
<td>53.2</td>
</tr>
<tr>
<td>70-74</td>
<td>52.4</td>
<td>74.9</td>
</tr>
<tr>
<td>75-79</td>
<td>76.8</td>
<td>105.0</td>
</tr>
<tr>
<td>80-84</td>
<td>121.3</td>
<td>158.7</td>
</tr>
<tr>
<td>85+</td>
<td>191.1</td>
<td>231.6</td>
</tr>
</tbody>
</table>

Note that, for any given age category, the death rate for married males never exceeds the death rate for single males. However, the overall death rate for married males (12.8) is 70% larger than the overall death rate for single males (7.5).

Explain mathematically how this can happen, and give the particular reason why it happens in this set of data.
(Question 1, continued)

(10 points) b) In an article on patients with coronary occlusion, the following data were presented for 330 patients who had had pre-existing anginal symptoms.

<table>
<thead>
<tr>
<th>Time</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day</td>
<td>4</td>
</tr>
<tr>
<td>2-6 days</td>
<td>12</td>
</tr>
<tr>
<td>1-4 weeks</td>
<td>20</td>
</tr>
<tr>
<td>1-2 months</td>
<td>28</td>
</tr>
<tr>
<td>3-6 months</td>
<td>54</td>
</tr>
<tr>
<td>7-12 months</td>
<td>45</td>
</tr>
<tr>
<td>1-2 years</td>
<td>50</td>
</tr>
<tr>
<td>2-3 years</td>
<td>23</td>
</tr>
<tr>
<td>3-5 years</td>
<td>38</td>
</tr>
<tr>
<td>5-10 years</td>
<td>36</td>
</tr>
<tr>
<td>10+ years</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>330</strong></td>
</tr>
</tbody>
</table>

Comment on the author's remark: "Occlusion occurs most frequently between 3 months and 2 years after onset of anginal symptoms".
2.

(6 points)  a) List the three most important types of statements in JCL and briefly explain the purpose of each (what does it do?).

(4 points)  b) Briefly describe a "standard data matrix." Illustrate with a figure.

(9 points)  c) Give a short description or definition of each of the following terms as used in data processing.

   (i) byte
   (ii) field
   (iii) logical record
   (iv) physical record (what is the distinction between logical and physical records?)

(6 points)  d) (i) For a sequential data set to be written on magnetic tape why would one generally prefer a blocksize of 8000 bytes to a blocksize of 80 bytes.

   (ii) What is the principal advantage of shorter blocksizes?
3. Suppose you have designed a before-after experiment with control group. In other words, you have matched two groups, pair by pair, and have before and after measures on both groups. Make use of the t test to test for the effectiveness of your experimental variable

(5 points) a) using only the "after" scores, ignoring the "before" scores,

(5 points) b) using the "before" and "after" scores of the experimental group only, and

(10 points) c) using all four sets of scores.

(5 points) Contrast the advantages and disadvantages of methods (a) and (b). What are the advantages of (c) over both (a) and (b)?

<table>
<thead>
<tr>
<th>Pair</th>
<th>Control group</th>
<th>Experimental group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>A</td>
<td>72</td>
<td>75</td>
</tr>
<tr>
<td>B</td>
<td>61</td>
<td>60</td>
</tr>
<tr>
<td>C</td>
<td>48</td>
<td>37</td>
</tr>
<tr>
<td>D</td>
<td>55</td>
<td>64</td>
</tr>
<tr>
<td>E</td>
<td>81</td>
<td>76</td>
</tr>
<tr>
<td>F</td>
<td>50</td>
<td>59</td>
</tr>
<tr>
<td>G</td>
<td>42</td>
<td>49</td>
</tr>
<tr>
<td>H</td>
<td>64</td>
<td>55</td>
</tr>
<tr>
<td>I</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>J</td>
<td>69</td>
<td>78</td>
</tr>
</tbody>
</table>

Note. Do not perform any computations but indicate what needs to be calculated for each of (a), (b), and (c).
4. In a certain probability situation, the probability of the occurrence of either event A or event B (or both) is 0.7, the conditional probability of A given B is 2/3, and the conditional probability of B given A is 4/5. Find the probability of each of the following events:

(6 points) a) A occurs,

(6 points) b) B occurs,

(7 points) c) both A and B occur,

(6 points) d) neither A nor B occurs.

EDITORIAL NOTE: Question #5 begins on the next page

6. Assume that for a population of animals you obtain the following mortality statistics per 1000. All animals who die, do so at the anniversary of their birth.

<table>
<thead>
<tr>
<th>Age (x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths at age x</td>
<td>0</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
</tr>
</tbody>
</table>

Calculate

(10 points) a) The life expectancy.

(5 points) b) The mean age of the stationary population.

(5 points) c) The probability that an animal approaching its second birthday will survive until age 4.

(5 points) d) The life table death rate.
5. "Q-Sort Methodology" is used in the social sciences to obtain from a person a description of himself or some other person. For example a "Q-Sort deck" may contain N=94 cards, where each card has a statement about a child's behavior such as "My child has temper tantrums almost every day." The child's parents, independently of one another, are then asked to "sort" these 94 statements as they pertain to their child. The 94 cards are to be sorted into the following categories, where the number of cards which end up in each category is fixed in advance.

<table>
<thead>
<tr>
<th>Category and Score for that Statement</th>
<th>Description of Category</th>
<th>Number of Cards in this Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Highly uncharacteristic of my child</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Quite uncharacteristic of my child</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Somewhat uncharacteristic of my child</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>Neither characteristic nor uncharacteristic of my child</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>Somewhat characteristic of my child</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>Quite characteristic of my child</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>Highly characteristic of my child</td>
<td>5</td>
</tr>
</tbody>
</table>

Hence, each parent will finish the Q-sort by identifying exactly 5 statements out of the 94 that are highly characteristic of the child, exactly 10 statements that are quite characteristic of the child, etc.
The "score" for statement \( j, j=1,2,\ldots,94 \), is determined by the category into which the parent puts statement \( j \). Let \( x_j \) and \( y_j \), \( j=1,2,\ldots,94 \) be the assigned score for statement \( j \) by the mother and father, respectively. Thus, the response of each parent can be described by a multivariate observation vector of 94 elements, where each element assumes the value of some integer in the set \{1,2,3,4,5,6,7\} - with the restriction that the frequency distribution of the 94 elements follows the specification in TABLE 1.

(5 points) a) Show that the average score is the same for each parent. That is, show that
\[
\bar{x} = \bar{y},
\]
where
\[
\bar{x} = \frac{1}{94} \sum_{j=1}^{94} x_j \quad \text{and} \quad \bar{y} = \frac{1}{94} \sum_{j=1}^{94} y_j.
\]

(5 points) b) Show that the variability of the scores for each parent is the same. That is, show that \( s_x^2 = s_y^2 \) where
\[
s_x^2 = \frac{1}{94} \sum_{j=1}^{94} (x_j - \bar{x})^2
\]
and
\[
s_y^2 = \frac{1}{94} \sum_{j=1}^{94} (y_j - \bar{y})^2.
\]
(15 points) c) One purpose of having both parents do the Q-sort is to compare the two Q sorts to find out if both parents view their child in similar or different ways. Two ways of measuring the similarity between the 2 Q sorts are: the correlation coefficient $r_{xy}$ and the sum of squared differences $D^2$. The formulas for these measures are:

$$r_{xy} = \frac{1}{N} \sum_{j=1}^{N} (x_j - \bar{x})(y_j - \bar{y})$$

$$D^2 = \sum_{j=1}^{N} (x_j - y_j)^2.$$ 

In this example, $N=94$.

Show that these two similarity measures are mathematically related by the equation

$$r = 1 - \frac{D^2}{2Ns^2}, \text{ for } N=94,$$

where $s^2 = s_x^2 = s_y^2$.

EDITORIAL NOTE: Question #6 is printed between Questions #4 and #5.
1. (15 points)  

The following data are from "Increases in Divorces - United States - 1967," Series 21, Number 20 of Vital and Health Statistics, published by National Center for Health Statistics, December 1970. The table below gives the age specific death rates for married and single males in the U. S. in the period 1959-1961. (Data for divorced and widowed males are available in the publication but not included here.)

Death Rates per 1000 Population, by Age and Marital Status: U. S. Males, 1959-61

<table>
<thead>
<tr>
<th>Age in Years</th>
<th>Married</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total, 15+</td>
<td>12.8</td>
<td>7.5</td>
</tr>
<tr>
<td>15-19</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>20-24</td>
<td>1.3</td>
<td>2.2</td>
</tr>
<tr>
<td>25-29</td>
<td>1.3</td>
<td>2.8</td>
</tr>
<tr>
<td>30-34</td>
<td>1.6</td>
<td>4.1</td>
</tr>
<tr>
<td>35-39</td>
<td>2.3</td>
<td>5.9</td>
</tr>
<tr>
<td>40-44</td>
<td>3.8</td>
<td>8.8</td>
</tr>
<tr>
<td>45-49</td>
<td>6.3</td>
<td>12.5</td>
</tr>
<tr>
<td>50-54</td>
<td>10.6</td>
<td>18.1</td>
</tr>
<tr>
<td>55-59</td>
<td>15.9</td>
<td>23.5</td>
</tr>
<tr>
<td>60-64</td>
<td>25.0</td>
<td>37.4</td>
</tr>
<tr>
<td>65-69</td>
<td>36.6</td>
<td>53.2</td>
</tr>
<tr>
<td>70-74</td>
<td>52.4</td>
<td>74.9</td>
</tr>
<tr>
<td>75-79</td>
<td>76.8</td>
<td>105.0</td>
</tr>
<tr>
<td>80-84</td>
<td>121.3</td>
<td>158.7</td>
</tr>
<tr>
<td>85+</td>
<td>191.1</td>
<td>231.6</td>
</tr>
</tbody>
</table>

Note that, for any given age category, the death rate for married males never exceeds the death rate for single males. However, the overall death rate for married males (12.8) is 70% larger than the overall death rate for single males (7.5).
(Question 1, continued)

Explain mathematically how this can happen, and give the particular reason why it happens in this set of data.

(10 points) b) In an article on patients with coronary occlusion, the following data were presented for 330 patients who had had pre-existing anginal symptoms.

Distribution of time between onset of symptoms and first attack of coronary occlusion

<table>
<thead>
<tr>
<th>Time</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day</td>
<td>4</td>
</tr>
<tr>
<td>2-6 days</td>
<td>12</td>
</tr>
<tr>
<td>1-4 weeks</td>
<td>20</td>
</tr>
<tr>
<td>1-2 months</td>
<td>28</td>
</tr>
<tr>
<td>3-6 months</td>
<td>54</td>
</tr>
<tr>
<td>7-12 months</td>
<td>45</td>
</tr>
<tr>
<td>1-2 years</td>
<td>50</td>
</tr>
<tr>
<td>2-3 years</td>
<td>23</td>
</tr>
<tr>
<td>3-5 years</td>
<td>38</td>
</tr>
<tr>
<td>5-10 years</td>
<td>36</td>
</tr>
<tr>
<td>10+ years</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>330</strong></td>
</tr>
</tbody>
</table>

Comment on the author's remark: "Occlusion occurs most frequently between 3 months and 2 years after onset of anginal symptoms".
2. Suppose you have designed a before-after experiment with control group. In other words, you have matched two groups, pair by pair, and have before and after measures on both groups. Make use of the t test to test for the effectiveness of your experimental variable.

(5 points) a) using only the "after" scores, ignoring the "before" scores,
(5 points) b) using the "before" and "after" scores of the experimental group only, and
(10 points) c) using all four sets of scores

(5 points) Contrast the advantages and disadvantages of methods (a) and (b). What are the advantages of (c) over both (a) and (b)?

<table>
<thead>
<tr>
<th>Pair</th>
<th>Control group</th>
<th>Experimental group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>A</td>
<td>72</td>
<td>75</td>
</tr>
<tr>
<td>B</td>
<td>61</td>
<td>60</td>
</tr>
<tr>
<td>C</td>
<td>48</td>
<td>37</td>
</tr>
<tr>
<td>D</td>
<td>55</td>
<td>64</td>
</tr>
<tr>
<td>E</td>
<td>81</td>
<td>76</td>
</tr>
<tr>
<td>F</td>
<td>50</td>
<td>59</td>
</tr>
<tr>
<td>G</td>
<td>42</td>
<td>49</td>
</tr>
<tr>
<td>H</td>
<td>64</td>
<td>55</td>
</tr>
<tr>
<td>I</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>J</td>
<td>69</td>
<td>78</td>
</tr>
</tbody>
</table>

Note. Do not perform any computations but indicate what needs to be calculated for each of (a), (b), and (c).
3. Assume that for a population of animals you obtain the following mortality statistics per 1000. All animals who die, do so at the anniversary of their birth.

<table>
<thead>
<tr>
<th>Age (x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths at age x</td>
<td>0</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
</tr>
</tbody>
</table>

Calculate

(10 points) a) The life expectancy.

(5 points) b) The mean age of the stationary population.

(5 points) c) The probability that an animal approaching its second birthday will survive until age 4.

(5 points) d) The life table death rate.
1. The following table gives the number of live births, infant deaths and infant death rates for the years 1965-1969 for Barry County, Michigan.

<table>
<thead>
<tr>
<th>Year</th>
<th>1965</th>
<th>66</th>
<th>67</th>
<th>68</th>
<th>69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live Births</td>
<td>594</td>
<td>595</td>
<td>626</td>
<td>626</td>
<td>626</td>
</tr>
<tr>
<td>Infant Deaths</td>
<td>14</td>
<td>17</td>
<td>5</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Infant Death Rate</td>
<td>23.6</td>
<td>28.6</td>
<td>8.0</td>
<td>14.4</td>
<td>24.0</td>
</tr>
</tbody>
</table>

(Per 1000)

a) Comment briefly (not more than 5-10 lines) on the variability observed in these rates.

b) Find a 95% confidence interval for the 1969 infant death rate.

c) If one assumes that the underlying infant death rate is constant over the period covered by this table, find a shorter 95% confidence interval for the infant death rate.
2. A certain manufacturing company claims that the average lifetime of its electric lights is 750 hours. A bulb is defective if it lasts less than 500 hours.

a) In a random sample of 30 light bulbs, three were observed to be defective. Is this sufficient evidence to dispute the manufacturer's claim that no more than 5% of the bulbs are defective?

b) In this same sample of 30, twelve bulbs burned longer than 750 hours. Under suitable assumptions (which you will state in your answer) test whether these data support the claim of an average life of 750 hours.

3. Let the capacity C of an environmental engineering system and the demand D upon it be random variables. The quantity \( P = k(C - D) \), where \( k \) is a known positive constant, is a measure of the performance of the system, whether it be "inadequate" (\( P < 0 \)) or "over-designed" (\( P \) large and positive).

a) Find the mean and variance of \( P \) in terms of the means and variances of \( C \) and \( D \) and the correlation coefficient \( \rho \) between \( C \) and \( D \).

b) Express the probability that the performance of a system is "inadequate" in terms of a probability statement concerning \( C \) and \( D \).

c) For a particular system, suppose that \( C \) and \( D \) are normal random variables with \( E(C) = 12, \ Var(C) = 9, E(D) = 2, Var(D) = 16 \), and \( \rho = 0 \). Find the probability of "inadequate" performance for this system.

4. Discuss the following quotation from A. J. Coale:

"We find at the end, then, that although the birth rate determines how old a population is, the death rate determines what the average birth rate in the long run must be. If prolonged life produces by its direct effects a younger population, it is nevertheless compatible only with an older population."
5. Let $X_{ik}$, $k=1,2,\ldots,K$ denote $K$ test scores for person $i$, $i=1,\ldots,N$.

We can define a distance $D_{ij}$ between person $i$ and person $j$ by the formula

$$D_{ij}^2 = \sum_{k=1}^{K} (X_{ik} - X_{jk})^2.$$ 

In some instances, it may be desirable to first remove each individual's elevation from his scores before computing the distance between individuals. That is, $X_{ik}$ is replaced by $X'_{ik} = X_{ik} - \bar{X}_i$, where

$$\bar{X}_i = \frac{1}{K} \sum_{k=1}^{K} X_{ik}.$$ 

The distance formula then becomes

$$D_{ij}'^2 = \sum_{k=1}^{K} (X'_{ik} - X'_{jk})^2.$$ 

In other instances, it may be desirable to remove both an individual's elevation and also his scatter before computing distances.

That is, $X_{ik}$ is replaced by $X''_{ik} = \frac{X_{ik}}{S_i}$, where

$$S_i = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (X_{ik} - \bar{X}_i)^2}.$$ 

and distances are computed as

$$D_{ij}''^2 = \sum_{k=1}^{K} (X''_{ik} - X''_{jk})^2.$$ 

Show that the following relationships hold among the different distance measures:

a) $D_{ij}'^2 = D_{ij}^2 - K(\bar{X}_i - \bar{X}_j)^2$ 

b) $D_{ij}''^2 = D_{ij}^2 - (S_i - S_j)^2$ 

$= \frac{1}{S_i S_j}$
6. This question pertains to the job setup which follows.

a) 1. What is the jobname?
   2. How much core storage will be allocated to this job?
   3. Name and briefly describe the programs or procedures used in this job.

b) In STEP1
   4. What is the DDNAME of the output data set?
   5. What does the '*' in the //SYSUT1 DD * card indicate?

c) In STEP2
   6. What does the 'A' in the //SYSOUT DD card refer to?
   7. What does '*.STEP1.SYSUT2' do?
   8. What does 'DISP=(OLD,PASS)' mean?
   9. What is the purpose of the cards 'SORTWK01' to 'SORTWK06'?
  10. What are the 'SORT' and 'END' cards called?
  11. What is the maximum number of records that can be processed?

d) In STEP3
  12. Where is the input data?
  13. In which positions of the record would variable 'Y' be found?
  14. How many subfiles are there and how many cases are in each?
  15. What SPSS procedure(s) is being used?
//BASIC JOB UNC.B.E1234,'B-JONES',MSGLEVEL=1,REGION=100K T=1,P=50
//STEP1 EXEC PGM=IEBGENER
//SYSPRINT DD SYSOUT=A
//SYSIN DD DUMMY
//SYSUT2 DD DSN=UNC.B.E1234.JONES.DATA,DISP=(OLD,PASS),UNIT=DISK
//SYSUT1 DD *

10014 3105750 1105 45.9
10021 893428 875 26.6
10033 750902 600 31.4
10042 1565414 1004 35.2
10053 2043917 1235 40.0
10061 909100 910 10.7
10073 1998433 1000 30.8
10083 2505488 2000 41.6
10092 1703706 1575 33.3
10102 987541 1010 15.5
/*
//STEP2 EXEC PGM=SORT
//SYSPRINT DD SYSOUT=A
//SORTLIB DD DSN=SYS1.SORTLIB,DISP=SHR
//SORTIN DD DSN=*..STEP1.SYSUT2,DISP=OLD,UNIT=DISK
//SORTOUT DD DSN=UNC.B.E1234.JONES.SORTDATA,DISP=(OLD,PASS),UNIT=DISK
//SORTWK01 DD UNIT=DISK,SPACE=(TRK,25,,CONTIG)
//SORTWK02 DD UNIT=DISK,SPACE=(TRK,25,,CONTIG)
//SORTWK03 DD UNIT=DISK,SPACE=(TRK,25,,CONTIG)
//SORTWK04 DD UNIT=DISK,SPACE=(TRK,25,,CONTIG)
//SORTWK05 DD UNIT=DISK,SPACE=(TRK,25,,CONTIG)
//SORTWK06 DD UNIT=DISK,SPACE=(TRK,25,,CONTIG)
//SYSIN DD *
SORT FIELDS=(5,1,CH,A),SIZE=10
END
/

//STEP3 EXEC SPSS

//G.TAPEIN DD DSN=*.STEP2.SORTOUT,DISP=OLD,UNIT=DISK

//G.SYSIN DD *
COLUMN 16
\X,Y,Z

VARIABLE LIST

SUBFILE LIST

INPUT MEDIUM

# OF CASES

INPUT FORMAT

VAR LABELS

PROCESS SBFILES

CONDESCRIPTIVE

OPTIONS

STATISTICS

READ INPUT DATA

PROCESS SBFILES

CONDESCRIPTIVE

STATISTICS

FINISH

END
/
//
1. a) State the assumptions of analysis of variance
   b) Complete the following ANOVA table and answer the questions given after the table.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>4</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>102</td>
<td>15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(i) What are your conclusions?
(ii) How many treatments were there?
(iii) Were there equal numbers of observations in each treatment?

