CHAPTER 11

Repeated Measures Designs

Repeated measures designs are designs for which, as the name suggests, repeated measurements are made on the experimental units, which are typically people as the design is often used in clinical work and in education. Certainly, repeated measures on the same subject are a necessary part of any design for studying rates of learning as a function of treatment effects. In clinical work, subjects may be measured over time after receiving some type of treatment.

At the outset, we should note that not much attention is given to repeated measures designs in most books on experimental design; exceptions are Hinkelmann and Kempthorne (2005), Giesbrecht and Gumpertz (2004), Kuehl (2000), and Winer, Brown, and Michels (1991). In particular, the latter has one chapter on single-factor designs with repeated measures and a chapter on multifactor designs with repeated measures. The best book that is devoted almost exclusively to repeated measures designs is probably Davis (2002). Other books on the subject include Lindsey (1999), Crowder and Hand (1990), Vonesh and Chinchilli (1997), Girden (1992), and Davidian and Giltinan (1995). Although somewhat outdated, Fleiss (1986) has two chapters on repeated measures designs. Review papers on the subject include Keselman, Algina, and Kowalchuk (2001), which is on the analysis of data from repeated measures designs, as is Everitt (1995), who provides illustrative examples.

There are both advantages and disadvantages associated with these designs. One advantage is that information on the time trend of the response variable is available under different treatment conditions. Furthermore, each subject serves as his or her own control, which permits the use of a smaller sample size. More specifically, subject-to-subject variation is not problematic when each subject receives all the treatments instead of each subject receiving only one treatment. This is especially important when the experimental unit/subject is a person because it can be difficult to select people who are homogeneous on characteristics that should be essentially constant for a given study.

Modern Experimental Design By Thomas P. Ryan

11.1 ONE FACTOR

The simplest example of a repeated measures experiment is when a group of people are measured on a certain characteristic both before and after some type of treatment, such as an industrial training program or a medication program. Of course this is a classic example of when a paired t-test would be used as the method of analysis.

A paired t-test cuts in half the number of observations used in the analysis, relative to a pooled t-test, since differences are formed. There are four components of the total variation when ANOVA is used, with the initial breakdown being between people and within people, with the latter broken down into treatments and residual.

Example 11.1

An experiment is conducted using nine experimental subjects in two breathing chambers, with each subject spending time in each chamber. One chamber had a high concentration of carbon monoxide, while the other did not. Our scientific interest is in establishing that exposure to carbon monoxide significantly increases respiration rate.

Subject	With CO	Without CO
1	30	30
2	45	40
3	26	25
4	24	23
5	34	30
6	51	49
7	46	41
8	32	35
9	30	28

Of course an F-test in ANOVA is nondirectional and the alternative hypothesis for this experiment is obviously directional, but we will use this as an example of the computations with each approach. (Of course with any experimentation the objective is generally to determine that some treatment is the best.)

The simplest *possible* model for the repeated measures design for this application is

$$Y_{ij} = \mu + A_i + B_j + \epsilon_{ij} \quad i = 1, 2, \ldots, 9 \quad j = 1, 2 \tag{11.1}$$

with μ denoting the mean, A_i the effect of the ith subject, and B_j the effect of the jth treatment. Of course the subject effect is considered to be a random effect. Notice that this is simply the model for two-factor ANOVA without interaction.

The model given by Eq. (11.1) may not be appropriate, however, and frequently won't be appropriate. In particular, observations made on the same experimental unit will frequently be dependent. This dependency might manifest itself in one or more

ways, such as carryover effects or correlation because the subjects respond differently to treatments, so that there will be correlation between treatments within subjects. Carryover effects and designs for carryover effects are discussed in Section 11.4. If much time elapses between the repeat measurements (which would be one way to try to avoid carryover effects), there might also be a period effect, perhaps caused by changing environmental conditions that serve as extraneous factors. Of course if there is a period effect, it would be confounded with the treatment effect. If experimental subjects were very slow to receive each treatment, there might be period effects within each treatment.

The bottom line is that careful thought must be given to the model in a repeated measures experiment and appropriate assumptions may not be easily determined.

The ANOVA table for the repeated measures data in this example is as follows.

Source	df	SS	MS	F
Between subjects	8	1289.78		
Within subjects	9	42.50		
Treatments	1	16.06	16.06	4.86
Residual	8	26.44	3.305	
Total	17	1332.28		

The between subjects and within subjects sum of squares are computed in the same way as in one-factor ANOVA if we regard subjects as the factor. Since there are multiple treatments for each subject, a treatment sum of squares can thus be broken out of the within subjects sum of squares. Another way to view this is to recognize that differences within subjects must be due at least in part to differences between treatments, with the remainder unaccounted for by the treatments.

Since we can analyze these data using a paired t-test, it stands to reason that the results must be numerically equivalent. That is, the numerical value of the t-statistic must be equal to $\sqrt{4.86}$. It can be shown that the value of the t-statistic is $1.89/0.857 = 2.205$ which, within rounding, does equal $\sqrt{4.86}$.

The analysis using a paired t-test is much simpler than the ANOVA approach since there are fewer computations to perform, but of course a paired t-test cannot be used if there are multiple factors or more than two levels of a single factor. (There will usually be more than two levels of a single factor.) Nevertheless, the paired t-test is certainly more intuitive than the ANOVA approach and should thus be used whenever possible.

The analysis of repeated measures data for one factor with more than two levels is straightforward, with the analysis being of the same general form as the analysis when there are only two levels. (Regarding the analysis of data from repeated measures designs, it is worth noting that this capability is absent in some of the leading software, such as MINITAB and JMP, with the latter providing capability only through a mixed model approach (i.e., not for a single factor or for multiple factors with other models).)

11.1.1 The Example in Section 2.1.2

In the example in Section 2.1.2, nine observations on each subject were averaged to produce a single number, which permitted the data to be analyzed as either a pooled t-test or its ANOVA equivalent. Although such condensing of data frequently occurs in practice, it would be better to use all the data. There might, for example, be a period effect and it would be useful to be able to test for this. We will see how a possible period effect can be isolated in subsequent sections in this chapter.

11.2 MORE THAN ONE FACTOR

The modeling and analysis that accompanies repeated measures designs is much more involved when there are multiple factors.

Example 11.2

Milliken (2004) gave an example with two factors in a semiconductor experiment, which was performed to investigate factors that influence the thickness of the oxide layer that develops on the surface of silicon wafers. The two factors that were used in the experiment were furnace temperature (900, 1000, and 1100° F) and position in the furnace (top, middle, and bottom). There is a factorial arrangement of levels because all nine combinations of furnace and position are used, with nine wafers from one lot used in each replicate of the experiment, with three wafers going into each of three furnaces. This process was then repeated four times with wafers from a different lot used in each replication. The three temperatures were randomly assigned to a furnace run but position obviously could not be randomly assigned to the part of the furnace, which was the experimental unit for the temperature factor, so the design is thus a repeated measures design. The model for the experiment is given by

$$Y_{ijk} = \mu + A_i + B_j + C_{ij} + D_k + AD_{ik} + E_{ijk}$$

with μ denoting the mean, A_i the temperature effect, B_j the lot effect, C_{ij} the furnace effect, D_k the furnace position effect, AD_{ik} the interaction between temperature and furnace position, and E_{ijk} the effect of positions within furnaces. The latter is the error term. The E_{ijk} are not independent because the positions are (obviously) not randomly assigned to the parts of the furnace. Consequently, the E_{ijk} are correlated on the third subscript, which of course corresponds to the position effect.

