1

Analysis of Stratified Data

This chapter discusses the analysis of clinical outcomes in the presence of influential covariates such as investigational center or patient demographics. The following analysis methods are reviewed:

- Stratified analyses of continuous endpoints using parametric methods based on fixed and random effects models as well as nonparametric methods.
- Simple randomization-based methods as well as more advanced exact and model-based methods for analyzing stratified categorical outcomes.
- Analysis of stratified time-to-event data using randomization-based tests and the Cox proportional hazards model.

The chapter also introduces statistical methods for studying the nature of treatment-by-stratum interactions in clinical trials.

1.1 Introduction

This chapter addresses issues related to adjustment for important covariates in clinical applications. The goal of an adjusted analysis is to provide an overall test of treatment effect in the presence of factors that have a significant effect on the outcome variable. Two different types of factors known to influence the outcome are commonly encountered in clinical trials: *prognostic* and *non-prognostic* factors (Mehrotra, 2001). Prognostic factors are known to influence the outcome variables in a systematic way. For instance, the analysis of survival data is always adjusted for prognostic factors such as a patient's age and disease severity because these patient characteristics are strongly correlated with mortality. By contrast, non-prognostic factors are likely to impact the trial's outcome but their effects do not exhibit a predictable pattern. It is well known that treatment differences vary, sometimes dramatically, across investigational centers in multicenter clinical trials. However, the nature of center-to-center variability is different from the variability associated with a patient's age or disease severity. Center-specific treatment differences are dependent on a large number of factors, e.g., geographical location, general quality of care, etc. As a consequence, individual centers influence the overall treatment difference in a fairly random manner and it is natural to classify the center as a non-prognostic factor.

Adjustments for important covariates are carried out using randomization- and model-based methods (Koch and Edwards, 1988; Lachin, 2000, Chapter 4). The idea behind randomization-based analyses is to explicitly control factors influencing the outcome variable while assessing the relationship between the treatment effect and outcome. The popular Cochran-Mantel-Haenszel method for categorical data serves as a good example of this approach. In order to adjust for a covariate, the sample is divided into strata that are relatively homogeneous with respect to the selected covariate. The treatment effect is examined separately within each stratum and thus the confounding effect of the covariate is eliminated from the analysis. The stratum-specific treatment differences are then combined to carry out an aggregate significance test of the treatment effect across the strata.

Model-based methods present an alternative to the randomization-based approach. In general, inferences based on linear or non-linear models are closely related (and often asymptotically equivalent) to corresponding randomization-based inferences. Roughly speaking, one performs regression inferences by embedding randomization-based methods into a modeling framework that links the outcome variable to treatment effect and important covariates. Once a model has been specified, an inferential method (most commonly the method of maximum likelihood) is applied to estimate relevant parameters and test relevant hypotheses. Looking at the differences between the two approaches to adjusting for covariates in clinical trials, it is worth noting that model-based methods are more flexible than randomization-based methods. For example, within a model-based framework, one can directly adjust for continuous covariates without having to go through an artificial and possibly inefficient process of creating strata.[1] Further, as pointed out by Koch et al. (1982), randomization- and model-based methods have been historically motivated by two different sampling schemes. As a result, randomization-based inferences are generally restricted to a particular study, whereas model-based inferences can be generalized to a larger population of patients.

There are two important advantages of adjusted analysis over a simplistic pooled approach that ignores the influence of prognostic and non-prognostic factors. First of all, adjusted analyses are performed to improve the power of statistical inferences (Beach and Meier, 1989; Robinson and Jewell, 1991; Ford, Norrie and Ahmadi, 1995). It is well known that, by adjusting for a covariate in a linear model, one gains *precision* which is proportional to the correlation between the covariate and outcome variable. The same is true for categorical and time-to-event data. Lagakos and Schoenfeld (1984) demonstrated that omitting an important covariate with a large hazard ratio dramatically reduces the efficiency of the score test in Cox proportional hazards models.

Further, failure to adjust for important covariates may introduce *bias*. Following the work of Cochran (1983), Lachin (2000, Section 4.4.3) demonstrated that the use of marginal unadjusted methods in the analysis of stratified binary data leads to biased estimates. The magnitude of the bias is proportional to the degree of treatment group imbalance within each stratum and the difference in event rates across the strata. Along the same line, Gail, Wieand and Piantadosi (1984) and Gail, Tan and Piantadosi (1988) showed that parameter estimates in many generalized linear and survival models become biased when relevant covariates are omitted from the regression.

Overview

Section 1.2 reviews popular ANOVA models with applications to the analysis of stratified clinical trials. Parametric stratified analyses in the continuous case are easily implemented using PROC GLM or PROC MIXED. The section also considers a popular nonparametric test for the analysis of stratified

[1] However, in fairness, it is important to note that modeling may require more assumptions; e.g., we may need to assume that the outcome variable and covariate are linearly related.

data in a non-normal setting. Linear regression models have been the focus of numerous monographs and research papers. The classical monographs of Rao (1973) and Searle (1971) provided an excellent discussion of the general theory of linear models. Milliken and Johnson (1984, Chapter 10), Goldberg and Koury (1990) and Littell, Freund and Spector (1991, Chapter 7) discussed the analysis of stratified data in an unbalanced ANOVA setting and its implementation in SAS.

Section 1.3 reviews randomization-based (Cochran-Mantel-Haenszel and related methods) and model-based approaches to the analysis of stratified categorical data. It covers both asymptotic and exact inferences that can be implemented in PROC FREQ, PROC LOGISTIC and PROC GENMOD. See Breslow and Day (1980), Koch and Edwards (1988), Lachin (2000), Stokes, Davis and Koch (2000) and Agresti (2002) for a thorough overview of categorical analysis methods with clinical trial applications.

Section 1.4 discusses statistical methods used in the analysis of stratified time-to-event data. The section covers both randomization-based tests available in PROC LIFETEST and model-based tests based on the Cox proportional hazards regression implemented in PROC PHREG. Kalbfleisch and Prentice (1980), Cox and Oakes (1984) and Collett (1994) gave a detailed review of classical survival analysis methods. Allison (1995), Cantor (1997) and Lachin (2000, Chapter 9) provided an introduction to survival analysis with clinical applications and examples of SAS code.

Section 1.5 introduces two popular tests for qualitative interaction developed by Gail and Simon (1985) and Ciminera et al. (1993). The tests for qualitative interaction help clarify the nature of the treatment-by-stratum interaction and identify patient populations that benefit the most from an experimental therapy. They can also be used in sensitivity analyses.

1.2 Continuous Endpoints

This section reviews parametric and nonparametric analysis methods with applications to clinical trials in which the primary analysis is adjusted for important covariates, e.g., multicenter clinical trials. Within the parametric framework, we will focus on fixed and random effects models in a frequentist setting. The reader interested in alternative approaches based on conventional and empirical Bayesian methods is referred to Gould (1998).

EXAMPLE: Multicenter Depression Trial

The following data will be used throughout this section to illustrate parametric analysis methods based on fixed and random effects models. Consider a clinical trial comparing an experimental drug with a placebo in patients with major depressive disorder. The primary efficacy measure was the change from baseline to the end of the 9-week acute treatment phase in the 17-item Hamilton depression rating scale total score (HAMD17 score). Patient randomization was stratified by center.

A subset of the data collected in the depression trial is displayed below. Program 1.1 produces a summary of HAMD17 change scores and mean treatment differences observed at five centers.

