The analysis of cross-over trials in animal behavior experiments: review and guide to the statistical literature

Ramón Díaz-Uriarte
(formerly at:
Departments of Zoology and Statistics
University of Wisconsin-Madison)
email: ligarto@navegalia.com

24 November, 2001

Abstract

Cross-over trials are frequently used in animal behavior experiments but are often analyzed incorrectly. In this report I review methods of analysis of cross-over trials in the context of animal behavior experiments. I group methods of analysis according to the type of response variable: non-parametric and robust methods for metric responses, parametric methods for metric responses —linear mixed-effects models—, models for categorical responses both non-parametric and parametric —extensions of generalized linear models—, censored observations —survival analysis–, and multivariate responses. Within-individual contrasts are explained in detail early on, as they are the basis of many different methods, from non-parametric to multivariate and survival-based models, and they offer a useful framework for extending the analysis of data from cross-over trials to situations where robust methods might be needed (e.g., permutation tests of censored multivariate responses). In this paper I also discuss some types of plots that are specific and particularly useful for cross-over trials. Before conducting a study, it is of paramount importance to consider both the design and analysis, since the type of response can affect the choice of design. Moreover, some types of responses can be very difficult to analyze, specially with small sample sizes, and can result in very low statistical power (in particular categorical and survival data), and might prompt us to redesign the experiment or consider measuring other responses.

Preface

Cross-over trials are frequently used in animal behavior (see Díaz-Uriarte 2002) as they allow us to conduct experiments with relatively small numbers of subjects that nonetheless achieve high statistical power by using each subject as its own control (Jones & Kenward 1989; Senn 1993a). Thus, cross-over designs are powerful tools when repeated testing of one subject is much simpler than recruitment of new subjects. However, cross-over experiments in animal behavior studies are usually analyzed incorrectly, as if they were matched pairs or "typical" repeated-measures designs, which they are not (see Díaz-Uriarte 2002, for details and examples). The main problems are failure to account for period and carry-over effects . The widespread use of inappropriate analysis could be the result of a lack of information about cross-over trials in statistical texts commonly used by behaviorists. The problem is compounded because in many behavioral experiments researchers often record data (such as categorical data or censored time to event data) that might not allow the use of standard parametric analysis, and frequently measure several response variables that ought to be analyzed with multivariate techniques.

The main objective of this book (or report, or tiny book) is to review the analysis of cross-over designs, with special emphasis on animal behavior experiments. This book should be of immediate and practical use for behaviorists and statistical consultants working with behaviorists, but also to users of cross-over trials in other disciplines. I review and show the connections among different methods that have recently appeared in the statistical literature and are particularly relevant to behavior researchers (e.g., multivariate responses and time to event data), but that are not covered in available texts (Jones & Kenward 1989; Senn 1993a; Ratkowsky et al 1993). On other topics (e.g., linear mixed-effects models) I provide practical discussion in the context of cross-over trials. Nonparametric and categorical data methods are considered in recent reviews of cross-over trials; I have included some new papers and eased the use of these methods by cross-referencing statistics textbooks and software packages. Small sample sizes, blocking, and among-subject treatments are all relevant to animal behavior experiments and are considered throughout the paper. I concentrate on methods that are available in major statistical packages (specially R, S-Plus and SAS; note that R is free GNU software that can be obtained from CRAN at http://cran.r-project.org and mirror sites; unless specified otherwise, S-Plus libraries are available from Statlib at http://lib.stat.cmu.edu and R packages from CRAN), or that can be implemented with a minimum amount of code writing. Finally, I emphasize randomization and permutation tests (e.g., Edgington, 1995;

Good 1994; Noreen 1989). Randomization tests , increasingly used in behavior and ecological research (e.g., Manly 1997; Crowley 1992), are a general alternative when parametric assumptions are not met, can be more powerful and flexible than traditional ”non-parametric” methods, and might be the most appropriate tests for many experimental settings (Ludbrook & Dudley 1998).

