# 1

# INTRODUCTION





The purpose of most research is to assess relationships among a set of variables, and choosing an appropriate statistical technique depends on the type of variables under investigation. Suppose we have a set of numerical values for a variable:

1. If each element of this set may lie only at a few isolated points, we have a discrete or categorical data set. In other words, a categorical variable is one for which the measurement scale consists of a set of categories; examples are race, sex, counts of events, or some sort of artificial grading.
2. If each element of this set may theoretically lie anywhere on the numerical scale, we have a continuous data set. Examples are blood pressure, cholesterol level, or time to a certain event such as death.

This text focuses on the analysis of categorical data and multivariate problems when at least three variables are involved. The first section of this chapter shows a simple example of real-life problems to which some of the methods described in this book can be applied. This example shows a potential complexity when the data involve more than two variables with a phenomenon known as effect modification. We will return to this example later when illustrating some methods of analysis for categorical data in Chapters 2, 3 and 4. The second section briefly reviews some likelihood-based statistical methods to be used in subsequent chapters with various

*Applied Categorical Data Analysis and Translational Research, Second Edition*, By Chap T. Le

regression models. The last section summarizes special features of this text, its objectives, and for whom it is intended; we also briefly points out special features of this second edition.

## 1.1 A PROTOTYPE EXAMPLE

Many research outcomes can be classified as belonging to one of two possible categories: for example, Presence and Absence, White and Nonwhite, Male and Female, Improved and Not Improved. Of course, one of these two categories is usually identified as of primary interest to the researcher; for example, Presence in the "Presence and Absence" classification, Nonwhite in the "White and Nonwhite" classification. We can, in general, relabel the two outcome categories as Positive (or $+$) and Negative (or $-$). An outcome is positive if the primary category is observed and is negative if the other category is observed. Health decisions are frequently based on the "proportion" of positive outcomes defined by

$$p = x/n$$

where $x$ in the above equation is the number of positive outcomes from the $n$ observations made on $n$ individuals: $0 \leq p \leq 1$ because $1 \leq \mathrm{x} \leq \mathrm{n}$. Proportion is a number used to describe a group of individuals according to a dichotomous (or binary) characteristic under investigation and the following example provides an illustration of its use in the health sciences.

Comparative studies are intended to show possible differences between two or more groups. Data for comparative studies may come from different sources, with the two fundamental designs being *retrospective* and *prospective*. Retrospective studies gather past data from selected cases and controls to determine differences, if any, in the exposure to a suspected risk factor. They are commonly referred to as *case–control studies*; a "case" is a person with the disease under investigation and a "control" is a person without that disease. In a case–control study, cases of a specific disease are ascertained as they arise from population-based registers or lists of hospital admissions and controls are sampled either as disease-free individuals from the population at risk, or as hospitalized patients having a diagnosis other than the one under study. The advantages of a retrospective study or case–control study are that it is economical and it is possible to obtain answers to research questions relatively quickly because the cases are already available. Major limitations are due to the inaccuracy of the exposure histories and uncertainty about the appropriateness of the control sample; these problems sometimes hinder retrospective studies and make them less preferred than prospective studies. The following example introduces a retrospective study concerning occupational health.

■ **Example 1.1** A case–control study was undertaken to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia (Blot et al., 1978). Cases (of lung cancer) were identified from these sources:

(i) diagnoses since 1970 at the single large hospital in Brunswick, (ii) diagnoses during 1975 and 1976 at three major hospitals in Savannah, and (iii) death certificates for the period 1970–1974 in the area.

Controls (or control subjects) were selected from admissions to the four hospitals and from death certificates in the same period for diagnoses other than lung cancer, bladder cancer, or chronic lung cancer. Data are tabulated separately for smokers and nonsmokers as follows:

| Smoking | Shipbuilding | Cases | Controls |
|---|---|---|---|
| No | Yes | 11 | 35 |
| | No | 50 | 203 |
| Yes | Yes | 84 | 45 |
| | No | 313 | 270 |

The exposure under investigation, ''Shipbuilding,'' refers to employment in shipyards during World War II. By a separate tabulation, with the first half of the table for nonsmokers and the second half for smokers, we treat *smoking* as a potential *confounder*. A confounder is a factor that may be an exposure by itself, not under investigation but related to the disease (in this case, lung cancer) and the exposure (shipbuilding); previous studies have linked smoking to lung cancer and construction workers are more likely to be smokers. The term *exposure* is used here to emphasize that employment in shipyards is a *risk factor*; however, the term would also be used in studies where the factor under investigation has beneficial effects.