Program 1.1 Depression trial data

```
data hamd17;
    input center drug $ change @@;
    datalines;
100 P 18 100 P 14 100 D 23 100 D 18 100 P 10 100 P 17 100 D 18 100 D 22
100 P 13 100 P 12 100 D 28 100 D 21 100 P 11 100 P  6 100 D 11 100 D 25
100 P  7 100 P 10 100 D 29 100 P 12 100 P 12 100 P 10 100 D 18 100 D 14
101 P 18 101 P 15 101 D 12 101 D 17 101 P 17 101 P 13 101 D 14 101 D  7
101 P 18 101 P 19 101 D 11 101 D  9 101 P 12 101 D 11 102 P 18 102 P 15
102 P 12 102 P 18 102 D 20 102 D 18 102 P 14 102 P 12 102 D 23 102 D 19
102 P 11 102 P 10 102 D 22 102 D 22 102 P 19 102 P 13 102 D 18 102 D 24
102 P 13 102 P  6 102 D 18 102 D 26 102 P 11 102 P 16 102 D 16 102 D 17
```

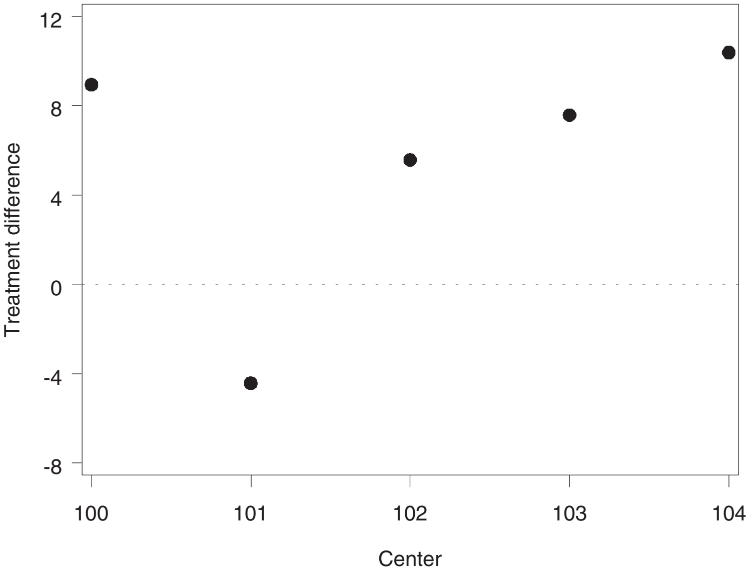
```
102 D  7 102 D 19 102 D 23 102 D 12 103 P 16 103 P 11 103 D 11 103 D 25
103 P  8 103 P 15 103 D 28 103 D 22 103 P 16 103 P 17 103 D 23 103 D 18
103 P 11 103 P -2 103 D 15 103 D 28 103 P 19 103 P 21 103 D 17 104 D 13
104 P 12 104 P  6 104 D 19 104 D 23 104 P 11 104 P 20 104 D 21 104 D 25
104 P  9 104 P  4 104 D 25 104 D 19
;
proc sort data=hamd17;
    by drug center;
proc means data=hamd17 noprint;
    by drug center;
    var change;
    output out=summary n=n  mean=mean std=std;
data summary;
    set summary;
    format mean std 4.1;
    label drug="Drug"
    center="Center"
    n="Number of patients"
    mean="Mean HAMD17 change"
    std="Standard deviation";
proc print data=summary noobs label;
    var drug center n mean std;
data plac(rename=(mean=mp)) drug(rename=(mean=md));
    set summary;
    if drug="D" then output drug; else output plac;
data comb;
    merge plac drug;
    by center;
    delta=md-mp;
axis1 minor=none label=(angle=90 "Treatment difference")
    order=(-8 to 12 by 4);
axis2 minor=none label=("Center") order=(100 to 104 by 1);
symbol1 value=dot color=black i=none height=10;
proc gplot data=comb;
    plot delta*center/frame haxis=axis2 vaxis=axis1 vref=0 lvref=34;
    run;
```

Output from Program 1.1

```
                   Number      Mean
                     of       HAMD17    Standard
Drug    Center    patients    change    deviation

 D        100        11        20.6         5.6
 D        101         7        11.6         3.3
 D        102        16        19.0         4.7
 D        103         9        20.8         5.9
 D        104         7        20.7         4.2
 P        100        13        11.7         3.4
 P        101         7        16.0         2.7
 P        102        14        13.4         3.6
 P        103        10        13.2         6.6
 P        104         6        10.3         5.6
```

Figure 1.1 The mean treatment differences in HAMD17 changes from baseline at the selected centers in the depression trial example



Output 1.1 lists the center-specific mean and standard deviation of the HAMD17 change scores in the two treatment groups. Further, Figure 1.1 displays the mean treatment differences observed at the five centers. Note that the mean treatment differences are fairly consistent at Centers 100, 102, 103 and 104. However, the data from Center 101 appears to be markedly different from the rest of the data.

As an aside note, it is helpful to remember that the likelihood of observing a similar treatment effect reversal by chance increases very quickly with the number of strata, and it is too early to conclude that Center 101 represents a true outlier (Senn, 1997, Chapter 14). We will discuss the problem of testing for *qualitative* treatment-by-stratum interactions in Section 1.5.

1.2.1 Fixed Effects Models

To introduce fixed effects models used in the analysis of stratified data, consider a study with a continuous endpoint comparing an experimental drug to a placebo across m strata (see Table 1.1). Suppose that the normally distributed outcome y_{ijk} observed on the kth patient in the jth stratum in the ith treatment group follows a two-way cell-means model

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}. \tag{1.1}$$

In the depression trial example, the term y_{ijk} denotes the reduction in the HAMD17 score in individual patients and μ_{ij} represents the mean reduction in the 10 cells defined by unique combinations of the treatment and stratum levels.

Table 1.1 A two-arm clinical trial with m strata

Stratum 1				Stratum m		
Treatment	Number of patients	Mean	...	Treatment	Number of patients	Mean
Drug	n_{11}	μ_{11}		Drug	n_{1m}	μ_{1m}
Placebo	n_{21}	μ_{21}		Placebo	n_{2m}	μ_{2m}

The cell-means model goes back to Scheffe (1959) and has been discussed in numerous publications, including Speed, Hocking and Hackney (1978) and Milliken and Johnson (1984). Let n_{1j} and n_{2j}

denote the sizes of the jth stratum in the experimental and placebo groups, respectively. Since it is uncommon to encounter empty strata in a clinical trial setting, we will assume there are no empty cells, i.e., $n_{ij} > 0$. Let n_1 and n_2 denote the number of patients in the experimental and placebo groups, and let n denote the total sample size, i.e.,

$$n_1 = \sum_{j=1}^{m} n_{1j}, \quad n_2 = \sum_{j=1}^{m} n_{2j}, \quad n = n_1 + n_2.$$

A special case of the cell-means model (1.1) is the familiar main-effects model with an interaction

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \tag{1.2}$$

where μ denotes the overall mean, the α parameters represent the treatment effects, the β parameters represent the stratum effects, and the $\alpha\beta$ parameters are introduced to capture treatment-by-stratum variability.

