Introduction

Hardly a day goes by at the Statistical Consulting Center at UNC-G that a student armed with computer output from one of the statistical packages doesn't wander by uttering the familiar phrase, "I have this computer printout but I don't know what it means." This admission of ignorance concerning whatever they have in hand also implies that the selection of the statistical treatment was in error and probably chosen in a questionable manner. After a few minutes with a consultant, the student soon realizes that the data warrants a different approach to the statistical analysis. But what about the student who never seeks advice on interpreting the output? After all, he or she was able to select and obtain complex statistics without much effort. Why not make an attempt at interpretation? It's all so easy. Ask your thesis chairman what statistics are most impressive, select a computer package that furnishes these statistics, follow the examples in the manual, and write up the results. Who needs a statistician?

Unfortunately, the advanced technology of statistical computing has produced students and researchers who have adopted this philosophy. Surprisingly, most students who are able to obtain but unable to interpret computer output have had graduate training in statistics and usually report doing quite well in terms of the grade they received. Students ambitious enough to take an additional stat computing course seem to be better equipped to perform their own interpretation but still have difficulty making the jump between the cognitive aspects of the statistical methodology and the mechanical aspects of statistical computing. The only way today's students can make this jump effectively is to be presented with the statistics and the computing together in an integrated manner. Teaching a statistics course without incorporating the computer is synonymous with teaching cellular biology without the use of a microscope. Today's technology has turned the computer into a tool of the trade for all researchers. It is therefore important for statistical educators to train future researchers and statisticians in the responsible use of this invaluable tool.

Statistical methods and statistical computing must be tied together in a meaningful experiential way if students are to understand the ideas and concepts taught in graduate level statistics courses. Fictitious data sets found at the end of textbook chapters are a poor substitute for what the student is likely to encounter in the future. A much more provocative approach involves the analysis of data obtained from published research. Published research not only provides data to analyze but provides information on how the experiment or observational study was conducted. The student is able to evaluate design considerations, methods and materials, and a variety of other factors that determine the validity of the statistics and the experiment in general. This is not possible with a collection of meaningless numbers found at the end of a textbook chapter.

Workbook Development and Content

In an effort to provide research oriented students with meaningful problems in biostatistics a collection of 16 laboratory problems were developed over a three year period. The 16 laboratory exercises presented in this book are based on major statistical concepts and methods that a student would encounter during a year of course work. Although the workbook was designed primarily for graduate students enrolled in service-type biostatistics courses, other faculty and staff have reported great success using the workbook in undergraduate courses and instructional seminars. Each lab utilizes SAS to either help the student understand basic statistical principles or solve specific statistical problems. Actual data from published research is utilized in every lab and in most cases the data set are small enough to be worked by hand if an instructor so desires. SAS syntax is described in detail for each exercise and answer sheets at the end of each lab were designed in such a way as to help the student organize results from the SAS listing into tabular form appropriate for publication. Most descriptions of SAS syntax are referenced to specific pages in the SAS User's Guides or Introductory Manual to encourage attempts at debugging before seeking outside advice. The first nine labs cover topics normally presented in an introductory level biostatistics course. The remaining seven labs cover
Intermediate level topics suitable for the second semester of a two semester biostatistics series. Each lab is organized into nine sections: (a) general overview and objectives, (b) suggested readings in the SAS manuals, (c) a short description of the data or experiment, (d) data reference, (e) the actual data and input information, (f) data modification and variable creation, (g) procedure statements, (h) special directions for completing the assignment, and (i) an answer sheet. The procedures section also gives a brief statistical explanation of the technique being used.

Table of Contents and Sources of Data

Introductory Labs

1. Introduction to SAS, JCL, and the Concept of Frequency and Percentage Histograms
   Data taken from:

2. Descriptive Statistics and the Normal Distribution
   Data taken from:
   Smith, H.L. & Willius, F.A. Adiposity of the heart: A clinical and pathological study of one hundred and thirty-six obese patients. Archives of Internal Medicine, 1933, 52, 910-932.

3. The Central Limit Theorem, Random Sampling, and Confidence Intervals
   Data taken from:

4. Elementary Probability and Comparisons of Proportions in a 2 X 2 Table
   Data taken from:

5. Analysis of an I X J Contingency Table
   Data taken from:

6. Comparison of Two Means in Independent Samples and the Mean Difference Between Observations in Paired Samples
   Data taken from:

7. One Way Analysis of Variance, Orthogonal Polynomials, and the Least Significant Difference
   Data taken from:

8. Bivariate Regression and Correlation
   Data taken from:

9. Nonparametric Methods: The Wilcoxon Signed Rank Test and the Kruskal-Wallis One Way Analysis of Variance
   Data taken from:

 Intermediate labs

10. Introduction to Multiple Linear Regression With Two Independent Variables

Data taken from:

11. Multiple Regression Analysis With More Than Two Independent Variables

Data taken from:
Knott et al. (see lab 10)

12. Analysis of Covariance With a Review of One Way Analysis of Variance

Data taken from:
Knott et al. (see lab 10)

note. Labs 10, 11, and 12 present multiple regression, analysis of variance, and analysis of covariance within the context of the general linear model. It was felt that the students would better understand the relationship between these techniques if the same data (study) were analyzed.