I review the analysis of data from cross-over designs according to the type of response variable (e.g., Agresti 1990, ch. 1). A metric or interval variable is one that has numerical distances between any two levels of the scale (e.g., length); arithmetic operations on the response are meaningful. One special type of metric variable is time to an event (examined later). Ordinal variables are categorical variables that have ordered levels (e.g., bad, fair, good), but differences, sums, and other algebraic operations on the ranks or levels are not meaningful. Nominal categorical variables have levels without natural orderings (e.g., Buddhist, Christian, Hindu). A particular type of categorical responses are binary outcomes (such as success/failure).

The first chapters are the central part of the book, and cover the analysis of data from cross-over experiments. Next I discuss plotting and graphical summaries in cross-over experiments. Then I mention sample size and missing data. I conclude with some recommendations on the use and analysis of cross-over experiments in animal behavior experiments. Since I will make frequent use of certain terms, and I will need to refer to some types of designs commonly used in cross-over trials, I have included a brief review of some terms and the analysis of cross-over trials in the appendices. The very serious problems with carry-over effects, how to deal with them and how to avoid them are not covered here; I give a brief discussion, with pointers to the literature, in Díaz-Uriarte (2002); the interested reader should consult Senn (1993 a, pp. 14-15, 52-54 and ch. 10) and the discussions in Grieve & Senn (1998), Jones & Wang (1998) and Koch (1998).

Future plans

I prepared the notes that eventually formed the skeleton of this book when I was faced with having to analyze several data sets from some behavioral experiments with lizards I had conducted as part of my dissertation (if you are curious, some of the research has been published as Díaz-Uriarte 1991, 2001). Sure enough, my data did not obey comfortable standards (e.g., I had multivariate censored responses from relatively small samples). That forced me to go the primary literature to figure out how to analyze these, and other, types of data. Eventually, I had taken notes and written several pages that could be of use to someone else (or myself in the future), and I decided to try and complete the task, and this is the result.

I think that the book as it stands now is potentially useful to researchers and applied statisticians, as I explain above. However, as some people who read previous versions have commented, the writing is sometimes dense and the book would benefit from commented examples. In the future, I plan to add examples to most of the sections of the book, and illustrate how to carry out the analyses using code from several stats packages. I have my own favorites, and the first one is R which, incidentally, can be used for virtually all of the analyses mentioned here. Another program that will deal with most of the analyses discussed is SAS. I plan to include commented input and output from the above two packages, and maybe also from

SPSS and MacAnova (http://www.stat.umn.edu/macanova/macanova.home.html), which can be used to carry out some, but not all, of the analyses discussed in this book. This version, therefore, is still a draft, since examples and code are missing; however, given that this book is now weekend work, completing it might take longer than I wish, and I thought it might be convenient to make this version available while I work on the longer one.

Acknowledgments

C. Lázaro-Perea provided advice, discussion, and comments on the book. A. R. Ives, B. Jones, J. K. Lindsey, C. A. Marler, E. V. Nordheim, and C. T. Snowdon provided comments of previous versions. E. V. Nordheim suggested that this was worth being written and has provided statistical discussion. I discussed with A. T. Galecki about multivariate mixed-effects models. The book by B. Jones and M. G. Kenward opened my eyes to cross-over trials, and it remains one of the best written and delightful stats book I have read. The initial preparation of the manuscript was partially supported by a John and Virginia Emlen Graduate Student Fellowship, Dept. Zoology, University of Wisconsin-Madison. Computerwise, most of my data analyses have been done with the open-source (and free) statistical package R (http://cran.r-project.org) and I want to thank the core development team and all the package writers for this amazing tool. Finally, I want to thank Samizdat Press and John Scales for their commitment to this type of publication and for his kind email reply and for providing a home for this book.

Contents

Chapter 1

Metric responses

1.1 Nonparametric and robust methods

1.1.1 Within-individual contrasts

With two treatments and dual sequences (see B) , a common way to carry out robust analyses (Jones & Kenward 1989, p. 60-65; 160; Hafner et al. 1988) is to use within-individual linear contrasts to reduce the data from each individual to a single number and then compare these numbers between sequences. The use of within-individual contrasts is the basis of many analyses of cross-over trials (including some multivariate analyses), and thus will be explained in detail.