The covariance structure must be determined. Milliken (2004) suggests testing a small number of reasonable covariance structures and selecting the one that seems most appropriate, pointing out that PROC Mixed in SAS provides several information criteria that can be used in determining which covariance structure best fits the data. (Crowder (2001) considered the analysis of data from repeated measures designs when the covariance structure has been misspecified.)

Milliken (2004) provided a thorough analysis of the data, analyzing the data for each of five covariance structures. The position × temperature (AD) interaction effect

is significant at the .05 level for four of those structures. Of those, the consequence of this is that the temperature and position main effects should not be computed and interpreted in the usual way. This especially applies to the analysis using the Toeplitz covariance structure since the p-value for temperature is .09.

Senn, Stevens, and Chaturvedi (2000) discussed when it can be advantageous to analyze repeated measures data using summary measures. Street, Eccleston, and Wilson (1990) gave some tables of optimal, small repeated measures designs and Kushner (1998) similarly studied optimal and efficient repeated measures designs under the assumption of uncorrelated observations. Carriére (1999) gave methods for analyzing data from repeated measures designs when there are missing observations. Kowalchuk, Keselman, Algina, and Wolfinger (2004) discussed the analysis of repeated measures data using mixed model adjusted F-tests.

11.3 CROSSOVER DESIGNS

The general idea is that there is a change (i.e., a "crossover") in the sequence in which treatments are applied to the subjects. An obvious question to ask is "Why the reversal?" Recall that in a randomized block design, the units are assumed to be homogeneous within blocks and each treatment is applied to only one experimental unit. Therefore, it shouldn't make any difference whether the ordering of treatments within a block is, say, $A\ B\ C$ or $C\ A\ B$. If there is expected to be a difference in the order in which the treatments are applied, however, such as a practice effect since each experimental subject receives more than one treatment, then obviously attention must be given to the sequencing.

The simplest crossover design is a design with two treatments and two periods, with half of the subjects receiving treatment A first followed by treatment B, and the other half receiving treatment B first followed by treatment A. For this design to work, there must be a washout period between adjacent treatment periods. Senn (1996) described the proper use of this design but did not advocate its usage. When there is no carryover effect (to be discussed shortly), the data can be analyzed using a two-sample t-test after first forming the differences $B_{1i} - A_{1i}, i = 1, 2, \ldots, n_1$, and $A_{2i} - B_{2i}, i = 1, 2, \ldots, n_2$, assuming n_1 subjects in the first group and n_2 subjects in the second group, with B_{1i} denoting the measurement obtained on the ith subject, the first period, and the "B" treatment, and similarly for the other three.

The primary motivation for using crossover designs is that they are economical relative to parallel group designs in which the subjects are randomly allocated among the treatments. Crossover designs eliminate most of the between-subject variation and have the same power to detect treatment differences as parallel group designs with far more subjects. This is important when there is a limited number of subjects, as will often be the case when human or animal subjects are used.

Crossover designs are used in various types of applications, especially sensory testing (see, e.g., Wakeling and MacFie, 1995; Kunert, 1998; Périnel and Pagès, 2004 and Kunert and Sailer, 2006), telecommunications (Lewis and Russell, 1998),

bioenvironmental and public health studies (Tudor, Koch, and Catellier, 2000), medical applications (Matthews, 1994), and dentistry (Claydon, Addy, Newcombe, and Moran, 2005). Early applications were in a variety of areas including weather modification experiments (Mielke, 1974), psychological experiments (Keppel, 1973), clinical trials (Grizzle, 1965), bioassay (Finney, 1965), tea-tasting experiments, crop experiments, cow-lactating experiments, and other applications. Methods of selecting crossover designs are discussed by Loughin, Johnson, Ives, and Nagaraja (2002).

Crossover designs have been used extensively in medical applications, especially in clinical trials for studying chronic ailments, and in other areas (see, e.g., Max, 2003; Senn, 2003; and Carriére, 1994). These designs have also been the primary designs used for bioavailability and bioequivalence studies. Unfortunately, Garcia, Benet, Arnau, and Cobo (2004) observed that in the 40 published crossover studies that they studied for the period 2000–2003, 18 of them did not give effect estimates and associated standard errors, thus preventing their use in any type of meta-analysis. This omission is in violation of the CONSORT (Consolidated Standards of Reporting Trials) recommendations. In comparing crossover designs and parallel designs, Garcia et al. (2004) found that parallel designs "... need, in order to achieve the same power, between 4 and 10 times more subjects than the corresponding cross-over design. ..."

The literature contains some slightly conflicting information about what a crossover design actually is. For example, Cochran and Cox (1957, p. 127) state that a crossover design "closely resembles" a Latin square design, whereas Montgomery (1996, p. 204) presents a particular crossover design as a set of Latin square designs, with the corresponding analysis. There has also been some confusion in the literature regarding the correct analysis of data from crossover designs. This is explained at the end of the Section 11.4.

It is clear, however, that a crossover design, which is also called a *changeover design*, is a type of repeated measures design since the subjects receive more than one treatment. Various researchers have extolled the benefits of crossover designs versus parallel (group) designs (i.e., completely randomized designs), for which each subject receives only one treatment. The main problem with parallel designs is generally the inhomogeneity of subjects that creates substantial between-subject variation and, consequently, can make it difficult to detect significant effects. This variation is either removed or largely reduced when treatment comparisons are made entirely or almost entirely within subjects. More specifically, crossover designs can have more power than parallel designs even when the latter have as many as 10 times the number of subjects (Louis, Lavori, Bailar, and Polansky, 1984).

Crossover designs are not without shortcomings, however, as there can be carryover effects from one treatment to the next since subjects are receiving more than one treatment. Furthermore, if treatments are spread out very far in time so as to eliminate or at least minimize carryover effects, experiments may require a long time to run. We continue our discussion of carryover effects after first looking at an example in which a crossover design was purportedly used.

Specifically, we consider the following example from the literature.