In an examination of the smokers in the above data set, the numbers of people employed in shipyards, 84 and 45, tell us little because the sizes of the two groups, cases and controls, are different. Adjusting these absolute numbers for the group sizes, we have the following:

(1a) For the controls,

$$\begin{aligned}\text{Proportion of exposure} &= 45/315 \\ &= 0.143 \text{ or } 14.3\%\end{aligned}$$

(2a) For the cases,

$$\begin{aligned}\text{Proportion of exposure} &= 84/397 \\ &= 0.212 \text{ or } 21.2\%\end{aligned}$$

The results reveal different exposure histories: the proportion of exposure among cases was higher than that among controls. It is not in any way yet a conclusive

proof, but it is a good clue indicating a possible relationship between the disease (lung cancer) and the exposure (employment in shipbuilding industry—a possible occupational hazard).

Similar examination of the data for nonsmokers shows that, by taking into consideration the numbers of cases and of controls, we have the following figures for employment:

(1b) For the controls:

$$\begin{aligned}\text{Proportion of exposure} &= 35/238 \\ &= 0.147 \text{ or } 14.7\%\end{aligned}$$

(2b) For the cases:

$$\begin{aligned}\text{Proportion of exposure} &= 11/61 \\ &= 0.180 \text{ or } 18.0\%\end{aligned}$$

Again, the results also reveal different exposure histories: the proportion of exposure among cases was higher than that among controls.

The above analyses also show that the difference (cases versus controls) between proportions of exposure among smokers, that is,

$$21.2\% - 14.3\% = 6.9\%$$

is different from the difference (cases versus controls) between proportions of exposure among nonsmokers, which is

$$18.0\% - 14.7\% = 3.3\%$$

The differences, 6.9% and 3.3%, are measures of the *strength of the relationship* between the disease and the exposure, one for each of the two strata—the two groups of smokers and nonsmokers, respectively. The above calculation shows that the possible effects of employment in shipyards (as a suspected risk factor) are different for smokers and nonsmokers. This difference of the two case–control differences (6.9% versus 3.3%), if confirmed, is called an ''interaction'' or an *effect modification*, where smoking alters the effect of employment in shipyards as a risk factor for lung cancer. In that case, smoking is not only a confounder, it is an *effect modifier,* which modifies the effects of shipbuilding (on the possibility of having lung cancer).

In some extreme examples, a pair of variables may even have their marginal association in a different direction from their partial association (the association between them as seen at each and every level of a confounder or effect modifier). This interesting phenomenon is called *Simpson's paradox*, which further emphasizes the analysis complexity when we have data involving more than two variables.

## 1.2 A REVIEW OF LIKELIHOOD-BASED METHODS

Problems in biological and health sciences are formulated mathematically by considering the data that are to be used for making a decision as the observed values of a certain random variable $X$. The distribution of $X$ is assumed to belong to a certain family of distributions specified by one or several parameters; a *parameter* can be defined as an (unknown) numerical characteristic of a population. The problem for decision makers is to decide on the basis of the data which members of the family could represent the distribution of $X$; that is, to predict or estimate the value of a certain parameter $\theta$ (or several parameters). The magnitude of a parameter often represents the effect of a risk or environmental factor, and knowing its value, even approximately, would shed some light on the impact of such a factor. The *likelihood function* $L(x;\theta)$ for a random sample ($x$'s) of size $n$ from the probability density function (or pdf) $f(x;\theta)$ is

$$L(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

The *maximum likelihood estimator* (MLE) of $\theta$ is the value $\hat{\theta}$ for which $L(x;\theta)$ is maximized. Calculus suggests setting the derivative of $L(x;\theta)$ with respect to $\theta$ equal to zero and solving the resulting equation.

Since

$$\frac{dL}{d\theta} = (L)\frac{d(\ln L)}{d\theta}$$

$dL/d\theta = 0$ if and only if $d(\ln L)/d\theta = 0$ because $L$ is never zero. Thus we can find the possible maximum of $L$ by maximizing $\ln L$; it is often easier to deal mathematically with a sum than with a product.

$$\ln L = \sum_{i=1}^{n} \ln\ f(x_i;\theta)$$

The MLE has a number of good properties, which we will state without proofs; readers can skip this entire section without having any discontinuity.