2. A physician observes the systolic blood pressure (SBP) of a patient and makes further tests if SBP > 160. Assume that in healthy people SBP is normally distributed with \( \mu = 120 \) and \( \sigma = 20 \). In diseased persons it normally distributed with \( \mu = 190 \) and \( \sigma = 30 \).

a) What is the probability that the physician makes further tests on a healthy person?

b) What is the probability that a diseased person is not detected by this procedure?

c) If 10% of the population is diseased, what is the probability that a person detected by this procedure is actually diseased?
3. Consider a population of gains in weights of swine during a 20 day period. These weight gains approximate a normal distribution with \( \mu = 30 \) and \( \sigma^2 = 100 \). Random samples of size \( n = 10 \) were repeatedly drawn (511 times) from this population and for each sample the sample variance, \( s^2 \), was computed. The following is a tabulation of these sample variances.

<table>
<thead>
<tr>
<th>class boundaries</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-30</td>
<td>12</td>
<td>0.023</td>
</tr>
<tr>
<td>30-50</td>
<td>47</td>
<td>0.092</td>
</tr>
<tr>
<td>50-70</td>
<td>92</td>
<td>0.180</td>
</tr>
<tr>
<td>70-90</td>
<td>93</td>
<td>0.182</td>
</tr>
<tr>
<td>90-110</td>
<td>72</td>
<td>0.141</td>
</tr>
<tr>
<td>110-130</td>
<td>73</td>
<td>0.143</td>
</tr>
<tr>
<td>130-150</td>
<td>42</td>
<td>0.082</td>
</tr>
<tr>
<td>150-170</td>
<td>29</td>
<td>0.057</td>
</tr>
<tr>
<td>170-190</td>
<td>26</td>
<td>0.051</td>
</tr>
</tbody>
</table>

- Draw a histogram of the sample variances.
- Draw a cumulative relative frequency polygon.
- Using the graph of the cumulative distribution, estimate the 10th, 50th, and 95th percentiles of this distribution.
- What is the theoretical 95th percentile of the distribution of \( s^2 \) and how does it compare with the estimate in part c.
4. In a sequence of four births, find
   a) The probability that there are two boys and two girls.
   b) The probability that at least two of the children will be girls.
   c) The probability that none of the children will be girls.
   d) The probability of exactly two boys given there is at least one boy.

Assume the probability of a girl is 1/2.
EXAMINATION FOR MPH AND MSPH CANDIDATES
(9-12 A.M., SATURDAY, MARCH 31, 1973)

1. Examination of 80 men revealed 40 with hypertension and 30 with elevated cholesterol, including 10 having both ailments. If a man is picked at random from the group, what is the probability that he is free of both ailments? What is the probability that he has both ailments if he has one of them?

2. In evaluating mental health programs, some of the following research designs might be used. Here, an O represents the process of observation or measurement, an X represents the exposure of a group of subjects to the experimental variable or "treatment", X's and O's on the same line refer to the same specific subjects, while parallel rows represent different groups of subjects. Temporal order is indicated by the left-to-right direction.

Discuss each design in terms of how it controls or fails to control for various sources of invalidity. Discuss where and for what purposes random assignment can be utilized.

1) X 0
2) O₁ X O₂
3) X O₁ O₂
4) O₁ X O₂ O₃ O₄
5) O₁ X O₂ O₃ O₄ X O₅ O₆
3. Under constant environmental conditions, the number of bacteria Q in a tank is known to increase proportionately to $e^{\lambda T}$, where T is the time and the parameter $\lambda$ is the growth rate (0<$\lambda$<1). If $k$ is the number of bacteria in the tank at time $T=0$, then

$$Q = ke^{\lambda T}, \quad Q \geq k.$$ 

a) If the time permitted for growth (as determined, say, by the time required for the charge of water in which the bacteria are living to pass through a filter) is a random variable with cumulative distribution function $F_T(t), \ t > 0$, find that particular t (say, $t^*$), which is a function of $\lambda$, k, and q, such that

$$\Pr(Q \leq q) = F_T(t^*).$$

b) If $F_T(t) = 1-e^{-t}$, $t > 0$, find $f_Q(q)$, the probability density function of Q.

c) For $F_T(t)$ as in (b), find $E(Q)$. [HINT: You do not need to know the answer to (b).]

d) How is $E(Q)$ related to $E(T^r)$?
4. This question concerns the job setup which follows.

a) To what account is the cost of this job charged?

b) How much time is allowed for the job?

c) How many "steps" does the job have?

d) In general, what is a "ddname"? (Check the best answer)
   - a name which is stored in every record of a data set
   - the symbolic name by which a program refers to a data set
   - the actual name of a particular data set
   - the name of the programmer who created the data set

e) In general, what is a "dsname"? (Check the best answer.)
   - a name which is stored in every record of a data set
   - the symbolic name by which a program refers to a data set
   - the actual name of a particular data set
   - the name of the programmer who created the data set

f) What do the ampersands ("&") mean on card #4?

g) What does the asterisk ("*") mean on card #24?

h) What is the function of card #8?

i) Does card #2 specify a program or a procedure to be executed? (Indicate which.)

j) How many logical records are in the data set described by card #14?

k) How many logical records are in each physical record (block) of the data set described on cards #4 and #5?

l) In general, what does it mean to "sort" a data set?

m) With respect to sorting, what is:
   - a "major" field?
   - a "minor" field?

n) Does a data set necessarily need to be in sorted order before it can be updated by a program such as the Biostatistics UPDATE II program? (Yes or No)
o) What is the function of the "MISSING VALUES" card (#30)?

p) What is the function of the "RECODE" cards (#31-33)?

q) If the "INPUT MEDIUM" card (#27) specified "CARD" (instead of "DISK"), where (at what point in the job deck) would the data for analysis be located?

r) What task (or analysis) does the "MARGINALS" card (#35) specify?

s) What task (or analysis) does the "CROSSTABS" card (#39) specify?

t) Name another statistical package (other than SPSS).
1//EXSAMF JOB UNCB.X566K,'SORANT.A-031024',M=1,T=(,59)
2//SRT EXEC DASORT
3//SORTIN DD DSN=UNCB.X566K.SORANT.UPDT,DISP=OLD
4//SORTOUT DD DSN=&&SORTED,DISP=(NEW,PASS),SPACE=(6400,(2,1)),*
5// UNIT=DISK,DCB=(RECFM=FB,LRECL=80,BLKSIZE=6400)
6//SYSIN DD *
7 SORT FIELDS=(11,1,CH,D,73,8,CH,A)
8/
9//UPD EXEC PGM=UPDATE2,REGION=110K
10//STEPLIB DD DSN=UNCB.BIOS.LIB,DISP=SHR
11//SYSPRINT DD SYSCUT=A
12//MASTER DD DSN=UNCB.X566K.SORANT.OLD,DISP=OLD
13//DETAIL DD DSN=&&SORTED,DISP=(OLD,DELETE)
14//SYSIN DD *
15CONTROL FIELD SPECS
16C 11 1 -1
17C 73 8 +1
18END CONTROL FIELD SPECS
19/*
20//REPORT DD DSN=UNCB.X566K.SORANT.NEW,DISP=(OLD,PASS),*
21/ DCB=(RECFM=FB,LRECL=80,BLKSIZE=6400)
22//TAB EXEC SPSS
23//TAPEIN DD DSN=*.UPD.REPORT,DISP=OLD
24//SYSIN DD *
25RUN NAME SAMPLE TABULATION OF EDUCATION, INCOME
26VARIABLE LIST GRP, ED, INC
27INPUT MEDIUM DISK
28# OF CASES ESTIMATED 200
29INPUT FORMAT FIXED (T11,F1.0,T21,F2.0,T31,F6.0)
30MISSING VALUES ED, INC (0)
31RECODE INC (1 THRU 4000 = 1) (4001 THRU 8000 = 2)
32 (8001 THRU 12000 = 3) (12001 THRU 16000 = 4)
33 (16001 THRU HIGHEST = 5)
34SELECT IF (GRP EQ 3)
35MARGINALS ED, INC
36OPTIONS 1
37STATISTICS ALL
38READ INPUT DATA
39CROSSTABS ED BY INC
40FINISH
41/*
42//
5. The Secretary of Health in country Z ordered his statistician to determine whether or not there was any evidence that ethnic group A had different size families than ethnic group B. The statistician selected an independent random sample of 25 and 30 year old women from the two groups and determined how many children each woman had. Means and standard errors were calculated and are shown in the following table:

<table>
<thead>
<tr>
<th>Statistic and Group</th>
<th>Age (years)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>Number of Women:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>B</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Mean Number of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.8</td>
<td>3.6</td>
</tr>
<tr>
<td>B</td>
<td>1.8</td>
<td>3.6</td>
</tr>
<tr>
<td>Standard Error of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the Mean:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.070</td>
<td>0.072</td>
</tr>
<tr>
<td>B</td>
<td>0.050</td>
<td>0.100</td>
</tr>
</tbody>
</table>

The statistician wrote a memorandum to the Secretary stating that:

"The mean number of children per woman is significantly higher in group A than in group B. The 95% confidence interval for the difference is 0.6 ± 0.2."

a) Do you agree with the statistician's conclusion? Explain your answer.
b) Briefly discuss the possibility that including older women in the study might change the results.
6. A highway-safety research investigator is interested in testing the acclaimed success of "air-bags" in the reduction of risk of death and serious injury to the driver in highway crashes. He feels that the air-bags would be most effective in head-on collisions. Suppose that the risk of a vehicle being involved in a crash during a year is 1/9. The chance of the crash being head-on is about 1/4 and the risk of death or serious injury to the driver in such a crash without air-bags is 1/4.

a) The number $X$ of deaths and serious injuries from head-on collisions during a year in a fleet of $N$ vehicles is assumed to be a Poisson random variable. Please comment on the basic philosophy underlying such an assumption. Also, under this assumption, what is $E(X)$ for a fleet of $N$ vehicles without air-bags?

b) It is planned to equip a fleet of $N$ vehicles for one year with air bags to test their effectiveness in reducing the risk of death and serious injury to the driver. Let $p$, $0 \leq p \leq 1/4$, denote the probability of death or serious injury for the driver in a head-on collision involving a vehicle equipped with air-bags. Consider the test of the null hypothesis

$$H_0: \quad p=1/4 \quad \text{versus} \quad H_a: \quad p=p_a<1/4.$$ 

Use the normal approximation to the Poisson distribution to determine the fleet size $N$ required for a test with Type I and Type II error probabilities $\alpha=\beta=0.05$. In particular, what fleet size is required when $p_a = 1/16$?
1. A diagnostic test for a disease D is often classified by its sensitivity and specificity. Sensitivity is the probability that the test gives a positive result when applied to a person who actually has D. Specificity is the probability that the test gives a negative result when applied to a person who is actually free of D. Of great importance to the usefulness of such a test is the probability that a tested person actually has D if the test gives a positive result. In order for this probability to be at least 90%, what must the lower bound on the prevalence of D be (i.e., what is the smallest allowable value of P(D)) if the diagnostic test has 99% sensitivity and 98% specificity?
2. a. i. From a population of size 50, ten elements are drawn by simple random sampling. Their associated values are:

15, 31, 16, 27, 22, 23, 25, 19, 20, 16.

Estimate the population mean $\mu$ and also the variance of the sampling distribution of your estimator.

ii. A population of size 50 is stratified into two subclasses of respective sizes 35 and 15. Five elements are drawn from each stratum by stratified random sampling, yielding:

Stratum I: 32, 40, 39, 42, 35

Stratum II: 60, 69, 64, 55, 63

Estimate the population mean and also the variance of the sampling distribution of your estimator. Is this proportional allocation?

b. Discuss the advantages and disadvantages of

i. simple random sampling

ii. stratified random sampling

iii. cluster random sampling

in terms of the interplay between precision and cost.

3. For the exponential distribution having probability density function

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x > 0,$$

the maximum likelihood estimator of $\mu$ is the mean $\bar{x}$ of a random sample of size $n$ from the population $f(x)$. When $n$ is reasonably large, the distribution of $\bar{x}$ can be satisfactorily approximated by a normal distribution. Using this approximation, derive a $100(1-\alpha)%$ confidence interval for $\mu$. Apply the result to obtain a 95% confidence interval when a sample of 49 yields $\bar{x} = 11.52$. 
4. Researchers believe that the prevalence ($Y$) of byssinosis in workers in textile manufacturing plants is linearly related to the mean daily cotton dust level ($X$) in mg/m$^3$.

a) Under the assumption that zero cotton dust level implies zero prevalence, write down an appropriate statistical model expressing $Y$ as a straight-line function of $X$. Make sure to define precisely all terms in the model you suggest.

b) Given $n$ pairs of data points $(X_i, Y_i)$, $i=1,2,...,n$, provide explicit expressions for the least squares estimates of all parameters associated with your model.

c) Using the above results, determine whether there is evidence of a significant relationship between $X$ and $Y$ based on the following set of data:

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>
5. The research designs often used in evaluating mental health programs may be open to several sources of error, including the following:

   a. **History** - extraneous events may occur simultaneously with the treatment.

   b. **Maturation** - the effect may be due to the passage of time alone.

   c. **Testing** - a "before" measure may stimulate change regardless of treatment.

   d. **Regression to Mean** - individuals with extreme scores at one time tend to score nearer the mean at another.

   e. **Selection** - different groups may be composed of individuals with differing degrees of susceptibility to treatment.

Give examples of various research designs which do and do not control for each of these sources of error. Be sure to discuss your choice of examples in detail.
6. The following table shows the distribution of the number of hospital outpatient attendances for males in a small town during a one-year period. For example, the table tells us that out of 35,008 males in the town 734 had 3 outpatient attendances during the year.

**Distribution of the number of hospital outpatient attendances for males over a one-year period**

<table>
<thead>
<tr>
<th>NUMBER OF OUTPATIENT ATTENDANCES</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>&gt;4</th>
<th>TOTAL</th>
<th>( \bar{X} )</th>
<th>( s^2 )</th>
<th>( S )</th>
</tr>
</thead>
<tbody>
<tr>
<td>27,411</td>
<td>4,339</td>
<td>1,539</td>
<td>734</td>
<td>985</td>
<td>35,008</td>
<td>0.43</td>
<td>1.26</td>
<td>1.12</td>
<td></td>
</tr>
</tbody>
</table>

It is desirable to develop a distributional form to describe this data as a first step in developing a mathematical model describing the delivery of health services. If we assume for each individual that

i) successive outpatient attendances occur independently,

ii) the probability of a given number of attendances is the same for all time intervals of the same length, and

iii) only one alternative occurs at a time,

then we have the basic assumptions for a Poisson Process. Under these assumptions, the distribution of the number of attendances for a fixed time period of one year will be Poisson; i.e.,

\[ \text{Pr}\{\text{i attendances for a given individual in one year}\} = e^{-\lambda} \frac{\lambda^i}{i!} \]

The parameter \( \lambda \) will be a measure of the tendency for an individual to use the outpatient service. If we further assume that all individuals have the same \( \lambda \), then the proportion of persons in the population with \( i \) outpatient attendances will be given by the Poisson distribution.
Investigate whether the Poisson distribution provides a suitable model for the above data. Have you any suggestions for an improved model?

It is suggested that you estimate the parameter of the Poisson distribution by the method of moments. Note that

\[ e^{-0.43} = 0.651. \]

For ease of computation, if

\[ p_i = e^{-\lambda} \frac{\lambda^i}{i!}, \]

then

\[ p_0 = e^{-\lambda}, \]
\[ p_1 = \lambda p_0, \]
\[ p_2 = \frac{1}{2} \lambda p_1, \]
\[ p_3 = \frac{1}{3} \lambda p_2, \]
\[ (p_4+p_5+\ldots) = 1 - p_0 - p_1 - p_2 - p_3. \]
1. Suppose that a proportion \( p \) of a large population has a particular disease, which can be detected by a blood test. Suppose that a random sample of \( n \) people is to be tested. This testing can be done in two ways:

i) Each of the \( n \) people is tested separately.

ii) The blood samples of the \( n \) people are mixed together and analyzed. If the test is negative, all of them are healthy (that is, just this one test is needed). If the test is positive, each of the \( n \) persons must be tested separately (that is, totally \( n+1 \) tests are needed).

a) For procedure (ii), give the probability distribution of \( X \), the number of tests required.

b) Show that

\[
E(X) = (n+1) - n(1-p)^n
\]

c) For \( n=2 \), show that the difference in the number of tests required by procedures (i) and (ii) on the average, that is

\[
2 - E(X)
\]

is positive for \( p < 1 - \frac{1}{\sqrt{2}} \).
2. Assume that you have been asked to serve as statistician for a Division of Preventive Health Services in a health department for a state with a population of approximately 6 million persons. The Division has responsibility for program development, operation, and evaluation for chronic disease, communicable disease, dental health, maternal and child health, and mental health. Specific responsibilities include identification of health needs in various population groups, study of factors associated with needs, determination of available manpower and facilities to help meet them, and provision and evaluation of needed services. The Division has been asked to include specific consideration of the growing hazards of our physical environment for the health of the people.

Your first assignment is to prepare a source book containing readily available data, and references to other data sources (including tapes, punch cards, or tabulations) that might be used in program development, operation, and evaluation.

List some of the major sources and types of data that you might include in this source book, and discuss some of the ways in which this information might be of use to the Division.
3. Nine animals were each starved for six days, each at a different humidity. The nine humidities \( X \) ranged from 0% to 93%, and weight loss \( Y \) was the response of interest. For the \( n=9 \) data points, the following quantities were calculated:

\[
\begin{align*}
\sum_{i=1}^{9} X_i &= 453.5, \\
\sum_{i=1}^{9} Y_i &= 54.2, \\
\sum_{i=1}^{9} (X_i - \bar{X})(Y_i - \bar{Y}) &= -441.82, \\
\sum_{i=1}^{9} (X_i - \bar{X})^2 &= 8,301.39, \\
\sum_{i=1}^{9} (Y_i - \bar{Y})^2 &= 24.13.
\end{align*}
\]

Assuming a model of the form

\[ Y = \beta_0 + \beta_1 (X - \bar{X}) + \epsilon, \]

a) Find the least squares estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) of \( \beta_0 \) and \( \beta_1 \).

b) Construct an appropriate analysis of variance table.

c) Obtain and interpret a 95% confidence interval for the true average value of \( Y \) when \( X=25 \).

4. In a certain measuring process, the distribution of measurements is normal with standard deviation 8. Because the mean has a tendency to shift upward from time to time, quality control personnel make periodic checks by taking random samples and comparing sample means against a prescribed standard. Each such check is a test of the hypothesis that the process mean is 100 against the alternative hypothesis that the process mean is greater than 100; the sample is composed of 16 measurements, and the 5% level of significance is used.

(a) Specify the quality control decision rule in terms of the sample mean.

(b) What is the probability of a correct decision if the process mean is truly 100?