Example 11.3

Lawson (1988) described the use of what he termed a crossover design in an industrial application. The design layout was as follows:

C	B	D	A	D	A	A	B	B	C	B	C	B	D	C	D
D	D	A	B	A	B	D	D	C	B	D	D	A	B	B	A
B	C	B	D	C	C	C	C	D	D	C	A	C	A	A	C
A	A	C	C	B	D	B	A	A	A	A	B	D	C	D	B

with the rows of the design representing the four furnaces, the columns designating replicates, and the letters representing four different mixtures of three reducing agents that were randomly assigned to the four furnaces *during each time period*. Of course this assignment of the mixtures to the furnaces causes the design to be unbalanced in that the different mixtures do not occur the same number of times in each furnace. In particular, mixture B is used only twice in the third furnace. This problem could have been avoided by using four 4×4 Latin squares.

Although this design layout was termed a crossover design by Lawson (1988), it actually has the basic characteristic of a randomized block design at each period, with the design becoming highly unbalanced when the "blocks" are formed by stringing the periods together. Of course in a randomized block design the experimental units within each block are considered to be homogeneous. This design should not be used if a furnace effect is expected so that furnace should be blocked on.

The raw data were not given by Lawson (1988), so a reanalysis of the data cannot be given here. Any such analysis would be flawed if there were a furnace effect, however. Indeed there was a furnace effect for the two dependent variables (1) average furnace efficiency in percent, and (2) percent of product lost in the slag, the p-value being less than .01 for each dependent variable.

The problem with a crossover design, as has been pointed out by various authors, is that it breaks down when there is carryover; that is, when there is a residual effect of a treatment such as the measurement on a subject in the second period influenced by the measurement on the same subject from the other treatment in the first period. As is illustrated by Giesbrecht and Gumpertz (2004, p. 261), any carryover effect is confounded with other effects. Another possible problem with crossover designs (and in general when the experimental unit is a person) is that subjects might drop out of the study. Low, Lewis, and Prescott (1999) assessed the robustness of crossover designs to subjects dropping out and Low, Lewis, and Prescott (2002) also addressed the dropout problem. Boon and Roes (1999) considered various design and analysis issues when crossover designs are used in phase I clinical trials. Federer and Kerschner (1998) compared different classes of crossover designs. Kempton, Ferris, and David (2001) considered optimal crossover designs under the assumption that carryover effects are proportional to direct effects, with the latter being the treatment effects. This paper motivated the work of Bailey and Kunert (2005). Stufken (1996) reviewed the optimality properties of crossover designs, and other work on optimal crossover

designs includes Kunert and Stufken (2002), Bose and Dey (2003), and Bose and Mukherjee (2003). Part A of Bose (2002), which is available on the Internet (see Bose, 2002), is based on Bose and Mukerjee (2003) and Part B is reproduced by Bose and Dey (2003).

Most of the optimality results in the literature are based on the model of Cheng and Wu (1980), which assumed that (1) carryover effects stop after the first period, (2) there is no interaction between the treatments applied in successive periods to the same subject, (3) the subject effects are fixed effects, and (4) the errors are independent with mean zero and constant variance. Unfortunately, as discussed by Bose (2002), one or more of these assumptions is likely to be violated in the type of applications of the design that have been used in recent years. In particular, as stated previously, subjects are generally assumed to be a random factor and problems with the second assumption will be discussed shortly. The independent errors assumption might also be shaky in certain applications. Donev (1998) considered crossover designs with correlated observations, and Matthews (1987) considered crossover designs with both carryover effects and autocorrelated observations.

As discussed by Bose (2002), optimal crossover designs are very sensitive to the model assumptions. Bose and Mukherjee (2000, 2003) did obtain optimal designs when some or all of these assumptions are relaxed, however. In particular, the second assumption has been criticized as being a weakness of the model, as discussed by Kunert (1987), for example. Sen and Mukerjee (1987) developed optimal designs for direct and carryover effects designs when the second assumption is relaxed, and assumed that each treatment has a different carryover effect for each treatment in the next period. Unfortunately, this assumption generally required too many parameters for the model to be useful. Hedayat and Afsarinejad (2002) offered a compromise model that assumed two carryover effects: one when a treatment follows itself and the other when the next treatment is any other treatment. These have been termed *self* and *mixed* carryover effects, respectively, in the literature, as in Hedayat and Afsarinejad (2002), and Kunert and Stufken (2002, 2005).

Hedayat and Stufken (2003) appropriately pointed out that a design that is efficient under various plausible models is preferable to one that is optimal under one model but performs poorly under certain other models. Their primary focus was performance under two models. Similarly, John, Russell, and Whitaker (2004) described an algorithm for constructing crossover designs that are efficient under a range of models.

A moderate amount of work has been performed to develop useful crossover designs. For example, Russell and Dean (1998) gave crossover designs when only a small number of subjects is available. See also Bate and Jones (2006) for the construction of crossover designs.

11.4 DESIGNS FOR CARRYOVER EFFECTS

When carryover is believed to exist, a crossover design should be balanced for carryover effects. A crossover design is balanced for (first-order) carryover effects if each treatment follows every other treatment equally often. The simplest designs of this

type are the ones introduced by Williams (1949), with Patterson (1952) presenting more general crossover designs for carryover effects. Carriére and Reinsel (1993) proved that the designs of Patterson (1952) are optimal for estimating direct treatment effects when there are two periods. John et al. (2004) gave an algorithm for constructing crossover designs that are efficient under various models for explaining carryover effects. One such model, a mixed effects model, was given by Putt and Chinchilli (1999), who assumed that carryover extended over only one period.

The designs given by Williams are, not surprisingly, called Williams squares and they are also referred to as column-complete Latin squares. One such Williams square is

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

with the letters representing the treatments, the columns representing the subjects, and the rows representing the treatment periods. Notice that each treatment comes first in one treatment period and in the other three periods, it is preceded by each of the other treatments. Thus, if there is a carryover effect from one treatment period to the next period, this design essentially adjusts for that carryover. For example, if there is a strong carryover from treatment A, the measurement using treatment B in the first treatment period will be affected, but the measurements using treatment C in the second time period and treatment D in the third time period should be similarly affected.

The design given above is not unique as Cochran and Cox (1957, p. 134) gave the following design, rearranged slightly here with the columns representing the treatment periods, which also has the required property.

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

Actually, more than one square should be used for the same reason that multiple Latin square designs should be used, as was discussed in Section 3.3.5. The designs presented by Williams (1949) were for an even number of treatments. Much more recently, Newcombe (1995, 1996) gave designs for an odd number of factors. For example, the design given for five factors and 15 subjects is

Subject	Sequence
1	ABDCE
2	BCEDA
3	CDAEB
4	DEBAC
5	EACBD
6	ACDBE

7	BDECA
8	CEADB
9	DABEC
10	EBCAD
11	ABEDC
12	BCAED
13	CDBAE
14	DECBA
15	EADCB

Although the designs given by Williams (1949) are Latin squares, the corresponding analysis methods that were provided break out the sum of squares for residual effects as well as a sum of squares for "direct effects." Those methods require that at least two squares be used, but more should actually be used because, for example, there will be only four degrees of freedom for the error term when there are three treatments and two squares. Of course there is the obvious question of whether or not this is necessary since the design adjusts for residual effects.