1. MLE is consistent.
2. If an efficient estimator exists, it is the MLE.
3. The MLE is asymptotically distributed as normal. The variance of this asymptotic distribution is given by the following formula:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{E\left\{-\dfrac{d^2 \ln L}{d\theta^2}\right\}}$$

in which $E\{.\}$ denotes the *expected value* and the value of the denominator, called Fisher's information matrix (in this case, it is a number), often needs to be estimated using the MLE of $\theta$.

This is an asymptotic distribution; that is, results are good for large samples only. If a closed-form solution does not exist, the iterative solution may be obtained by first solving for an additive correction,

$$\Delta\hat{\theta} = -\left(\frac{d \ln L}{d\theta}\right) \Big/ \left(\frac{d^2 \ln L}{d\theta^2}\right)$$

using numerical values of the derivatives. The iterative solution by this Newton–Raphson method would proceed as follows:

*Step 1*: Provide an initial value of $\hat{\theta}$, denoted by $\hat{\theta}^{(0)}$.

*Step 2*: Determine the value of $\Delta\hat{\theta}$ by evaluating the derivatives at $\hat{\theta}^{(0)}$.

*Step 3*: Add $\Delta\hat{\theta}$ to the initial value to obtain a new value for $\hat{\theta}$, that is, $\theta^{(1)} = \theta^{(0)} + \Delta\theta$.

*Step 4*: Repeat Steps 2 and 3 using $\hat{\theta}^{(1)}$ to obtain $\hat{\theta}^{(2)}$ and stop when results from successive steps are very close (below certain previously set threshold).

After a final solution has been obtained, an estimate of its variance is then given by

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{-\dfrac{d^2 \ln L}{d\theta^2}}$$

where the second derivative is evaluated using the value of the MLE of $\theta$.

For example, we have for a binomial distribution with unknown probability $\pi$

$$L(x;\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$$

$$\ln L(x;\pi) = \ln\binom{n}{x} + x\ln\pi + (n-x)\ln(1-\pi)$$

$$\frac{d}{dx}\ln L(x;\pi) = \frac{x}{\pi} - \frac{n-x}{1-\rho}$$

$$-\frac{d^2}{dp^2}\ln L(x;p) = \frac{x}{\pi^2} + \frac{n-x}{(1-\pi)^2}$$

$$E\left\{-\frac{d^2}{dp^2}\ln L(x;\pi)\right\} = \frac{np}{\pi^2} + \frac{n-np}{(1-\pi)^2}$$
$$= \frac{n}{\pi(1-\pi)}$$

The results are

$$\hat{\pi} = p$$
$$= \frac{x}{n} \text{ (sample proportion)}$$
$$\text{Var}(p) = \frac{1}{E\left\{-\frac{d^2}{dp^2}\ln L(x;\pi)\right\}}$$
$$= \frac{\pi(1-\pi)}{n}$$

Consider the case of a two-parameter model with probability density function $f(.)$. Let $L(x;\, \theta_1,\, \theta_2)$ be the likelihood function defined from a random sample $\{x_i\}$ by

$$L(x;\theta_1,\theta_2) = \prod_{i=1}^{n} f(x_i;\, \theta_1,\theta_2)$$

The MLEs $\theta_1$ and $\theta_2$ are the values $\hat{\theta}_1$ and $\hat{\theta}_2$ of $\theta_1$ and $\theta_2$ for which $L(x;\, \theta_1,\, \theta_2)$ is maximized. These estimators are obtained by solving the following equations:

$$\frac{\delta}{\delta\theta_1}\ln L = 0 \quad \text{and} \quad \frac{\delta}{\delta\theta_2}\ln L = 0$$

If closed-form solutions do not exist, the iterative solutions to these equations may be obtained by the Newton–Raphson method, which is similar to that for the above one-parameter model.

