Chapter 1
Introduction

1.1 Overview

Data analysts often encounter response measures that are categorical in nature; their outcomes reflect categories of information rather than the usual interval scale. Frequently, categorical data are presented in tabular form, known as contingency tables. Categorical data analysis is concerned with the analysis of categorical response measures, regardless of whether any accompanying explanatory variables are also categorical or are continuous. This book discusses hypothesis testing strategies for the assessment of association in contingency tables and sets of contingency tables. It also discusses various modeling strategies available for describing the nature of the association between a categorical outcome measure and a set of explanatory variables.

An important consideration in determining the appropriate analysis of categorical variables is their scale of measurement. Section 1.2 describes the various scales and illustrates them with data sets used in later chapters. Another important consideration is the sampling framework that produced the data; it determines the possible analyses and the possible inferences. Section 1.3 describes the typical sampling frameworks and their ramifications. Section 1.4 introduces the various analysis strategies discussed in this book and describes how they relate to one another. It also discusses the target populations generally assumed for each type of analysis and what types of inferences you are able to make to them. Section 1.5 reviews how the SAS System handles contingency tables and other forms of categorical data. Finally, Section 1.6 provides a guide to the material in the book for various types of readers, including indications of the difficulty level of the chapters.

1.2 Scale of Measurement

The scale of measurement of a categorical response variable is a key element in choosing an appropriate analysis strategy. By taking advantage of the methodologies available for the particular scale of measurement, you can choose a well-targeted strategy. If you do not take the scale of measurement into account, you may choose an inappropriate strategy that could lead to erroneous conclusions. Recognizing the scale of measurement and using it properly are very important in categorical data analysis.

Categorical response variables can be

- dichotomous
- discrete counts
- nominal
- ordinal
- grouped survival times

Dichotomous responses are those that have two possible outcomes—most often they are yes and no. Did the subject develop the disease ? Did the voter cast a ballot for the Democratic or Republican candidate ? Did the student pass the exam ? For example, the objective of a clinical trial for a new medication for colds is whether patients obtained relief from their pain-producing ailment. Consider Table 1.1, which is analyzed in Chapter 2, ‘‘The 2 $\times$ 2 Table.’’

Table 1.1 Respiratory Outcomes

Treatment	Favorable	Unfavorable	Total
Placebo	16	48	64
Test	40	20	60

The placebo group contains 64 patients, and the test medication group contains 60 patients. The columns contain the information concerning the categorical response measure: 40 patients in the Test group had a favorable response to the medication, and 20 subjects did not. The outcome in this example is thus dichotomous, and the analysis investigates the relationship between the response and the treatment.

Frequently, categorical data responses represent more than two possible outcomes, and often these possible outcomes take on some inherent ordering. Such response variables have an *ordinal* scale of measurement. Did the new school curriculum produce little, some, or high enthusiasm among the students ? Does the water exhibit low, medium, or high hardness ? In the former case, the order of the response levels is clear, but there is no clue as to the relative distances between the levels. In the latter case, there is a possible distance between the levels: medium might be twice the hardness of low, and high might be three times the hardness of low. Sometimes the distance is even clearer: a 50% potency dose versus a 100% potency dose versus a 200% potency dose. All three cases are examples of ordinal data.

An example of an ordinal measure occurs in data displayed in Table 1.2, which is analyzed in Chapter 8, ‘‘Logistic Regression I: Dichotomous Response.’’ A clinical trial investigated a treatment for rheumatoid arthritis. Male and female patients were given either the active treatment or a placebo; the outcome measured was whether they showed marked, some, or no improvement at the end of the clinical trial. The analysis uses a modeling technique called the proportional odds model to assess the relationship between the response variable and gender and treatment.

Table 1.2 Arthritis Data

		Improvement			
Sex	Treatment	Marked	Some	None	Total
Female	Active	16	5	6	27
Female	Placebo	6	7	19	32
Male	Active	5	2	7	14
Male	Placebo	1	0	10	11

Note that categorical response variables can often be managed in different ways. You could combine the Marked and Some columns in Table 1.2 to produce a dichotomous outcome: No Improvement versus Improvement. Grouping categories is often done during an analysis if the resulting dichotomous response is also of interest.

Categorical response variables sometimes contain *discrete counts*. Instead of falling into categories that are labeled (yes, no) or (low, medium, high), the outcomes are numbers themselves. Was the litter size 1, 2, 3, 4, or 5 members ? Did the house contain 1, 2, 3, or 4 air conditioners ? While the usual strategy would be to analyze the mean count, the assumptions required for the standard linear model are often not met with discrete counts that have small range; the counts may not be distributed normally.