(c) What is the probability of a correct decision if the process mean is truly 105?
5. Goal Attainment Scaling is a model which is currently in use in the evaluation of mental health programs. A patient entering a mental health service will be scored at follow-up on \( k \) variables \( X_1, X_2, \ldots, X_k \), where \( k \) may be different for different patients. The variables \( X_1, X_2, \ldots, X_k \) are constructed and scaled so that each one can be considered to have a normal distribution with mean zero and variance one. An overall score is determined for each patient by the formula

\[
T = 50 + 10 \left\{ \frac{\sum_{i=1}^{k} w_i X_i - E\left( \sum_{i=1}^{k} w_i X_i \right)}{\left[ \text{Var} \left( \sum_{i=1}^{k} w_i X_i \right) \right]^{1/2}} \right\}
\]

where \( w_1, w_2, \ldots, w_k \) are constants reflecting the relative importance of \( X_1, X_2, \ldots, X_k \).

a) Show that \( E\left( \sum_{i=1}^{k} w_i X_i \right) = 0 \).

b) What is \( E(T) \), \( \text{Var}(T) \), and what is the distribution of \( T \)?

c) If \( \rho_{ij} \) is the correlation between \( X_i \) and \( X_j \) and if \( \rho_{ij} = \rho \) for all \( i \neq j \), show that \( T \) can be written in the form

\[
T = 50 + 10 \left\{ \frac{\sum_{i=1}^{k} w_i X_i}{\left[ (1-\rho) \sum_{i=1}^{k} w_i^2 + \rho \left( \sum_{i=1}^{k} w_i \right)^2 \right]^{1/2}} \right\}
\]
6. A group of 100 rabbits is being used in a nutrition study. A pre-study weight is recorded for each rabbit. The average of these weights is 3.1 pounds. After two months, the experimenter wants to obtain a rough approximation to the average weight of the 100 rabbits. A simple random sample of n=10 rabbits is selected and each rabbit is weighed. The pre-study weights and current weights for the sample are presented below:

<table>
<thead>
<tr>
<th>RABBIT NUMBER</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE-STUDY WT. (X_i)</td>
<td>3.2</td>
<td>3.0</td>
<td>2.9</td>
<td>2.8</td>
<td>2.8</td>
<td>3.1</td>
<td>3.0</td>
<td>3.2</td>
<td>2.9</td>
<td>2.8</td>
</tr>
<tr>
<td>CURRENT WT. (Y_i)</td>
<td>4.1</td>
<td>4.0</td>
<td>4.1</td>
<td>3.9</td>
<td>3.7</td>
<td>4.1</td>
<td>4.2</td>
<td>4.1</td>
<td>3.9</td>
<td>3.8</td>
</tr>
<tr>
<td>DIFF. d_i=Y_i-X_i</td>
<td>0.9</td>
<td>1.0</td>
<td>1.2</td>
<td>1.1</td>
<td>0.9</td>
<td>1.0</td>
<td>1.2</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

\[ \sum_{i=1}^{10} Y_i = 39.9, \quad \sum_{i=1}^{10} X_i = 29.7, \quad \sum_{i=1}^{10} X_i Y_i = 118.67, \quad \sum_{i=1}^{10} d_i = 10.2 \]

\[ \sum_{i=1}^{10} Y_i^2 = 159.43, \quad \sum_{i=1}^{10} X_i^2 = 88.43, \quad \sum_{i=1}^{10} d_i^2 = 10.52 \]

i) Using the \( d_i \)'s, estimate the average current weight of the population of 100 rabbits in the study and place a bound on the error of estimation.

ii) Discuss an alternative method to that used in (i) to estimate the average current weight of the 100 rabbits (still using information on pre-study weight), and then comment on the relative merits of the different methods.
1. A study was conducted to determine if method A of teaching reading to first grade students was superior to the presently used method, method B. Twelve pairs of students were identified and one member of each pair was taught using method A while the second member was taught by method B. After three months of training, a reading test was given to the students. For 9 of the 12 pairs, the student taught by method A scored higher than his counterpart taught by method B.

(a) State the null and alternative hypotheses for the purposes of this study.

(b) Compute the exact probability of the occurrence of this result or worse under the null hypothesis.

(c) If there had been 64 pairs, and the results had been 48 students taught by method A scored higher, what conclusion would you draw for $\alpha = .05$, using large sample theory.

(d) If you had been asked prior to the experiment "How many pairs of students are needed for this study?", what additional information would you need?
2. Two dentists A and B made a survey of the condition of the teeth of 200 children in a village. Dentist A selects a simple random sample of 20 children and counts the number of decayed teeth for each child, with the following results:

<table>
<thead>
<tr>
<th>Child Number</th>
<th>Number of Decayed Teeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
\Sigma y_i = 42
\]

\[
\Sigma y_i^2 = 252
\]

Dentist B, using the same dental techniques, examines all 200 children, recording merely those who have no decayed teeth. He finds 60 children with no decayed teeth.

(10) a) Estimate the total number of decayed teeth in the village children using Dentist A's results. Is your estimate unbiased? Provide an estimate of the variance of this estimate of the total number of decayed teeth.

(10) b) Consider the above random sample as a sample of 8 from the 60 children Dentist B found with no decayed teeth and a sample of 12 from the 140 children with one or more decayed teeth. Provide an alternate estimate of the total number of decayed teeth in the village children. Is this estimate unbiased? Provide an estimate of the variance of this estimate.

(5) c) Which procedure is more precise? How do you explain this?
3. A random sample of size \( n \) is taken from a population having the exponential distribution defined by the probability density function

\[
f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x > 0,
\]

wherein the mean is \( \mu \) and variance is \( \mu^2 \).

(20) a) For the large-sample case, apply the Central Limit Theorem to show how an approximate confidence interval (confidence coefficient \( 1 - \alpha \)) can be based on the sample mean as the only sample parameter required.

(5) b) What is the specific 90% confidence interval of this kind in the case where a random sample of 64 observations yields \( \bar{x} = 24.12 \)?

4. An executive is willing to hire a secretary who has applied for a position unless the applicant averages more than one error per typed page. A random sample of five pages is selected from some material typed by this secretary and the errors per page are: 3, 4, 3, 1, 2.

(10) a) Assuming that the number of errors per page has a Poisson distribution with parameter \( \lambda \), give an expression for the exact p-value for a test of \( H_0: \lambda = 1 \) versus the alternative \( H_a: \lambda > 1 \).

(10) b) Suppose that the executive looks at a random sample of 225 pages of the secretary's work and finds 252 errors. Approximate the p-value for the same test as in part a.

(5) c) With \( \alpha = 0.05 \), what is the executive's decision based on the results in part b?

5. Briefly describe a typical research data management system for a project with sufficient data flow to warrant computerized data management. Be sure your description includes each major component of such a system, including all major data processing phases from data collection through statistical analysis, and includes brief descriptions of measures taken to assure data quality, data security, and data confidentiality.
6. Consider the following descriptions of two populations, A and B.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-19</td>
<td>32</td>
<td>4</td>
<td>144</td>
<td>18</td>
</tr>
<tr>
<td>20-39</td>
<td>27</td>
<td>5</td>
<td>89</td>
<td>19</td>
</tr>
<tr>
<td>40-59</td>
<td>22</td>
<td>6</td>
<td>51</td>
<td>16</td>
</tr>
<tr>
<td>60+</td>
<td>19</td>
<td>8</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>23</td>
<td>301</td>
<td>61</td>
</tr>
</tbody>
</table>

(5) (a) Calculate the crude death rate, and the death rates specific to the enumerated age groups for populations A and B.

(5) (b) Calculate the directly age-standardized death rate for population B, using population A as the standard.

(5) (c) Explain the reasons for the change in the death rates for population B obtained under (a) and (b).

(5) (d) Assume that no information is available on deaths by age in population B. Explain a procedure to calculate a standardized death rate for population B, using population A as the standard.

(5) (e) Write a short note on the merits and demerits of standardized death rates.

EDITORIAL NOTE: Two tables were appended to this examination:

a) Binomial Coefficients $C(N,X)$ for $N=1(1)20$, $X=1(1)10$

b) Standard normal distribution function $\phi(x)$ for $x=0(.05)3$
1. (a) Massachusetts and Mississippi each claimed a lower tuberculosis death rate than the other in 1950. Using the following data, show how both claims can be substantiated, and explain the paradox.

Population and Number of Tuberculosis Deaths, by Color, in Massachusetts and Mississippi, 1950

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>White</th>
<th>Non White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massachusetts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>4,690,514</td>
<td>4,611,503</td>
<td>79,001</td>
</tr>
<tr>
<td>TB deaths*</td>
<td>1,005</td>
<td>942</td>
<td>63</td>
</tr>
<tr>
<td>Mississippi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>2,178,914</td>
<td>1,188,632</td>
<td>990,282</td>
</tr>
<tr>
<td>TB deaths*</td>
<td>573</td>
<td>191</td>
<td>382</td>
</tr>
</tbody>
</table>

*Deaths from all forms of tuberculosis

(b) Given that the United States had a population of 151,325,798 in 1950, of which 89.31% was White, calculate suitably standardized tuberculosis mortality rates for Massachusetts and Mississippi.

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) Binomial
(b) Poisson
(c) Multinomial
(d) Normal

Also, if \( \text{Var}(X) = 50 \), \( \text{Var}(X + Y) = 80 \) and \( \text{Var}(X - Y) = 40 \), find

(e) \( \text{Var}(Y) \) and \( \text{Cov}(X,Y) \)

3. (a) Define simple random sampling with and without replacement. Compare and contrast, briefly. Explain requisite formulae for estimates produced and standard errors.

(b) What is stratified random sampling and in what circumstances is it a preferred procedure? Explain requisite formulae for estimates produced and standard errors.
4. 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>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>
</tr>
<tr>
<td>2</td>
<td>161</td>
</tr>
<tr>
<td>3</td>
<td>169</td>
</tr>
<tr>
<td>4</td>
<td>186</td>
</tr>
<tr>
<td>5</td>
<td>183</td>
</tr>
<tr>
<td>6</td>
<td>184</td>
</tr>
<tr>
<td>7</td>
<td>268</td>
</tr>
<tr>
<td>8</td>
<td>184</td>
</tr>
<tr>
<td>9</td>
<td>180</td>
</tr>
<tr>
<td>10</td>
<td>188</td>
</tr>
<tr>
<td>11</td>
<td>184</td>
</tr>
<tr>
<td>12</td>
<td>182</td>
</tr>
<tr>
<td>13</td>
<td>179</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

\[
\begin{align*}
\sum X &= 2405 \\
\sum Y &= 8619 \\
\sum XY &= 1633249 \\
\sum X^2 &= 453517 \\
\sum Y^2 &= 5892485 \\
\frac{\sum X}{N} &= 185 \\
\frac{\sum Y}{N} &= 663 \\
\sum X^2 - \left(\frac{\sum X}{N}\right)^2 &= 8592, \\
\sum XY - \frac{(\sum X)(\sum Y)}{N} &= 38734, \\
\sum Y^2 - \left(\frac{\sum Y}{N}\right)^2 &= 178088.
\end{align*}
\]

(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.
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 $5.0 \text{mg/100ml}$ and a standard deviation of $1.0 \text{mg/100ml}$, and in diseased persons it has a mean of $8.5 \text{mg/100ml}$ and a standard deviation of $1.0 \text{mg/100ml}$.

A test $T$ for the presence of gout is to classify those persons with a serum uric acid level of at least $7.0 \text{mg/100ml}$ 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 the tabulated values of the standard normal distribution.

(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 are a single observation on the serum uric acid $x$ of the person to be classified.

(c) What is the rejection region of the test $T$?

(d) What is the size of the test (expressed in terms of tabulated values of the standard normal distribution)?

(e) 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.
DEPARTMENT OF BIOSTATISTICS

Examination for MPH and MSPH Candidates

(9:30 a.m.-1:30 p.m., Saturday, April 8, 1978)

1. Suppose you were going to carry out a survey of a rural area of North Carolina of about 5000 families to establish the need for a health clinic in the area. The geographic area corresponded approximately to the service area of two telephone exchanges (a telephone exchange corresponds to the first three digits of a seven digit telephone number), and a telephone directory covering these exchanges was available.

   a) Outline briefly some of the advantages and disadvantages of carrying out the survey by telephone.

   b) Suppose you established that the telephone listings for the two exchanges covered 90% of the 5000 families that comprised the target population. Cost considerations prevented personal interviews and so the survey population was taken to be the 4500 families with listed telephone numbers. Further, suppose that the survey budget was sufficient to allow 300 families to be telephoned with adequate follow-up. Assuming an overall response rate of 80%, with what precision would a proportion 0.5 be estimated given that the sample was selected by simple random sampling?

   c) Consider the possibility of stratifying by telephone exchange and selecting a simple random sample from each exchange. How might such a procedure influence the precision of your estimate? Explain your answer.

   d) Suppose you decide to select a stratified random sample, each exchange comprising one stratum. You are particularly interested in estimating the overall proportion of families with a problem of accessing health care. The data available from the survey is as follows.

<table>
<thead>
<tr>
<th>Stratum 1 First Telephone exchange</th>
<th>Stratum 2 Second Telephone exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size (N)</td>
<td>1500</td>
</tr>
<tr>
<td># respondents =</td>
<td></td>
</tr>
<tr>
<td>Sample size (n)</td>
<td>80</td>
</tr>
<tr>
<td># of households in sample with access problem</td>
<td>10</td>
</tr>
</tbody>
</table>

   Estimate the overall proportion of families with a problem accessing health care.

   What is the standard error of your estimate?
The table gives the results of an experiment to compare 4 treatments with a control.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>12, 14, 15, 20, 24</td>
<td>85</td>
<td>17</td>
</tr>
<tr>
<td>#1</td>
<td>5, 8, 10, 17</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>#2</td>
<td>4, 7, 8, 13</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>#3</td>
<td>13, 16, 19, 24</td>
<td>72</td>
<td>18</td>
</tr>
<tr>
<td>#4</td>
<td>7, 10, 10, 17</td>
<td>44</td>
<td>11</td>
</tr>
</tbody>
</table>

a) State a linear model for this experiment, define all the terms in it, and estimate all its parameters.

b) Construct the standard analysis of variance table and test the hypothesis of no differences among treatments and control (α = 5%).

c) The original intention was to compare each treatment individually with the control. Use the Bonferroni (addition) method to decide which treatments are significantly different from the control, with an overall error rate not to exceed 5%.

d) After seeing the data, the experimenter decides to compare the treatments with each other. By Scheffe's method, which pairwise differences are significant, with an overall error rate not exceeding 5%?

c) Same as (d), but use Tukey's method.
3. A regression analysis is considered in which the sampling units are the counties of a state. The dependent variable is a five-year age-race-sex adjusted lung cancer mortality rate; the investigators wish to find determinants of lung cancer. In planning the study, the investigators list the following as some of the independent variables of interest to be included in the model (there may be other variables also):

\[ X_1 = \% \text{ urban of county population} \]
\[ X_2 = \% \text{ rural of county population} \]
\[ X_3 = -1, 0, 1 \text{ if the county is in the South, Central or North region} \]
\[ X_4 = \text{mean per capita income of the county} \]
\[ X_5 = \% \text{ white of county population} \]

Critique this; redefine variables where necessary in commenting on:

(a) Dependent variable
(b) \( X_1 \)
(c) \( X_2 \)
(d) \( X_3 \)
(e) \( X_4 \)
(f) \( X_5 \)
(g) The basic design with respect to the research problem of finding determinants of lung cancer.
4. The number of machine malfunctions per shift at a factory is recorded for 200 shifts and the following data are obtained:

<table>
<thead>
<tr>
<th>No. of malfunctions (X)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of shifts</td>
<td>77</td>
<td>68</td>
<td>38</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>200</td>
</tr>
</tbody>
</table>

a) What is a reasonable probability model for this type of data? Justify your answer with suitable reasons for choosing the particular model.

b) What is the maximum likelihood estimator of $\mu$, the true average number of malfunctions per shift?

c) Test if the model in (a) adequately describes the data.

d) The management claims that $\mu = 1$ while the workers insist that $\mu > 1$. Test whether you agree to accept the management's claim based on the above data.

5. Suppose we observe the number of defective items in a certain production process. Let $p$ be the probability of obtaining a defective item on any single trial.

a) In a sample of $n_1 = 150$ items, $r_1 = 6$ were found to be defective. Find the 90 percent confidence interval for $p$.

b) Two other samples from the same process were obtained with the following results

Sample 2: $n_2 = 100$   $r_2 = 5$
Sample 3: $n_3 = 200$   $r_3 = 10$.

Find the "best" estimator of $p$ and the 90 percent confidence interval for $p$ using the information from these samples.
1. A city transportation department is conducting a survey to determine the gasoline usage of its residents. Stratified random sampling is used and the four city wards are treated as the strata. The amount of gasoline purchased in the last week is recorded for each household sampled. The strata sizes and the summary information obtained from the sample are:

<table>
<thead>
<tr>
<th>STRATA</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum size</td>
<td>3750</td>
<td>3272</td>
<td>1387</td>
<td>2475</td>
</tr>
<tr>
<td>Sample size</td>
<td>50</td>
<td>45</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Sample mean (in gallons)</td>
<td>12.6</td>
<td>14.5</td>
<td>18.6</td>
<td>13.8</td>
</tr>
<tr>
<td>Sample variance</td>
<td>2.8</td>
<td>2.9</td>
<td>4.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>

a) Estimate the mean weekly gasoline usage per household for the city population and construct a 95% error bound for your estimate.

b) Estimate the width of a 95% error bound for the estimator in simple random sampling (ignoring stratification when n = 155).

c) Do you agree with the surveyor that stratification has led to some reduction in the sampling error of the estimation (over simple random sampling)?
2. A forester seeking information on basic dimensions obtains the following measurements of the diameter 4.5 feet above the ground and the height of 12 sugar maple trees.

<table>
<thead>
<tr>
<th>Diameter x (in inches)</th>
<th>Height y (in feet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9, 1.2, 2.9, 3.1, 3.3, 3.9, 4.3, 6.2, 9.6, 12.6, 16.1, 25.8</td>
<td>18, 26, 32, 36, 44.5, 35.6, 40.5, 57.5, 67.3, 84, 67, 87.5</td>
</tr>
</tbody>
</table>

The forester wished to determine if the diameter measurements can be used to predict tree height.

a) Plot the scatter diagram and comment on the appropriateness of a straight line relation.

b) Determine an appropriate linearizing transformation. In particular, try $x' = \log x$ and $y' = \log y$.

c) Fit a straight line regression to the transformed data.

d) What proportion of the variability is explained by the fitted model?

e) Find a 95% confidence bond for predicted height when $x = 10$.

[Computations up to the second decimal place should be good enough.]
3. An experiment is conducted to determine the soil moisture deficit resulting from varying amounts of residual timber left after cutting trees in a forest. The three treatments are Treatment I: no timber left; Treatment II: 2000 bd. ft. left, and Treatment III: 8000 bd ft. left (board feet is a particular unit of measurement of timber volume). The measurements of moisture deficit are given below:

**MOISTURE DEFICIT IN SOIL**

<table>
<thead>
<tr>
<th>Treatment I</th>
<th>Treatment II</th>
<th>Treatment III</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.52</td>
<td>1.63</td>
<td>2.56</td>
</tr>
<tr>
<td>1.38</td>
<td>1.82</td>
<td>3.32</td>
</tr>
<tr>
<td>1.29</td>
<td>1.35</td>
<td>2.76</td>
</tr>
<tr>
<td>1.48</td>
<td>1.03</td>
<td>2.63</td>
</tr>
<tr>
<td>1.63</td>
<td>2.30</td>
<td>2.12</td>
</tr>
<tr>
<td>1.46</td>
<td>1.45</td>
<td>2.78</td>
</tr>
</tbody>
</table>

a) Does any significant difference exist among the mean moisture deficits for the three treatments?

b) Obtain simultaneous confidence regions for the mean differences between individual pairs of treatments.

c) Use (b) to provide a multiple comparison test for paired differences.