Williams designs can be analyzed by using the "Williams" option in the CROSSDES package in *R* (see Section 11.7), and Hinkelmann and Kempthorne (2005, p. 700) illustrate the analysis of data from a Williams design using SAS PROC GLM.

One question that arises is whether or not a preliminary test for carryover effects should be performed, with the outcome of the test influencing the form of the analysis, similar to letting interactions that aren't significant become part of the error term so as to increase the degrees of freedom for the error term. The general recommendation is that there should be no preliminary test for carryover effects.

Whether or not there is any carryover effect depends on what is being measured. Cochran and Cox (1957) gave an example of a crossover design to compare the computing speeds of two machines. One person used each of two machines to compute a sum of squares of 27 observations. There were 10 such datasets, which comprised the replications of the design. Since it was felt that the second computation of a sum of squares might be faster than the first computation, the order in which the machines were used was reversed in half of the replications.

Thus, the design was of the following form:

A	B	A	B	B	A	B	A	A	B
B	A	B	A	A	B	A	B	B	A

with the columns representing the replications.

The analysis of such a design is similar to the analysis of a Latin square design in that row and column effects are isolated and separated from error. Sums of squares for rows and columns are computed in the usual way and treatments are tested against error.

A variation of this design (as described by Cochran and Cox, 1957) is to break up the above configuration into five 2×2 Latin square designs. The advantage of this is that row-to-row variation can be broken up into that variation within squares. This is

an advantage if the replicates are such that row-to-row variation within the replicates can be expected to differ and thus differ from the overall row-to-row variation. Of course the column-to-column variation becomes that variation within squares when the Latin squares are used.

Sometimes carryover effects can result from factors not being reset. (Recall the discussion of factors not being reset in Section 4.20.) Giesbrecht and Gumpertz (2004, p. 257) gave an example of an actual experiment that involved fuel additives and the engineers could never be sure that all of the fuel additive from one test had been removed from the metal surfaces in the carburetor by the time the next test was performed. Consequently, a design that was balanced for carryover effects was used.

A safe strategy would be to proceed as if carryover effects are present and to analyze the data accordingly. Weerahandi (2004, p. 268) stated that at least the treatment effects, carryover effects, and period effects should be modeled when a crossover design is used. Period effects may not be estimable, however, as is illustrated in the following clinical trials example.

Example 11.4

Weerahandi (2004) gave an example of a clinical trial experiment that was originally given by Senn (2002). The objective of the experiment was to compare two treatments, a single inhaled dose of 200 μg Salbutamol (S) and a single inhaled dose of 12 μg Formoterol (F). The response variable was the peak expiratory flow, a measure of lung function There were 13 children involved in the experiment, each of whom had moderate or severe asthma. There were two periods involved, with a washout period of one day between those two periods. Seven children received the sequence FS over the two periods, and the other six children received the sequence SF.

Our objective, of course, is to determine if there is a difference in the two means, μ_{F} and μ_{S}, and if so, to estimate that difference. The first question we might ask is "Is this a good design for accomplishing that objective?" Since there are two groups and two periods, we can use a 2×2 table to represent the parameters that are involved in each cell. If we let $A = F$ and $B = S$, so as to use the conventional A and B notation, we then have the following tabular representation with the rows representing the groups and the columns representing the periods.

$\mu_{\mathrm{A}} + \gamma$	$\mu_{\mathrm{B}} + \gamma + \pi + \lambda_{\mathrm{A}}$
$\mu_{\mathrm{B}} - \gamma$	$\mu_{\mathrm{A}} - \gamma + \pi + \lambda_{\mathrm{B}}$

with μ_{A} and μ_{B} denoting the two means of interest, γ the sequence effect (if it exists), λ_{A} and λ_{B} the carryover effects for each treatment from the first to the second period, and π the period effect.

We see immediately that we cannot form a linear combination of the four cells to estimate μ_{A} and μ_{B} individually, and also we cannot estimate their difference

unless we assume $\lambda_A = \lambda_B = \lambda$. With this assumption, we can obviously estimate $\delta = \mu_A - \mu_B$ as $\widehat{\delta} = \frac{1}{2}(\overline{Y}_{11} + \overline{Y}_{22}) - \frac{1}{2}(\overline{Y}_{12} + \overline{Y}_{21})$, with $\overline{Y}_{ij}$ denoting the average observation in the ith row and jth column of the table.

Which other parameters are estimable? In order to estimate γ, we need to use a linear combination for which the coefficients will be $+1$ for the first row and -1 for the second row so that the other parameters will drop out, still assuming of course that $\lambda_A = \lambda_B = \lambda$. Doing so will produce a coefficient of 4 for γ. Thus, γ is estimated as $\widehat{\gamma} = \frac{1}{4}(\overline{Y}_{11} + \overline{Y}_{12} - \overline{Y}_{21} - \overline{Y}_{22})$.

Are any other parameters estimable? We observe that we cannot estimate either π or λ individually, although their sum could be estimated as $\frac{1}{2}(\overline{Y}_{12} + \overline{Y}_{22} - \overline{Y}_{11} - \overline{Y}_{21})$. Thus, the two effects are confounded.

Since we are interested in comparing the two means, we would expect that we would be able to construct a t-test. This would not be a routine t-test, however, as it couldn't be an independent sample t-test because the data aren't independent. It also could not be a paired t-test because there are two sequences, not one. There is also not a single variance since there are two periods and two groups of patients.

There is, however, an analogy with a paired t-test because with that test the variance of the difference $Y_1 - Y_2$ is equivalent to $\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$, with σ_{12} denoting the covariance between Y_1 and Y_2 and the first two terms denoting the variance of each. Here the covariance is between the periods and since there are two groups, the variances and covariance must be assumed to be the same for the two groups. Otherwise, a standard analysis approach could not be performed. With this assumption, the analysis is performed by merging the data from the two groups.

Let $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$. With n_1 observations in the first group and n_2 observations in the second group, σ^2 is thus estimated by pooling the data from the two groups, completely analogous to what is done with a pooled t-test. Since $\text{Var}(\overline{Y}_{22} - \overline{Y}_{21}) = \sigma^2/n_2$ and $\text{Var}(\overline{Y}_{11} - \overline{Y}_{12}) = \sigma^2/n_1$ and the two differences are independent, it follows that $\text{Var}(\widehat{\delta}) = \text{Var}\{1/2((\overline{Y}_{22} - \overline{Y}_{21}) - (\overline{Y}_{11} - \overline{Y}_{12}))\} = \frac{\sigma^2}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Therefore, with δ_0 denoting the hypothesized value of δ, the t-statistic is

$$t = \frac{\widehat{\delta} - \delta_0}{\frac{s}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has $n_1 + n_2 - 2$ degrees of freedom.