For example, researchers examining respiratory disease in children visited children in different regions two times and determined whether they showed symptoms of respiratory illness. The response measure was whether the children exhibited symptoms in 0, 1, or 2 periods. Table 1.3 contains these data, which are analyzed in Chapter 12, ‘‘Weighted Least Squares.’’

Table 1.3 Colds in Children

		Periods with Colds			
Sex	Residence	0	1	2	Total
Female	Rural	45	64	71	180
Female	Urban	80	104	116	300
Male	Rural	84	124	82	290
Male	Urban	106	117	87	310

The table represents a crossclassification of gender, residence, and number of periods with colds. The analysis is concerned with modeling mean colds as a function of gender and residence.

If you have more than two outcome categories, and there is no inherent ordering to the categories, you have a nominal measurement scale. Which of four candidates did you vote for in the town council election ? Do you prefer the beach, mountains, or lake for a vacation ? There is no underlying scale for such outcomes and no apparent way in which to order them.

Consider Table 1.4, which is analyzed in Chapter 5, ‘‘The $s \times r$ Table.’’ Residents in one town were asked their political party affiliation and their neighborhood. Researchers were interested in whether political affiliation could predict neighborhood. Unlike ordinal

response levels, the classifications Bayside, Highland, Longview, and Sheffeld lie on no conceivable underlying scale. However, you can still assess whether there is association in the table, which is done in Chapter 5.

Table 1.4 Distribution of Parties in Neighborhoods

	Neighborhood			
Party	Bayside	Highland	Longview	Sheffeld
Democrat	221	160	360	140
Independent	200	291	160	311
Republican	208	106	316	97

Finally, another type of response variable in categorical data analysis is one that represents *grouped survival times*. With survival data, you are tracking the number of patients with certain outcomes (possibly death) over time. Often, the times of the condition are grouped together so that the response variable represents the number of patients who fail during a specific time interval. Such data are called *grouped survival times*. For example, the data displayed in Table 1.5 are from Chapter 15, ‘‘Categorized Time-to-Event Data.’’ A clinical condition is treated with an active drug for some patients and with a placebo for others. The response categories are whether there are recurrences, no recurrences, or whether the patients left the study. The entries correspond to the time intervals 0–1 years, 1–2 years, and 2–3 years, which make up the rows of the table.

Table 1.5 Life Table Format for Clinical Condition Data

Controls				
Interval	No Recurrences	Recurrences	Withdrawals	At Risk
0–1 Years	50	15	9	74
1–2 Years	30	13	7	50
2–3 Years	17	7	6	30
Active				
Interval	No Recurrences	Recurrences	Withdrawals	At Risk
0–1 Years	69	12	9	90
1–2 Years	59	7	3	69
2–3 Years	45	10	4	59

1.3 Sampling Frameworks

Categorical data arise from different sampling frameworks. The nature of the sampling framework determines the assumptions that can be made for the statistical analyses and in turn influences the type of analysis that can be applied. The sampling framework also determines the type of inference that is possible. Study populations are limited to target populations, those populations to which inferences can be made, by assumptions justified by the sampling framework.

Generally, data fall into one of three sampling frameworks: historical data, experimental data, and sample survey data. Historical data are observational data, which means that the study population has a geographic or circumstantial definition. These may include all the

occurrences of an infectious disease in a multicounty area, the children attending a particular elementary school, or those persons appearing in court during a specified time period. Highway safety data concerning injuries in motor vehicles is another example of historical data.

Experimental data are drawn from studies that involve the random allocation of subjects to different treatments of one sort or another. Examples include studies where types of fertilizer are applied to agricultural plots and studies where subjects are administered different dosages of drug therapies. In the health sciences, experimental data may include patients randomly administered a placebo or treatment for their medical condition.

In sample survey studies, subjects are randomly chosen from a larger study population. Investigators may randomly choose students from their school IDs and survey them about social behavior; national health care studies may randomly sample Medicare users and investigate physician utilization patterns. In addition, some sampling designs may be a combination of sample survey and experimental data processes. Researchers may randomly select a study population and then randomly assign treatments to the resulting study subjects.

The major difference in the three sampling frameworks described in this section is the use of randomization to obtain them. Historical data involve no randomization, and so it is often difficult to assume that they are representative of a convenient population. Experimental data have good coverage of the possibilities of alternative treatments for the restricted protocol population, and sample survey data have very good coverage of some larger population.