The within-individual contrasts are linear functions of the observations of each subject; the contrasts' coefficients are the same for all sequences, and the sum of the contrasts' coefficients adds to zero. The estimator of the effect of interest is the difference between (the mean of the) within-individual contrasts of the two sequences. For example, in the 2x2 design the contrast for treatment effects is the difference between the measures in the first and second periods; we obtain the estimator of treatment effects as the difference between the mean contrasts from sequences 1 and 2 (see Jones & Kenward 1989 —pp. 23-28— and Senn 1993a —pp. 42-44; also Díaz-Uriarte 2002).

Contrasts are chosen so that they isolate the effects we are interested in (e.g., treatment effects). In the 2x2 design the Hills-Armitage analysis (explained in Jones & Kenward 1989, Senn 1993a; see also Díaz-Uriarte 2002) is an example of the within-individual contrasts logic. In more complicated designs, there can be several possible linear contrasts for a particular effect, but the estimators with the smallest variance are the Ordinary Least Squares (OLS) estimators (Hafner et al. 1988). The design matrix **X** (with one row per cell mean) that is used to obtain the OLS estimators includes subject, treatment, and period effects and, if appropriate, carry-over effects (see Ratkowsky et al. 1993, for examples with cross-over designs). Senn (1993a, p. 238-248) shows how to obtain the estimators without using matrices. Although these OLS estimators are, strictly, only optimal for uniform covariance structures, with other covariance structures the estimators are less efficient but are still unbiased (see Jones & Kenward 1989) and will not result in increased Type I error rates. These estimators will all take the form of a difference between groups of contrasts among the periods and all the valid contrasts must have the same form in the two sequences.

Once we obtain the contrasts, we compare them between the two sequences. As a test statistic we can use the difference between the two sequences of the mean (within sequence) of the within-individual contrasts. The p-value for this test can be obtained from a randomization test (e.g., Edgington 1995), an independent t-test, or a Wilcoxon rank sum (=Mann-Whitney) test (e.g., Jones & Kenward 1989, p. 51-60; Senn 1993a, p. 93); Tudor & Koch (1994) use the quadratic statistic given by their eq. 2.8 instead of a t-test. If using a Wilcoxon rank sum test, the ranking is done after the linear contrasts are applied (i.e., we do not rank the original data). Covariates (if their value remains the same over all periods of an individual) and other factors can also be examined by using as a response variable the within-individual contrasts (instead of the original values themselves) in a linear model that includes the covariates (e.g., Hafner et al. 1988). If using randomization tests, the restrictions in the allocation of subjects (e.g., same number of subjects to each sequence) should be taken into account.

In some designs (e.g., ABBA,BAAB), the variance of the estimator of treatment effects is smaller when no carry-over effects are included in the model. We can start with an OLS estimator from a design matrix that includes carry-over effects, and if the test of carry-over effects clearly indicates that these effects are unlikely, we could obtain a new OLS estimator of treatment effects from a design matrix that includes no carry-over effect (e.g., Hafner et al., 1988).

An analysis based on within-individual contrasts is robust in the sense that it makes no assumptions about the covariance structure (Jones & Kenward 1989, p. 65, 160, 283; Hafner et al. 1988), although the analysis does assume that the responses of different animals are independent. However, in particular in designs with many periods, power is lost with respect to, say, a linear mixed-effects model when assumptions of the mixed-model are met.

The use of contrasts can be understood in a randomization test context. Under the null hypothesis of no treatment effects, an individual that was assigned to sequence AB would have yielded the same pair of values if it had been assigned to sequence BA, because individuals are assigned randomly to sequences. Thus, the difference between periods 1 and 2 should be the same regardless of sequence assignment (note that possible period effects are thus taken into account). Contrasts must be the same regardless of sequence: under the null hypothesis, a linear combination of an individual's responses must remain unchanged —i.e., the estimate must be invariant under permutations of the observations. For most designs we cannot test for period effects using a randomization test: as periods are not randomized, the order of observations must remain the same in all possible random assignments of subjects to sequences (Shen & Quade 1983). In fact, the OLS estimator for period will differ depending on the sequence, and we cannot devise a randomization test to examine period effects.

Transformations of data can affect the results of nonparametric and randomization methods. Before conducting any analyses, we should consider the appropriate scale for the data; e.g., if the effect of treatment will be to increase the response in one treatment by a multiple of the response under the other treatment (i.e., a multiplicative effect) then we will probably want to log-transform the data before any tests. Notice, however, that interpretation of results from parametric and non-parametric tests can differ (e.g., Conover 1980; Johnson, 1995; Stewart-Oaten 1995; Seaman & Jaeger 1990).