Stratified data can be analyzed using several SAS procedures, including PROC ANOVA, PROC GLM and PROC MIXED. Since PROC ANOVA supports balanced designs only, we will focus in this section on the other two procedures. PROC GLM and PROC MIXED provide the user with several analysis options for testing the most important types of hypotheses about the treatment effect in the main-effects model (1.2). This section reviews hypotheses tested by the Type I, Type II and Type III analysis methods. The Type IV analysis will not be discussed here because it is different from the Type III analysis only in the rare case of empty cells. The reader can find more information about Type IV analyses in Milliken and Johnson (1984) and Littell, Freund and Spector (1991).

Type I Analysis

The Type I analysis is commonly introduced using the so-called $R()$ notation proposed by Searle (1971, Chapter 6). Specifically, let $R(\mu)$ denote the reduction in the error sum of squares due to fitting the mean μ, i.e., fitting the reduced model

$$y_{ijk} = \mu + \varepsilon_{ijk}.$$

Similarly, $R(\mu, \alpha)$ is the reduction in the error sum of squares associated with the model with the mean μ and treatment effect α, i.e.,

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}.$$

The difference $R(\mu, \alpha) - R(\mu)$, denoted by $R(\alpha|\mu)$, represents the additional reduction due to fitting the treatment effect after fitting the mean and helps assess the amount of variability explained by the treatment accounting for the mean μ. This notation is easy to extend to define other quantities such as $R(\beta|\mu, \alpha)$. It is important to note that $R(\alpha|\mu)$, $R(\beta|\mu, \alpha)$ and other similar quantities are independent of restrictions imposed on parameters when they are computed from the normal equations. Therefore, $R(\alpha|\mu)$, $R(\beta|\mu, \alpha)$ and the like are uniquely defined in any two-way classification model.

The Type I analysis is based on testing the α, β and $\alpha\beta$ factors in the main-effects model (1.2) in a sequential manner using $R(\alpha|\mu)$, $R(\beta|\mu, \alpha)$ and $R(\alpha\beta|\mu, \alpha, \beta)$, respectively. Program 1.2 computes the F statistic and associated p-value for testing the difference between the experimental drug and placebo in the depression trial example.

Program 1.2 Type I analysis of the HAMD17 changes in the depression trial example

```
proc glm data=hamd17;
    class drug center;
    model change=drug|center/ss1;
    run;
```

Output from Program 1.2

```
Source              DF       Type I SS     Mean Square   F Value   Pr > F

drug                 1     888.0400000     888.0400000     40.07   <.0001
center               4      87.1392433      21.7848108      0.98   0.4209
drug*center          4     507.4457539     126.8614385      5.72   0.0004
```

Output 1.2 lists the F statistics associated with the DRUG and CENTER effects as well as their interaction (recall that `drug|center` is equivalent to `drug center drug*center`). Since the Type I analysis depends on the order of terms, it is important to make sure that the DRUG term is fitted first. The F statistic for the treatment comparison, represented by the DRUG term, is very large ($F = 40.07$), which means that administration of the experimental drug results in a significant reduction of the HAMD17 score compared to the placebo. Note that this *unadjusted analysis* ignores the effect of centers on the outcome variable.

The $R()$ notation helps clarify the structure and computational aspects of the inferences; however, as stressed by Speed and Hocking (1976), the notation may be confusing, and precise specification of the hypotheses being tested is clearly more helpful. As shown by Searle (1971, Chapter 7), the Type I F statistic for the treatment effect corresponds to the following hypothesis:

$$H_I: \quad \frac{1}{n_1}\sum_{j=1}^{m} n_{1j}\mu_{1j} = \frac{1}{n_2}\sum_{j=1}^{m} n_{2j}\mu_{2j}.$$

It is clear that the Type I hypothesis of no treatment effect depends both on the true within-stratum means and the number of patients in each stratum.

Speed and Hocking (1980) presented an interesting characterization of the Type I, II and III analyses that facilitates the interpretation of the underlying hypotheses. Speed and Hocking showed that the Type I analysis tests the simple hypothesis of no treatment effect

$$H: \quad \frac{1}{m}\sum_{j=1}^{m} \mu_{1j} = \frac{1}{m}\sum_{j=1}^{m} \mu_{2j}$$

under the condition that the β and $\alpha\beta$ factors are both equal to 0. This characterization implies that the Type I analysis ignores center effects and it is prudent to perform it when the stratum and treatment-by-stratum interaction terms are known to be negligible.

The standard ANOVA approach outlined above emphasizes hypothesis testing and it is helpful to supplement the computed p-value for the treatment comparison with an estimate of the average treatment difference and a 95% confidence interval. The estimation procedure is closely related to the Type I hypothesis of no treatment effect. Specifically, the "average treatment difference" is estimated in the Type I framework by

$$\frac{1}{n_1}\sum_{j=1}^{m} n_{1j}\overline{y}_{1j\cdot} - \frac{1}{n_2}\sum_{j=1}^{m} n_{2j}\overline{y}_{2j\cdot}.$$

It is easy to verify from Output 1.1 and Model (1.2) that the Type I estimate of the average treatment difference in the depression trial example is equal to

$$\begin{aligned}\widehat{\delta} = \widehat{\alpha}_1 - \widehat{\alpha}_2 &+ \left(\frac{11}{50} - \frac{13}{50}\right)\widehat{\beta}_1 + \left(\frac{7}{50} - \frac{7}{50}\right)\widehat{\beta}_2 \\ &+ \left(\frac{16}{50} - \frac{14}{50}\right)\widehat{\beta}_3 + \left(\frac{9}{50} - \frac{10}{50}\right)\widehat{\beta}_4 + \left(\frac{7}{50} - \frac{6}{50}\right)\widehat{\beta}_5\end{aligned}$$

$$
\begin{aligned}
&+\frac{11}{50}\widehat{(\alpha\beta)}_{11}+\frac{7}{50}\widehat{(\alpha\beta)}_{12}+\frac{16}{50}\widehat{(\alpha\beta)}_{13}+\frac{9}{50}\widehat{(\alpha\beta)}_{14}+\frac{7}{50}\widehat{(\alpha\beta)}_{15}\\
&-\frac{13}{50}\widehat{(\alpha\beta)}_{21}-\frac{7}{50}\widehat{(\alpha\beta)}_{22}-\frac{14}{50}\widehat{(\alpha\beta)}_{23}-\frac{10}{50}\widehat{(\alpha\beta)}_{24}-\frac{6}{50}\widehat{(\alpha\beta)}_{25}\\
={}&\widehat{\alpha}_1-\widehat{\alpha}_2-0.04\widehat{\beta}_1+0\widehat{\beta}_2+0.04\widehat{\beta}_3-0.02\widehat{\beta}_4+0.02\widehat{\beta}_5\\
&+0.22\widehat{(\alpha\beta)}_{11}+0.14\widehat{(\alpha\beta)}_{12}+0.32\widehat{(\alpha\beta)}_{13}+0.18\widehat{(\alpha\beta)}_{14}+0.14\widehat{(\alpha\beta)}_{15}\\
&-0.26\widehat{(\alpha\beta)}_{21}-0.14\widehat{(\alpha\beta)}_{22}-0.28\widehat{(\alpha\beta)}_{23}-0.2\widehat{(\alpha\beta)}_{24}-0.12\widehat{(\alpha\beta)}_{25}.
\end{aligned}
$$

To compute this estimate and its associated standard error, we can use the ESTIMATE statement in PROC GLM as shown in Program 1.3.