Note that the unit of randomization can be a single subject or a cluster of subjects. In addition, randomization may be applied to subsets, called strata or blocks, with equal or unequal probabilities. In sample surveys, this can lead to more complicated designs, such as stratified random samples, or even multistage cluster random samples. In experimental design studies, such considerations lead to repeated measurements (or split-plot) studies.

1.4 Overview of Analysis Strategies

Categorical data analysis strategies can be classified into those that are concerned with hypothesis testing and those that are concerned with modeling. Many questions about a categorical data set can be answered by addressing a specific hypothesis concerning association. Such hypotheses are often investigated with randomization methods. In addition to making statements about association, you also may want to describe the nature of the association in the data set. Statistical modeling techniques using maximum likelihood estimation or weighted least squares estimation are employed to describe this variation in terms of a parsimonious statistical model.

Most often the hypothesis of interest is whether association exists between the rows of a contingency table and its columns. The only assumption that is required is randomized allocation of subjects, either through the study design (experimental design) or through the hypothesis itself (necessary for historical data). In addition, particularly for the use of historical data, you often want to control for other explanatory variables that may have influenced the observed outcomes.

1.4.1 Randomization Methods

Table 1.1, the respiratory outcomes data, contains information obtained as part of a randomized allocation process. The hypothesis of interest is whether there is an association between treatment and outcome. For these data, the randomization is accomplished by the study design.

Table 1.6 contains data from a similar study. The main difference is that the study was conducted in two medical centers. The hypothesis of association is whether there is an association between treatment and outcome, controlling for any effect of center.

Table 1.6 Respiratory Improvement

Center	Treatment	Yes	No	Total
1	Test	29	16	45
1	Placebo	14	31	45
Total		43	47	90
2	Test	37	8	45
2	Placebo	24	21	45
Total		61	29	90

Chapter 2, ''The 2×2 Table,'' is primarily concerned with the association in 2×2 tables; in addition, it discusses measures of association, that is, statistics designed to evaluate the strength of the association. Chapter 3, ''Sets of 2×2 Tables,'' discusses the investigation of association in sets of 2×2 tables. When the table of interest has more than two rows and two columns, the analysis is further complicated by the consideration of scale of measurement. Chapter 4, ''Sets of $2 \times r$ and $s \times 2$ Tables,'' considers the assessment of association in sets of tables where the rows (columns) have more than two levels.

Chapter 5 describes the assessment of association in the general $s \times r$ table, and Chapter 6, ''Sets of $s \times r$ Tables,'' describes the assessment of association in sets of $s \times r$ tables. The investigation of association in tables and sets of tables is further discussed in Chapter 7, ''Nonparametric Methods,'' which discusses traditional nonparametric tests that have counterparts among the strategies for analyzing contingency tables.

1.4.2 Modeling Strategies

Often, you are interested in describing the variation in your data with a statistical model. In the continuous data setting, you frequently fit a model to the expected mean response. However, with categorical outcomes, there are a variety of response functions that you can model. Depending on the response function that you choose, you may use weighted least squares or maximum likelihood methods to estimate the model parameters.

Perhaps the most common response function modeled for categorical data is the logit. If you have a dichotomous response and represent the proportion of those subjects with an event (versus no event) outcome as p, then the logit can be written

$$\log\left(\frac{p}{1-p}\right)$$

Logistic regression is a modeling strategy that relates the logit for a set of explanatory variables to a linear model. One of its benefits is that estimates of odds ratios, important measures of association, can be obtained from the parameter estimates. Maximum likelihood estimation is used to provide those estimates.

Chapter 8, ''Logistic Regression I: Dichotomous Response,'' discusses logistic regression for a dichotomous outcome variable. Chapter 9, ''Logistic Regression II: Polytomous Response,'' discusses logistic regression for the situation where there are more than two outcomes for the response variable. Logits called generalized logits can be analyzed when the outcomes are nominal. And logits called cumulative logits can be analyzed when the outcomes are ordinal. Chapter 10, ''Conditional Logistic Regression,'' describes a specialized form of logistic regression that is appropriate when the data are highly stratified or arise from matched case-control studies.