[Computations up to the third decimal place should be good enough.]
4. A biologist wishes to investigate the incidence of a particular disease in a certain tribal population. Let $p$ be the fraction of the population affected by this disease.

a) Individual members are to be randomly chosen and examined one at a time until exactly $m \geq 1$ affected persons are found in the sample. How many persons ($M$) are expected to be examined using this method of sampling when the units are drawn with replacement? What is the variance of $M$?

b) A second biologist decides to draw a random sample of size $n$ and to examine and count the number of persons ($X$) affected by the disease. Derive the formulae for the expectation and variance of $X$ when sampling is made

(a) with replacement; and (b) without replacement.
5. Surgery was performed on glaucoma patients in order to reduce intraocular pressure (IOP) and thereby provide relief from the symptoms of glaucoma. In some cases, medication was provided to the patients; these patients would take the drug periodically over the days following surgery. Assume that the patients receiving medication were selected at random. At a certain time following surgery, the patients were reexamined and their post-operation IOP measurements were taken and recorded with their pre-operation IOP measurements. The surgery plus medicine group is composed of 15 patients (26 eyes) and the surgery alone group is composed of 5 patients (9 eyes). The pre-surgery IOP, the follow-up IOP, and the absolute and percent changes in IOP are given in the following tables.

a) Summarize the data in a manner interpretable to physicians who are interested in whether or not the medication had any benefits in addition to the benefits of surgery regarding IOP. Provide appropriate counts, means, standard deviations, standard errors, ranges, or whatever you think appropriate.

b) Perform at least one appropriate test of significance on the data in addressing the question of interest (see part a). Consider both parametric and nonparametric tests in planning the analysis.

c) What additional information, if any, pertinent to exploring the data would you attempt to get in discussions with the physicians?
<table>
<thead>
<tr>
<th>Patient No.</th>
<th>IOP Before Surgery</th>
<th>Follow-up IOP</th>
<th>IOP Change</th>
<th>Percent Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>RE = 26</td>
<td>27</td>
<td>+1</td>
<td>+3.8</td>
</tr>
<tr>
<td></td>
<td>LE = 40</td>
<td>25</td>
<td>-15</td>
<td>-37.5</td>
</tr>
<tr>
<td>3.</td>
<td>LE = 28</td>
<td>19</td>
<td>-9</td>
<td>-32.1</td>
</tr>
<tr>
<td>4.</td>
<td>RE = 25</td>
<td>19</td>
<td>-6</td>
<td>-24</td>
</tr>
<tr>
<td>5.</td>
<td>RE = 26</td>
<td>19</td>
<td>-7</td>
<td>-26.9</td>
</tr>
<tr>
<td></td>
<td>LE = 22</td>
<td>19</td>
<td>-3</td>
<td>-13.6</td>
</tr>
<tr>
<td>6.</td>
<td>RE = 30</td>
<td>20</td>
<td>-10</td>
<td>-33.3</td>
</tr>
<tr>
<td></td>
<td>LE = 40</td>
<td>17</td>
<td>-23</td>
<td>-57.5</td>
</tr>
<tr>
<td>7.</td>
<td>RE = 40</td>
<td>14</td>
<td>-26</td>
<td>-65.0</td>
</tr>
<tr>
<td></td>
<td>LE = 58</td>
<td>18</td>
<td>-40</td>
<td>-69.0</td>
</tr>
<tr>
<td>8.</td>
<td>RE = 30</td>
<td>18</td>
<td>-12</td>
<td>-40.0</td>
</tr>
<tr>
<td></td>
<td>LE = 40</td>
<td>18</td>
<td>-22</td>
<td>-55.0</td>
</tr>
<tr>
<td>10.</td>
<td>RE = 36</td>
<td>22</td>
<td>-14</td>
<td>-38.9</td>
</tr>
<tr>
<td></td>
<td>LE = 39</td>
<td>24</td>
<td>-15</td>
<td>-38.5</td>
</tr>
<tr>
<td>11.</td>
<td>RE = 40</td>
<td>22</td>
<td>-18</td>
<td>-45.0</td>
</tr>
<tr>
<td></td>
<td>LE = 40</td>
<td>24</td>
<td>-16</td>
<td>-40.0</td>
</tr>
<tr>
<td>14.</td>
<td>RE = 24</td>
<td>22</td>
<td>-2</td>
<td>-8.3</td>
</tr>
<tr>
<td></td>
<td>LE = 25</td>
<td>18</td>
<td>-7</td>
<td>-28.0</td>
</tr>
<tr>
<td>15.</td>
<td>RE = 30</td>
<td>20</td>
<td>-10</td>
<td>-33.3</td>
</tr>
<tr>
<td>20.</td>
<td>LE = 30</td>
<td>19</td>
<td>-11</td>
<td>-36.7</td>
</tr>
<tr>
<td>21.</td>
<td>RE = 28</td>
<td>19</td>
<td>-9</td>
<td>-32.1</td>
</tr>
<tr>
<td></td>
<td>LE = 30</td>
<td>21</td>
<td>-9</td>
<td>-30.0</td>
</tr>
<tr>
<td>24.</td>
<td>RE = 35</td>
<td>35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LE = 38</td>
<td>27</td>
<td>-11</td>
<td>-28.9</td>
</tr>
<tr>
<td>25.</td>
<td>RE = 46</td>
<td>30</td>
<td>-16</td>
<td>-34.8</td>
</tr>
<tr>
<td></td>
<td>LE = 32</td>
<td>31</td>
<td>-1</td>
<td>-3.1</td>
</tr>
</tbody>
</table>

*RE and LE denote the right and left eye respectively.*
### Surgery Alone

<table>
<thead>
<tr>
<th>Patient No.</th>
<th>IOP Before Surgery</th>
<th>Follow-up IOP</th>
<th>IOP Change</th>
<th>Percent Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>RE = 34 LE = 53</td>
<td>17 8</td>
<td>-17 -45</td>
<td>-50.0 -84.9</td>
</tr>
<tr>
<td>9.</td>
<td>RE = 22 LE = 30</td>
<td>14 16</td>
<td>-8 -14</td>
<td>-36.4 -46.7</td>
</tr>
<tr>
<td>17.</td>
<td>RE = 26</td>
<td>13</td>
<td>-13</td>
<td>-50.0</td>
</tr>
<tr>
<td>18.</td>
<td>RE = 26 LE = 25</td>
<td>22 10</td>
<td>-4 -15</td>
<td>-15.4 -60.0</td>
</tr>
<tr>
<td>23.</td>
<td>RE = 21 LE = 20</td>
<td>18 17</td>
<td>-3 -3</td>
<td>-14.3 -15.0</td>
</tr>
</tbody>
</table>
SECTION A

The following questions apply to stratified sampling:

3 points a) Briefly describe stratified (simple) random sampling.

3 points b) What makes stratified random sampling a type of probability sampling?

6 points c) Briefly discuss the set of guidelines which you would use in forming strata in this type of design.

3 points d) Briefly describe the relationship between an epsem (or self-weighting) design and proportionate stratified sampling.

10 points e) Noting that the estimated variance of \( \bar{y}_o = \frac{1}{n} \sum_j y_j \) from a simple random sample of size \( n \) from a population of size \( N \) is

\[
\text{var}(\bar{y}_o) = \left( \frac{1-f}{n} \right) s^2
\]

where

\[
s^2 = \frac{\sum_j (y_j - \bar{y}_o)^2}{n-1} \quad \text{and} \quad f = \left( \frac{n}{N} \right),
\]

derive an estimator for the variance of the estimated mean from a stratified random sample,

\[
\bar{y}_{wo} = \frac{1}{H} \sum_h W_h \bar{y}_h
\]

where \( W_h = \left( \frac{N_h}{N} \right) \), \( N_h \) is the number of elements in the \( h \)-th stratum of the population, \( N = \sum_h N_h \), \( H \) is the total number of strata,

\[
\bar{y}_h = \frac{1}{n_h} \sum_j y_{nj}
\]

and \( n_h \) is the number of elements in the sample from the \( h \)-th stratum.
Question 2  Consider the regression model
\[ Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + e_i, \quad i = 1, \ldots, n. \]
Suppose that you want to test
\[ H_0: \beta_2 = 0 \quad \text{vs.} \quad H_1: \beta_2 > 0. \]

5 points  a) State the usual assumptions you need to make for this purpose.

5 points  b) Derive the expression for the mean square due to error.

5 points  c) Write down the formula for the test statistics.

5 points  d) Suppose that you want to test also
\[ H_0^*: \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_1^*: (\beta_1, \beta_2) \neq 0. \]

5 points  e) What are the critical regions for the testing problems in (c) and (d)?

Question 3  A large machine consists of 50 components. Past experience has shown that a particular component will fail during an 8-hour shift with probability .1. The equipment will work if no more than one component fails during an 8-hour shift.

7 points  a) Calculate the probability that the machine will work throughout an entire 8-hour shift, assuming the binomial distribution is applicable. Carefully define any notation and state the assumptions required for the valid application of the binomial procedure.

4 points  b) State the general situation and assumptions which are required for the Poisson model to be applicable to this situation. Give the Poisson density function.

5 points  c) The Poisson distribution may be used to approximate the binomial distribution when \( n \) is "large" and \( p \) is "small". Use the Poisson distribution to find an approximation to the probability computed in part (a).

5 points  d) Calculate an approximation to the probability obtained in part (a) by using the normal approximation to the binomial distribution. State the assumptions under which the approximation is reasonable.

4 points  e) Compare the accuracy and usefulness of these three calculations.
SECTION B

Question 4  Suppose that $X$ has a chi-square distribution

$$f_X(x) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{\nu/2 - 1} e^{-x/2}, \quad x > 0, \quad \nu > 0,$$

and $Y$ has a beta distribution

$$f_Y(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} y^{\alpha - 1} (1 - y)^{\beta - 1}, \quad \alpha > 0, \quad \beta > 0$$

$$0 < y < 1.$$

6 points a) Obtain formulae for $E(X^r)$ and $E(Y^r)$, where $r$ is a positive integer.

7 points b) If $r < \left[\frac{\nu}{2}\right]$, show that

$$E(X^{-r}) = \left[\frac{1}{(\nu - 2)(\nu - 4) \cdots (\nu - 2r)}\right]^{-1}.$$

6 points c) Find $E(Y^{-r})$. What condition must $r$ satisfy?

6 points d) Assume that $X$ and $Y$ are independent, and let $Z = X/Y$. Find $E(Z)$.

Question 5  Suppose that lifetime $T(T)$ of a certain mechanical device has a Weibull distribution with PDF

$$f_T(t) = \frac{c}{\theta} t^{c-1} e^{-\left(\frac{t}{\theta}\right)^c}, \quad \theta > 0, \quad c > 0, \quad t > 0$$

12 points a) Obtain the formula for cumulative distribution function, $F_T(t)$, and evaluate the expected proportion of failures

(i) before time $\tau$;

(ii) after time $\tau$.

13 points b) Suppose that $N$ items were put on test at $t = 0$, and $n$ were observed to fail before time $\tau$. Suppose that exact failure times were recorded; let $t_i$ denote the failure time of the $i$-th item among those items which failed ($i = 1, \ldots, n$). Construct the likelihood function.

* $[s]$ denotes the largest integer $\leq s$.

EDITORIAL NOTE: Two tables were appended to this examination:

a) Standard normal distribution function $\Phi(x)$ for $x = -3(0.01)3$

b) Natural logarithms $\ln(x)$ for $x = 1(0.01)10$
M.P.H. and M.S.P.H. students are to answer any two questions during the two-hour period (1 pm - 3 pm). M.S. students are to answer three questions of which not more than 2 should be from Section A - time period 1 pm - 4 pm.

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

SECTION A

Question 1 A survey of 320 families, each with 5 children, revealed the following distribution:

<table>
<thead>
<tr>
<th>No. of girls</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of families</td>
<td>18</td>
<td>56</td>
<td>110</td>
<td>88</td>
<td>40</td>
<td>8</td>
<td>320</td>
</tr>
</tbody>
</table>

a) Is the result consistent with the hypothesis that male and female births are equally probable?

15 points Test this hypothesis at the significance level $\alpha = 0.05$, $\alpha = 0.01$

10 points b) What is the maximum likelihood estimate of the probability of a female birth?
Question 2  A. Briefly describe or explain the following terms:

(a) OS
(b) TSO
(c) Track (on magnetic disk)
(d) Block
(e) Logical Record
(f) byte
(g) JCL

7 points

B. Compare and contrast:

6 points

(a) OS dataset
(b) SAS database
(c) SAS dataset

(That is, demonstrate that you know what each of these terms means and the differences between them.)

B. Compare and contrast:

6 points

(a) OS dataset
(b) SAS database
(c) SAS dataset

(That is, demonstrate that you know what each of these terms means and the differences between them.)

C. The printout in Figure 1 was produced by the PROC PRINT statement in line 190 of the SAS program shown in Figure 2. Show what will be printed as a result of

12 points

(a) the PROC PRINT statement in line 230 in Figure 2. (Be careful to write out the entire output to be produced by SAS except titles and page headings.)
(b) Show what will be printed as a result of the PROC MEANS (Figure 2, line 250) and related statements.
(c) Show what will be printed as a result of the PROC PRINT statement in line 440 of Figure 2.
(d) Show what will be printed as a result of the PROC PRINT statement in line 530 of Figure 2.

Figure 1. Printout Produced by the PROC PRINT Statement in Line 190 of Figure 2.

<table>
<thead>
<tr>
<th>CBS</th>
<th>NAME</th>
<th>SEX</th>
<th>AGE</th>
<th>HEIGHT</th>
<th>WEIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALFRED</td>
<td>M</td>
<td>14</td>
<td>69</td>
<td>112</td>
</tr>
<tr>
<td>2</td>
<td>ALICE</td>
<td>F</td>
<td>13</td>
<td>56</td>
<td>84</td>
</tr>
<tr>
<td>3</td>
<td>DEFNADETTE</td>
<td>F</td>
<td>13</td>
<td>65</td>
<td>98</td>
</tr>
<tr>
<td>4</td>
<td>BARBADA</td>
<td>F</td>
<td>14</td>
<td>62</td>
<td>102</td>
</tr>
<tr>
<td>5</td>
<td>JAMES</td>
<td>M</td>
<td>12</td>
<td>57</td>
<td>83</td>
</tr>
</tbody>
</table>
Figure 2. SAS Program for Problem 2

```sas
00050 // EXEC SAS
00060 //*PW=EXAM
00070 //READ DD *
00080 M 14 69 112 ALFRED
00090 F 13 56 84 ALICE
00100 F 13 65 92 HERMANETTE
00110 F 14 62 102 BARBARA
00120 M 12 57 83 JAMES
00130 //SYSIN DD *
00140 DATA STUDENT1;
00150 INFILE RNK;
00160 LENGTH NAME $ 20 SEX $ 1;
00170 INPUT SEX AGE HEIGHT WEIGHT NAME;
00180 00190 PROC PRINT; * THIS PRODUCES FIGURE 2;
00200 00210 PROC SORT; BY SEX AGE HEIGHT;
00220 00230 PROC PRINT; * PROBLEM 3(A);
00240 00250 PROC MEANS MEAN N; * PROBLEM 3(B);
00260 00270 BY SEX;
00280 00290 VAR AGE HEIGHT;
00300 00310 PROC SORT; BY NAME SEX AGE;
00320 00330 DATA CHANGES;
00340 00350 LENGTH NAME $ 20 SEX $ 1;
00360 00370 MISSNIG _;
00380 00390 INPUT SEX AGE HEIGHT WEIGHT NAME;
00400 00410 CARDS;
00420 00430 F 13 50 ALICE
00440 M 12 89 JAMES
00450 M 14 ALFRED
00460 F 14 62 BARBARA
00470 F 14 61 BARBARA
00480 F 12 59 84 JANE
00490 ; * END OF CHANGF DATA;
00500 PROC PRINT; * PROBLEM 3(C);
00510 00520 PROC SORT; BY NAME SEX AGE;
00530 DATA STUDENT2;
00540 UPDATE STUDENT1 CHANGES;
00550 00560 00570 IF HEIGHT > 0;
00580 00590 00600 PROC PRINT; * PROBLEM 3(D);
```

Question 3  A sample survey of 800 adults is conducted in a large city in order to determine citizen attitudes toward national health insurance. The sampling frame for the survey is a list of adult taxpayers from which four strata are formed corresponding to each of four major sections of the city. The design calls for selecting a simple random sample within each stratum. General data for the survey sample as well as results for an opinion question on national health insurance are as follows:

<table>
<thead>
<tr>
<th>Stratum (h)</th>
<th>(Inner City)</th>
<th>(Blue Collar)</th>
<th>(Middle Class Suburbs)</th>
<th>(Affluent Suburbs)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of adults ((N_h))</td>
<td>20,000</td>
<td>50,000</td>
<td>20,000</td>
<td>10,000</td>
<td>100,000</td>
</tr>
<tr>
<td>Number of adults interviewed in the sample ((n_h))</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>Number of interviewed sample adults who are in favor of national health insurance (x_h)</td>
<td>190</td>
<td>180</td>
<td>40</td>
<td>30</td>
<td>440</td>
</tr>
</tbody>
</table>

5 points a) Calculate a stratified estimate of the proportion \((P)\) of adults in the city who favor national health insurance.

5 points b) Calculate an estimate of the variance of the estimate produced in Part (a).

5 points c) A colleague argues that since a "random sample" of adults has been chosen and since simple random sampling was used, the estimate and its variance can be calculated as if a simple sample of \(n = 800\) adults had been selected. Calculate the estimate and its variance as the colleagues suggests.

5 points d) Comment briefly on the difference between your analysis and your colleague's analysis.

5 points e) If you decide to use Neyman allocation for a similar survey on national health insurance in the future, how would you then allocate a sample of \(n = 800\) adults to the same four strata?
SECTION B

Question 4  Read the attached article by Topoff and Mirenda from a recent issue of *Science*. Note that from the data of Table 1 the authors were led to calculate a chi-squared statistic. Discuss the statistical appropriateness and the numerical accuracy of their analysis.

EDITORIAL NOTE: The article referred to, including Table 1, was reproduced and appended to this examination. It originally appeared in *Science* 207: 1099-1100 (7 March 1980).

Question 5  Ten observations on three independent variables \((X_1, X_2, X_3)\) and one dependent variable were obtained as follows:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7</td>
<td>5.5</td>
<td>6.0</td>
<td>2.3</td>
</tr>
<tr>
<td>2.2</td>
<td>6.2</td>
<td>6.8</td>
<td>1.8</td>
</tr>
<tr>
<td>2.7</td>
<td>5.5</td>
<td>6.0</td>
<td>2.7</td>
</tr>
<tr>
<td>3.0</td>
<td>6.2</td>
<td>7.3</td>
<td>2.7</td>
</tr>
<tr>
<td>3.2</td>
<td>7.9</td>
<td>7.1</td>
<td>2.8</td>
</tr>
<tr>
<td>2.2</td>
<td>6.2</td>
<td>6.8</td>
<td>2.2</td>
</tr>
<tr>
<td>2.7</td>
<td>5.5</td>
<td>6.0</td>
<td>2.4</td>
</tr>
<tr>
<td>4.0</td>
<td>9.0</td>
<td>7.8</td>
<td>1.6</td>
</tr>
<tr>
<td>3.0</td>
<td>6.2</td>
<td>7.3</td>
<td>2.9</td>
</tr>
<tr>
<td>4.0</td>
<td>9.0</td>
<td>7.8</td>
<td>2.2</td>
</tr>
</tbody>
</table>

It is postulated that the model

\[
E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_{12} X_1^2 + \varepsilon_{32} X_3^2 + \varepsilon_{33} X_3^3
\]

is reasonable.

5 points  (i)  Inspect the model and the data and comment on the appropriateness of the model.

6 points  (ii)  Determine an estimate of the variance of the random error, \(\sigma^2\), for these data.

8 points  (iii)  By plotting the data or otherwise, select a model with few parameters that appears appropriate and fit that model. Comment on the fit.

6 points  (iv)  Does \(R^2\) tell you the same thing as a direct test of lack of fit based on an estimate of pure error?
**INSTRUCTIONS:**

a) This is a closed book examination.

b) M.P.H. students are to answer any two questions during the two hour time period (10 am – 12 noon). M.S.P.H. and M.S. students are to answer three questions of which not more than 2 should be from Group A – time period 9 am – 12 noon.

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

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

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

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

**GROUP A**

**Question 1**

a) Define probability sampling from finite populations.

b) Briefly discuss the advantages of probability samples over nonprobability samples.

c) The recreation department in a large city hires you to design a cluster sample of families for an interview survey. The most important survey objective is to produce a 95% confidence interval for the proportion (P) of families who are aware of city recreation programs. The expected "half-width" of the confidence interval for the estimated proportion (p) will be about*

\[ d = 1.96 \sqrt{\text{Var}(p)} \]

You decide to select a simple random sample of blocks. Your past experience tells you that the design effect will be about 4. You also know that P is about 0.3. Ignoring the finite population correction, how large a sample of families would you recommend so that \( d = 0.05 \)?

*Recall that a 95% confidence interval for the estimated proportion (p) produced in your analysis will be of the form:

- **Lower Limit:** \( p - 1.96 \sqrt{\text{Var}(p)} \)
- **Upper Limit:** \( p + 1.96 \sqrt{\text{Var}(p)} \)
Question 2  Let X and Y have the joint density function
\[ f_{X,Y}(x,y) = 2(1-y)^{-2}, \quad 0 < y < x < 1. \]

a) Find the marginal distribution of X.

b) Find the conditional distribution of Y given that X = x.

c) The quantity \( E(Y|X = x) \) is often called the regression function of Y on x. Find \( E(Y|X = x) \).

d) For n pairs of data points \((x_i, y_i), i = 1, \ldots, n, \) derive the formula for the least squares estimator of \( \theta \) using your answer to part (c).