Program 1.3 Type I estimate of the average treatment difference in the depression trial example

```
proc glm data=hamd17;
    class drug center;
    model change=drug|center/ss1;
    estimate "Trt diff"
        drug 1 -1
        center -0.04 0 0.04 -0.02 0.02
        drug*center 0.22 0.14 0.32 0.18 0.14 -0.26 -0.14 -0.28 -0.2 -0.12;
    run;
```

Output from Program 1.3

```
                                               Standard
Parameter                    Estimate             Error    t Value    Pr > |t|

Trt diff                   5.96000000        0.94148228       6.33      <.0001
```

Output 1.3 displays an estimate of the average treatment difference along with its standard error, which can be used to construct a 95% confidence interval associated with the obtained estimate. The t test for the equality of the treatment difference to 0 is identical to the F test for the DRUG term in Output 1.2. One can check that the t statistic in Output 1.3 is equal to the square root of the corresponding F statistic in Output 1.2. It is also easy to verify that the average treatment difference is simply the difference between the mean changes in the HAMD17 score observed in the experimental and placebo groups without any adjustment for center effects.

Type II Analysis

In the Type II analysis, each term in the main-effects model (1.2) is adjusted for all other terms with the exception of higher-order terms that contain the term in question. Using the $R()$ notation, the significance of the α, β and $\alpha\beta$ factors is tested in the Type II framework using $R(\alpha|\mu,\beta)$, $R(\beta|\mu,\alpha)$ and $R(\alpha\beta|\mu,\alpha,\beta)$, respectively.

Program 1.4 computes the Type II F statistic to test the significance of the treatment effect on changes in the HAMD17 score.

Program 1.4 Type II analysis of the HAMD17 changes in the depression trial example

```
proc glm data=hamd17;
    class drug center;
    model change=drug|center/ss2;
    run;
```

Output from Program 1.4

```
Source              DF     Type II SS    Mean Square   F Value   Pr > F

drug                 1    889.7756912    889.7756912     40.15   <.0001
center               4     87.1392433     21.7848108      0.98   0.4209
drug*center          4    507.4457539    126.8614385      5.72   0.0004
```

We see from Output 1.4 that the F statistic corresponding to the DRUG term is highly significant ($F = 40.15$), which indicates that the experimental drug significantly reduces the HAMD17 score after an adjustment for the center effect. Note that, by the definition of the Type II analysis, the presence of the interaction term in the model or the order in which the terms are included in the model do not affect the inferences with respect to the treatment effect. Thus, dropping the DRUG*CENTER term from the model generally has little impact on the F statistic for the treatment effect (to be precise, excluding the DRUG*CENTER term from the model has no effect on the numerator of the F statistic but affects its denominator due to the change in the error sum of squares).

Searle (1971, Chapter 7) demonstrated that the hypothesis of no treatment effect tested in the Type II framework has the following form:

$$H_{II}: \quad \sum_{j=1}^{m} \frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}\mu_{1j} = \sum_{j=1}^{m} \frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}\mu_{2j}.$$

Again, as in the case of Type I analyses, the Type II hypothesis of no treatment effect depends on the number of patients in each stratum. It is interesting to note that the variance of the estimated treatment difference in the jth stratum, i.e., Var $(\overline{y}_{1j\cdot} - \overline{y}_{2j\cdot})$, is inversely proportional to $n_{1j}n_{2j}/(n_{1j}+n_{2j})$. This means that the Type II method averages stratum-specific estimates of the treatment difference with weights proportional to the precision of the estimates.

The Type II estimate of the average treatment difference is given by

$$\left(\sum_{j=1}^{m} \frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}\right)^{-1} \sum_{j=1}^{m} \frac{n_{1j}n_{2j}}{n_{1j}+n_{2j}}(\overline{y}_{1j\cdot} - \overline{y}_{2j\cdot}). \tag{1.3}$$

For example, we can see from Output 1.1 and Model (1.2) that the Type II estimate of the average treatment difference in the depression trial example equals

$$\begin{aligned}
\widehat{\delta} = \widehat{\alpha}_1 - \widehat{\alpha}_2 + &\left(\frac{11\times 13}{11+13} + \frac{7\times 7}{7+7} + \frac{16\times 14}{16+14} + \frac{9\times 10}{9+10} + \frac{7\times 6}{7+6}\right)^{-1} \\
\times &\left(\frac{11\times 13}{11+13}\widehat{(\alpha\beta)}_{11} + \frac{7\times 7}{7+7}\widehat{(\alpha\beta)}_{12} + \frac{16\times 14}{16+14}\widehat{(\alpha\beta)}_{13} + \frac{9\times 10}{9+10}\widehat{(\alpha\beta)}_{14} + \frac{7\times 6}{7+6}\widehat{(\alpha\beta)}_{15}\right. \\
&\left. - \frac{11\times 13}{11+13}\widehat{(\alpha\beta)}_{21} - \frac{7\times 7}{7+7}\widehat{(\alpha\beta)}_{22} - \frac{16\times 14}{16+14}\widehat{(\alpha\beta)}_{23} - \frac{9\times 10}{9+10}\widehat{(\alpha\beta)}_{24} - \frac{7\times 6}{7+6}\widehat{(\alpha\beta)}_{25}\right) \\
= \widehat{\alpha}_1 - \widehat{\alpha}_2 &+ 0.23936\widehat{(\alpha\beta)}_{11} + 0.14060\widehat{(\alpha\beta)}_{12} + 0.29996\widehat{(\alpha\beta)}_{13} + 0.19029\widehat{(\alpha\beta)}_{14} \\
&+ 0.12979\widehat{(\alpha\beta)}_{15} - 0.23936\widehat{(\alpha\beta)}_{21} - 0.14060\widehat{(\alpha\beta)}_{22} - 0.29996\widehat{(\alpha\beta)}_{23} \\
&- 0.19029\widehat{(\alpha\beta)}_{24} - 0.12979\widehat{(\alpha\beta)}_{25}.
\end{aligned}$$

Program 1.5 computes the Type II estimate and its standard error using the ESTIMATE statement in PROC GLM.

Program 1.5 Type II estimate of the average treatment difference in the depression trial example

```
proc glm data=hamd17;
    class drug center;
    model change=drug|center/ss2;
    estimate "Trt diff"
        drug 1 -1
        drug*center 0.23936 0.14060 0.29996 0.19029 0.12979
            -0.23936 -0.14060 -0.29996 -0.19029 -0.12979;
    run;
```

Output from Program 1.5

```
                                            Standard
Parameter                   Estimate           Error    t Value    Pr > |t|

Trt diff                  5.97871695      0.94351091       6.34      <.0001
```

Output 1.5 shows the Type II estimate of the average treatment difference and its standard error. As in the Type I framework, the t statistic in Output 1.5 equals the square root of the corresponding F statistic in Output 1.4, which implies that the two tests are equivalent. Note also that the t statistics for the treatment comparison produced by the Type I and II analysis methods are very close in magnitude, $t = 6.33$ in Output 1.3 and $t = 6.34$ in Output 1.5. This similarity is not a coincidence and is explained by the fact that patient randomization was stratified by center in this trial. As a consequence, n_{1j} is close to n_{2j} for any $j = 1, \ldots, 5$ and thus $n_{1j}n_{2j}/(n_{1j} + n_{2j})$ is proportional to n_{1j}. The weighting schemes underlying the Type I and II tests are almost identical to each other, which causes the two methods to yield similar results. Since the Type II method becomes virtually identical to the simple Type I method when patient randomization is stratified by the covariate used in the analysis, one does not gain much from using the randomization factor as a covariate in a Type II analysis. In general, however, the standard error of the Type II estimate of the treatment difference is considerably smaller than that of the Type I estimate and therefore the Type II method has more power to detect a treatment effect compared to the Type I method.