Remember that the analysis is based on the assumption that $\widehat{\delta}$ is approximately normally distributed and $\lambda_A = \lambda_B = \lambda$. The second assumption cannot be tested with the design that is used. Note in particular that $\lambda_A - \lambda_B = 4\gamma$ (from the 2×2 table), so the effects cannot be separated. In general, analyses should not be based on assumptions that cannot be tested. In this instance, if λ_A and λ_B differ considerably, the t-test result will be impacted.

Could the design have been improved? In particular, could the problem of having to assume $\lambda_A = \lambda_B$ and not being able to test the assumption have been avoided? This can be accomplished by using four groups instead of two and using *AB*, *BA*, *AA*, and *BB* as the four sequences in the four groups, respectively. This design is due to

Balaam (1968). As shown by Weerahandi (2004, p. 278), expressions for $\widehat{\lambda}_{\mathrm{A}}$ and $\widehat{\lambda}_{\mathrm{B}}$ can be easily obtained. These estimators are functions of both $\widehat{\mu}_{\mathrm{A}}$ and $\widehat{\mu}_{\mathrm{B}}$, however. A simpler expression is obtainable if we look at $\widehat{\lambda}_{\mathrm{A}} - \widehat{\lambda}_{\mathrm{B}}$, which can be written as $\frac{1}{2}(\overline{Y}_{12} + \overline{Y}_{32} - \overline{Y}_{22} - \overline{Y}_{42})$, with $\overline{Y}_{ij}$ as defined previously except that this time there are four rows, corresponding to the four groups. Although $\widehat{\mu}_{\mathrm{A}}$ and $\widehat{\mu}_{\mathrm{B}}$ are unbiased, different weights are assigned to the cell means in the 4×2 table in the process of producing the unbiasedness.

It is possible in certain types of experiments, such as crop experiments, for carryover effects to extend over more than one or two periods. Patterson (1968) introduced serial designs for such applications.

Motivated by the use of crossover designs in the telecommunications industry, Lewis and Russell (1998) presented crossover designs when there are carryover effects from two factors. They gave methods for constructing crossover designs that are based on the Latin squares of Williams (1949) and of Russell (1991). The reader is also referred to Senn (2002) for additional reading on crossover designs and extensive discussion of carryover effects, to Raghavarao (1990) for the use of crossover designs in industry, and to Chapter 19 in Hinkelmann and Kempthorne (2005). Other recommended sources on crossover designs include Patterson (1951, 1973), Afsarinejad (1983), Fletcher (1987), Russell (1991), Ratkowsky, Evans, and Alldredge (1993), Wakeling and MacFie (1995), Vonesh and Chinchilli (1997), Goad and Johnson (2000), and Weerahandi (2004). In particular, Jones and Kenward (2003) gave a historical perspective as they traced the early history of crossover trials through crop-rotation experiments, feeding trials, and bioassay, in addition to giving a comprehensive catalog of designs and downloadable SAS programs. Dallal (2001) is also recommended for computer analysis of crossover designs and for other information. In particular, he points out that care must be exercised in reading the literature on crossover designs, noting for example that an error in Grizzle (1965) that was later corrected in Grizzle (1974) is apparently still misleading people, as noted by Grieve (1982) and as is evident from some later publications. Possible time effects have also been considered by various authors, and one proposed solution is a switchback design, which requires an extra period so as to allow for a switchback to the treatment used in the first time period. Accordingly, a simple switchback design would be of the form $A \to B \to A$ and $B \to A \to B$ (see, e.g., Oman and Seiden, 1988).

11.5 HOW MANY REPEATED MEASURES?

An obvious question to address is, "How many repeated measurements are to be made on each experimental unit?" This issue has been considered by various authors, including Vickers (2003), who addressed the issue for medical applications, such as patients being evaluated every three months after thoracic surgery. Of course power is increased as the number of repeated measurements is increased, but there is a point of diminishing returns.

11.6 FURTHER READING

The optimality and efficiency of crossover designs has been studied over a period of four decades, with the history traced by Hedayat and Yang (2004), who refer readers to Stufken (1996) for additional references. See also, Hedayat and Yang (2003, 2005) and Hedayat, Stufken, and Yang (2005) for optimal crossover designs. Practitioners who are not particularly interested in the optimality results of crossover designs may wish to read Stufken (1996), which is an expository article. Although the analysis of data from a crossover design is usually numerical, Pontius and Schantz (1994) applied the sliding-square plot idea of Rosenbaum (1989) for the graphical analysis of data from a two-period crossover design.

Missing data in small repeated measures designs can be problematic and can cause a considerable loss of power for detecting significant effects. Huang and Carriére (2006) developed a strategy for multiple imputations and compared analysis using multiple imputations with non-imputation incomplete data methods, such as that given by Carriére (1999). Unfortunately, they found that the non-imputation methods were superior to the use of multiple imputations.

Other papers that may be of interest include Hedayat and Afsarinejad (1975, 1978), Marshall and Jackson (1993), Senn and Lambrou (1998), Dean, Lewis, and Chang (1999), and Prescott (1999).

11.7 SOFTWARE

One problem that faces the user of crossover designs and other repeated measurement designs is that software that has this capability seems, at first glance, not to be as plentiful as software for more common designs, such as factorial designs. For example, Design-Expert 7.0 does not list this capability, nor does Release 14 of MINITAB. Both software packages can be used to analyze repeated measures data, however. (For MINITAB, see http://www.minitab.com/support/answers/answer.aspx?ID=413 for an explanation.)

The help file material in JMP indicates that the latter can be used to analyze repeated measures data, using either a univariate, mixed model approach or a multivariate approach.

The most commonly used statistical procedures for repeated measures data undoubtedly include SAS PROC GLM and SPSS GLM. SAS PROC MIXED is necessary for modeling the covariance structure, however, rather than simply assuming a certain structure. Clearly, the former will generally be the preferred approach. JMP can also be used to analyze repeated measures data. The construction of balanced carryover designs, including those of Williams, Patterson, and Afsarinejad, using R, the freeware version of S-Plus, is available using the CROSSDES package (see http://www.ibiblio.org/pub/languages/R/CRAN/doc/packages/crossdes.pdf). Cross Over 1.0, discussed in Jones and Kenward (2003), is another software package for crossover designs. Useful information on the analysis of crossover design data with various statistical software packages is given by Dallal (2001), who points out that such data

can be analyzed by any software that has the capability for repeated measures analysis of variance.

It should be noted that these software procedures are for the analysis of repeated measurement data; the user of these designs must first select an appropriate design.

11.8 SUMMARY

Repeated measurement designs are used in various applications, especially experiments involving people. This is one topic that is generally discussed more extensively in research articles than in books, with Davis (2002) and Winer et al. (1991) being notable exceptions. With some experimental designs, such as screening designs, it isn't necessary to formulate a tentative model. With other designs, only a good idea of the types of effects that are likely to be significant is needed. Repeated measurement designs present some unique problems, however, as the possibility of carryover effects and possible or probable lack of independence should affect decisions regarding the design and the subsequent analysis.