The variance–covariance matrix of the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ can be obtained from the Fisher's information matrix, which is defined as

$$I = \begin{bmatrix} E\left(-\frac{\delta^2}{\delta\theta_1^2}\right)\ln L & E\left(-\frac{\delta^2}{\delta\theta_1\delta\theta_2}\ln L\right) \\ E\left(-\frac{\delta^2}{\delta\theta_1\delta\theta_2}\ln L\right) & E\left(-\frac{\delta^2}{\delta\theta_2^2}\right)\ln L \end{bmatrix}$$

In obtaining numerical variance–covariance estimates, all expected values of the partial derivatives are replaced by numerical evaluations of those partial derivatives

using MLE values for the parameters or values from the last iteration if iterative solutions are required.

$$\begin{bmatrix} \widehat{\mathrm{Var}}(\hat{\theta}_1) & \widehat{\mathrm{Cov}}(\hat{\theta}_1, \hat{\theta}_2) \\ \widehat{\mathrm{Cov}}(\hat{\theta}_1, \hat{\theta}_2) & \widehat{\mathrm{Var}}(\hat{\theta}_2) \end{bmatrix} = \begin{bmatrix} -\dfrac{\delta^2}{\delta\theta_1^2}\ln L & \dfrac{\delta^2}{\delta\theta_1\delta\theta_2}\ln L \\ -\dfrac{\delta^2}{\delta\theta_1\delta\theta_2}\ln L & -\dfrac{\delta^2}{\delta\theta_2^2}\ln L \end{bmatrix}^{-1}$$

Of course, the maximum likelihood procedure, as explained for the two-parameter model, can be easily generalized to models with more than two parameters.

As an example of two-parameter models, let us consider a random sample of size $n$, $\{x_i\}$, from the normal distribution with mean $\mu$ and variance $\theta = \sigma^2$. We have

$$\begin{aligned} \theta &= \sigma^2 \\ L(x;\mu,\theta) &= \prod_{i=1}^{n} \frac{1}{\theta^{1/2}\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\theta}\right] \\ \ln L(x;\mu,\theta) &= -\frac{n}{2}\ln\,\theta - \frac{n}{2}\ln(2\pi) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2 \\ \frac{\delta}{\delta\mu}\ln L &= \frac{1}{\theta}(x_i-\mu)^2 \\ \frac{\delta}{\delta\theta}\ln L &= \frac{1}{2\theta}\left\{-n + \frac{1}{\theta}\sum (x_i-\mu)^2\right\} \end{aligned}$$

The results are

$$\begin{aligned} \hat{\mu} &= \bar{x} \\ \hat{\sigma}^2 &= \hat{\theta} \\ &= \frac{1}{n}(x_i-\mu)^2 \\ &= \frac{n-1}{n}s^2 \end{aligned}$$

From these derivatives, we find

$$\begin{aligned} \frac{\delta^2}{\delta\mu^2}\ln L &= -n/\theta \\ \frac{\delta^2}{\delta\mu\,\delta\theta}\ln L &= -\frac{1}{\theta^2}\left\{\sum x_i - n\mu\right\} \\ \frac{\delta^2}{\delta\theta^2}\ln L &= \frac{n}{2\theta^2} - \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

And from their expected values, we can easily derive the variances and covariance:

$$\begin{aligned} \mathrm{Var}(\bar{x}) &= \frac{\sigma^2}{n} \\ \mathrm{Var}(\hat{\sigma}^2) &= \frac{2\sigma^4}{n} \\ \mathrm{Cov}(\bar{x}, \hat{\sigma}^2) &= 0 \end{aligned}$$

A multiple regression model involves many parameters, the unknown regression coefficients, $\beta$. Once we have fit such a multiple regression model and obtained estimates for the various parameters of interest using the above method, we want to answer questions about the contributions of various factors to the prediction of the response variable. There are three types of questions:

1. *An Overall Test.* Taken collectively, does the entire set of explanatory or independent variables contribute significantly to the prediction of the response? The null hypothesis for this test may be stated as: "all $k$ independent variables considered together do not explain the variation in the responses." In other words, the null hypothesis is

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Two likelihood-based statistics can be used to test this *global* null hypothesis; each has an asymptotic chi-squared distribution with $k$ degrees of freedom under the null hypothesis $H_0$:

(a) *Likelihood Ratio Test*.

$$X^2_{\mathrm{LR}} = 2[\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\mathbf{0})]$$

(b) *Score Test*.

$$X^2_{\mathrm{S}} = \left[\frac{\delta}{\delta\beta}\ln L(\mathbf{0})\right]\left[-\frac{\delta^2}{\delta\beta^2}\ln L(\mathbf{0})\right]^{-1}\left[\frac{\delta}{\delta\beta}\ln L(\mathbf{0})\right]$$

Both statistics are provided by most standard computer programs such as SAS and they are asymptotically equivalent yielding identical statistical decisions most of the time.