As demonstrated by Speed and Hocking (1980), the Type II method tests the simple hypothesis

$$H: \quad \frac{1}{m}\sum_{j=1}^{m}\mu_{1j} = \frac{1}{m}\sum_{j=1}^{m}\mu_{2j}$$

when the $\alpha\beta$ factor is assumed to equal 0 (Speed and Hocking, 1980). In other words, the Type II analysis method arises naturally in trials where the treatment difference does not vary substantially from stratum to stratum.

Type III Analysis

The Type III analysis is based on a generalization of the concepts underlying the Type I and Type II analyses. Unlike these two analysis methods, the Type III methodology relies on a reparametrization of the main-effects model (1.2). The reparametrization is performed by imposing certain restrictions on the parameters in (1.2) in order to achieve a full-rank model. For example, it is common to assume that

$$\sum_{i=1}^{2}\alpha_i = 0, \quad \sum_{j=1}^{m}\beta_j = 0,$$

$$\sum_{i=1}^{2}(\alpha\beta)_{ij} = 0, \quad j = 1, \ldots, m, \quad \sum_{j=1}^{m}(\alpha\beta)_{ij} = 0, \quad i = 1, 2. \tag{1.4}$$

Once the restrictions have been imposed, one can test the α, β and $\alpha\beta$ factors using the R quantities associated with the obtained reparametrized model (these quantities are commonly denoted by R^*).

The introduced analysis method is more flexible than the Type I and II analyses and allows one to test hypotheses that cannot be tested using the original R quantities (Searle, 1976; Speed and Hocking, 1976). For example, as shown by Searle (1971, Chapter 7), $R(\alpha|\mu, \beta, \alpha\beta)$ and $R(\beta|\mu, \alpha, \alpha\beta)$ are not meaningful when computed from the main-effects model (1.2) because they are identically equal to 0. This means that the Type I/II framework precludes one from fitting an interaction term before the main effects. By contrast, $R^*(\alpha|\mu, \beta, \alpha\beta)$ and $R^*(\beta|\mu, \alpha, \alpha\beta)$ associated with the full-rank reparametrized model can assume non-zero values depending on the constraints imposed on the model parameters. Thus, each term in (1.2) can be tested in the Type III framework using an adjustment for all other terms in the model.

The Type III analysis in PROC GLM and PROC MIXED assesses the significance of the α, β and $\alpha\beta$ factors using $R^*(\alpha|\mu, \beta, \alpha\beta)$, $R^*(\beta|\mu, \alpha, \alpha\beta)$ and $R^*(\alpha\beta|\mu, \alpha, \beta)$ with the parameter restrictions given by (1.4). As an illustration, Program 1.6 tests the significance of the treatment effect on HAMD17 changes using the Type III approach.

Program 1.6 Type III analysis of the HAMD17 changes in the depression trial example

```
proc glm data=hamd17;
    class drug center;
    model change=drug|center/ss3;
    run;
```

Output from Program 1.6

```
Source              DF    Type III SS    Mean Square   F Value   Pr > F

drug                 1    709.8195519    709.8195519     32.03   <.0001
center               4     91.4580063     22.8645016      1.03   0.3953
drug*center          4    507.4457539    126.8614385      5.72   0.0004
```

Output 1.6 indicates that the results of the Type III analysis are consistent with the Type I and II inferences for the treatment comparison. The treatment effect is highly significant after an adjustment for the center effect and treatment-by-center interaction ($F = 32.03$).

The advantage of making inferences from the reparametrized full-rank model is that the Type III hypothesis of no treatment effect has the following simple form (Speed, Hocking and Hackney, 1978):

$$H_{III}: \quad \frac{1}{m}\sum_{j=1}^{m}\mu_{1j} = \frac{1}{m}\sum_{j=1}^{m}\mu_{2j}.$$

The Type III hypothesis states that the simple average of the true stratum-specific HAMD17 change scores is identical in the two treatment groups. The corresponding Type III estimate of the average treatment difference is equal to

$$\frac{1}{m}\sum_{j=1}^{m}(\overline{y}_{1j\cdot} - \overline{y}_{2j\cdot}).$$

It is instructive to contrast this estimate with the Type I estimate of the average treatment difference. As was explained earlier, the idea behind the Type I approach is that individual observations are weighted equally. By contrast, the Type III method is based on weighting observations according to the size of each stratum. As a result, the Type III hypothesis involves a direct comparison of stratum means and is not affected by the number of patients in each individual stratum. To make an analogy, the Type I analysis corresponds to the U.S. House of Representatives, where the number of representatives from each state is a function of the state's population. The Type III analysis can be thought of as a statistical equivalent of the U.S. Senate, where each state sends along two Senators.

Since the Type III estimate of the average treatment difference in the depression trial example is given by

$$\begin{aligned}\widehat{\delta} = \widehat{\alpha}_1 - \widehat{\alpha}_2 + \frac{1}{5}\big[&\widehat{(\alpha\beta)}_{11} + \widehat{(\alpha\beta)}_{12} + \widehat{(\alpha\beta)}_{13} + \widehat{(\alpha\beta)}_{14} + \widehat{(\alpha\beta)}_{15} \\ &- \widehat{(\alpha\beta)}_{21} - \widehat{(\alpha\beta)}_{22} - \widehat{(\alpha\beta)}_{23} - \widehat{(\alpha\beta)}_{24} - \widehat{(\alpha\beta)}_{25}\big],\end{aligned}$$

we can compute the estimate and its standard error using the following ESTIMATE statement in PROC GLM.

Program 1.7 Type III estimate of the average treatment difference in the depression trial example

```
proc glm data=hamd17;
    class drug center;
    model change=drug|center/ss3;
    estimate "Trt diff"
        drug 1 -1
        drug*center 0.2 0.2 0.2 0.2 0.2 -0.2 -0.2 -0.2 -0.2 -0.2;
    run;
```

Output from Program 1.7

```
                                            Standard
Parameter                     Estimate         Error    t Value    Pr > |t|

Trt diff                    5.60912865    0.99106828       5.66      <.0001
```

Output 1.7 lists the Type III estimate of the treatment difference and its standard error. Again, the significance of the treatment effect can be assessed using the t statistic shown in Output 1.7 since the associated test is equivalent to the F test for the DRUG term in Output 1.6.

Comparison of Type I, Type II and Type III Analyses

The three analysis methods introduced in this section produce identical results in any balanced data set. The situation, however, becomes much more complicated and confusing in an unbalanced setting and one needs to carefully examine the available options to choose the most appropriate analysis method. The following comparison of the Type I, II and III analyses in PROC GLM and PROC MIXED will help the reader make more educated choices in clinical trial applications.

Type I Analysis

The Type I analysis method averages stratum-specific treatment differences with each observation receiving the same weight, and thus the Type I approach ignores the effects of individual strata on the outcome variable. It is clear that this approach can be used only if one is not interested in adjusting for the stratum effects.