1.1.2 Blocking, between-subject treatments, and more than two sequences

When experiments are carried out in blocks (e.g., weeks, age groups, or locations), analyses that use randomization tests can be applied as before, but the randomization tests must preserve the restricted randomization used in the experiment (e.g., Edgington 1995; p. 131; Noreen 1989, p. 28; Maritz 1995, p. 191). The test statistic is computed from all data together for each permutation, but the random reallocation is restricted to within-blocks. Designs that involve both among and within-individual level treatments can be analyzed with the approach above, although care is required in the selection of the test statistic and the specification of the underlying model (e.g., interactions between the among and within-individual treatments should generally be considered). An alternative is to use the extended Mantel-Haenszel test (e.g., Agresti 1990, p. 283, Koch & Edwards 1988, p. 418; for cross-over Tudor & Koch 1994, p. 358 and 375; there are several tests which contain the words "Mantel-Haenszel"; the test referred to here is applicable to ordinal response variables). With small sample sizes, this statistic's approximate chi-square distribution (1 d.f.) is not appropriate, and the p-value should be determined with a randomization test.

Designs made by pairs of dual sequences can be analyzed like a blocked design, but now each sub-design will have its corresponding OLS estimator. The analyses using t-tests involve obtaining a combined estimator of the treatment difference and its variance, and are shown in Jones & Kenward (1989, p. 171 and ff.). Alternatively, with randomization tests, the testing procedure would be analogous to a blocked design, where each sub-design constitutes a block (e.g., Tudor & Koch 1994, p. 376); however, in contrast to the blocked design, here the test statistic is computed separately for each of the designs, and later combined (after weighting by sample size of each sub-design). For the AA,BB,AB,BA design see Elswick & Uthoff (1989; also Tudor & Koch 1994, p. 374).

1.1.3 More than two treatments

Non-parametric tests of designs for three or more treatments are more complicated. Senn (1993a, p. 144-152) presents a test that can be applied to designs with the appropriate structure (e.g., Table B.2 b); the procedure is analogous to the one used for designs made of dual sequences (see paragraph above), where we test differences between pairs of treatments by arranging sequences in pairs where the two treatments appear in interchanged periods (analogous to dual designs). For each pair, we obtain the statistic by forming the appropriate within-individual contrasts. We then combine the statistics over all pairs of sequences using a weighted sum. This is another example of the extended Mantel-Haenszel test, and can be analyzed as such (Senn 1993a, p. 150; Koch & Edwards 1988). Application of this test requires a particular (and somewhat restrictive) design; if we suspect we will use nonparametric methods, we should design the trial to conform to this structure in advance.

For designs that do not have this structure, Peace & Koch (1993) present a more general test, which is based on obtaining sequence differences of period contrasts, so as to isolate the effects of interest (e.g., pairwise differences between treatments). This method requires relatively large sample sizes and that the different sequences have the same number of subjects; allocation of subjects to sequences during the execution of the experiment should be done by blocks (with number of subjects per block an integer multiple of the number of sequences). A randomization test for a three-period, three-treatment trial is shown in Shen & Quade (1983); it can handle missing data, but assumes uncorrelated errors.

Tests for three-treatment, three-period designs that consist of replicated sets of two Williams squares (see B) are shown in Bellavance & Tardiff (1995). These tests are based on a non-parametric test of a randomized block design (a procedure similar to, but more efficient than, Friedman's test); it assumes that correlation of errors across time does not change, and it can not be extended to more than three treatments. For the s-treatment, s-period (s$\geq$3) Williams square design, Ohrvick (1998) presents tests for treatment effects (and procedures for multiple comparisons); these tests also assume that correlation of errors across time does not change.