An extensive, although somewhat outdated, review of repeated measurement designs was given by Afsarinejad (1990).

REFERENCES

Afsarinejad, K. (1983). Balanced repeated measurements designs. *Biometrika*, **70**, 199–204.

Afsarinejad, K. (1990). Repeated measurements designs — a review. *Communications in Statistics — Theory and Methods*, **19**(11), 3985–4028.

Bailey, R. A. and J. Kunert (2006). On optimal cross-over designs when carry-over effects are proportional to direct effects. *Biometrika*, **93**(3). (available at http://www.designtheory.org/library/preprints/rabjkpre.pdf)

Balaam, L. N. (1968). A two-period design with t^2 experimental units. *Biometrics*, **24**, 61–73.

Bate, S. T. and B. Jones (2006). The construction of nearly balanced and nearly strongly balanced uniform cross-over designs. *Journal of Statistical Planning and Inference*, **136**(9), 3248–3267.

Boon, P. C. and K. C. B. Roes (1999). Design and analysis issues for crossover designs in phase I clinical trials. *Journal of Biopharmaceutical Statistics*, **9**, 109–128.

Bose, M. (2002). Crossover designs: Analysis and optimality using the calculus for factorial arrangements. In *Design Workshop Lecture Notes*, International Statistical Institute, Kolkata, India, November 25–29, pp. 183–192. (available at http://www.isid.ac.in/~ashish/workshop/mausumiw3.pdf.)

Bose, M. and A. Dey (2003). Some small and efficient cross-over designs under an additive model. *Utilitas Mathematica*, **63**, 173–182.

Bose, M. and B. Mukherjee (2000). Cross-over designs in the presence of higher order carryovers. *Australian and New Zealand Journal of Statistics*, **42**, 235–244.

Bose, M. and B. Mukherjee (2003). Optimal crossover designs under a general model. *Statistics and Probability Letters*, **62**(4), 413–418.

Carriére, K. C. (1994). Crossover designs for clinical trials. *Statistics in Medicine*, **13**, 1063–1069.

Carriére, K. C. (1999). Methods of repeated measures data analysis with missing values. *Journal of Statistical Planning and Inference*, **77**, 221–236.

Carriére, K. C. and G. C. Reinsel (1993). Optimal two-period repeated measures designs with two or more treatments. *Biometrika*, **80**, 924–929.

Cheng, C.-S. and C.-F. Wu (1980). Balanced repeated measures designs. *Annals of Statistics*, **8**, 1272–1283.

Claydon, N. C., M. Addy, R. Newcombe, and J. Moran (2005). The prevention of plaque re-growth by toothpastes and solutions containing block copolymers with and without polypeptide. *Journal of Clinical Periodontology*, **32**(6), 545–548.

Cochran, W. G. and G. M. Cox (1957). *Experimental Designs*, 2nd ed. New York: Wiley.

Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society, Series B*, **63**, 55–62.

Crowder, M. J. and D. J. Hand (1990). *Analysis of Repeated Measures*. London: Chapman and Hall.

Dallal, G. E. (2001). The computer-aided analysis of crossover studies. (available at http://www.tufts.edu/∼gdallal/crossovr.htm)

Davidian, M. and D. M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. Boca Raton, FL: CRC Press.

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Dean, A. M., S. M. Lewis, and J. Y. Chang (1999). Nested changeover designs. *Journal of Statistical Planning and Inference*, **77**, 337–351.

Donev, A. N. (1998). Crossover designs with correlated observations. *Journal of Biopharmaceutical Statistics*, **8**, 249–262.

Everitt, B. S. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, **44**, 113–135.

Federer, W. T. and R. P. Kerschner (1998). Comparison of classes of changeover designs. *Journal of Combinatorics, Information, and System Sciences*, **23**, 379–391.

Finney, D. J. H. (1965). Crossover designs in bioassay. *Proceedings of the Royal Society*, **145B**, 42–61.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.

Fletcher, D. J. (1987). A new class of change-over designs for factorial experiments. *Biometrika*, **74**, 649–654.

Garcia, R., M. Benet, C. Arnau, and E. Cobo (2004). Efficiency of the cross-over design: An empirical investigation. *Statistics in Medicine*, **23**(24), 3773–3780.

Giesbrecht, F. G. and M. L. Gumpertz (2004). *Planning, Construction, and Statistical Analysis of Comparative Experiments*. Hoboken, NJ: Wiley.

Girden, E. R. (1992). *ANOVA: Repeated Measures*. London, UK: SAGE Publications.

Goad, C. L. and D. E. Johnson (2000). Crossover tests: A comparison of ANOVA tests and alternative analyses. *Journal of Agricultural, Biological, and Environmental Statistics*, **5**, 69–87.

Grieve, A. P. (1982). Correspondence: The two-period changeover design in clinical trials. *Biometrics*, **38**, 517.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, **21**, 467–480.

Grizzle, J. E. (1974). Correction. *Biometrics*, **30**, 727.

Hedayat, A. S. and K. Afsarinejad (1975). Repeated measurements designs, I. In *A Survey of Statistical Design and Linear Models*, pp. 229–242. (J. N. Srivastava, ed.). Amsterdam: North-Holland.

Hedayat, A. S. and K. Afsarinejad (1978). Repeated measurements designs, II, *The Annals of Statistics*, **6**, 619–628.

Hedayat, A. S. and K. Afsarinejad (2002). Repeated measurement designs for a model with self and mixed carryover effects. *Journal of Statistical Planning and Inference*, **106**, 449–459.

Hedayat, A. S. and J. S. Stufken (2003). Optimal and efficient crossover designs under different assumptions about the carryover effects. *Journal of Biopharmaceutical Statistics*, **13**, 519–528.

Hedayat, A. S. and M. Yang (2003). Universal optimality of balanced crossover designs. *The Annals of Statistics*, **31**, 978–983. (see http://statistics.unl.edu/faculty/yang/crossover.pdf)

Hedayat, A. S. and M. Yang (2004). Universal optimality for selected crossover designs. *Journal of the American Statistical Association*, **99**, 461–466.

Hedayat, A. S. and M. Yang (2006). Optimal and efficient crossover designs for comparing test treatments with a control treatment. *The Annals of Statistics*, **33**, 915–943.

Hedayat, A. S., J. Stufken, and M. Yang (2006). Optimal and efficient crossover designs when subject effects are random. *Journal of the American Statistical Association*, **101**, 1031–1038.

Hinkelmann, K. and O. Kempthorne (2005). *Design and Analysis of Experiments, Volume 2: Advanced Experimental Design*. Hoboken, NJ: Wiley.

Huang, R. and Carriére, K. C. (2006). Comparison of methods for incomplete repeated measures data analysis in small samples. *Journal of Statistical Planning and Inference*, **136**(1), 235–247.