Type II Analysis

The Type II approach amounts to comparing weighted averages of within-stratum estimates among the treatment groups. The weights are inversely proportional to the variances of stratum-specific estimates of the treatment effect, which implies that the Type II analysis is based on an optimal weighting scheme when there is no treatment-by-stratum interaction. When the treatment difference does vary across strata, the Type II test statistic can be viewed as a weighted average of stratum-specific treatment differences with the weights equal to sample estimates of certain population parameters. For this reason, it is commonly accepted that the Type II method is the preferred way of analyzing continuous outcome variables adjusted for prognostic factors (Fleiss, 1986; Mehrotra, 2001).

Attempts to apply the Type II method to stratification schemes based on nonprognostic factors (e.g., centers) have created much controversy in the clinical trial literature. Advocates of the Type II approach maintain that centers play the same role as prognostic factors, and thus it is appropriate to carry out Type II tests in trials stratified by center as shown in Program 1.4 (Senn, 1998; Lin, 1999). Note that the outcome of the Type II analysis is unaffected by the significance of the interaction term. The interaction analysis is run separately as part of routine sensitivity analyses such as the assessment of treatment effects in various subsets and the identification of outliers (Kallen, 1997; Phillips et al., 2000).

Type III Analysis

The opponents of the Type II approach argue that centers are intrinsically different from prognostic factors. Since investigative sites actively recruit patients, the number of patients enrolled at any given center is a rather arbitrary figure, and inferences driven by the sizes of individual centers are generally difficult to interpret (Fleiss, 1986). As an alternative, one can follow Yates (1934) and Cochran (1954a), who proposed to perform an analysis based on a simple average of center-specific estimates in the presence of a pronounced interaction. This unweighted analysis is equivalent to the Type III analysis of the model with an interaction term (see Program 1.6).

It is worth drawing the reader's attention to the fact that the described alternative approach based on the Type III analysis has a number of limitations:

- The Type II F statistic is generally larger than the Type III F statistic (compare Output 1.4 and Output 1.6) and thus the Type III analysis is less powerful than the Type II analysis when the treatment difference does not vary much from center to center.
- The Type III method violates the marginality principle formulated by Nelder (1977). The principle states that meaningful inferences in a two-way classification setting are to be based on the main effects α and β adjusted for each other and on their interaction adjusted for the main effects. When one fits an interaction term before the main effects (as in the Type III analysis), the resulting test statistics depend on a totally arbitrary choice of parameter constraints. The marginality principle implies that the Type III inferences yield uninterpretable results in unbalanced cases. See Nelder (1994) and Rodriguez, Tobias and Wolfinger (1995) for a further discussion of pros and cons of this argument.
- Weighting small and large strata equally is completely different from how one would normally perform a meta-analysis of the results observed in the strata (Senn, 2000).
- Lastly, as pointed out in several publications, sample size calculations are almost always done within the Type II framework; i.e., patients rather than centers are assumed equally weighted. As a consequence, the use of the Type III analysis invalidates the sample size calculation method. For a detailed power comparison of the weighted and unweighted approaches, see Jones et al. (1998) and Gallo (2000).

Type III Analysis with Pretesting

The described weighted and unweighted analysis methods are often combined to increase the power of the treatment comparison. As proposed by Fleiss (1986), the significance of the interaction term is assessed first and the Type III analysis with an interaction is performed if the preliminary test has yielded a significant outcome. Otherwise, the interaction term is removed from the model and thus the treatment effect is analyzed using the Type II approach. The sequential testing procedure recognizes the power advantage of the weighted analysis when the treatment-by-center interaction appears to be negligible.

Most commonly, the treatment-by-center variation is evaluated using an F test based on the interaction mean square; see the F test for the DRUG*CENTER term in Output 1.6. This test is typically carried out at the 0.1 significance level (Fleiss, 1986). Several alternative approaches have been suggested in the literature. Bancroft (1968) proposed to test the interaction term at the 0.25 level before including it in the model. Chinchilli and Bortey (1991) described a test for consistency of treatment differences across strata based on the noncentrality parameter of an F distribution. Ciminera et al. (1993) stressed that tests based on the interaction mean square are aimed at detecting *quantitative interactions* that may be caused by a variety of factors such as measurement scale artifacts. To alleviate the problems associated with the traditional pretesting approach, Ciminera et al. outlined an alternative method that relies on *qualitative interactions*; see Section 1.5 for more details.

When applying the pretesting strategy, one needs to be aware of the fact that pretesting leads to more frequent false-positive outcomes, which may become an issue in pivotal clinical trials. To stress this point, Jones et al. (1998) compared the described pretesting approach with the controversial practice of pretesting the significance of the carryover effect in crossover trials, a practice that is known to inflate the false-positive rate.

1.2.2 Random Effects Models

A popular alternative to the fixed effects modeling approach described in Section 1.2.1 is to explicitly incorporate random variation among strata in the analysis. Even though most of the discussion on center effects in the ICH guidance document "Statistical principles for clinical trials" (ICH E9) treats center as a fixed effect, the guidance also encourages trialists to explore the heterogeneity of the treatment effect across centers using mixed models. The latter can be accomplished by employing models with random stratum and treatment-by-stratum interaction terms. While one can argue that the selection of centers is not necessarily a random process, treating centers as a random effect could at times help statisticians better account for between-center variability.

Random effects modeling is based on the following mixed model for the continuous outcome y_{ijk} observed on the kth patient in the jth stratum in the ith treatment group:

$$y_{ijk} = \mu + \alpha_i + b_j + g_{ij} + \varepsilon_{ijk}, \tag{1.5}$$

where μ denotes the overall mean, α_i is the fixed effect of the ith treatment, b_j and g_{ij} denote the random stratum and treatment-by-stratum interaction effects, and ε_{ijk} is a residual term. The random and residual terms are assumed to be normally distributed and independent of each other. We can see from Model (1.5) that, unlike fixed effects models, random effects models account for the variability across strata in judging the significance of the treatment effect.

Applications of mixed effects models to stratified analyses in a clinical trial context were described by several authors, including Fleiss (1986), Senn (1998) and Gallo (2000). Chakravorti and Grizzle (1975) provided a theoretical foundation for random effects modeling in stratified trials based on the familiar randomized block design framework and the work of Hartley and Rao (1967). For a detailed overview of issues related to the analysis of mixed effects models, see Searle (1992, Chapter 3). Littell et al. (1996, Chapter 2) demonstrated how to use PROC MIXED in order to fit random effects models in multicenter trials.

Program 1.8 fits a random effects model to the HAMD17 data set using PROC MIXED and computes an estimate of the average treatment difference. The DDFM=SATTERTH option in Program 1.8 requests that the degrees of freedom for the F test be computed using the Satterthwaite formula. The Satterthwaite method provides a more accurate approximation to the distribution of the F statistic in random effects models than the standard ANOVA method (it is achieved by increasing the number of degrees of freedom for the F statistic).