In logistic regression, the objective is to predict a response outcome from a set of explanatory variables. However, sometimes you simply want to describe the structure of variation in a set of variables for which there are no obvious outcome or predictor variables. This occurs frequently for sociological studies. The loglinear model is a traditional modeling strategy for categorical data and is appropriate for describing the variation in such a set of variables. It is closely related to logistic regression, and the parameters in a loglinear model are also estimated with maximum likelihood estimation. Chapter 14, ''Loglinear Models,'' discusses the loglinear model, including several typical applications.

Some application areas have features that lead to special statistical techniques being developed for them. One of these areas for categorical data is bioassay analysis. Bioassay is the process of determining the potency or strength of a reagent or stimuli based on the response it elicits in biological organisms. Logistic regression is a technique often applied in bioassay analysis, where its parameters take on specific meaning. Chapter 11, ''Quantal Bioassay Analysis,'' discusses the use of categorical data methods for quantal bioassay.

Besides the logit, other useful response functions that can be modeled include proportions, means, and measures of association. Weighted least squares estimation is a method of analyzing response functions such as these, based on large sample theory. These methods are appropriate when you have sufficient sample size and when you have a randomly selected sample, either implicitly through study design or explicitly via assumptions concerning the representativeness of the data. Not only can you model a variety of useful functions, but weighted least squares estimation also provides a useful framework for the analysis of repeated categorical measurements, particularly those limited to a small number of repeated values.

Chapter 12, ''Weighted Least Squares,'' addresses modeling categorical data with weighted least squares methods, and Chapter 13, ''Modeling Repeated Measurements Data,'' discusses these techniques as applied to the analysis of repeated measurements data. Also described in Chapter 13 is the use of generalized estimating equations in the analysis of repeated measurements.

Finally, another special application area for categorical data analysis is the analysis of grouped survival data. Chapter 15 discusses some features of survival analysis that are pertinent to grouped survival data, including how to model them using a model called the piecewise exponential model. Since the Poisson regression model is employed to fit this

model, the chapter also provides an overview of Poisson regression.

1.5 Working with Tables in the SAS System

This section discusses some considerations of managing tables with the SAS System. If you are already familiar with the FREQ procedure, you may want to skip this section.

Many times, categorical data are presented to the researcher in the form of tables, and other times, they are presented in the form of case record data. SAS procedures can handle either type of data. In addition, many categorical data are ordinal, so that the order of the levels of the rows and columns takes on special meaning. There are numerous ways that you can specify a particular order to SAS procedures.

Consider the following SAS DATA step that inputs the data displayed in Table 1.1.

```
data respire;
   input treat $ outcome $ count ;
   cards;
   placebo f 16
   placebo u 48
   test    f 40
   test    u 20
   ;
proc freq;
   weight count;
   tables treat*outcome;
run;
```

The data set RESPIRE contains three variables: TREAT is a character variable containing values for treatment, OUTCOME is a character variable containing values for the outcome (f for favorable and u for unfavorable), and COUNT contains the number of observations that have the respective TREAT and OUTCOME values. Thus, COUNT effectively takes values corresponding to the cells of Table 1.1. The PROC FREQ statements request that a table be constructed using TREAT as the row variable and OUTCOME as the column variable. By default, PROC FREQ orders the values of the rows (columns) in alphanumeric order. The WEIGHT statement is necessary to tell the procedure that the data are count data, or frequency data; the variable listed in the WEIGHT statement contains the values of the count variable.

Output 1.1 contains the resulting frequency table.

Output 1.1 Frequency Table

```
                TABLE OF TREAT BY OUTCOME

TREAT     OUTCOME

Frequency|
Percent  |
Row Pct  |
Col Pct  |f       |u       |  Total
---------+--------+--------+
placebo  |     16 |     48 |     64
         |  12.90 |  38.71 |  51.61
         |  25.00 |  75.00 |
         |  28.57 |  70.59 |
---------+--------+--------+
test     |     40 |     20 |     60
         |  32.26 |  16.13 |  48.39
         |  66.67 |  33.33 |
         |  71.43 |  29.41 |
---------+--------+--------+
Total          56       68      124
              45.16    54.84   100.00
```

Suppose that a different sample produced the numbers displayed in Table 1.7.

Table 1.7 Respiratory Outcomes

Treatment	Favorable	Unfavorable	Total
Placebo	5	10	15
Test	8	20	28

These data may be stored in case record form, which means that each individual is represented by a single observation. You can also use this type of input with the FREQ procedure. The only difference is that the WEIGHT statement is not required.