13. Analysis of a 2 X 3 Factorial

Data taken from:

14. Randomized Block Analysis of Variance

Data taken from:

15. Utilizing Orthogonal Polynomials to Analyze Single Sample Repeated Measures Experiments

Data taken from:
Liang, C., & Hood, W.B. Comparison of cardiac output responses to 2,4-dinitrophenol induced hypermetabolism and muscular work. The Journal of Clinical Investigation, 1973, 52, 2283-2292.

16. Univariate Repeated Measures Analysis: Three Between Factors and Eleven Within Factors (Split-Plot Format)

Data taken from:

Evaluation

Although no informal evaluation of the effectiveness of these labs has ever been attempted, informal evaluations by the students who used the labs indicate they were very helpful. Over a three year period not one student expressed a negative opinion about the labs in general or the relevance of learning SAS. When asked, "What did you feel were the strong points of the course," many students mentioned the labs. Typical responses were:

- The SAS labs, which gave practical applications.
- The labs, which really reinforced and clarified the lecture and textbook readings. Very good study aids.
- The applicability of the labs to actual research.
- The realistic lab assignments.
- I think the labs helped enormously in understanding the material.
- The labs, which provided practical applications I am likely to use in the future.
- The laboratory exercises and learning the SAS computer package.
- Usefulness of the labs to my discipline.
- The labs which really supported the lectures.
When asked, "Do you think the lab materials presented in this course will be beneficial to you in your future professional experiences?" Ninety-five per cent of over 125 students said, "Yes". This very informal evaluation does suggest that the students appreciate the applicability of the labs and find them very supportive of class work. Also, the main objective of integrating methods, computing, and application, which initially motivated the development of the labs, seems to have been met.

Lab materials are available from the author for a small fee which covers the expense of copying and postage. All of the labs are currently being reformatted and proofed and will be available during the latter part of 1985. The labs are currently copyrighted by the author. The appendix contains a complete listing of lab #4.

*SAS is the registered trademark of SAS Institute Inc., Cary, NC, USA.

Mailing Address

David A. Ludwig
Dept. of Mathematics
University of North Carolina
at Greensboro
Greensboro, NC 27412

Appendix

Problem Set #5
15 Points

Introduction to R x C Contingency Table

Overview. The concept of independence, introduced in lab 4, is extended to the analysis of R x C contingency tables. Statistical results and conclusions obtained by the authors are verified by reanalyzing the data. Specific questions concerning treatment effects are answered by partitioning the main table. The concept of statistical hypothesis testing is also incorporated into the analysis.

Suggested Reading Before Attempting This Lab.

The Data. The effects of environmental stress on an unborn fetus is an important question in many areas of medical research. The following experiment was conducted to investigate various forms of stress as catalysts to audiogenic seizures after birth. An abbreviated description of the methods is as follows:

Treatments consisted of intraperitoneal injection of the mother on days 10, 11 and 12 of gestation, of either a) 30 mg of 3,4,5-trihydroxytyramine in 1 ml of 0.5N NaOH, 0.5% HCl, and 0.9% NaCl; b) solvent alone; or c) a sham treatment, in which a needle was inserted into the abdomen and held in place with nothing injected. A fourth group served as an unhandled control.

Twenty-three days after birth, each mouse was placed into a galvanized wash tub to which was attached a four inch bell. After the mice were acclimatized to the chamber, the bell was rung for 30 seconds. During this time the mice were observed and their seizure responses were scored. Responses lay on a continuum from mild running through clonic (twitching and partial) seizures followed by recovery to tonic (rigid and complete) seizures. A total of 423 mice were available for experimentation.


The Data Set. A summary of the results by treatment and response are listed below. Input this 4 x 4 table into SAS as was done in lab 4.
Data Modification and Variable Creation. This lab does not require any data modification or variable creation.

Procedure Statements. Use PROC FREQ to generate a 4 x 4 table (crosstabulation) of treatment by response. You will soon realize that entering a table into PROC FREQ is different than creating a table from raw data as you did in Lab 6. In Lab 6, the number of occurrences were counted across each individual and this count used as the observed frequency.