Program 1.8 Analysis of the HAMD17 changes in the depression trial example using a random effects model

```
proc mixed data=hamd17;
    class drug center;
    model change=drug/ddfm=satterth;
    random center drug*center;
    estimate "Trt eff" drug 1 -1;
    run;
```

Output from Program 1.8

```
        Type 3 Tests of Fixed Effects

              Num     Den
Effect         DF      DF    F Value    Pr > F

drug            1    6.77       9.30    0.0194

                          Estimates

                       Standard
Label      Estimate       Error      DF    t Value    Pr > |t|

Trt eff      5.7072      1.8718    6.77       3.05      0.0194
```

Output 1.8 displays the F statistic ($F = 9.30$) and p-value ($p = 0.0194$) associated with the DRUG term in the random effects model as well as an estimate of the average treatment difference. The estimated treatment difference equals 5.7072 and is close to the estimates computed from fixed effects models. The standard error of the estimate (1.8718) is substantially greater than the standard error of the estimates obtained in fixed effects models (see Output 1.6). This is a penalty one has to pay for treating the stratum and interaction effects as random, and it reflects lack of homogeneity across the five strata in the depression data. Note, for example, that dropping Center 101 creates more homogeneous strata and, as a consequence, reduces the standard error to 1.0442. Similarly, removing the DRUG*CENTER term from the RANDOM statement leads to a more precise estimate of the treatment effect with the standard error of 1.0280.

In general, as shown by Senn (2000), fitting main effects as random leads to lower standard errors; however, assuming a random interaction term increases the standard error of the estimated treatment difference. Due to the lower precision of treatment effect estimates, analysis of stratified data based on models with random stratum and treatment-by-stratum effects has lower power compared to a fixed effects analysis (Gould, 1998; Jones et al., 1998).

1.2.3 Nonparametric Tests

This section briefly describes a nonparametric test for stratified continuous data proposed by van Elteren (1960). To introduce the van Elteren test, consider a clinical trial with a continuous endpoint measured in m strata. Let w_j denote the Wilcoxon rank-sum statistic for testing the null hypothesis of no treatment effect in the jth stratum (Hollander and Wolfe, 1999, Chapter 4). Van Elteren (1960) proposed to combine stratum-specific Wilcoxon rank-sum statistics with weights inversely proportional to stratum sizes. The van Elteren statistic is given by

$$u = \sum_{j=1}^{m} \frac{w_j}{n_{1j} + n_{2j} + 1},$$

where $n_{1j} + n_{2j}$ is the total number of patients in the jth stratum. To justify this weighting scheme, van Elteren demonstrated that the resulting test has asymptotically the maximum power against a broad range of alternative hypotheses. Van Elteren also studied the asymptotic properties of the testing procedure and showed that, under the null hypothesis of no treatment effect in the m strata, the test statistic is asymptotically normal.

As shown by Koch et al. (1982, Section 2.3), the van Elteren test is a member of a general family of Mantel-Haenszel mean score tests. This family also includes the Cochran-Mantel-Haenszel test for categorical outcomes discussed later in Section 1.3.1. Like other testing procedures in this family, the van Elteren test possesses an interesting and useful property that its asymptotic distribution is not directly affected by the size of individual strata. As a consequence, one can rely on asymptotic p-values even in sparse stratifications as long as the total sample size is large enough. For more information about the van Elteren test and related testing procedures, see Lehmann (1975), Koch et al. (1990) and Hosmane, Shu and Morris (1994).

EXAMPLE: Urinary Incontinence Trial

The van Elteren test is an alternative method of analyzing stratified continuous data when one cannot rely on standard ANOVA techniques because the underlying normality assumption is not met. As an illustration, consider a subset of the data collected in a urinary incontinence trial comparing an experimental drug to a placebo over an 8-week period. The primary endpoint in the trial was a percent change from baseline to the end of the study in the number of incontinence episodes per week. Patients were allocated to three strata according to the baseline frequency of incontinence episodes.[2]

Program 1.9 displays a subset of the data collected in the urinary incontinence trial and plots the probability distribution of the primary endpoint in the three strata.

Program 1.9 Distribution of percent changes in the frequency of incontinence episodes in the urinary incontinence trial example

```
data urininc;
    input therapy $ stratum @@;
    do i=1 to 10;
        input change @@;
        if (change^=.) then output;
    end;
    drop i;
    datalines;
Placebo  1  -86  -38   43 -100  289    0  -78   38  -80  -25
Placebo  1 -100 -100  -50   25 -100 -100  -67    0  400 -100
Placebo  1  -63  -70  -83  -67  -33    0  -13 -100    0   -3
Placebo  1  -62  -29  -50 -100    0 -100  -60  -40  -44  -14
Placebo  2  -36  -77   -6  -85   29  -17  -53   18  -62  -93
Placebo  2   64  -29  100   31   -6 -100  -30   11  -52  -55
Placebo  2 -100  -82  -85  -36  -75   -8  -75  -42  122  -30
Placebo  2   22  -82    .    .    .    .    .    .    .    .
Placebo  3   12  -68 -100   95  -43  -17  -87  -66   -8   64
Placebo  3   61  -41  -73  -42  -32   12  -69   81    0   87
Drug     1   50 -100  -80  -57  -44  340 -100 -100  -25  -74
Drug     1    0   43 -100 -100 -100 -100  -63 -100 -100 -100
Drug     1 -100 -100    0 -100  -50    0    0  -83  369  -50
Drug     1  -33  -50  -33  -67   25  390  -50    0 -100    .
Drug     2  -93  -55  -73  -25   31    8  -92  -91  -89  -67
Drug     2  -25  -61  -47  -75  -94 -100  -69  -92 -100  -35
Drug     2 -100  -82  -31  -29 -100  -14  -55   31  -40 -100
Drug     2  -82  131  -60     .   .    .    .    .    .    .
Drug     3  -17  -13  -55  -85  -68  -87  -42   36  -44  -98
Drug     3  -75  -35    7  -57  -92  -78  -69   -21 -14    .
;
```

[2] This clinical trial example will be used here to illustrate a method for the analysis of non-normally distributed endpoints in the presence of a categorical stratification variable. One can think of other ways of analyzing the urinary incontinence data that may be more appropriate in this setting. For example, one can consider redefining the primary outcome variable since a variable based on percent change from baseline makes an inefficient use of data. Further, categorizing continuous data leads to loss of power and thus the analysis described above will be inferior to an analysis which uses the baseline frequency of incontinence episodes as a continuous covariate. Yet another sensible approach is based on fitting a model that accounts for the discrete nature of incontinence episodes, e.g., a Poisson regression model for counts.

```
proc sort data=urininc;
    by stratum therapy;
proc kde data=urininc out=density;
    by stratum therapy;
    var change;
proc sort data=density;
    by stratum;
* Plot the distribution of the primary endpoint in each stratum;
%macro PlotDist(stratum,label);
axis1 minor=none major=none value=none label=(angle=90 "Density")
    order=(0 to 0.012 by 0.002);
axis2 minor=none order=(-100 to 150 by 50)
    label=("&label");
symbol1 value=none color=black i=join line=34;
symbol2 value=none color=black i=join line=1;
data annotate;
    xsys="1"; ysys="1"; hsys="4"; x=50; y=90; position="5";
    size=1; text="Stratum &stratum"; function="label";
proc gplot data=density anno=annotate;
    where stratum=&stratum;
    plot density*change=therapy/frame haxis=axis2 vaxis=axis1 nolegend;
    run;
    quit;
%mend PlotDist;
%PlotDist(1,);
%PlotDist(2,);
%PlotDist(3,Percent change in the frequency of incontinence episodes);
```

The output of Program 1.9 is shown in Figure 1.2. We can see from Figure 1.2 that the distribution of the primary outcome variable is consistently skewed to the right across the three strata. Since the normality assumption is clearly violated in this data set, the analysis methods described earlier in this section may perform poorly. The magnitude of treatment effect on the frequency of incontinence episodes can be assessed more reliably using a nonparametric procedure. Program 1.10 computes the van Elteren statistic to test the null hypothesis of no treatment effect in the urinary incontinence trial using PROC FREQ. The statistic is requested by including the CMH2 and SCORES=MODRIDIT options in the TABLE statement.