Question 3  Suppose that 60% of a particular breed of mice exhibit aggressive behavior when injected with a given dose of a stimulant. An experimenter will apply the stimulant to 3 mice one after another and will observe the presence or absence of aggressive behavior in each case.

a) List the sample space for the experiment. (Use A to denote aggressive and N to denote nonaggressive).

b) Assuming that the behaviors of different mice are independent, determine the probability of each elementary outcome.

c) Find the probability that
  (i) two or more mice will be aggressive,
  (ii) exactly two mice will be aggressive, and
  (iii) the first mouse will be nonaggressive while the other two will be aggressive.
d) What can you say about the exact distribution of the signed rank statistic when \( H_0 \) holds and \( n \) is small? What approximation would you recommend when \( n \) is large?

**Group B**

**Question 4** Let \( f_x(x) = \theta^{-1}e^{-x/\theta}, x > 0, \theta > 0 \) be the pdf of an exponential distribution.

a) Show that, in fact, \( X \) is distributed as \((\theta/2)\chi^2_2\) where \( \chi^2_2 \) has the chi square distribution with 2 degrees of freedom.

b) Based on a random sample \( X_1, \ldots, X_n \) of size \( n \), find the maximum likelihood estimator of \( \theta \).

c) Construct the likelihood ratio test for testing the null hypothesis \( H_0: \theta = \theta_0 \) against the alternative \( \theta < \theta_0 \).

d) Derive the power function of the test.

**Question 5** Let \( X_1, \ldots, X_n \) be \( n \) independent and identically distributed random variables with an unknown (but continuous) distribution function \( F \) and \( \theta \) be the median of \( F \). Suppose that one wants to test for

\[ H_0: \theta = \theta_0 \text{ (Specified) against } H_1: \theta > \theta_0. \]

a) Write down the expression for the sign test statistic for this testing problem and its exact distribution under \( H_0 \). What approximation to this distribution would you recommend when \( n \) is large?

b) What additional assumption do you need to make to use the Wilcoxon signed rank statistic for this testing problem?

c) Compute the first two moments of the Wilcoxon signed-rank statistic (under \( H_0 \)).
PART II

INSTRUCTIONS:

a) This is an open book examination.

b) M.P.H. and M.S.P.H. students are to answer any two questions during the two-hour period (1 pm - 3 pm). M.S. students are to answer three questions of which not more than 2 should be from Group A - time period 1 pm - 4 pm.

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

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

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

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

GROUP A

Question 1

The mean drying time of a brand of spray paint is known to be 90 seconds. The research division of the company that produces this paint contemplates that adding a new chemical ingredient to the paint will accelerate the drying process. To investigate this conjuncture, the paint with the chemical additions is sprayed on 15 surfaces and the drying times are recorded. The mean and standard deviations computed from these measurements are 86 seconds and 5.6 seconds respectively.

(a) Do these data provide strong evidence that the mean drying time is reduced by the addition of the new chemical?

(b) Construct a 98% confidence interval for the mean drying time of the paint with the chemical additive.

(c) Suppose that the actual standard deviation for the drying time does not change with the addition of the new chemical and is known to be equal to 6 seconds. Given this additional information, what would be your conclusions in (a) and (b)?

(d) Suppose that it is also conjectured that the standard deviation of the drying time decreases with the addition of the new chemical. Do these data provide a strong evidence for that?
The state welfare agent is in the process of sampling unemployment data in his state. The state is divided into 4 regions each with approximately the same population. Each region is in turn divided into 750 equal-sized sampling units. From each region five sampling units are selected and sampled intensively. The percentage unemployment for one such test is given below.

Region A: 4.2 4.4 4.5 5.0 5.1
Region B: 3.7 3.9 4.1 4.4 4.5
Region C: 4.8 5.0 5.1 5.2 5.2
Region D: 3.1 3.5 3.6 3.7 3.9

(a) Is this a simple random sampling? Why or why not?

(b) Calculate the mean unemployment rate for each region and use them to estimate the mean rate for the entire sample of 20 observations.

(c) Compute the mean using the entire 20 observations. Does it differ from the answer in part (b)? Why or why not?

(d) What procedure would be necessary if the regions and sampling units were of different sizes in terms of population?
Question 3

A) Briefly describe the purpose(s) of:

(a) JCL

(b) DATA step of SAS

(c) PROC step of SAS

B) Define, and describe the relationships between:

OS file (dataset)

SAS database

SAS dataset

(Examples of corresponding JCL and SAS code may be useful.)

C) In less than 1 page, outline the basic steps of the process required to create a SAS dataset that is ready for statistical analysis. Suppose the input dataset is a raw, unchecked dataset stored on disk. The SAS dataset is also to be stored on disk.

D) Write out the job or jobs, including JCL and SAS code, needed to do the following on a dataset with the format given in Figure 1. Use UNC.B.E.99U as the account number and MEXAM as the password.

(a) Create a SAS dataset, stored on on-line disk, called VEHACC that includes all the variables, but just the reportable cases. Use the variable names given in capitals in the format. Label Height and Weight.

(b) Print 10 observations.

(c) Create a variable called AGE_GP with the following codes:

<table>
<thead>
<tr>
<th>AGE_GP values</th>
<th>AGE values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 thru 10</td>
</tr>
<tr>
<td>2</td>
<td>11 thru 24</td>
</tr>
<tr>
<td>3</td>
<td>25 thru 54</td>
</tr>
<tr>
<td>4</td>
<td>55 and over</td>
</tr>
<tr>
<td>.</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

(d) Plot Height vs. Weight for males and females separately.

(e) Create cross-tabulation tables of Sex by Restraint, Sex by Injury, and Injury by Restraint.
Figure 1

Format for a vehicle-oriented accident file

**DSN=UNC.B.E999U.MASTERS.VEH.RAW**

There are approximately 1000 records

<table>
<thead>
<tr>
<th>Column</th>
<th>Information</th>
</tr>
</thead>
</table>
| 1      | Accident Reporting Type - ATYPE  
         | 1 Non-reportable  
         | 2 On private property  
         | 3 Reportable  |
| 2-7    | Accident Case Number - ID |
| 8      | Injury Class - INJ  
         | 1 Not injured  
         | 2 Class C injury  
         | 3 Class B injury  
         | 4 Class A injury  
         | 5 Killed  
         | 6 Not stated  |
| 9      | Restraint Used - BELT  
         | 1 No belt  
         | 2 Lap and shoulder belt  
         | 3 Child restraint  
         | 4 Not stated  |
| 10     | Race - RACE  
         | 1 White  
         | 2 Black  
         | 3 Indian  
         | 4 Other  
         | 5 Not stated  |
| 11     | Sex - SEX  
         | 1 Male  
         | 2 Female  
         | 3 Not stated  |
| 12     | Age - AGE  
         | 01-97 Actual age  
         | 98 Older than 97  
         | 99 Not stated  |
| 13-14  | Height - HT  
         | 01-98 Actual height in inches  
         | 99 Not stated  |
| 15-17  | Weight - WT  
         | 001-998 Actual weight in pounds  
         | 999 Not stated  |
GROUP B

Question 4  Suppose we have conducted an experiment to estimate the weight gain for a sample of 14 dairy cows as a result of a one week exposure to a feed additive. One statistical question of interest is to test whether the weight gain is zero. Another major objective is to estimate the weight gain.

The 14 dairy cows comprised three different breeds; there were 7 Holsteins, 5 Jerseys, and 2 Guernseys. The data collected were as follows:

<table>
<thead>
<tr>
<th>Weight Gain in lbs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holsteins</td>
</tr>
<tr>
<td>-7, 5, -1, 3, 1, 6, 0</td>
</tr>
<tr>
<td>Jerseys</td>
</tr>
<tr>
<td>2, -2, 1, 7, 2</td>
</tr>
<tr>
<td>Guernseys</td>
</tr>
<tr>
<td>9, 2</td>
</tr>
</tbody>
</table>

1. Suppose we were interested in testing \( H_0: \text{mean weight gain} = 0 \), and we had in mind a target population of dairy cows in which Holsteins, Jerseys, and Guernseys were in the ratio 7:5:2. Specify the test you would carry out. Compute the test and state the significance level. Provide a corresponding estimate of mean weight gain.

2. Suppose we were interested in testing \( H_0: \text{mean weight gain} = 0 \), and we were primarily interested in a target population of dairy cows in which Holsteins, Jerseys, and Guernseys were in equal proportions. Specify the test you would carry out. Compute the test. Provide a corresponding estimate of mean weight gain.

3. Are tests in questions 1 and 2 the same? Comment.

4. Suppose now that we were interested in testing \( H_0: \text{mean weight} = 0 \) and we had a target population in which Holsteins, Jerseys, and Guernseys were in the ratio \( W_H:W_J:W_G \). Specify an appropriate test.
Question 5  

In measuring the various constituents of cow's milk, it is of interest to determine how protein (Y) is related to fat (x₁) and solid-nonfat (x₂). Samples of 10 cows were taken and the following data were obtained:

<table>
<thead>
<tr>
<th>Observations</th>
<th>Protein (Y)</th>
<th>Fat (x₁)</th>
<th>Solid Non-Fat (x₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.75</td>
<td>4.74</td>
<td>9.50</td>
</tr>
<tr>
<td>2</td>
<td>3.19</td>
<td>3.66</td>
<td>8.56</td>
</tr>
<tr>
<td>3</td>
<td>2.99</td>
<td>4.27</td>
<td>8.54</td>
</tr>
<tr>
<td>4</td>
<td>3.46</td>
<td>4.03</td>
<td>8.62</td>
</tr>
<tr>
<td>5</td>
<td>3.27</td>
<td>3.51</td>
<td>9.35</td>
</tr>
<tr>
<td>6</td>
<td>3.27</td>
<td>3.97</td>
<td>8.39</td>
</tr>
<tr>
<td>7</td>
<td>2.78</td>
<td>3.23</td>
<td>7.87</td>
</tr>
<tr>
<td>8</td>
<td>3.59</td>
<td>3.79</td>
<td>9.33</td>
</tr>
<tr>
<td>9</td>
<td>3.16</td>
<td>3.36</td>
<td>8.86</td>
</tr>
<tr>
<td>10</td>
<td>3.65</td>
<td>3.64</td>
<td>9.21</td>
</tr>
</tbody>
</table>

Assume the linear regression of Y on x₁, x₂ and

(i) Find the least squares estimates of the (partial) regression coefficients.

(ii) We are interested in testing the null hypothesis that both these partial regression coefficients are equal to 0. Test this hypothesis at the significance level α = 0.05. State clearly the assumptions you need to make in this context.
GROUP A

Question 1.

You are called upon to assist the health department in a large city with the design of a local household survey. The survey's principal objective will be to estimate the proportion of households in which the person usually responsible for preparing the meals is aware of the importance of a balanced dietary intake by members of the household. Between 20 and 40 percent of the local households are thought to be aware of this. The health department recognizes the importance of a high response rate, but has only a modest amount of money to do the survey. Moreover, the survey will probably have to be conducted by staff of the health department who collectively have little survey experience.
a. *Briefly* discuss the relative merits of the three methods of data collection being considered for the survey: self-administered questionnaire by mail, telephone interview, and personal interview.

b. If either telephone or personal interviewing is the selected method, two-stage cluster sampling will be used to select the sample of households. The design effect in either case is expected to be about 1.5. Briefly describe what a "design effect" is and what things contribute to its size.

c. In the event that the two-stage design is used, determine the number of completed household interviews which would be required to yield a coefficient of variation of 10 percent. You can ignore the finite population correction.

Hint: Recall that the coefficient of variation for the estimator \( \hat{p} \) of the true population \( p \) is

\[
CV(\hat{p}) = \left[ \frac{Var(\hat{p})}{\hat{p}} \right]^{\frac{1}{2}}
\]

**Question 2.**

Let \( U \) be a random variable with p.d.f.

\[ f(u) = 1, \quad 0 < u < 1 \]

(a) Find the p.d.f. of \( X = -\lambda \log(1-U) \).

(b) Find \( E(X) \).

(c) Let \( Z \) be a random variable (independent of \( U \)) with p.d.f.

\[ g(z) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2} z^2), \quad -\infty < z < \infty \]

Let \( Y = e^{-Z} \) and \( W = -Y \log(1-U) \).

Find \( E(W) \).
Question 3.

An executive is willing to hire a secretary who has applied for a position unless a significance test indicates that she averages more than one error per typed page. A random sample of five pages is selected from some typed material by this secretary and the errors per page are: 3, 4, 3, 1, 2.

(a) Assuming that the numbers of errors, \( \lambda \), say, per page has Poisson distribution, what decision will be made? Use significance level \( \alpha \leq 0.05 \).

(b) Calculate the power when the average number of errors per page, \( \lambda \), is equal to 2.

(c) Suppose that the executive looks at a random sample of 225 pages of the secretary's work and finds 252 errors. What would be his decision, using \( \alpha \approx 0.05 \)?

Individual terms, \( e^{-m \frac{i}{i!}} \), of the Poisson distribution

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.36788</td>
<td>.13534</td>
<td>.00674</td>
<td>.00005</td>
</tr>
<tr>
<td>1</td>
<td>.36788</td>
<td>.27067</td>
<td>.03369</td>
<td>.00045</td>
</tr>
<tr>
<td>2</td>
<td>.18394</td>
<td>.27067</td>
<td>.08422</td>
<td>.00227</td>
</tr>
<tr>
<td>3</td>
<td>.06131</td>
<td>.18045</td>
<td>.14037</td>
<td>.00757</td>
</tr>
<tr>
<td>4</td>
<td>.01533</td>
<td>.09022</td>
<td>.17547</td>
<td>.01892</td>
</tr>
<tr>
<td>5</td>
<td>.00307</td>
<td>.03609</td>
<td>.17547</td>
<td>.03783</td>
</tr>
<tr>
<td>6</td>
<td>.00051</td>
<td>.01203</td>
<td>.14623</td>
<td>.06306</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>.00344</td>
<td>.10445</td>
<td>.09008</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>.00086</td>
<td>.06528</td>
<td>.11260</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>.03627</td>
<td>.12511</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>-</td>
<td>.01813</td>
<td>.12511</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>.00824</td>
<td>.11374</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>.00343</td>
<td>.09478</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>.00132</td>
<td>.07291</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-</td>
<td>.00047</td>
<td>.05208</td>
</tr>
</tbody>
</table>

Normal percentile point

\[ T_{.95} = 1.645 \]

\[ T_{.90} = 1.282 \]

\[ T_{.975} = 1.960 \]
Question 4.

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 $n (> 1)$ 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 $\text{Var}(\hat{\theta}_n)$.

Question 5.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be $n$ independent bivariate observations from a continuous bivariate distribution $F(x, y)$, $-\infty < x, y < \infty$. Let $H_0$ be the null hypothesis that $X$ and $Y$ are independent.

(a) Define $\tau = 2\Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1$ and show that under $H_0$, $\tau = 0$.

(b) Obtain the symmetric and unbiased estimator ($t_n$) of $\tau$ based on the $n$ observations and deduce the expressions for $E(t_n|H_0)$ and $\text{Var}(t_n|H_0)$.

(c) What can you say about the large sample distribution of $n^{\frac{1}{2}}t_n$, when $H_0$ holds?

(d) What modifications to $t_n$ would you suggest to accommodate possible ties among the $X$'s and/or the $Y$'s?
PART II

INSTRUCTIONS:

a) This is an *open book* examination.

b) M.P.H. and terminal M.S.P.H. students are to answer *any two* questions during the two-hour period (2 pm - 4 pm). M.S. and M.S.P.H. students desiring to proceed to the doctoral program are to answer *three* questions of which not more than 2 should be from Group A - time period 2 pm - 5 pm.

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

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

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

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

Q.1. An investigator is planning a study to evaluate a new medication for the treatment of hypertension. She knows from past experience that for patients with hypertension the mean diastolic blood pressure is 105 mm., the standard deviation is 15 mm. and the correlation between two measurements is 0.7.

(a) Find the sample size needed if she uses each patient as his own control. Assume $\alpha = 0.05$, $\beta = 0.1$, a one-sided test and that she wants to detect a change in blood pressure of 10 mm.

(b) Suppose she takes a group of patients and randomly divides them into two groups. She will give one group the new drug and the other group will get no treatment. If she will compare the change in the blood pressure in the treated group to that in the control group, what sample size does she need? Use the same assumptions as above.

(c) Discuss the relative merits of the two designs.

(d) Suppose the design in part (b) is chosen and that a total of 10 patients will be used. Use the attached table of random numbers to prepare a randomization schedule such that 5 patients will be assigned to treatment and 5 to control. Please give details about how the table is used so that the grader can reconstruct your schedule.

EDITORIAL NOTE: An attached table presented 3000 random digits in 60 rows of 50 digits each.
Q.2. An experiment was conducted to determine whether selenium supplementation is associated with reduced incidence of benign ovarian tumors in pregnant cows. One treatment group and one control group, of approximately equal sizes (N = 25 in each) were used. Each cow in the treatment group received the same amount of selenium, injected once, a fixed number of weeks before the end of pregnancy.

To verify that the treatment raised the blood levels of two important proteins throughout pregnancy, for each cow blood samples were taken before injection and after the end of pregnancy. The concentrations of the proteins were determined at each of these two times for each cow. The important questions of interest here are whether blood levels were similar in the two treatment groups before treatment, whether these blood levels changed between treatment and the end of pregnancy within each treatment group, and whether the change was greater in the treated than in the control group if the latter also had a change. (If the treatment is effective, blood levels should increase.)

For each protein, the investigators determined whether differences existed by the use of two-way ANOVA (with factors treatment, and time when blood drawn), followed by the use of Tukey's multiple comparison method with $P = .05$.

(a) Evaluate the method of analysis. Are the required assumptions met? Does the analysis answer the questions of interest?

(b) If you find the current analysis inappropriate, propose a better one, showing how to answer the primary questions with level $P = .05$. If you find the current analysis appropriate, discuss the use of Tukey's multiple comparison test vs. some other method to answer the primary questions.

Q.3.
A. Briefly describe the purpose(s) of:
   (1) JCL
   (2) DATA step of SAS
   (3) PROC step of SAS

B. Define, and describe the relationships between:
   OS file (dataset)
   SAS database
   SAS dataset
   (Examples of corresponding JCL and SAS code may be useful.)

C. List the major components of a large modern computer (CPU, etc.), and briefly describe the functions of each component and describe the relationship among them. A simple diagram may help you in organizing your answer.

D. List each type of JCL statement and briefly describe the function of each. Write a valid job (or jobs) including at least one example of each type of statement.
Q.4. A dentist who was responsible for dental care of cerebral palsied children in a state institution wanted to determine whether he should recommend that electric toothbrushes be purchased for routine use by the patients. He wanted to be as objective as possible in arriving at a decision, and decided to consult with a statistician about designing an experiment to determine whether short term improvement in oral hygiene could be demonstrated. In answering the major question, "Should the purchase of electric toothbrushes be recommended for this institution?", there are other considerations over and above any real improvement in oral hygiene of patients which should be taken into account but these were ignored in designing the study.

**Study Design and Conduct of Trial**

It was decided that the study should be designed to determine whether brushing with electric toothbrushes resulted in "cleaner teeth" than brushing with regular toothbrushes during a two week period. First, a search of the literature for measures of tooth cleanliness resulted in a decision to use the debris index, which is an average of debris scores for six teeth,* as the response variable, i.e., the variable which is to be altered (hopefully) by "treatment". Next, factors which could (potentially) influence results were listed:

1. Age
2. Race
3. Sex
4. Degree of ability to care for teeth (brushes own teeth or brushed by nurse)
5. Initial level of cleanliness (debris index)
6. Placebo effects:
   (a) Attitudes and actions of children and nurses
   (b) Attitude and actions of examining dentist

Because only 35 children were available for the initial examination it was impractical to stratify (or control) on all of these variables. However, randomization, a way of balancing the effect of variables which cannot be controlled, was used.

The study was carried out as follows:

1. Each child was examined by the dentist and a pre-trial debris index was determined using the following 3x5 form for recording.