The following statements create a SAS data set for these data and invoke PROC FREQ for case record data. The @@ symbol in the INPUT statement means that the data lines contain multiple observations.

```
data respire;
    input treat $ outcome $ @@ ;
    cards;
placebo f placebo f  placebo f
placebo f placebo f
placebo u placebo u  placebo u
placebo u placebo u  placebo u
placebo u placebo u  placebo u
placebo u
test    f test    f  test    f
test    f test    f  test    f
test    f test    f
test    u test    u  test    u
test    u test    u  test    u
test    u test    u  test    u
test    u test    u  test    u
```

```
test    u test    u  test    u
test    u test    u  test    u
test    u test    u
    ;
proc freq;
    tables treat*outcome;
run;
```

Output 1.2 displays the resulting frequency table.

Output 1.2 Frequency Table

```
                TABLE OF TREAT BY OUTCOME

              TREAT     OUTCOME

              Frequency|
              Percent  |
              Row Pct  |
              Col Pct  |f       |u       |  Total
              ---------+--------+--------+
              placebo  |      5 |     10 |     15
                       |  11.63 |  23.26 |  34.88
                       |  33.33 |  66.67 |
                       |  38.46 |  33.33 |
              ---------+--------+--------+
              test     |      8 |     20 |     28
                       |  18.60 |  46.51 |  65.12
                       |  28.57 |  71.43 |
                       |  61.54 |  66.67 |
              ---------+--------+--------+
              Total          13       30       43
                          30.23    69.77   100.00
```

In this book, the data are generally presented in count form.

When ordinal data are considered, it becomes quite important to ensure that the levels of the rows and columns are sorted correctly. By default, the data are going to be sorted alphanumerically. If this isn't suitable, then you need to alter the default behavior.

Consider the data displayed in Table 1.2. IMPROVE is the outcome variable, and the values marked, some, and none are listed in decreasing order. Suppose that the data set ARTHRIT is created with the following statements.

```
data arthrit;
   length treat $7. sex $6. ;
   input sex $ treat $ improve $ count @@ ;
   cards ;
female active  marked 16 female active  some 5 female active  none  6
female placebo marked  6 female placebo some 7 female placebo none 19
male   active  marked  5 male   active  some 2 male   active  none  7
male   placebo marked  1 male   placebo some 0 male   placebo none 10
;
run;
```

If you invoked PROC FREQ for this data set and used the default sort order, the levels of the columns would be ordered marked, none, and some, which would be incorrect. One

way to change this default sort order is to use the ORDER=DATA option in the PROC FREQ statement. This specifies that the sort order is the same order in which the values are encountered in the data set. Thus, since 'marked' comes first, it is first in the sort order. Since 'some' is the second value for IMPROVE encountered in the data set, then it is second in the sort order. And 'none' would be third in the sort order. This is the desired sort order. The following PROC FREQ statements produce a table displaying the sort order resulting from the ORDER=DATA option.

```
proc freq order=data;
    weight count;
    tables treat*improve;
 run;
```

Output 1.3 displays the frequency table for the crossclassification of treatment and improvement for these data; the values for IMPROVE are in the correct order.

Output 1.3 Frequency Table from ORDER=DATA option

```
                TABLE OF TREAT BY IMPROVE

TREAT     IMPROVE

Frequency|
Percent  |
Row Pct  |
Col Pct  |marked  |some    |none    |  Total
---------+--------+--------+--------+
active   |     21 |      7 |     13 |     41
         |  25.00 |   8.33 |  15.48 |  48.81
         |  51.22 |  17.07 |  31.71 |
         |  75.00 |  50.00 |  30.95 |
---------+--------+--------+--------+
placebo  |      7 |      7 |     29 |     43
         |   8.33 |   8.33 |  34.52 |  51.19
         |  16.28 |  16.28 |  67.44 |
         |  25.00 |  50.00 |  69.05 |
---------+--------+--------+--------+
Total          28       14       42       84
              33.33    16.67    50.00   100.00
```

Other possible values for the ORDER= option include FORMATTED, which means sort by the formatted values. The ORDER= option is also available with the CATMOD, LOGISTIC, and GENMOD procedures. For information on the ORDER= option for the FREQ procedure, refer to the *SAS/STAT User's Guide, Version 6, Fourth Edition*. This option is used frequently in this book.