2. *Test for the Value of a Single Factor.* Does the addition of one particular variable of interest add significantly to the prediction of response over and above that achieved by other independent variables? Let us assume that we now wish to test whether the addition of one particular independent variable of interest adds sig-

nificantly to the prediction of the response over and above that achieved by other factors already present in the model. The null hypothesis for this test may be stated as: ''factor $X_i$ does not have any value added to the prediction of the response given that other factors are already included in the model.'' In other words,

$$H_0\colon \beta_i = 0$$

To test such a null hypothesis, one can perform a likelihood ratio chi-squared test, with one degree of freedom, similar to that for the above global hypothesis:

$$X^2_{\mathrm{LR}} = 2[\ln L(\hat{\boldsymbol{\beta}}; \text{all } X\text{'s}) - \ln L(\hat{\boldsymbol{\beta}}; \text{all other } X\text{'s with } X_i \text{ deleted})]$$

A much easier alternative method is to use.

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

where $\hat{\beta}_i$ is the corresponding estimated regression coefficient and $SE(\hat{\beta}_i)$ is the estimate of the standard error of $\hat{\beta}_i$, both of which are printed by standard packaged computer programs. In performing this test, we refer the value of the $z$ statistic to percentiles of the standard normal distribution.

3. *Test for Contribution of a Group of Variables.* Does the addition of a group of variables add significantly to the prediction of response over and above that achieved by other independent variables? This testing procedure addresses the more general problem of assessing the additional contribution of two or more factors to the prediction of the response over and above that made by other variables already in the regression model. In other words, the null hypothesis is of the form

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

To test such a null hypothesis, one can perform a likelihood ratio chi-squared test, with $m$ degrees of freedom:

$$X^2_{\mathrm{LR}} = 2[\ln L(\hat{\boldsymbol{\beta}}; \text{all } X\text{'s}) - \ln L(\hat{\boldsymbol{\beta}}; \text{all other } X\text{'s with } X\text{'s under investigation deleted})]$$

This ''multiple contribution'' procedure is very useful for assessing the importance of potential explanatory variables. In particular, it is often used to test whether a similar group of variables, such as ''demographic characteristics,'' is important for the prediction of the response; these variables have some trait in common. Another application would be a collection of powers and/or product terms (referred to as interaction variables). It is often of interest to assess the interaction effects collectively before trying to consider individual interaction terms in a model as previously

suggested. In fact, such use reduces the total number of tests to be performed and this, in turn, helps to provide better control of overall Type I error rates, which may be inflated due to multiple testing.

In many applications, we wish to identify from many available factors a small subset of factors that relate significantly to the outcome, for example, the disease under investigation. In that identification process, of course, we wish to avoid a large Type I (or false positive) error. In a regression analysis, a Type I error corresponds to including a predictor that has no real relationship to the outcome; such an inclusion can greatly confuse the interpretation of the regression results. In a standard multiple regression analysis, this goal can be achieved by using a strategy that adds into or removes from a regression model one factor at a time according to a certain order of relative importance. Therefore the two important steps are:

1. Specifying a criterion or criteria for selecting a model. The selection is often based on the likelihood ratio chi-squared statistic.
2. Specifying a strategy for applying the chosen criterion or criteria. Such a strategy is concerned with whether a particular variable should be added to a model or whether any variable should be deleted from a model at a particular stage of the process (stepwise regression). As computers became more accessible and more powerful, these practices became more popular.

## 1.3 INTERVAL ESTIMATION FOR A PROPORTION

Recall the following results on the maximum likelihood estimation of an unknown probability or proportion $\pi$:

$$\begin{aligned}
\hat{\pi} &= p \\
&= \frac{x}{n} \quad \text{(sample proportion)} \\
\text{Var}(p) &= \frac{1}{E\left\{-\dfrac{d^2}{dp^2}\ln L(x;\pi)\right\}} \\
&= \theta \\
&= \frac{\pi(1-\pi)}{n}
\end{aligned}$$

Consider the usual estimate of variance of $p$:

$$\text{var}(p) = \frac{p(1-p)}{n}$$

We can see that

$$\begin{aligned} E\{\text{var}(p)\} &= \frac{1}{n}\{E(p)-E(p^2)\} \\ p &= \frac{\sum x_i}{n};\ x_i = 0/1 \\ p^2 &= \frac{\{\sum x_i^2 + 2\sum x_i x_j\}}{n^2} \\ E(p) &= \pi \\ E(p^2) &= \frac{\{\pi + (n-1)\pi^2\}}{n} \\ E\{\text{var}(p)\} &= \left\{\frac{n-1}{n}\right\}\frac{\pi(1-\pi)}{n} \\ &= \left\{1-\frac{1}{n}\right\}\pi \end{aligned}$$

The results have the following meaning:

1. var($p$) is a biased estimate of Var($p$); an unbiased estimate of Var($p$) is $p(1-p)/(n-1)$, with denominator $(n-1)$ similar to the sample variance of a continuous sample.
2. However, var($p$), with denominator $n$, is asymptotically unbiased and its use is popular.