2. The children were stratified by sex(ward), and degree of disability, i.e., divided into four groups:
   (a) Male - brushes own teeth
   (b) Male - assisted by nurse
   (c) Female - brushes own teeth
   (d) Female - assisted by nurse

3. Within each group children were randomly assigned to one of two brushing groups:
   (a) Electric toothbrush
   (b) Regular tooth care

The assignments were not disclosed to the dentist, i.e., he was "blind" as to type of care each child received.

4. A list of children assigned to the two groups was posted in each ward and the nurses supervised (and assisted where necessary) to see that assignments were followed.

5. At the end of the two week trial another examination was made by the dentist, who followed the same procedure as in the pre-trial examination to determine a debris index for each child. Results were recorded on another 3x5 card without reference to results of the original examination.

6. The results were matched with those from the first examination. Actual results are shown on the following page.

PROBLEM

(a) Test the statistical significance of the decline in debris index observed in each group, and with electric toothbrushing as compared with regular.

(b) Write a brief report, aimed at the dentist and the director of the state institution, describing the results and their analysis.
### Results of Two Week Trial of Regular and Electric Tooth Brushing, in Cerebral Palsy Hospital Patients

**Regular Tooth Brush**

<table>
<thead>
<tr>
<th>Child Number</th>
<th>Debris Index Before</th>
<th>Debris Index After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.0</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>9</td>
<td>1.5</td>
<td>0.8</td>
<td>-0.7</td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
<td>1.5</td>
<td>+0.5</td>
</tr>
<tr>
<td>16</td>
<td>1.5</td>
<td>Disch.</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>0.5</td>
<td>1.0</td>
<td>+0.5</td>
</tr>
<tr>
<td>28</td>
<td>2.5</td>
<td>2.2</td>
<td>-0.3</td>
</tr>
<tr>
<td>31</td>
<td>2.2</td>
<td>1.2</td>
<td>-1.0</td>
</tr>
<tr>
<td>34</td>
<td>1.7</td>
<td>1.0</td>
<td>-0.7</td>
</tr>
<tr>
<td>4</td>
<td>2.5</td>
<td>2.2</td>
<td>-0.3</td>
</tr>
<tr>
<td>14</td>
<td>0.7</td>
<td>0.5</td>
<td>-0.2</td>
</tr>
<tr>
<td>39</td>
<td>2.0</td>
<td>2.0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1.8</td>
<td>0.8</td>
<td>-1.0</td>
</tr>
<tr>
<td>24</td>
<td>1.8</td>
<td>0.8</td>
<td>-1.0</td>
</tr>
<tr>
<td>26</td>
<td>1.5</td>
<td>1.8</td>
<td>+0.3</td>
</tr>
<tr>
<td>29</td>
<td>0.7</td>
<td>0.5</td>
<td>-0.2</td>
</tr>
<tr>
<td>32</td>
<td>1.0</td>
<td>0.7</td>
<td>-0.3</td>
</tr>
<tr>
<td>33</td>
<td>2.8</td>
<td>2.3</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

**Electric Tooth Brush**

<table>
<thead>
<tr>
<th>Child Number</th>
<th>Debris Index Before</th>
<th>Debris Index After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.3</td>
<td>+0.3</td>
</tr>
<tr>
<td>3</td>
<td>2.2</td>
<td>Disch.</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>0.8</td>
<td>Disch.</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>1.3</td>
<td>Disch.</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>1.5</td>
<td>1.2</td>
<td>-0.3</td>
</tr>
<tr>
<td>38</td>
<td>1.7</td>
<td>1.5</td>
<td>-0.2</td>
</tr>
<tr>
<td>25</td>
<td>2.5</td>
<td>1.0</td>
<td>-1.5</td>
</tr>
<tr>
<td>30</td>
<td>2.3</td>
<td>Disch.</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>2.0</td>
<td>0.8</td>
<td>-1.2</td>
</tr>
<tr>
<td>23*</td>
<td>2.2</td>
<td>0.8</td>
<td>-1.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Child Number</th>
<th>Debris Index Before</th>
<th>Debris Index After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>1.7</td>
<td>1.3</td>
<td>-0.4</td>
</tr>
<tr>
<td>5</td>
<td>0.9</td>
<td>Disch.</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>1.3</td>
<td>+0.3</td>
</tr>
<tr>
<td>7</td>
<td>2.0</td>
<td>1.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>17</td>
<td>2.5</td>
<td>2.0</td>
<td>-0.5</td>
</tr>
<tr>
<td>19</td>
<td>2.2</td>
<td>0.7</td>
<td>-1.5</td>
</tr>
<tr>
<td>10</td>
<td>1.0</td>
<td>1.2</td>
<td>+0.2</td>
</tr>
<tr>
<td>35</td>
<td>2.9</td>
<td>Disch.</td>
<td>-</td>
</tr>
</tbody>
</table>

**Sum**

<table>
<thead>
<tr>
<th>Debris Index</th>
<th>Sum Before</th>
<th>Sum After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.7</td>
<td>20.3</td>
<td>-5.9</td>
<td></td>
</tr>
</tbody>
</table>

**Sum of squares**

<table>
<thead>
<tr>
<th>Debris Index</th>
<th>Sum Before</th>
<th>Sum After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.53</td>
<td>31.81</td>
<td>6.17</td>
<td></td>
</tr>
</tbody>
</table>

(N corrected)

<table>
<thead>
<tr>
<th>Child Number</th>
<th>Debris Index</th>
<th>Sum of squares</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

*Child #23 was originally assigned to regular care group but transferred by nurses to electric toothbrush group.*

**Note:** "Disch." refers to children who were discharged before post-trial examination.
Q.5. Consider the following data (n=5) for a dependent variable Y and an independent variable X:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

\[ \Sigma X = 29, \quad \Sigma Y = 58, \quad \Sigma XY = 414, \quad \Sigma X^2 = 207, \quad \Sigma Y^2 = 832 \]

(i) Fit the simple linear regression model

\[ E(Y) = \beta_0 + \beta_1 X \]

(ii) Theory strongly suggests \( \beta_0 = 0 \). So fit the model \( E(Y) = \beta_1 X \).

(iii) Are there any differences between models (i) and (ii) as regards numerical results? If so, please specify.

(iv) Suppose two further (X,Y) points, (0,1) and (0,4), were collected. What impact would these have on the estimate of \( \beta_1 \) for model (ii)?

(v) Would the impact of these two data points be the same for model (i)? Explain in one or two brief sentences.

(vi) For model (ii), define

\[ h_i = \frac{X_i^2}{\sum_{k=1}^{n} X_k^2} \]

(denominator is summing over the data points for X). Take \( h_i \) to be the leverage that the observation \( Y_i \) has on the predicted value \( \hat{Y} \). This leverage is exerted through the spacing of the X values (i.e., the design), not through the actual observed value of \( Y_i \). In general \( \hat{Y}_i \) is a linear combination of \( Y \)'s where the coefficients are \( h \)-like terms.

Write a few brief sentences interpreting the formula for \( h_i \) and comment if the definition of \( h_i \) is compatible with your answer to part (iv).

Note: If \( h_i = 0 \), the observed value \( Y_i \) has no influence whatever on the predicted value \( \hat{Y}_i \). At the other extreme if \( h_i = 1 \), the predicted value \( \hat{Y}_i \) will always be the observed value \( Y_i \).

(vii) For model (i), define

\[ h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{k=1}^{n} (X_k - \bar{X})^2} \]

Again, interpret the formula for \( h_i \) in one or two sentences and comment whether the definition is compatible with your answer to part (v).
BASIC MASTER LEVEL WRITTEN EXAMINATION IN BIOSTATISTICS

PART II

Special Offer: May 29, 1982

INSTRUCTIONS:

a) This is an open book examination.

b) M.P.H. and M.S.P.H. students are to answer any two questions during the two-hour period (1:30 pm - 3:30 pm). M.S. students are to answer three questions of which not more than 2 should be from Group A - time period 1:30 - 4:30 pm.

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

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

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

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

EDITORIAL NOTE: This "special offer" was identical with the regular Part II given on 12 April 1981 (see pages 87-92).
MPH WRITTEN EXAMINATION IN BIOSTATISTICS

April 10, 1983
(1 PM - 5 PM)

INSTRUCTIONS:

a) This is an open-book "in class" examination.

b) Answer from each part (I and II) any two of the 3 questions which follow. (Thus 4 answers in all)

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

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

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

PART I

Q.1 Let $X_1, X_2, \ldots, X_n$ be a random sample from the density function

$$f_X(x) = (\alpha+1)x^\alpha, \quad 0 < x < 1, \quad \alpha > -1$$

a) Find a general expression for $E(X^r)$, and then use your result to find $V(X)$.

b) Find an estimator for $\alpha$ using the method of moments.

c) Find the maximum likelihood estimator of $\alpha$. 
Q.2 An epidemiologist has reason to believe that the remission time \( T \) (in years) for a certain rare form of cancer of the navel (i.e., "bellybutton") after a special chemotherapy treatment can be modelled using the cumulative distribution function

\[
F_T(x) = 1 - e^{-x}, \quad x > 0.
\]

a) What is the average remission time based on this distribution?

b) What proportion of all patients can be anticipated to be in remission for at least 6 months?

c) For five independently selected patients, what is the probability that no more than three of them will stay in remission for at least 6 months? (JUST SET UP YOUR ANSWER)

d) What remission time will be exceeded by half of all patients based on the assumed remission time distribution?

Q.3 With \( i = 1, 2, \ldots, n \) indexing the counties of a state, let \( x_i \) be the number of sudden infant deaths (SIDs) from among \( B_i \) births in the \( i \)-th county during a calendar year.

a) Give an estimator for the SID syndrome rate in the \( i \)-th county.

b) Assuming the number of SIDs in each county is independently and Poisson distributed, construct a weighted average estimator of a common SID syndrome rate with the data from all the counties.

c) Again if the number of SIDs is Poisson distributed, derive the maximum likelihood estimator (MLE) for a common SID syndrome rate with the data from all the counties.

d) What is the variance of the MLE estimator of the common SID syndrome rate?

PART II

Q.4 Assume that \( n=100 \) individuals participate in a study. A response variable \( Y \) is measured, along with a continuous factor \( X_1 \) and the presence or absence of a factor \( X_2 \). The individuals come from four groups (A,B,C,D), coded as \( X_3 \).

a) Suppose that a regression model is fit, predicting \( Y \) from the three main effects, all two-way interactions, and the three-way interaction. Fill in the summary ANOVA table and the degrees of freedom in the detailed table. The second table gives the extra sums of squares as each factor is added to the model (the SAS Type I sums of squares).

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td>1500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected total</td>
<td></td>
<td>2340</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. What is the predicted value of \( Y \) for a subject in Group A with \( X_1 = 10 \) if \( X_2 \) is absent? If \( X_1 = 10 \) and \( X_2 \) is present?

ii. What is the predicted change in \( Y \) if \( X_1 \) changes from 10 to 5 in an individual in Group A with \( X_2 \) present?

e) Suppose that the mean values for \( Y \) for \( X_2 \) present and absent, in each group, are as shown in the following table. Assess informally whether there is an interaction between group and the presence or absence of \( X_2 \). Do not do an hypothesis test.

<table>
<thead>
<tr>
<th>Group</th>
<th>( X_2 ) Absent</th>
<th>( X_2 ) Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>B</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

b) Show that the three-way interaction term can be deleted from the model.

c) Test whether all the two-way interaction terms could simultaneously be deleted from the model not containing the three-way interaction.

d) Assume that \( X_2 \) is coded as 0 (absent) or 1 (present), and that a model is fit using only the data from group A. This model contains main effects for \( X_1 \) and \( X_2 \), and the \( X_1 X_2 \) interaction term, with parameter estimates for \( X_1 X_2 \) of \(-2\), \(+10\), and \(+4\), respectively. The estimate of the overall mean is 50.
Q.5

Some or all of the JCL statements below contain one or more errors. Circle each error, the whole error, and nothing but the error, and briefly explain what is wrong. Consider each statement separately. Assume that blanks within the lines occur only where the symbol "_" appears. None of the errors involves "1" versus "0" or "0" versus "0".

1. //JOB_unc.B.E1234,Hosking
2. //EXEC_PGM=COPY,WASTE=YES
3. //PHRED DD UNIT=DISK,DISP=(NEW,DELETE),VOL=REF=UNCCC.OFFLINE,
   SPACE=(TRK,(10,20)),DCB=(RECFM=FB,BLKSIZE=600,LRECL=250),
   DSN=UNC.B.E1234.JONES.X
4. //COPYDISKS_JOB_unc.B.E1234,Brown,T=(45),M=0
5. //EXEC_PGM=COPY,PARM=LIST
6. //OUTPUT_DD DSN=UNC.B.E9999.JONES.MYSTUFF,UNIT=TAPE,DISP=(NEW,CATLG),
   RING=IN,LABEL=(3,SL)
7. //JOE_DD UNIT=DISK,DSN=UNC.B.E1234.DATA.407,DISP=OLD,DCB=(RECFM=FB,
   LRECL=80,BLKSIZE=6000)
8. //SYSOUT_DD DSN=UNC.B.E1234.HELMS.STUFF,
   VOL=REF=UNCCC.ONLINE,
   DCB=(RECFM=VB,BLKSIZE=6000,LRECL=5996),
   SPACE=(TRK,(10,5,RLSE)),UNIT=DISK,
   DISP=(NEW,CATLG,DELETE)
9. //INPUT_DD DSN=UNC.B.E5001.SMITH.INDATA.ONTAPE,LABEL=(3,SL)
   DISP=OLD
10. //JOB#1_JOB_unc.B.E1234_SMITH,REGION=999K,MSGLEVEL=(1,0),TIME=5
b) Describe the processing of the DATA step listed below. Your answer should indicate a description of both the compilation and the execution phase, and a detailed description of the SAS data set created. Your answer should include the following terms/concepts:

- input buffer
- compilation
- execution
- program data vector
- variable name
- history information
- length
- type
- data matrix
- missing values
- EBCDIC/floating point representation

DATA WORK.ONE;
  LENGTH NAME $10 QUIZ1-QUIZ3 4;
  INPUT NAME $1-10
    QUIZ1 12-14
    QUIZ2 16-18
    QUIZ3 20-22
  ;
  AVERAGE=MEAN(QUIZ1, QUIZ2, QUIZ3);
  OUTPUT;
  RETURN;
  CARDS;
Q.6 For the period 1 July 1974 through 30 June 1978, North Carolina experienced a sudden infant death (SID) rate of two per thousand live births.

a) For a county having 3000 live births and 12 SIDS in the same calendar period, calculate the P-value for the possibility that the SID rate in this county is greater than that of North Carolina. By analogy to a binomial situation, use the normal distribution to approximate a presumed underlying Poisson model.

b) Let $\lambda$ denote the SID rate for a county with 3000 live births during the above specified calendar period and consider the null hypothesis $H_0: \lambda = 0.002$, i.e., $\lambda$ is two SIDS per thousand live births. Determine an upper tail critical region with a level of significance $\alpha = 0.05$. Again use a normal approximation.

c) Following part (b), calculate the power corresponding to the alternative hypothesis $H_a: \lambda = 0.006$.

d) Comment briefly on the appropriateness of the normal approximation to the Poisson distribution for the above calculations. Sketch how you would proceed in parts (a), (b), and (c) if the normal approximation were not appropriate.
INSTRUCTIONS:

a) This is an *open book* examination.

b) Students are to answer *any two* questions.

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

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

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

f) You are required to answer *only what is asked in the questions* and *not all you know about the topics.* Also, *do not answer* all the three questions.

1. Suppose that you are asked to help design the sample for a survey of taxpayers in a town with \( N = 2000 \) taxpayers. Town officials want to estimate the proportion (\( P \)) of taxpayers who favor building a small community hospital and we suspect that \( 0.4 < P < 0.6 \). You consider two sampling designs. One is simple random sampling from a list of taxpayers and the other is two-stage cluster sampling with blocks as the PSU. You are able to determine that the coefficient of variation for the estimate (\( \hat{p} \)) should be 5 percent.

   a. Recalling that for simple random samples,
      
      \[
      \text{Var}(\hat{Y}_o) = \frac{1-f}{n} \sigma^2,
      \]

      where \( f = n/N \) and \( n \) is the sample size, what would your recommendation for \( n \) be if simple random sampling was the design you chose?

   b. What is the "finite population correction" and how much does it affect your recommendation in part (a)?

   c. Would your recommended sample size be larger or smaller if the two-stage cluster sample was used instead of the simple random sample? Briefly explain your answer.
2. A test for the presence or absence of gout is based on the serum uric acid level \( X \) in the blood. Assume that \( X \) is a normally distributed random variable. In healthy individuals, \( X \) has a mean of 5.0 mg/100 ml and a standard deviation of 1.0 mg/100 ml; in diseased persons, \( X \) has a mean of 8.5 mg/100 and a standard deviation of 1.0 mg/100 ml.

A certain test \( T \) for the presence or absence of gout classifies those persons with a serum uric acid level of at least 7.0 mg/100 ml as having gout and classifies those persons with a serum uric acid level less than 7.0 mg/100 ml as not having gout.

a. In epidemiologic terminology, the "sensitivity" of a test for a disease is defined to be the probability that the test detects the disease when it is actually present. What is the sensitivity of the test \( T \) as defined above? (NOTE: A numerical answer is required.)

b. The "specificity" of a test is defined to be the probability of the test not indicating the presence of disease when there is no disease. What is the specificity of the test \( T \) as defined above? (NOTE: A numerical answer is required.)

c. If a randomly selected group of \( n = 1,000 \) individuals, all of which are disease-free, are examined using the test \( T \), what is the approximate probability that no more than 20\% of them will be incorrectly classified as having gout? (NOTE: A numerical answer is required.)

d. If actually 100 out of the 1,000 disease-free individuals in part (c) are incorrectly classified as having gout, find a 95\% confidence interval for the true underlying specificity of the test \( T \).

3. Let \( \theta \) be a parameter and \( \hat{\theta} \) be an estimator of \( \theta \). Define the mean squared error of \( \hat{\theta} \).

b. Let \( MSE (\hat{\theta}) \) be the mean squared error of \( \hat{\theta} \); show that
\[
MSE (\hat{\theta}) = \text{Variance} (\hat{\theta}) + [\text{Bias} (\hat{\theta})]^2.
\]

c. Suppose \( Y_1, Y_2 \) denotes a sample size of 2 from an exponential distribution with pdf
\[
f(y) = e^{-\frac{1}{\theta}y} \quad y > 0; \theta > 0.
\]
Find the MSE of \( \hat{\theta} = \frac{Y_1 + 2Y_2}{3} \).
PART II

INSTRUCTIONS:

a) This is an open book examination.

b) M.P.H. students are to answer any two questions during the two-hour period (1:30 pm - 3:30 pm). M.S. students are to answer three questions of which not more than 2 should be from Group A - time period 1:30 pm - 4:30 pm.

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

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

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

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

Group A

1. A randomized clinical trial for hypertension is to be designed to compare four treatment modalities in a parallel group design. The physician in charge of the trial wants to know the correct sample size to use. Review of past trials shows that the standard deviation for diastolic blood pressure phase V of entering patients is around 6 mm Hg. The physician feels that differences between treatment groups of 3 mm Hg must be detectable by the experiment.

Perform some power calculations using the attached charts to help identify appropriate sample sizes for the following situations:

(I) If any two treatments differ by 3 mm Hg, then the 4 treatment group comparison with 3 d.f. for the numerator of the F-test should be significant.

(II) If any two treatments differ by 3 mm Hg, then the pairwise difference based on the t-test should be significant.

(III) If any two treatments differ by 3 mm, then a Bonferroni adjusted pairwise difference (based on the fact that six pairwise differences are being inspected) should be significant.

EDITORIAL NOTE. The "attached charts" were those given on pages 115 and 116 in E.S. Pearson and H.O. Hartley, "Charts of the power function for analysis of variance tests, derived from the non-central F-distribution", Biometrika 38 (1951) 112-130.
2. A teacher wishes to determine the value of providing a manual and/or certain notes to his classes. He has 48 students, whom he distributes at random among 4 different groups, placing 12 in each. The assignment of teaching aid combinations to groups is also done at random. After the course is over, all students who are still enrolled take the same exam, with results as shown below.