Often, you want to analyze sets of tables. For example, you may want to analyze the crossclassification of treatment and improvement for both males and females. You do this in PROC FREQ by using a three-way crossing of the variables SEX, TREAT, and IMPROVE.

```
proc freq order=data;
    weight count;
    tables sex*treat*improve / nocol nopct;
run;
```

The two rightmost variables in the TABLES statement determine the rows and columns of the table, respectively. Separate tables are produced for the unique combination of values of the other variables in the crossing. Since SEX has two levels, one table is produced for males and one table is produced for females. If there were four variables in this crossing, with the two variables on the left having two levels each, then four tables would be produced, one for each unique combination of the two leftmost variables in the TABLES statement.

Note also that the options NOCOL and NOPCT are included. These options suppress the printing of column percentages and cell percentages, respectively. Since generally you are interested in row percentages, these options are often specified in the code displayed in this book.

Output 1.4 contains the two tables produced with the preceding statements.

Output 1.4 Producing Sets of Tables

```
          TABLE 1 OF TREAT BY IMPROVE
              CONTROLLING FOR SEX=female

TREAT     IMPROVE

Frequency|
Row Pct  |marked  |some    |none    |  Total
---------+--------+--------+--------+
active   |     16 |      5 |      6 |     27
         |  59.26 |  18.52 |  22.22 |
---------+--------+--------+--------+
placebo  |      6 |      7 |     19 |     32
         |  18.75 |  21.88 |  59.38 |
---------+--------+--------+--------+
Total          22       12       25       59


          TABLE 2 OF TREAT BY IMPROVE
            CONTROLLING FOR SEX=male

TREAT     IMPROVE

Frequency|
Row Pct  |marked  |some    |none    |  Total
---------+--------+--------+--------+
active   |      5 |      2 |      7 |     14
         |  35.71 |  14.29 |  50.00 |
---------+--------+--------+--------+
placebo  |      1 |      0 |     10 |     11
         |   9.09 |   0.00 |  90.91 |
---------+--------+--------+--------+
Total          6        2       17       25
```

This section reviewed some of the basic table management necessary for using the FREQ procedure. Other related options are discussed in the appropriate chapters.

1.6 Using This Book

This book is intended for a variety of audiences, including novice readers with some statistical background (solid understanding of regression analysis), those readers with

substantial statistical background, and those readers with background in categorical data analysis. Therefore, not all of this material will have the same importance to all readers. Some chapters include a good deal of tutorial material, while others have a good deal of advanced material. This book is not intended to be a comprehensive treatment of categorical data analysis, so some topics are mentioned briefly for completeness and some other topics are emphasized because they are not generally well documented.

The data used in this book come from a variety of sources and represent a wide breadth of application. However, due to the biostatistical background of all three authors, there is a certain inevitable weighting of biostatistical examples. Most of the data come from practice, and the original sources are cited when this is true; however, due to confidentiality concerns and pedalogical requirements, some of the data are altered or created. However, they still represent realistic situations.

Chapters 2–4 are intended to be accessible to all readers, as is most of Chapter 5. Chapter 6 is an integration of Mantel-Haenszel methods at a more advanced level, but scanning it is probably a good idea for readers interested in the topic. In particular, the discussion about the analysis of repeated measurements data with extended Mantel-Haenszel methods is useful material for all readers comfortable with the Mantel-Haenszel technique.

Chapter 7 is a special interest chapter relating Mantel-Haenszel procedures to traditional nonparametric methods used for continuous data outcomes.

Chapters 8 and 9 on logistic regression are intended to be accessible to all readers, particularly Chapter 8. The last section of Chapter 8 describes the statistical methodology more completely for the advanced reader. Most of the material in Chapter 9 should be accessible to most readers. Chapter 10 is a specialized chapter that discusses conditional logistic regression and requires somewhat more statistical expertise. Chapter 11 discusses the use of logistic regression in analyzing bioassay data.

Chapter 12 discusses weighted least squares and is written at a somewhat higher statistical level than Chapters 8 and 9, but most readers should find this material useful, particularly the examples.

Chapters 13–15 discuss advanced topics and are necessarily written at a higher statistical level. Chapter 13 describes the analysis of repeated measurements data using weighted least squares and includes a section discussing the use of generalized estimating equations for repeated measurements data. The opening sections introduce repeated measurements analysis and discuss a basic example; this material is intended to be accessible to a wide range of readers. Chapter 14 discusses loglinear model analysis and Chapter 15 discusses the analysis of categorized time-to-event data.

This book describes statistical techniques and discusses their implementation with the SAS System. In some instances, statistics are computed by hand or other software is mentioned so that the methodological information presented is complete. All examples were executed with Release 6.10 of the SAS System; features new in that release are pointed out. A few features upcoming in Release 6.11 are also described.