In summary, we have an approximate 95% confidence interval for a population proportion $\pi$:

$$p \pm 1.96\, SE(p)$$

where the standard error of the sample proportion, $SE(p)$, is calculated as

$$SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

■ **Example 1.2** Consider the problem of estimating the prevalence of malignant melanoma in 45–54-year-old women in the United States. Suppose a random sample of ($n = 5000$) women is selected from this age group and ($x = 28$) are found to have the disease. Our point estimate for the prevalence of this disease is 0.0056 (=28/5000); its standard error is

$$\begin{aligned} SE(p) &= \sqrt{\frac{(0.0056)(1-0.0056)}{5000}} \\ &= 0.0011 \end{aligned}$$

Therefore a 95% confidence interval for the prevalence $\pi$ of malignant melanoma in 45–54-year-old women in the United States is given by

$$0.0056 \pm (1.96)(0.0011) = (0.0034, 0.0078) \text{ or } (0.34\%, 0.78\%)$$

## 1.4 ABOUT THIS BOOK

This book is intended to meet the needs of practitioners and students in applied fields by covering major, updated methods in the analysis of categorical data. It is also intended to meet the needs of clinicians and students in biomedical sciences with some basic introduction to a reemerging field called ''translational research.'' It is written for beginning graduate students in biostatistics, epidemiology, and environmental health, as well as for biomedical research workers. As a book for biostatistics and statistics students, it is designed to offer some details for a better understanding of the various procedures as well as the relationships among different methods. However, the mathematics have been kept to an absolute minimum. As a book for students in applied fields and as a reference book for practicing biomedical research workers, this book is very application oriented. It introduces applied research areas and a large number of real-life examples, most of which are completely solved with samples of computer programs.

The book is divided into nine chapters including this introductory chapter.

Chapter 2 covers basic methods and applications of two-way contingency tables including etiologic fraction, the evaluation of ordinal risks, and the Mantel–Haenszel method. Compared to the first edition, the first section (on screening tests) is moved and expanded to form a new chapter, Chapter 8.

Chapter 3 is devoted to loglinear models; topics covered include the selection of the best model for three-way tables, and selection of a model for higher-dimensional tables, with or without the identification of a dependent variable.

Chapter 4 is focused on logistic regression models, both binary and ordinal responses. Topics covered include the stepwise procedure, measures of goodness-of-fit, and the use of logistic models for different designs. Compared to the first edition, the new edition represents a major overhaul of Chapter 4: (i) we added a new introductory Section 4.1, ''Modeling a Probability,'' to include other models such as probit; (ii) we moved old Section 4.2.5, ''ROC Curve,'' to the new Chapter 8; and (iii) we added a new Section 4.5, ''Quantal Bioassays,'' an important topic in translational research.

Chapter 5 covers similar topics as those in Chapters 2–4, but for matched designs, singly or multiply, including the conditional logistic regression model.

Chapter 6 covers analytical methods for count data including the Poisson regression model. Topics covered in this chapter include overdispersion and how to fit overdispersed models.

Chapter 7, ''Categorical Data and Translational Research,'' is a new chapter. Topics covered represent the core material of translational research—early phase clinical trials. These topics include, among others, the standard design, the sequential monitoring of toxicity, and one-stage and two-stage designs for Phase II clinical trials.

Chapter 8, ‘‘Categorical Data and Diagnostic Medicine’’, is another new addition. Topics covered include examples and description of the disease screening process, some basic issues, the ROC curve and the corresponding optimization problem, and the roles of covariates.

Chapter 9 presents a brief introduction to survival analysis and Cox’s regression model. This inclusion is partly to show the difference between categorical data and survival data, and partly to serve as a brief introduction to the field of survival analysis, which is an important part of translational research.

In each of the nine chapters, numerous examples are provided for illustration.