<table>
<thead>
<tr>
<th>Manual</th>
<th>Notes</th>
<th>N</th>
<th>Data</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
<td>60, 64, 67, 68, 68, 69, 71, 73, 75</td>
<td>68.33</td>
<td>4.53</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>10</td>
<td>32, 41, 44, 47, 48, 54, 54, 64, 65, 73</td>
<td>52.20</td>
<td>12.42</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>11</td>
<td>41, 44, 47, 49, 51, 54, 56, 59, 61, 61, 73</td>
<td>54.18</td>
<td>9.16</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>7</td>
<td>44, 51, 55, 59, 59, 66, 69</td>
<td>57.57</td>
<td>8.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>3</td>
<td>1456</td>
<td>485.3</td>
</tr>
<tr>
<td>Within</td>
<td>33</td>
<td>2831</td>
<td>85.8</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>4287</td>
<td>119.1</td>
</tr>
</tbody>
</table>

(These data are from W.C. Guenther, Analysis of Variance, page 44.)

a) Assuming that differences in average grade may be attributed solely to the different combinations of teaching aids, write out a linear model for this experiment. Estimate all its parameters.

b) Consider two simplifications of the model:
1) Eliminate interaction, making it additive.
2) Combine all groups except Yes/Yes.

Show that the data will support the second simplification but not the first.

c) Present an argument for transforming these data. What would be a good transformation to try? (Do not actually perform any analysis with transformed data.)

d) Criticize the experimental design, and/or the use of standard analysis of variance with it. Can you suggest improvements?
3. Your answers to Question 3 must be submitted on these sheets. NO MATERIAL ON ANY OTHER SHEET WILL BE CONSIDERED!

You may use the back of the sheet if necessary, but enough space has been left under each part to provide the expected answer for that part.

Parts (a)-(e) of this question are worth three points each. In each case, the DATA step shown is executed reading one or both of the SAS data sets shown below (note that both are sorted BY ID). For each step, list the data part of the data set created (include variable names as shown below), and answer the questions presented.

<table>
<thead>
<tr>
<th>data set ONE</th>
<th>data set TWO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>X</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>ID</td>
<td>X</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Output Dataset

a) DATA A;
    SET TWO;
    Y=SUM(X,Z);
    IF (X LT Z) THEN OUTPUT;
    RETURN;

How many times is the DATA step executed?  

__________
b) DATA B;
   SET TWO ONE(RENAME=(Y=Z));
   IF (X GE 4);
   Q=X+Z;
   DROP X;
   OUTPUT;
   RETURN;

   What variables are in the PDV?

   Output Dataset


c) DATA D;
   MERGE TWO ONE;
   BY ID;
   IF LAST.ID THEN OUTPUT;
   RETURN;

   Output Dataset


How many observations would be in the data set if the OUTPUT statement was replaced by a DELETE statement (following the IF/THEN)?
d) DATA D;
SET TWO;
DROP X Z;
VAR='X';
VAL=X;
OUTPUT;
VAR='Y';
VAL=Y;
OUTPUT;
RETURN;

What is the length of VAL? Why?

---

e) DATA F;
SET ONE;
NUM=PUT(ID,WORDS5.);
AVG=(X+Y)/2;
MEAN=MEAN(X,Y);
DROP ID X Y;
OUTPUT;
RETURN;

What is the length of NUM? Why?
Parts (f) and (g) are worth five points each.

In parts (f) and (g) of this problem, you are only asked to compile the descriptor part of the data set, not to execute it or create the data part.

For each of the DATA steps in parts (f) and (g) below, fill in the information SAS would store in the descriptor part of the data sets being created. Sketch the Program Data Vector that would be created, indicating the type and length of each variable in it. When appropriate, indicate the existence and size of input and/or output buffers.

f) DATA F;
   INPUT NAME $ FNAME $ SATM SATV;
   SATTOT=SATM+SATV;
   IF(SATTOT GT 1000) THEN STATUS='ADMIT';
   ELSE STATUS='REJECT';
   FILE PRINT;
   PUT NAME $ 1-10 STATUS $ 12-2U;
   OUTPUT;
   RETURN;
CARDS;

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work.D</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Length</th>
<th>Informat</th>
<th>Format</th>
<th>Label</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Length</th>
<th>Informat</th>
<th>Format</th>
<th>Label</th>
</tr>
</thead>
</table>
DATA G;
INFILE IN LRECL=50;
INPUT #1 @1 NAME $20.
   @21 BDATE MMDDYY8.
   @31 DDATE MMDDYY8.
#2 @1 CAUSE1 $6.
   @11 CAUSE2 $6.
;
FORMAT BDATE DDATE DATE.;
LENGTH SURVTIME 4;
SURVTIME=DDATE-BDATE;
OUTPUT;
RETURN;

PDV
Name:
Type:
Length:

WORK.E
Name:
Type:
Length:
Informat:
Format:
Label:
For the period 1 July 1974 through 30 June 1978, North Carolina experienced a sudden infant death (SID) rate of two per thousand live births.

(a) For a county having 3000 live births and 12 SIDs in the same calendar period, calculate the P-value for the possibility that the SID rate in this county is greater than that of North Carolina. Use the normal distribution to approximate a presumed underlying Poisson model.

(b) Let $\lambda$ denote the SID rate for a county with 3000 live births during the above specified calendar period and consider the null hypothesis $H_0: \lambda = 0.002$, i.e., $\lambda$ is two SIDs per thousand live births. Determine an upper-tail critical region with a level of significance $\alpha = 0.05$.

(c) Following part (b), calculate the power corresponding to the alternative hypothesis $H_a: \lambda = 0.006$.

(d) Comment briefly on the appropriateness of the normal approximation to the Poisson distribution for the above calculations. Sketch how you would proceed in parts (a), (b), and (c) if the normal approximation were not appropriate.

**APPENDIX**

| Table 1. Normal Probability Distribution Function (Probabilities That Given Standard Normal Variables Will Not Be Exceeded—Lower Tail) $N_z(-z)$. Also $N_z(z) = 1 - N_z(-z)$. |
|---|---|---|---|---|---|---|---|---|---|
| $-z$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
| .0 | .50000 | .49601 | .49202 | .48803 | .48405 | .48006 | .47608 | .47210 | .46812 | .46414 |
| .1 | .46017 | .45620 | .45224 | .44828 | .44433 | .44038 | .43644 | .43251 | .42858 | .42465 |
| .2 | .42074 | .41683 | .41294 | .40905 | .40517 | .40129 | .39743 | .39358 | .38974 | .38591 |
| .3 | .38209 | .37828 | .37448 | .37070 | .36693 | .36317 | .35942 | .35569 | .35197 | .34827 |
| .4 | .34458 | .34090 | .33724 | .33360 | .32997 | .32636 | .32276 | .31918 | .31561 | .31207 |

EDITORIAL NOTE. The table above has been abridged from one extended to $Z = 3.99$ which was attached to the original examination.
Question 1

Let the joint density function of two continuous random variables $X$ and $Y$ be

$$f_{X,Y}(x,y) = \frac{y + x\theta}{\theta^2}; \quad 0 < x < 1, \quad 0 < y < \theta.$$  

a) Show that the marginal distribution of $X$ is

$$f_X(x) = \frac{1}{2} + x, \quad 0 < x < 1.$$  

b) Show that the conditional distribution of $Y$ given $X = x$ is

$$f_Y(y|x) = \frac{2(y + x\theta)}{\theta^2(1 + 2x)}, \quad 0 < y < \theta.$$  

c) $E(Y|X = x)$ is often referred to as the "regression equation" of $Y$ on $X$. For this problem, show that

$$E(Y|X = x) = \theta z,$$

where $z = \frac{(2 + 3x)}{3(1 + 2x)}.$

d) Suppose that it is desired to estimate the parameter $\theta$ using $n$ pairs of data points $(x_i, Y_i)$, $i = 1, 2, \ldots, n$. A reasonable criterion for choosing an estimator $\hat{\theta}$ of $\theta$ is to choose $\hat{\theta}$ to minimize the quantity

$$\sum_{i=1}^{n} (Y_i - \theta z_i)^2.$$  

Show that the choice for $\hat{\theta}$ which minimizes the above expression is

$$\hat{\theta} = \frac{\sum_{i=1}^{n} z_i Y_i}{\sum_{i=1}^{n} z_i^2}.$$  

e) For the $n = 3$ pairs of $(X, Y)$ data points $(-1, 1), (0, 0)$ and $(1, 3)$, what is the value of $\hat{\theta}$?
Question 2

Suppose that a survey is done to estimate a rate \( P^* \) which can be expressed as a constant \( K \) times a proportion \( p \). Assuming simple random sampling (SRS), the appropriate estimator of \( P^* \) is \( p^* = Kp \), where \( p \) is the corresponding sample proportion.

Ignoring the finite population correction, verify the following properties of \( p^* \) for an element sample of size \( n \) from a population with \( N \) elements:

a) True Variance: \( \text{Var}(p^*) = \frac{NP^*(K-p^*)}{n(N-1)} \)

b) Coefficient of Variation: \( \text{CV}(p^*) = \frac{N(K-P^*)}{n(N-1)P^*} \)

c) Assuming that \( K = 10^3 \), \( P^* = 20 \), and \( N = 10^6 \) (so that the finite population correction can be ignored), how large a sample would be needed in the survey to yield a standard error of \( 5 \) for the estimator \( p^* \)?

Question 3

Let \( Y_1, Y_2, \ldots, Y_n \) denote a random sample from the probability density function:

\[
f(y) = \theta y^{\theta-1}, \quad 0 < y < 1; \quad \theta > 0.
\]

a) Find the maximum likelihood estimator for \( \theta \).

b) Find an estimate of \( \theta \) by the method of moments.

c) Show that \( \bar{Y} \) is a consistent estimator of \( \theta/(\theta+1) \).
INSTRUCTIONS:

a) This is an open book examination.

b) M.P.H. students are to answer any two questions during the three-hour period (2 PM - 5 PM). M.S. students are to answer three questions -- time period 1 PM - 5 PM.

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

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

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

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

Question 1.

Two faculty members independently graded the responses to an exam question answered by seven students. Each grader assigned a score between 0 and 25 to each response. Professor One stored his grades in a SAS dataset in a catalogued SAS data library stored on online disk at TCC. The PROC CONTENTS output for this dataset is attached. Professor Two stored his grades on a scrap of paper, reproduced below. Write a program, including all necessary JCL and SAS statements to do the following:

1) Produce a scatter plot of Professor One's score versus Professor Two's score, using the student's ID letter as the plotting symbol. Title and label the plot appropriately.

2) Produce a report listing each student whose two scores differ by five or more points. The listing should include the students ID, the two scores, and the difference between the scores.

NOTES: In your JCL use the account code UNC.B.E1234, a programmer name of SMITH, and a password of SECRET. You may assume that each student has exactly one record in each "file", and that no values are missing in either "file".
### CONTENTS OF SAS DATA SET ALPHA.EXAM

**TRACKS USED**: 2  **SUBEXTENSIONS**: 1  **OBSERVATIONS**: 7  
CREATED BY CS JOB LOGO  
DATE: 23-4081-621232  AT 16:23 WEDNESDAY, MARCH 20, 1985  
BY SAS RELEASE 9.4  
FILENAME=UT.EL.E512Y.CSE.EXAMPLE  
LIBNAME=19864  
LIBNAME=13  OBSERVATIONS PER TRACK=1466  
GENERATED BY DATA  

### ALPHA-HEBI LIST OF VARIABLES

<table>
<thead>
<tr>
<th>#</th>
<th>VARIABLE</th>
<th>TYPE</th>
<th>LENGTH</th>
<th>POSITION</th>
<th>FORMAT</th>
<th>INFORMAT</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>CHAR</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td>STUDENT IDENTIFICATION</td>
</tr>
<tr>
<td>2</td>
<td>GRADE</td>
<td>NUM</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
<td>EXAM SCORE FOR PROF. ONE</td>
</tr>
</tbody>
</table>

**Student Scores**

- Student 1: Score 22  
- Student 2: Score 14
Question 2.

Provide the P-value for the following comparison of two drugs, where the new drug is expected to be at least as effective as the old one.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Improved</th>
<th>Not Improved</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Drug</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Old Drug</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>13</td>
<td>20</td>
</tr>
</tbody>
</table>

a) Use Pearson's chi-square test, with and without Yates correction.

b) Use Fisher's Exact test.

c) Use the normal approximation to Fisher's Exact test.

d) Summarize briefly these results for the comparison of the two drugs.

Question 3.

a) What is the difference between regression with an intercept and regression through the origin?

b) Perform an appropriate regression analysis for the following data, X is carrier and Y the dependent variable.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
</tr>
</tbody>
</table>

\[ \Sigma X = 28 \quad \Sigma Y = 56 \quad \Sigma XY = 348 \]
\[ \Sigma X^2 = 174 \quad \Sigma Y^2 = 700 \]

c) Repeat the regression analysis of Part b using centered data.

d) Is there a difference between regression through the origin and regression using centered data? Explain.

e) Suppose a sixth data point for Part b was taken and it was X = 0, Y = 1. Which of the procedures you have considered (regression with intercept, regression through origin, regression with centered data) become inapplicable? Explain.
Question 4.

Consider a population composed of the following \( N = 30 \) elements:

\[ Y_1 = 1; \ Y_2 = 2, \ldots, Y_{30} = 30. \]

a) Compute the true variance of the mean of a simple random sample of size \( n = 6 \).

b) If the population were divided into \( H = 6 \) strata as follows:

\[ (Y_1, Y_2, Y_3, Y_4, Y_5); (Y_6, Y_7, Y_8, Y_9, Y_{10}); \ldots; (Y_{26}, Y_{27}, Y_{28}, Y_{29}, Y_{30}), \]

what is the true variance of the estimate of the population mean for a stratified simple random sample of \( n = 6 \) chosen by selecting \( n_h = 1 \) element in each stratum?

c) Verify whether or not the design in part (b) is ESEM.

HINT: Recall that \( \sum_{i=1}^{N} i = N(N+1)/2 \) and \( \sum_{i=1}^{N} i^2 = N(N+1)(2N+1)/6 \).
QUESTION 1

We want to investigate whether drugs A and B taken orally have any long term carcinogenic side effects. The response variable will be the number of malignant tumors, treated as a continuous variable. 150 mice are divided into three groups of 50 mice each (Control, A and B), and fed the appropriate diets for two years.

There is reason to believe that the carcinogenic effect is influenced by caloric intake of the mice over the duration of the experiment. Although we are not interested in this "caloric effect" per se, we need to take it into account. Assume that by weighing leftover mouse chow, a fairly good estimate can be provided for caloric intake.

We want to test the hypothesis that there is no difference between mice on the three diets.

1) What assumption needs to be made in order to use covariate analysis? Please state it in terms of the variables in the problem.

2) Write down the model needed to test the assumption in (1). State the null hypothesis for this test in terms of the parameters of your model. Define all your variables!

3) If you feel you cannot make the necessary assumption for covariate analysis, what different hypothesis could you test to distinguish between the three groups of mice? State the hypothesis in terms of the parameters of your model. Give the F-statistic used to test the hypothesis in terms of sums of squares associated with the variables in your model. Include degrees of freedom.

4) If the assumption in (1) is not rejected (so that covariate analysis can be used), write down a model that could be used to test whether or not there is a difference between the three diets. Define all variables and write the null hypothesis in terms of the parameters of your model.

5) Give the F-statistic used to test the hypothesis in (4) in terms of sums of squares associated with variables in your model. Include the degrees of freedom for the numerator and denominator.
QUESTION 2

Consider the following population of 1,000 individuals who have been divided into the following strata:

<table>
<thead>
<tr>
<th>h</th>
<th>( W_h )</th>
<th>( S^2_h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>25</td>
</tr>
</tbody>
</table>

where \( W_h \) is the proportion of the population falling in the \( h \)th stratum and \( S^2_h \) is the stratum-specific element variance. A proportionate stratified simple random sample of individuals is to be chosen for a survey whose dual purpose is to estimate the population mean per individual (\( \bar{Y} \)) as well as the difference in means between the first and second strata (i.e., \( D = \bar{Y}_1 - \bar{Y}_2 \)).

a. Briefly describe the steps you would follow in selecting the sample.

b. How large an overall sample (i.e., for all three strata combined) would be needed for the estimator (\( d \)) of \( D \) to have a standard error of 2?

c. Assuming per-unit interviewing costs to be equal among strata, what would be the relative sizes of the three stratum sampling rates \( (f_n = n_h/N_h) \) if optimum stratum allocation were used?

QUESTION 3

Let

\[ f_Y(y) = \theta e^{-\theta y}, \quad y > 0, \]

be the density function for a continuous random variable \( Y \). Let \( Y_1, Y_2, \ldots, Y_n \) be a random sample of size \( n \) from this continuous distribution.

a. What is the joint distribution (i.e., the likelihood function) of \( Y_1, Y_2, \ldots, Y_n \)?

b. Show that the maximum likelihood estimator of \( \theta \) is

\[ \hat{\theta} = (\bar{Y})^{-1}, \]

where \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \).

c. Show that \( \hat{\theta} \) is a sufficient statistic for \( \theta \).

d. Find an appropriate large sample 95% confidence interval for \( \theta \) if \( n = 100 \) and \( \bar{Y} = 4 \).
INSTRUCTIONS:

a) This is an open book examination.

b) Answer any two questions.

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

d) Put your CODE LETTER, (not your name), on each page.

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

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

QUESTION 1

a. Briefly describe or explain (in one or two sentences) the following terms:

(1) OS
(2) TSO
(3) Track (on magnetic disk)
(4) Block
(5) Parity bit
(6) Field
(7) JCL
(8) EBCDIC
(9) Collating sequence
(10) Backup

b. In a brief paragraph, possibly using diagrams or Figures, compare and contrast each group of terms:

(1) OS file, SAS data library, SAS dataset
(2) Logical record, physical record
(3) JCL step, SAS step
(4) DSN, ddname
(5) Program data vector, input buffer
QUESTION 2

A random sample of 230 individuals was typed for the MN blood group system, with results as follows:

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM ((f_1))</td>
<td>120</td>
</tr>
<tr>
<td>MN ((f_2))</td>
<td>101</td>
</tr>
<tr>
<td>NN ((f_3))</td>
<td>9</td>
</tr>
<tr>
<td>Total ((T))</td>
<td>230</td>
</tr>
</tbody>
</table>

The maximum likelihood estimates of the allele frequencies of M and N are given by

\[
\hat{p} = \text{freq}(M) = \frac{2f_1 + f_2}{2T}
\]

\[
\hat{q} = \text{freq}(N) = 1 - \hat{p}
\]

\[
\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{2T} = \text{Var}(\hat{q})
\]

At equilibrium, we expect the genotypic frequencies to be in the proportion

\[p^2 : 2pq : q^2\], \(MM:MN:NN\).

Is there evidence that the system deviates from equilibrium conditions?

QUESTION 3

Two exams were given in a course. The first exam had average score 50 and standard deviation 10. The second exam had average score 40 and standard deviation 15. The correlation between the two scores was +.4. If a student scored 60 points on the first exam, what score would you expect the student to have on the second exam? How good is your estimate?
INSTRUCTIONS:

a) This is an open book examination.

b) Students are to answer any two questions during the two hour time period (10 am - 12 noon).

c) Put the answers to different questions on separate set of papers.

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

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

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

NOTE: Question #1 is on the next page.

Question 2.

Let $X_1, X_2, \ldots, X_n$ be a random sample from the density function

$$f_X(x) = (a + 1)x^a, \quad 0 < x < 1 \text{ and } a > -1.$$ 

a) Show that $f_X(x)$ is a valid density function.

b) Find $E(X)$ and $V(X)$.

c) Find an estimator of $a$ using the method of moments.

d) Write down the joint distribution of $X_1, X_2, \ldots, X_n$.

e) Find the maximum likelihood estimator of $a$. 

Question 1.