Program 1.10 Analysis of percent changes in the frequency of incontinence episodes using the van Elteren test

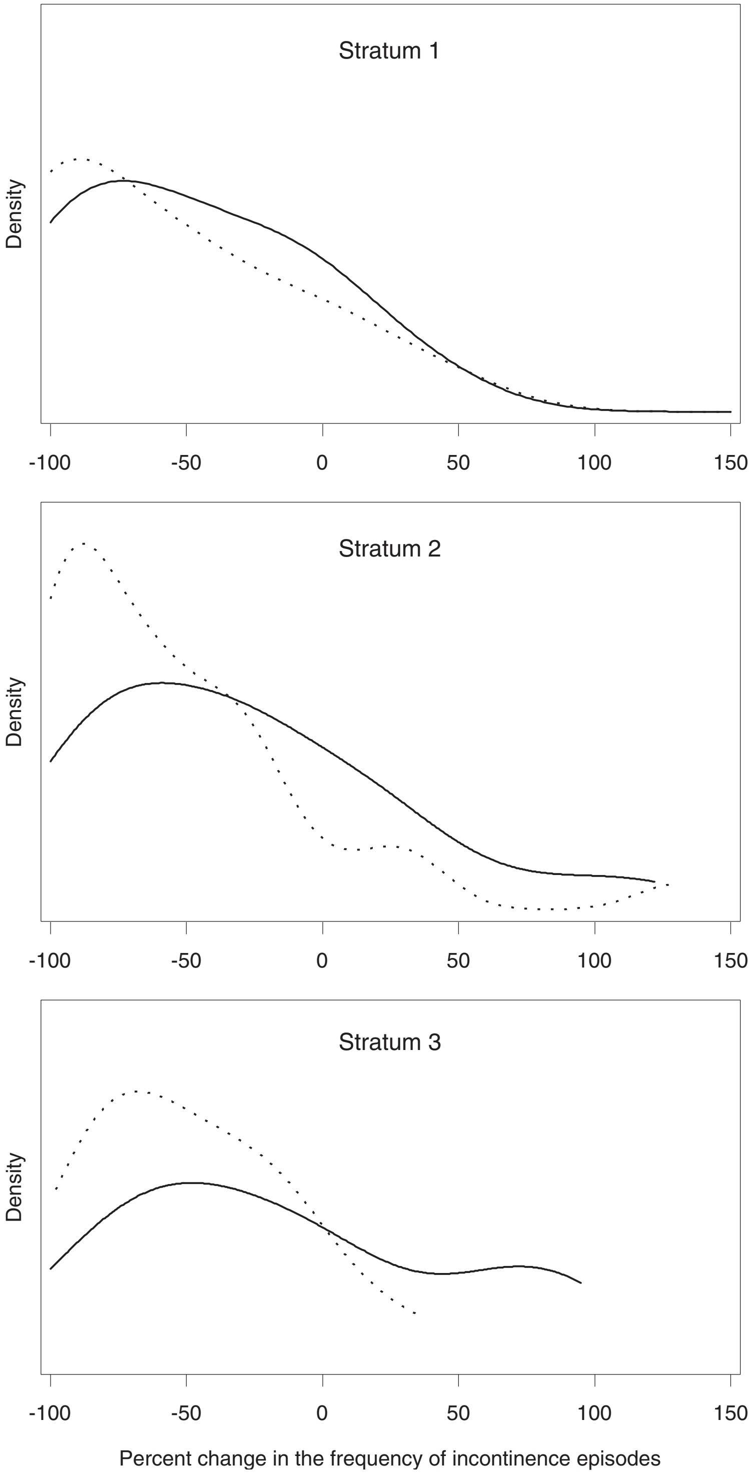
```
proc freq data=urininc;
    ods select cmh;
    table stratum*therapy*change/cmh2 scores=modridit;
    run;
```

Output from Program 1.10

```
Summary Statistics for therapy by change
Controlling for stratum
  Cochran-Mantel-Haenszel Statistics (Modified Ridit Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      6.2505    0.0124
    2        Row Mean Scores Differ     1      6.2766    0.0122
```

Figure 1.2 The distribution of percent changes in the frequency of incontinence episodes in the experimental (· · ·) and placebo (—) groups by stratum in the urinary incontinence trial



Output 1.10 lists two statistics produced by PROC FREQ (note that extraneous information has been deleted from the output using the ODS statement). The van Elteren statistic corresponds to the row mean scores statistic labeled "Row Mean Scores Differ" and is equal to 6.2766. Since the asymptotic p-value is small ($p = 0.0122$), we conclude that administration of the experimental drug resulted in a significant reduction in the frequency of incontinence episodes. To compare the van Elteren test with the Type II

and III analyses in the parametric ANOVA framework, Programs 1.4 and 1.6 were rerun to test the significance of the treatment effect in the urinary incontinence trial. The Type II and III F statistics were equal to 1.4 ($p = 0.2384$) and 2.15 ($p = 0.1446$), respectively. The parametric methods were unable to detect the treatment effect in this data set due to the highly skewed distribution of the primary endpoint.

1.2.4 Summary

This section discussed parametric and nonparametric methods for performing stratified analyses in clinical trials with a continuous endpoint. Parametric analysis methods based on fixed and random effects models are easy to implement using PROC GLM (fixed effects only) or PROC MIXED (both fixed and random effects).

PROC GLM and PROC MIXED support three popular methods of fitting fixed effects models to stratified data. These analysis methods, known as Type I, II and III analyses, are conceptually similar to each other in the sense that they are all based on averaging stratum-specific estimates of the treatment effect. The following is a quick summary of the Type I, II and III methods:

- Each observation receives the same weight when a Type I average of stratum-specific treatment differences is computed. Therefore, the Type I approach ignores the effects of individual strata on the outcome variable.
- The Type II approach is based on a comparison of weighted averages of stratum-specific estimates of the treatment effect, with the weights being inversely proportional to the variances of these estimates. The Type II weighting scheme is optimal when there is no treatment-by-stratum interaction and can also be used when treatment differences vary across strata. It is generally agreed that the Type II method is the preferred way of analyzing continuous outcome variables adjusted for prognostic factors.
- The Type III analysis method relies on a direct comparison of stratum means, which implies that individual observations are weighted according to the size of each stratum. This analysis is typically performed in the presence of a significant treatment-by-stratum interaction. It is important to remember that Type II tests are known to have more power than Type III tests when the treatment difference does not vary much from stratum to stratum.

The information about treatment differences across strata can also be combined using random effects models in which stratum and treatment-by-stratum interaction terms are treated as random variables. Random effects inferences for stratified data can be implemented using PROC MIXED. The advantage of random effects modeling is that it helps the statistician better account for between-stratum variability. However, random effects inferences are generally less powerful than inferences based on fixed effects models. This is one of the reasons why stratified analyses based on random effects models are rarely performed in a clinical trial setting.

A stratified version of the nonparametric Wilcoxon rank-sum test, known as the van Elteren test, can be used to perform inferences in a non-normal setting. It has been shown that the asymptotic distribution of the van Elteren test statistic is not directly affected by the size of individual strata and therefore this testing procedure performs well in the analysis of a large number of small strata.