1.2 Linear mixed-effects models

The distinguishing features of cross-over designs (e.g., Jones & Kenward 1989; Lindsey 1993) are time-changing covariates (the most obvious one being the within-individual treatment; other within-individual covariates might also change over time) and potentially correlated observations within individuals. Covariates can easily be considered in linear mixed-effects models, and these models can also be used to analyze complex experimental designs. Traditionally, cross-overs (and other repeated measures designs) were analyzed with split-plot ANOVA. With more than two periods, however, the split-plot analysis makes restrictive and potentially unrealistic assumptions about the covariance structure (the so-called sphericity condition that, for example, implies that differences between responses in any two periods have the same variance). There are ways to deal with these restrictive assumptions (e.g., Diggle et al. 1994; Crowder & Hand 1990), but it is generally more satisfactory to directly model the covariance structure using linear mixed-effects models (see Pinheiro & Bates 2000; Littell et al. 1996; also Verbeke & Molenberghs, 1997; Bennington & Thayne 1994; Lindsey 1993). Mixed models are ideally suited for cross-over experiments as the latter include both fixed effects (treatment, period, carry-over) and random effects (the subjects or animals). Moreover, software for linear mixed models allows flexible modeling of the covariance structure, deal much better with unbalanced data than traditional ANOVA, and allow use of covariates that change both at the within and among-individual level. Additionally, mixed models can recover information about treatment effects available between subjects

(Littell et al. 1996), which can be important in cross-over designs with unbalance (Brown & Kempton 1994), either from missing data or by design —e.g., partially balanced designs. Finally, linear mixed-models are natural for examining questions of repeatability and individual differences (an important topic in animal behavior —e.g., DeWitt et al. 1999; Aragaki & Meffert 1998; and references therein), as they make it possible to test the relevance of the among-individual variance component.

Linear mixed models can be fitted using, for example, R and S-Plus (package nlme), SAS (PROC MIXED), as well as MacAnova with the mixed function BMDP, Genstat, and others. Examples with cross-over trials are presented in Vonesh & Chinchilli (1997, ch. 4), Littell et al. (1996, pp. 392 & ff.), Lindsey (1993; pp. 136 & ff.). Aside from the modeling of covariance structure and variance heterogeneity, mixed models have many similarities with the usual linear models. An overview of the theory of linear mixed models can be found in Pinheiro & Bates (2000) and Littell et al. (1996) (see also Davidian & Giltinan 1995, ch. 3). General strategies for model building are discussed in Pinheiro & Bates (2000) and Diggle et al. (1994; specially ch. 4 and 5) (see also Verbeke & Molenberghs 1997); in the context of cross-over designs, see Vonesh & Chinchilli (1997, ch. 4). Diagnostic plots of fitted models are covered in detail in Pinheiro & Bates (2000; see also Verbeke & Molenberghs 1997). Mixed models present some difficulties with selecting the appropriate degrees of freedom to use when testing fixed effects (Brown & Kempton 1994 —but with large F-values the differences in d.f. are inconsequential), and can be questionable with small sample sizes (in particular for the effect on estimation of the covariance matrix).

Because of the problems with carry-over effects, there has been disagreement about the appropriate parameterization of the 2x2 design (e.g., see Ratkowsky et al. 1993, ch. 3). One parameterization, based on Jones & Kenward (1989, p. 30) is

$$y_{ijk} = \mu + \lambda_i + s_{ij} + \text{other.random } + \pi_k + \tau_{d[i,k]} + \text{other.fixed} + e_{ijk} \tag{1.1}$$

where in the fixed effects part μ is the intercept, λ is the carry-over (which in this parameterization is equivalent to a sequence effect), π is the effect of period k, τ is the direct treatment effect of the treatment given in period k of sequence group i, s are independent and identically distributed (i.i.d.) N(0,σ_s^2) are the random effects of individual j in sequence i, and e i.i.d. N(0,σ^2) are the within individual errors. All random effects are independent of each other. "Other.fixed" refers to other fixed effects (covariates like body weight or temperature), and "other.random" refers to other random effects (e.g., blocks). A problematic aspect of this parameterization for the 2x2 design is the inclusion of the carry-over effects (see discussion above).