John, J. A., K. G. Russell, and D. Whitaker (2004). CrossOver: An algorithm for the construction of efficient crossover designs. *Statistics in Medicine*, **23**(17), 2645–2658.

Jones, B. and M. G. Kenward (2003). *Design and Analysis of Cross-over Trials*, 2nd ed. New York: Chapman and Hall.

Kempton, R. A., S. J. Ferris, and O. David (2001). Optimal change-over designs when carry-over effects are proportional to direct effects of treatments. *Biometrika*, **88**, 391–399.

Keppel, G. (1973). *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice-Hall.

Keselman, H. J., J. Algina, and R. K. Kowalchuk (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, **54**, 1–20.

Kowalchuk, R. K., H. J. Keselman, J. Algina, and R. D. Wolfinger (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement*, **64**(2), 224–242.

Kuehl, R. O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*, 2nd ed. New York: Duxbury.

Kunert, J. (1987). An example of universal optimality in a full-rank model. *Metrika*, **34**, 217–223.

Kunert, J. (1998). Sensory experiments as crossover studies. *Food Quality and Preference*, **9**, 243–253.

Kunert, J. and J. Stufken (2002). Optimal crossover designs in a model with self and mixed carryover effects. *Journal of the American Statistical Association*, **97**, 898–906.

Kunert, J. and J. Stufken (2005). Optimal crossover designs for two treatments in the presence of mixed and self carryover effects. Technical Report No. 2005-11, Department of Statistics, University of Georgia.

Kunert, J. and O. Sailer (2006). On nearly balanced trials for sensory trials. *Food Quality and Preference*, **17**(3–4), 219–227.

Kushner, H. B. (1998). Optimal and efficient repeated-measurements designs for uncorrelated observations. *Journal of the American Statistical Association*, **93**, 1176–1187.

Lawson, J. S. (1988). A case study of effective use of statistical experimental design in a smoke stack industry. *Journal of Quality Technology*, **20**(1), 51–62.

Lewis, S. M. and K. G. Russell (1998). Crossover designs in the presence of carryover effects from two factors. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **47**, 379–391.

Lindsey, J. K. (1999). *Models for Repeated Measurements*, Vol. 19. Oxford, UK: Oxford University Press.

Loughin, T. M., D. E. Johnson, S. E. Ives, and T. G. Nagaraja (2002). Methods for selecting crossover designs with applications to an experiment with two factors in a split plot. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 143–156.

Louis, T. A., P. W. Lavori, J. C. Bailar, III, and M. Polansky (1984). Crossover and self-controlled designs in clinical research. *New England Journal of Medicine*, **310**(1), 24–31.

Low, J. L., S. M. Lewis, and P. Prescott (1999). Assessing robustness of crossover designs to subjects dropping out. *Statistics and Computing*, **9**, 219–227.

Low, J. L., S. M. Lewis, and P. Prescott (2002). An application of Pólya Theory to crossover designs with dropout. *Utilitas Mathematica*, **63**, 129–142.

Marshall, R. J. and R. T. Jackson (1993). Analysis of case-crossover design. *Statistics in Medicine*, **12**, 2333–2341.

Matthews, J. N. S. (1987). Optimal crossover designs for the comparison of two treatments in the presence of carryover effects and autocorrelated errors. *Biometrika*, **74**, 311–320.

Matthews, J. N. S. (1994). Modeling and optimality in the design of crossover studies for medical applications. *Journal of Statistical Planning and Inference*, **42**, 89–108.

Max, M. B. (2003). The design of clinical trials for treatment of pain. In *Interactive Textbook on Clinical Symptom Research*, Chap. 1 (M. B. Max and J. Lynn, eds.) (Available at http://symptomresearch.nih.gov/tablecontents.htm.)

Mielke, P. W., Jr. (1974). Square rank test appropriate to weather modification cross-over design. *Technometrics*, **16**, 13–16.

Milliken, G. A. (2004). Mixed models and repeated measures: Some illustrative industrial examples. In *Handbook of Statistics, Vol. 22: Statistics in Industry*, Chap. 5 (R. Khattree and C. R. Rao, eds.). Amsterdam, The Netherlands: Elsevier Science B.V.

Montgomery, D. C. (1996). *Design and Analysis of Experiments*, 4th ed. New York: Wiley.

Newcombe, R. G. (1995). Residual effect of chlorhexidine gluconate in 4-day plaque regrowth crossover trials, and its implications for study design. *Journal of Periodontal Research*, **30**, 319–324.

Newcombe, R. G. (1996). Sequentially balanced three-squares cross-over designs. *Statistics in Medicine*, **15**(20), 2143–2147.

Oman, S. D. and E. Seiden (1988). Switch-back designs. *Biometrika*, **75**(1), 81–89.

Patterson, H. D. (1951). Change-over trials. *Journal of the Royal Statistical Society, Series B*, **13**, 256–271.

Patterson, H. D. (1952). The construction of balanced designs for experiments involving sequences of treatments. *Biometrika*, **39**, 32–48.

Patterson, H. D. (1968). Serial factorial design. *Biometrika*, **55**(1), 67–81.

Patterson, H. D. (1973). Quenouille's changeover designs. *Biometrika*, **60**, 33–45.

Périnel, E. and J. Pagès (2004). Optimal nested crossover designs in sensory analysis. *Food Quality and Preference*, **15**, 439–446.

Pignatiello, J. (1984). SPC questions, queries, and quandaries: Two-stage nested designs. *ASQ Statistics Division Newsletter*, **6**(1).

Pontius, J. S. and R. M. Schantz (1994). Graphical analyses of a two-period crossover design. *The American Statistician*, **48**, 249–253.

Prescott, P. (1999). Construction of uniform-balanced cross-over designs for any odd number of treatments. *Statistics in Medicine*, **18**, 265–272.

Putt, M. and V. N. Chinchilli (1999). A mixed effects model for the analysis of repeated measures cross-over studies. *Statistics in Medicine*, **18**, 3037–3058.

Raghavarao, D. (1990). Crossover designs in industry. In *Statistics, Textbooks and Monographs*, Vol. 109. New York: Marcel Dekker, Chap. 18, pp. 517–530.

Ratkowsky, D. A., M. A. Evans, and J. R. Alldredge (1993). *Cross-over Experiments: Design, Analysis and Applications*. New York: Marcel Dekker.

Rosenbaum, P. R. (1989). Exploratory plots for paired data. *The American Statistician*, **43**, 108–109.

Russell, K. G. (1991). The construction of good change-over designs when there are fewer subjects than treatments. *Biometrika*, **78**, 305–313.

Russell, K. G. and A. M. Dean (1998). Factorial cross-over designs with few subjects. *Journal of Combinatorics, Information and System Sciences*, **23**, 209–235.

Sen, M. and R. Mukerjee (1987). Optimal repeated measures designs under interaction. *Journal of Statistical Planning and Inference*, **17**, 81–91.