When a table is entered (i.e. the number of occurrences has already been tabulated), SAS must be told that cell frequencies and not raw data are being used to construct the requested table. This is done with a WEIGHT statement (e.g. page 516). The WEIGHT statement identifies the variable that contains the cell frequencies and tells SAS to use the value of this variable as the cell count. The WEIGHT statement follows the TABLES statement and the general form of the procedure is as follows:

```
PROC FREQ;
TABLES (SPECIFY CROSSTABULATION)/(OPTIONS);
WEIGHT (VARIABLE NAME THAT IDENTIFIES FREQUENCIES);
```

You will note that when using a WEIGHT statement, the number of data lines will always equal the number of cells in the contingency table. Also specify the following options:

a. EXPECTED
b. DEVIATION
c. CELLCHI2
d. CHISQ

Assignment. Answer the questions on the attached answer sheet. Cut the 4 x 4 table from the computer output and staple it to the back of the answer sheet. Do not hand in any other computer output.

1. The first question that needs to be addressed is that of complete independence in the overall table. Suppose prior to the analysis a TYPE I error rate of 0.05 is set. Based on this cutoff, what would you conclude about the independence of treatment and response. Explain your conclusion in three or four sentences. (1 point)

2. Examine the cell $X^2$ values. Which two cells have the largest $X^2$ values? Complete the table below. (1 point)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Cell $X^2$ Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. What percentage of the total $X^2$ is contributed by these two cells? (1 point)

4. In four or five sentences, attempt to explain what is occurring in this experiment. Remember, the researchers feel that handling the damp will have an effect and not the injection they are given. Look at all the cell $X^2$ values and the residuals before answering this question. Use percentages, residuals or other statistics to support your conclusion. (2 points)
5. The authors claim that the frequency of the response is identical in the first three groups. To check this statement, one can extract the following from overall table by combining the last three columns and labelling it "some response".

<table>
<thead>
<tr>
<th>Treatment</th>
<th>No response</th>
<th>Some response</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Thielylamine</td>
<td>21</td>
<td>75</td>
<td>96</td>
</tr>
<tr>
<td>2) Solvent</td>
<td>15</td>
<td>88</td>
<td>103</td>
</tr>
<tr>
<td>3) Sham</td>
<td>21</td>
<td>81</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>244</td>
<td>303</td>
</tr>
</tbody>
</table>

What were the frequencies of some response in the three groups? (1 point)

1)  
2)  
3)  

Are these frequencies identical? Yes or No

Perhaps what the authors mean is that the frequencies observed could have arisen by chance alone in a situation where the true underlying frequencies do not depend on the treatment. If we make that assumption, find the expected values and the value of $X^2$ that can be utilized to assess the strength of the evidence for the assumption of independence between treatment and seizure. Work by hand. (3 points)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Obs.</th>
<th>Exp.</th>
<th>Obs.</th>
<th>Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Thielylamine</td>
<td>23</td>
<td></td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>2) Solvent</td>
<td>15</td>
<td></td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>3) Sham</td>
<td>23</td>
<td></td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td></td>
<td>244</td>
<td></td>
</tr>
</tbody>
</table>

$X^2 = \frac{(O - E)^2}{E}$

d.f. =

Probability (P) of an equal or greater value of $X^2$ based on chi-square approximation given that the underlying situation is one of independence:

Do you agree with the authors' implication that there is no compelling evidence to refute the idea that these two factors, type of injection and seizure, are unrelated? Yes or No.

6. Many other questions could be asked of these data. One would be to assess the evidence that true underlying frequency of seizures in the combined treatment groups differ from that of the control. Calculate the $X^2$, give the approximate probability of an equal or larger value and make a statement about whether the result supports the idea of independence of handling and seizure. Work by hand. (3 points)

<table>
<thead>
<tr>
<th>No response</th>
<th>Some response</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>All treated</td>
<td>Obs =</td>
<td>Obs =</td>
</tr>
<tr>
<td></td>
<td>Exp =</td>
<td>Exp =</td>
</tr>
<tr>
<td>Unhandled Control</td>
<td>Obs =</td>
<td>Obs =</td>
</tr>
<tr>
<td></td>
<td>Exp =</td>
<td>Exp =</td>
</tr>
<tr>
<td>Totals</td>
<td>Obs =</td>
<td>Obs =</td>
</tr>
<tr>
<td></td>
<td>Exp =</td>
<td>Exp =</td>
</tr>
</tbody>
</table>

Proportion responding: treated control

$X^2 = \frac{(O - E)^2}{E}$

d.f. =

Prob of equal or larger value =

Conclusion (2 points)