Consider the linear model $Y = \mu(x_1, x_2) + \varepsilon$ where $\mu(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$. Suppose the experimental data is as follows:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>-1</td>
<td>-13</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-12</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

(a) Find the design matrix $X$.
(b) Find $X'X$.
(c) Find $(X'X)^{-1}$.
(d) Find the least square estimates of $\beta_1$ and $\beta_2$.
(e) Determine $s$, where $s^2$ is the usual unbiased estimate of $\sigma^2$.
(f) Find the two-sided 95% confidence interval for $\beta_1$ and $\beta_2$.
(g) Find the two-sided 95% confidence interval for $\beta_1 + \beta_2$ and $\beta_1 - \beta_2$.
(h) Determine approximate p-values for two-sided tests of each of the following hypotheses.
   $H$: $\beta_1 = 0$; $H$: $\beta_2 = 0$; $H$: $\beta_1 = \beta_2$.
(i) Calculate the F statistic and determine the proper numbers of degrees of freedom for the usual F test of the hypothesis $H$: $\beta_1 = 0$ and $\beta_2 = 0$.
(j) Determine a two-sided 95% confidence interval for $\mu(2,3)$ and a two-sided 95% prediction interval for a new response $Y$ corresponding to $x_1 = 3$ and $x_2 = 2$. 
Question 3.

A scientist hypothesizes the following esoteric theory for cellular damage in humans due to radiation exposure. A particular cell in an individual is "hit" by radiation X times during the course of the individual's lifetime, with the number X of such hits having the Poisson distribution

\[ p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \ldots, \infty. \]

For any such hit, there is a fixed probability p that some basic structural change will occur in the cell; also, the occurrences (or not) of structural changes for different hits are assumed to be mutually independent.

The random variable of interest is Y, the total number of structural changes that the particular cell undergoes during the course of the individual's lifetime.

a) What is \(\Pr(Y = y|X = x_0)\)? In other words, what is the conditional distribution of Y given that a particular cell experiences exactly \(x_0\) hits?

b) Using the result in part (a), show that the unconditional distribution of Y, namely \(\Pr(Y = y)\), is Poisson with mean \(\lambda p\).

c) The scientist further hypothesizes that a particular cell will become cancerous if it undergoes at least k structural changes. Given the result in part (b), what is the probability that a particular cell will become cancerous?

[NOTE: You do not need to be able to work parts (a) and (b) in order to answer this question.]
Let \( \bar{Y} \) denote the average number of children ever born to women aged 15-49 years in some population where a survey has been done. Three stage cluster sampling with stratification in the first stage was used to select the sample of women, 8,025 of whom participated in the survey. Properly computed estimates of \( \bar{Y} \), its standard error, and its "design effect" are 2.84, 4.24x10^-2 and 1.33, respectively.

a) Ignoring the finite population correction, estimate the element variance (\( S^2 \)) for the \( Y \)-variable from these results.

b) Calculate the coefficient of variation for the estimate of \( \bar{Y} \).

c) Suppose that, in adding values of the \( Y \)-variable over all sample respondents and dividing by 8,025, you get 2.67. Assuming that no computational errors are involved, what does this tell you about the estimator used to compute 2.84 above and about the sampling design of this survey?

d) Suppose that simple random sampling had been used in this survey. What is your estimate of the sample size that would have been needed to achieve the same level of sampling error as reported above?
Question 1.

The weights (kg) of 10 people before and after going on a high carbohydrate diet for 3 months are shown in the following table:

<table>
<thead>
<tr>
<th>Before</th>
<th>64</th>
<th>71</th>
<th>64</th>
<th>69</th>
<th>76</th>
<th>53</th>
<th>52</th>
<th>72</th>
<th>79</th>
<th>68</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>61</td>
<td>72</td>
<td>63</td>
<td>67</td>
<td>72</td>
<td>49</td>
<td>54</td>
<td>72</td>
<td>74</td>
<td>66</td>
</tr>
</tbody>
</table>

You want to know if, overall, there was significant change in weight. Let \( Y_{kj} \) denote the weight of the \( j \)th person at time \( k \) where \( k = 1 \) before the diet, \( k = 2 \) afterwards and \( j = 1, \cdots, 10 \). Also, let \( \beta_1 \) = mean weight before the diet and \( \beta_2 \) = mean weight afterwards.

(i) Write the weights by using a linear model:

\[
Y = X\beta + \varepsilon
\]

where \( Y' = (Y_{1,1}, \cdots, Y_{1,10}, Y_{2,1}, \cdots, Y_{2,10}) \), \( \beta' = (\beta_1, \beta_2) \).

and \( \varepsilon' = (\varepsilon_1, \cdots, \varepsilon_{20}) \) consists of independent normal \((0, \sigma^2)\).

(a) Find the design matrix \( X \).

(b) Test the hypothesis \( H_0 \) that there was no change in weight, that is, \( \beta_1 = \beta_2 \).

(ii) Let \( D_j = Y_{1j} - Y_{2j} \), for \( j = 1, \cdots, 10 \). If \( H_0 \) is true then \( E(D_j) = 0 \), so another test is to compare the models

\[
D_j = \mu + \varepsilon_j
\]

and

\[
D_j = \varepsilon_j
\]

assuming \( \varepsilon_j \) are independent normal \((0, \sigma^2)\) for \( j = 1, \cdots, 10 \). Use this set up to test the hypothesis \( H_0: \mu = 0 \), that is, there was no change in weight.

(iii) Discuss the assumptions you made for the analyses you made in (i) and (ii). Which analysis was more appropriate? Why?
Question 3.

A. Briefly describe or explain the following terms:
   (a) OS
   (b) TSO
   (c) Track on a magnetic disk
   (d) Block
   (e) Logical Record
   (f) Byte
   (g) JCL

B. Compare and contrast: (That is, demonstrate that you know what each of these terms means and what are the differences among them.)
   (a) OS dataset
   (b) SAS database (SAS data library)
   (c) SAS dataset

C. The printout in Figure 1 was produced by the PROC PRINT statement in line 00140002 of the SAS program shown in Figure 2. Show what will be printed as a result of
   (a) the PROC PRINT statement in line 00160002.
   (b) the PROC MEANS in statement 00180002 and related statements.
   (c) the PROC PRINT statement in line 00380002.
   (d) the PROC PRINT statement in line 00470002.

Figure 1

<table>
<thead>
<tr>
<th>OBS</th>
<th>NAME</th>
<th>SEX</th>
<th>AGE</th>
<th>HEIGHT</th>
<th>WEIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALFRED</td>
<td>M</td>
<td>14</td>
<td>69</td>
<td>112</td>
</tr>
<tr>
<td>2</td>
<td>ALICE</td>
<td>F</td>
<td>13</td>
<td>56</td>
<td>84</td>
</tr>
<tr>
<td>3</td>
<td>BERNADETTE</td>
<td>F</td>
<td>13</td>
<td>65</td>
<td>98</td>
</tr>
<tr>
<td>4</td>
<td>BARBARA</td>
<td>F</td>
<td>14</td>
<td>62</td>
<td>102</td>
</tr>
<tr>
<td>5</td>
<td>JAMES</td>
<td>M</td>
<td>12</td>
<td>57</td>
<td>83</td>
</tr>
</tbody>
</table>
Figure 2

```sas
//MSEXAM1 JOB UNCG.B.E5111,GALLAGHER,MSGLEVEL=(1,1)
//*PW-GLASS
//STEP1 EXEC SAS
//REX DD *
M 14 69 112 ALFRED
F 13 56 84 ALICE
F 13 65 98 BERNADETTE
F 14 62 102 BARBARA
M 12 57 83 JAMES
//SYSIN DD *
DATA STUDENT1;
  INFILE REX;
  LENGTH NAME $ 20 SEX $ 1;
  INPUT SEX AGE HEIGHT WEIGHT NAME;
PROC PRINT; * This produces Figure 1;
PROC SORT; BY SEX AGE HEIGHT;
PROC PRINT; * Problem C(a);
PROC MEANS MEAN N; * Problem C(b);
          BY SEX;
          VAR AGE HEIGHT;
PROC SORT; BY NAME SEX AGE;
DATA CHANGES;
          LENGTH NAME $ 20 SEX $ 1;
          MISSING _;
          INPUT SEX AGE HEIGHT WEIGHT NAME;
CARDS;
   F 13 50   . ALICE
   M 12   . 89 JAMES
   M 14   . ALFRED
   F 14   1   . BARBARA
   F 16 62 102 BARBARA
   F 12 59 84 JANE
* End of change data;
PROC PRINT; * Problem C(c);
PROC SORT; BY NAME SEX AGE;
DATA STUDENT2;
  UPDATE STUDENT1 CHANGES;
        BY NAME SEX AGE;
        IF HEIGHT > 0;
PROC PRINT; * Problem C(d);
```
Question 4.

Suppose that the regression function of \( Y \) on a real-valued \( x \) is continue; suppose also that it is linear on \( x < 0 \) and \( x \geq 0 \) separately, but with possibly different slopes on these two intervals. That is, we can write

\[
Y = \mu(x; \beta) + \varepsilon
\]

where

\[
\mu(x; \beta) = \beta_1 + \beta_2 x, \quad x < 0,
\]

\[
= \beta_1 + \beta_3 x, \quad x \geq 0.
\]

An experiment is conducted in which one measurement of \( Y \) is made at each of 21 settings \( x_i, i = 1, \ldots, 21 \), which start at -10 and increase by one unit at a time (so that \( x_{21} = 10 \)). Assume \( Y_1, \ldots, Y_{21} \) are independent and normally distributed with constant variance \( \sigma^2 > 0 \).

(a) Determine the design matrix \( X \) and \( XX \).

The inverse of the \( X'X \) matrix is

\[
(X'X)^{-1} = \frac{1}{783475} \begin{pmatrix}
148225 & 21175 & -21175 \\
21175 & 5060 & -3025 \\
-21175 & -3025 & 5060
\end{pmatrix}
\]

Use this explicit part in solving the remaining parts of the problem. Also use the exact distributions rather than normal approximation.

(b) Determine explicitly, in terms of \( \hat{\beta}_3 \) and \( \hat{\sigma} \), the two-sided 95% confidence interval for \( \beta_3 \).

(c) Determine explicitly the two-sided 95% confidence interval for \( \beta_3 - \beta_2 \).

(d) Describe explicitly in terms of \( \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma} \) how to obtain the p-value for the two-sided test of the hypothesis that \( \beta_3 - \beta_2 = 0 \); or equivalently, that the regression function is globally linear.

(e) Describe explicitly in terms of \( \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma} \) how to obtain the p-value for the two-sided test of the hypothesis that \( \beta_3 = \beta_2 = 0 \); i.e., that the regression function is globally constant.

(f) Let \( s^2 = \frac{1}{21} \sum_{i=1}^{21} (Y_i - \bar{Y})^2 \) denote the sample variance for the \( Y_i \)'s. Describe explicitly in terms of \( s \) and \( \hat{\sigma} \) how to obtain the p-value for the two-sided test of the hypothesis stated in (f).
PROBLEM 1:

The statures of 18 pairs of brothers were measured to the nearest inch, and the following set of results was obtained:

<table>
<thead>
<tr>
<th>65,67</th>
<th>66,68</th>
<th>72,73</th>
</tr>
</thead>
<tbody>
<tr>
<td>69,67</td>
<td>70,74</td>
<td>65,71</td>
</tr>
<tr>
<td>70,71</td>
<td>65,68</td>
<td>66,66</td>
</tr>
<tr>
<td>67,68</td>
<td>68,69</td>
<td>69,64</td>
</tr>
<tr>
<td>71,72</td>
<td>70,70</td>
<td>69,71</td>
</tr>
<tr>
<td>68,67</td>
<td>70,68</td>
<td>67,63</td>
</tr>
</tbody>
</table>

State a model for these data, perform the analysis, and infer that "the statures of brothers tend to be similar". (This is the conclusion drawn from these data by one of the standard textbooks in statistics.) It is given that the sum of the 36 statures is 2464 and that the sum of their squares is 168878.
PROBLEM 2:

The wireworm population of a field is estimated by taking 4-inch cylindrical cores at random points in the field and counting the number of wireworms in each core. Suppose a mean of 3.52 wireworms, with a standard deviation of 3.15, is found in a sample of 50 cores.

a) Assuming normality, how closely can the true population be determined with 95% certainty?

b) How reasonable is the assumption of normality? Suggest something more reasonable, and test its fit.

PROBLEM 3:

Suppose that $X_1, X_2, ..., X_n$ are a random sample from

$$f_X(x) = 2x, 0 < x < 1.$$  

a) What is the exact form of the joint distribution of $X_1, X_2, ..., X_n$?

b) If $S = \frac{1}{n} \sum_{i=1}^{n} X_i$, find $E(S)$ and $V(S)$.

c) For $n = 1800$, find values for $a$ and $b$ such that

$$\Pr(a < S < b) = 0.95.$$  

(BE SURE TO STATE CLEARLY ANY ASSUMPTIONS YOU MAKE.)

d) For $n = 2$, what is the exact value of

$$\Pr(X_1 < 2X_2)?$$
INSTRUCTIONS:

a) This is an open book examination.

b) M.P.H. students are to answer any two questions during the two-hour period (11:00am - 2:00pm). M.S. students are to answer three questions of which not more than 2 should be from Group A - time period 11:00am - 3:00pm.

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

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

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

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

THE SCENARIO:

Imagine that you find yourself in the following situation:

Having been graduated by the School of Public Health here as a brand new MPH in Biostatistics, you accept a good paying job with a clinical research group that is long on clinical experience and acumen (that is, they know medicine very well indeed) but is frighteningly ignorant of the data collection and data management techniques that will be required for the field study of disease and sanitation in rural institutional settings for which they have funding. As a result of your all-morning meeting with the group on Monday, May 16th, you realize that your group intends to charge you with responsibility for everything that goes wrong with any part of the data collection and management as well as dramatically underfund those activities. You instantly regret bitterly not having found time to take BIOS 213. Thinking as brilliantly as a graduate of this department should, you plan to devise in your lunch hour before the afternoon meeting a presentation to educate your clinician colleagues to both provide adequate funds for the data collection and management as well as send you to the short course on Data Collection, Data Management, and Data Analysis for Clinical Trials at Johns Hopkins on June 9-10.

Here is your plan: by sheer dumb luck you have in your backpack a copy of the BIOS 213 course outline (partially reproduced below). You decide to pick the one data collection/management topic on the outline about which you know the most and prepare the most scholarly and well illustrated discussion of the topic that you can. You realize that your exposition must simultaneously teach your colleagues

- that data collection/management requires as much (albeit different) skill and professionalism as do their professions,
- that the complexity and number of tasks that must be done correctly require the allocation of adequate funds and resources, and
- that you know enough about the topic to be right when you say that, if you are expected to be responsible for the data management, as a bare minimum you require the additional training that the Johns Hopkins program offers.
ASSIGNMENT:
Select at least one topic from the list which follows.
Prepare a convincing discussion aimed at intelligent scientists who are, unfortunately, ignorant of data management problems.
Point out the problems,
the techniques for handling the problems,
the consequences of mishandling the problems, and
the cost in time, money, and personnel.

Please do your best to remember that the possible graders of this problem have all worked in an office with a W.C. Fields poster with the slogan which said (in paraphrase) "If you can't dazzle them with brilliance, baffle them with baloney."; your population of graders is almost certain to recognize baloney immediately. The standard by which you will be measured is that of an intelligent, thinking holder of a Master's degree who has not taken BIOS 213.

AN EXTRACT FROM THE BIOS 213 COURSE OUTLINE:

Research Data Management Methodology

1. Data acquisition
2. Data entry
3. Data editing
4. Database processing
5. Data communications
6. Database closure
7. Data retrieval, reporting, and analysis
8. Data inventory systems
9. Security and confidentiality
10. Documentation
11. Archiving
Problem 2:

It can be seen from the scatter plot (Graph 2) of Data set 1 that the data consists of two groups. The problem of testing the equality of the intercepts (i.e. the regression function is globally linear) of these two groups can be formulated as follows:

1. Construct a linear model:

\[
Y = X \beta + \epsilon
\]

where

\[
\begin{align*}
N \times 1 & \quad N \times 3 \quad 3 \times 1 \\
1 & \quad \beta_1 & \quad \beta_2
\end{align*}
\]

so that the mean of the response variable \( Y_i \) is

\[
\mu(x_i) = \beta_0 x_i + \beta_1 \text{ if } x_i < 0; \\
= \beta_0 x_i + \beta_2 \text{ if } x_i \geq 0.
\]

What is the design matrix \( X \)?

The matrix \( X'X \) is:

\[
\begin{bmatrix}
26.19052 & -9.05199 & 10.94220 \\
-9.05199 & 9.00000 & 0.00000 \\
10.94220 & 0.00000 & 11.00000
\end{bmatrix}
\]

and its inverse \( (X'X)^{-1} \):

\[
\begin{bmatrix}
.16125 & .16218 & -.16040 \\
.16218 & .27423 & -.16133 \\
-.16040 & -.16133 & .25047
\end{bmatrix}
\]

2. Determine the least squares estimates of \( \beta_1 \) and \( \beta_2 \).

3. Construct confidence intervals for \( \beta_1 \), \( \beta_2 \) and \( (\beta_1 - \beta_2) \).

4. Test the hypothesis \( H_0: \beta_1 = \beta_2 \) at \( \alpha = 0.05 \).

5. Comment briefly on the assumptions you used.
Data 2:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.07741</td>
<td>-1.09981</td>
</tr>
<tr>
<td>.84213</td>
<td>-1.23418</td>
</tr>
<tr>
<td>.84409</td>
<td>-1.42824</td>
</tr>
<tr>
<td>-2.50611</td>
<td>4.18857</td>
</tr>
<tr>
<td>-1.17731</td>
<td>3.30818</td>
</tr>
<tr>
<td>-.58900</td>
<td>2.73544</td>
</tr>
<tr>
<td>1.69406</td>
<td>-1.80055</td>
</tr>
<tr>
<td>.84626</td>
<td>-.99331</td>
</tr>
<tr>
<td>.73733</td>
<td>-1.35074</td>
</tr>
<tr>
<td>-.57602</td>
<td>3.39548</td>
</tr>
<tr>
<td>-1.75298</td>
<td>3.10252</td>
</tr>
<tr>
<td>.84241</td>
<td>-.32413</td>
</tr>
<tr>
<td>1.42200</td>
<td>-1.08596</td>
</tr>
<tr>
<td>1.73195</td>
<td>-2.19157</td>
</tr>
<tr>
<td>-.30844</td>
<td>2.78678</td>
</tr>
<tr>
<td>-1.07501</td>
<td>3.08156</td>
</tr>
<tr>
<td>-.31674</td>
<td>2.32838</td>
</tr>
<tr>
<td>-.75038</td>
<td>2.73648</td>
</tr>
<tr>
<td>.50426</td>
<td>-.82217</td>
</tr>
<tr>
<td>.40030</td>
<td>-.59237</td>
</tr>
</tbody>
</table>

Graph 2
PROBLEM 3:

The accompanying paper reports two studies: the first using live subjects, and the second using cadavers.

1) 18 points

Summarize, and comment on, the statistical aspects of the FIRST study. What research objective(s) did the authors have? What was their study design (an experiment, or not? if so, of what type?) Was it well-conceived for achieving the objective(s)? What methods of statistical analysis were used, and were they appropriate? What statistical conclusions were reached, and were they justified?

2) 7 points

Comment similarly on the SECOND study. (Assume that the summary statistics given at the bottom of Table I were calculated correctly.)

Note that we are more interested in the first study than the second, and are awarding points accordingly.

EDITORIAL NOTE.

Group B

PROBLEM 4:

For the period 1 July 1974 through 30 June 1978, North Carolina experienced a sudden infant death (SID) rate of two per thousand live births.

(a) For a county having 3000 live births and 12 SIDs in the same calendar period, calculate the P-value for the possibility that the SID rate in this county is greater than that of North Carolina. Use the normal distribution to approximate a presumed underlying Poisson model.

(b) Let $\lambda$ denote the SID rate for a county with 3000 live births during the above specified calendar period and consider the null hypothesis $H_0: \lambda = 0.002$, i.e., $\lambda$ is two SIDs per thousand live births. Determine an upper tail critical region with a level of significance $\alpha = 0.05$.

(c) Following part (b), calculate the power corresponding to the alternative hypothesis $H_a: \lambda = 0.006$.

(d) Comment briefly on the appropriateness of the normal approximation to the Poisson distribution for the above calculations. Sketch how you would proceed in parts (a), (b), and (c) if the normal approximation were not appropriate.