A parameterization that can be extended to models with more than two periods is

$$y_{ijk} = \mu + \xi_i + \text{other.fixed} + \text{other.random} + s_{ij} + \pi_k + \tau_{d[i,k]} + \lambda_{d[i,k-1]} + e_{ijk} \tag{1.2}$$

where everything is as above, but we have added ξ as the effect of sequence. e i.i.d. N($\mathbf{0}$,$\mathbf{R}$) is the random error associated with the m-th period measurement of subject k from sequence i, where $\mathbf{R}$ is the within-individual covariance matrix and is the same across levels of i, j, k. All random effects are independent of each other. Here we can include both sequence and carry over effects. When there are more than two periods, the covariance structure should always be modeled appropriately. I have included in the later model a sequence effect; this is not done by Jones & Kenward (1989) or Senn (1993a), but it appears in Vonesh & Chinchilli (1997, ch. 4; see also Lindsey 1993, p. 15 and 135). We will generally want to include a term for sequence for three reasons. First, when fitting mixed models it is convenient to start with a "saturated model" to estimate the covariance structure (Diggle et al. 1994, ch. 4). Second, the sequence effect, if significant, might alert us to potential problems with the model; a significant sequence effect might result from bad luck during the randomization of subjects to sequences, but it could also be the result of higher order treatment*period and treatment*carry-over interactions not included in the model (see also Elswick & Uthoff 1989). Third, in some cases sequence effects might be what are affected by among-subject treatments (i.e., we will find significant interactions sequence by among-subject treatment).

When modeling period effects it might be appropriate to initially model them as a categorical variable (as the effect of period might plateau), but it might be possible to obtain a simpler model by using polynomial contrasts and sequentially eliminating the higher-order terms, which could result in a model with just a linear trend with time. Moreover, modeling period as a continuous variable eliminates the confounding of period with carry-over (Ratkowsky et al. 1993). Finally, although a typical strategy of model building is generally employed (Jones & Kenward 1989, but see Senn 1993a), where non-significant terms are dropped from the model, the correct approach with non-significant carry-over effects is debated (e.g., Jones & Kenward 1989, p. 150).

There are some differences in the literature on how to code the carry-over term. For example, suppose that our design has treatments A, B, C; we will need a carry-over column in our data with levels A, B, C, and 0 (Crowder & Hand 1990, p. 107), as the first period has no previous treatment (but this means that the first period and carry-over 0 are completely confounded). This is the approach used by Senn (1993a) and Littell et al. 1996 (p. 392). However, in SAS we will not be able to obtain estimates (e.g., LSMEANS statement); thus, Littell et al. (1996) recode carry-over, creating one dummy variable per treatment which has a 1 if that treatment was in the previous period, and 0 otherwise; this has no effect on the p-values, but allows to obtain estimates. We can also use dummy variables for both period and carry-over that avoid over-parameterizations (see e.g., Diggle et al. 1994, p. 156). In the example of three-periods and three-treatments, for period we use two dummies (say, x1 and x2), which take value 0 on the first period, and for carry-over we also use two dummies (say, x3 and x4), which take value 0 for previous treatment A; note that we do not need to code for the no-carry-over of the first period, as this corresponds to x1=0 and x2=0. This third coding strategy should produce similar results as the first two. The first two approaches do not work with nlme (S-Plus and R) if period is coded as a categorical variable, as we end up with a singular design matrix; however, the third will work in both SAS and S-Plus and R. Littell et al. (1991, p. 206) use a different method, which can yield different results from the above one. Ratkowsky et al. (1993) propose making the first carry-over (0) equal to one of the other treatments; this, however, is not recommended as results from mixed models depend on which other treatment is set as the carry-over in the first period.

The d.f. that our analyses will yield should be examined during the design period, and also serve as a check of the software output (but beware that Satterthwaite's approximation —see details, for instance, in Little et al. 1996, pp. 37 & ff. and pp. 146 & ff.— might yield different d.f. in unbalanced designs). Following Jones & Kenward (1989) (p. 141), for a design with s sequences and p periods we will have (sp-1) d.f. that can be divided in (s-1) d.f. between groups, (p-1) between periods, and (s-1)(p-1) for the group*period effects (more will be available if period is modeled as a continuous variable). The latter (group*period d.f.) are the d.f. which relate to the effects of interest, specially treatment effects, treatment*period interactions, and carry-over effects. We can partition these d.f. in several different ways, but we will always be limited by the total (s-1)(p-1) d.f. (or more if period is continuous). Jones & Kenward (1989) discuss how some terms (in particular carry-over and treatment*period) might be aliased, which can affect the interpretation of treatment effects (see also Koch et al. 1983). With between-subject treatments, some of the d.f. will be used to account for interactions such as treatment*among-subject treatment, period*among-subject, etc.