Senn, S. (1996). The AB/BA crossover: How to perform the two-stage analysis if you can't be persuaded that you shouldn't. In *Liber Amicorum Roel van Strik* (B. Hansen and M. De Ridder, eds.) Rotterdam: Erasmus University, pp. 93–100. (available online at http://www.senns.demon.co.uk/ROEL.pdf)

Senn, S. (2002). *Cross-over Trials in Clinical Research*, 2nd ed. New York: Wiley.

Senn, S. (2003). Within-patient studies: Cross-over trials and *n*-of-1 studies. In *Interactive Textbook on Clinical Symptom Research*, Chap. 6 (M. B. Max and J. Lynn, eds.). (Available at http://symptomresearch.nih.gov/tablecontents.htm.)

Senn, S and D. Lambrou (1998). Robust and realistic approaches to carry-over. *Statistics in Medicine*, **17**, 2849–2864.

Senn, S., L. Stevens, and N. Chaturvedi (2000). Repeated measures in clinical trials: Simple strategies for analysis using summary measures. *Statistics in Medicine*, **19**, 861–877.

Street, D. J., J. A. Eccleston, W. H. Wilson (1990). Tables of small optimal repeated measurements designs. *Australian Journal of Statistics*, **32**, 345–359.

Stufken, J. (1996). Optimal crossover designs. In *Handbook of Statistics*, Vol. 13: *Design and Analysis of Experiments*, pp. 63–90 (S. Ghosh and C. R. Rao, eds.). Amsterdam: North Holland.

Thaler, J. S. (1999). Induced resistance in agricultural crops: Effects of jasmonic acid on herbivory and yield in tomato plants. *Environmental Entomology*, **28**(1), 30–31.

Tudor, G. E., G. G. Koch, and D. Catellier (2000). Statistical methods for crossover designs in bioenvironmental and public health studies. In *Handbook of Statistics*, Vol. 18: *Bioenvironmental and Public Health Statistics*, pp. 571–614 (P. K. Sen and C. R. Rao, eds.) Amsterdam: North Holland.

Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology*, **3**, 1–22.

Vonesh, E. F. and V. M. Chinchilli (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wakeling, I. N. and H. J. H. MacFie (1995). Designing consumer trials balanced for first and higher orders of carryover effect when only a subset of k samples from t may be tested. *Food Quality and Preference*, **6**, 299–308.

Weerahandi, S. (2004). *Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models*. Hoboken, NJ: Wiley.

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, **2**, 149–168.

Winer, B. J., D. R. Brown, and K. M. Michels (1991). *Statistical Principles in Experimental Design*, 3rd ed. New York: McGraw-Hill.

EXERCISES

11.1 There were two 4×4 Latin square designs given in Section 11.4 that would each be suitable for use when carryover effects from use of a crossover design are expected. Is the following Latin square design also suitable for that purpose? Why, or why not?

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

11.2 (Harder problem) Show algebraically that the analysis of repeated measures data with two levels of one factor has the same corresponding relationship as one-factor ANOVA and an independent sample t-test for independent data. That is, the square of the t-statistic in a paired t-test is equal to the F-statistic for the treatments effect, as was illustrated in Section 10.1.

11.3 VOICE.MTW is one of the sample data files that comes with the MINITAB software. A professor in a theater department wanted to determine if a voice training class improved a performer's voice quality. There were 10 students who took the class and six judges rated each student on a scale of 1–6, with 6 being the best score. The judges scores were as follows:

Student #									
1		2		3		4		5	
Before	After	Before	After	Before	After	Before	After	Before	After
5	3	2	2	5	3	4	3	5	4
4	5	3	3	5	4	5	4	6	5
4	5	3	1	6	3	5	5	6	6
5	6	3	2	4	5	4	3	6	5
4	5	2	2	2	3	3	3	4	3
5	5	3	3	5	5	4	4	5	5

6		7		8		9		10	
Before	After	Before	After	Before	After	Before	After	Before	After
3	4	2	2	2	2	5	2	3	3
3	3	2	2	4	4	2	3	3	3
4	3	1	1	5	2	2	2	3	2
4	3	4	3	5	4	5	3	4	4
3	3	2	2	4	3	3	2	3	4
4	4	2	3	5	4	4	4	4	4

Analyze the data and determine whether or not the voice training appears to have been worthwhile.

11.4 Repeated measures can occur with virtually any type of design, so the term "repeated measures design" should perhaps be viewed as a misnomer. Thaler (1999, references) used repeated measures in agricultural experimentation. This paper is available online at www.botany.utoronto.ca/ResearchLabs/thalerLab/Environmental%20Entomology.pdf. Read the article and determine on what the repeated measures occur. You will notice mention of other experimentation without repeated measures. Comment on the author's experimental design approach. In particular, would you suggest that anything different be done in future experimentation of this type?

11.5 (Harder problem) Pignatiello (1984, see references) described the following scenario that was presented to him by an engineer working for a military contractor.

Five parts were randomly sampled from each of two lots and the value of a quality characteristic was recorded. This process was repeated so that each of the 10 parts was measured twice, the rationale for doing so being that individual readings vary due to measurement error. Assume that the two lots are the only lots that are of interest. The data are given below:

	Lot 1					Lot 2				
Parts	1	2	3	4	5	1	2	3	4	5
Measurements	10.2	8.4	9.2	10.5	11.3	10.9	14.1	11.8	12.8	13.3
	9.1	10.4	9.1	10.1	10.4	12.6	11.5	11.4	10.7	12.5

These data were analyzed as having come from a nested design with the error sum of squares, with 10 df, computed as one would compute an error sum of squares in the presence of replication. Would you have treated the error term differently? Specifically, is there a repeated measures aspect to the data, and if so, does the repeated measures aspect dictate that the data be analyzed differently from the analysis of replicated design points without repeated measurements? Explain. Analyze the data and draw appropriate conclusions.

11.6 Assume that an experiment is performed to determine if there are differences in two types of athletic shoes, A and B, relative to running speeds, with the distance being 50 yards. Twelve individuals are recruited for the study and the times are as follows:

Subject	Times	
1	B (6.3)	A (6.1)
2	B (5.9)	A (6.0)
3	A (6.0)	B (5.8)
4	B (6.5)	A (6.5)
5	A (5.7)	B (5.6)
6	A (5.9)	B (6.0)
7	A (6.1)	B (6.2)
8	B (6.0)	B (5.9)
9	A (6.3)	B (6.1)
10	B (6.2)	A (6.1)
11	A (6.4)	B (6.2)
12	B (6.1)	A (6.0)

(a) Analyze the data under the assumption that there are no carryover effects.
(b) Considering the nature of the experiment, do you believe that carryover effects should exist? Why, or why not? If you believe that they should exist, what would you recommend?

11.7 Consider Exercise 3.31 in Chapter 3. What complications do repeated measures present relative to the study in which a 3×3 Latin square design was used (which were apparently not addressed in that study)?