Chapter 2

Categorical data

Categorical data are among the most difficult to analyze in cross-over designs; at the same time this is an area of very active statistical research. I start discussing several nonparametric-like methods, first for binary responses and next for ordinal outcomes. Later I review methods that are explicitly model-based.

2.1 Nonparametric-like methods

For the 2x2 trial with binary response, there are two main tests for treatment effects (see Jones & Kenward 1989, p. 89-105; Senn 1993a, p. 106-109; Crowder & Hand 1990, p. 109-110; Fidler 1984), and (as usual) these tests are appropriate for treatment effects in the absence of differential carry-over effects. Both tests are based on comparing scores for individuals in the two periods; each subject yields a pair of responses, cd, which means response c in period 1 and response d in period 2; thus, we can have pairs 00, 11, 01, 10 (the last two outcomes are referred to as showing a preference). The Mainland-Gart test uses only information form the 10 and 01 outcomes, comparing the number of each of these outcomes between the two sequences using, for example, Fisher's exact test. Prescott's test is equivalent to scoring profile 01 as -1, profile 10 as +1, and profiles 00 and 11 as 0, and comparing the mean profile between the two sequences using a randomization t-test (which is equivalent to using an exact conditional test for linear trend on the 2 x 3 contingency table —this is different from an exact test for independence). If the software package reports one-sided p-values for exact conditional tests for contingency tables we will want to double that p-value. The Mainland-Gart test does not depend on the random allocation of subjects to sequences, whereas Prescott's test does, but in virtually all behavioral ecology experiments subjects will have been allocated to sequences randomly. Moreover, Prescott's test is generally more sensitive than the Mainland-Gart test. Thus, Prescott's test is likely to be the more useful of the two. However, tests of binary response data in 2x2 trials tend not be very powerful (i.e., they are not very sensitive to treatment differences), and this can be aggravated if only a few subjects in each sequence show a preference (i.e., are either 01 or 10). Becker & Balagtas (1993) present a test that can can be slightly more powerful than Prescott's test, but is also more complicated.

For binary responses and designs with three or more treatments and a particular structure (e.g., Table B.2), Senn (1993a, p. 153-155) proposes a method analogous to the one described above for non-parametric analyses of metric responses with more than two treatments; this method can be applied with both Mainland-Gart's and Prescott's tests.

With ordered categorical data, Senn (1993a, p. 109-113; for a detailed example see also Senn 1993 b, and discussion by Ezzet & Whitehead 1991 1993) presents a simple method based on a heuristic argument for the 2x2 design. For each subject, we reduce the data from the two periods to another ordered categorical response (e.g., if in period 1 an individual was in good condition whereas in period 2 it was in very good condition, the value for this individual becomes "improve"). We are left with ordinal data for each sequence, and differences between the two sequence groups are an indication of treatment effects. We can compare the two sequence groups using, e.g., proportional odd models (Agresti 1990, pp. 323-331). These methods might be questionable in trials with small sample sizes.

Alternatively, for ordinal data, Tudor & Koch (1994, p. 359-361) present several tests based on Wilcoxon's rank sum statistic; these tests involve differences between ranks within periods (in contrast to the other non-parametric tests where ranking was done over the whole sample). These statistics are easy to compute; with small samples, the p-value can be obtained from the permutation distribution. A more complicated approach is presented in Brunner & Newmann (1987) who use different tests based on alternative schemes of ranking the observations.

For 2-treatment, 2-sequences (and $\geq$2 periods) designs, Jung & Koch (1999) present a development of methods discussed in Tudor & Koch (1994, p. 361-362) based on Mann-Whitney measures of association. In each period, these statistics estimate the probability of a larger response of a randomly selected member from one of the groups relative to a randomly selected member of the other group. This method allows stratification and inclusion of covariates and only requires moderate sample sizes ($\geq$ 10 individuals per sequence); the method is slightly complicated to apply (although Jung & Koch 1999, present three detailed examples of application), but is useful for ordinal response variables and continuous asymmetric distributions (with possible outliers). Nonparametric methods for ordinal data with three or more treatments are not well developed.

2.2 Explicitly model-based methods

The methods in the previous section are specific for certain types of responses and/or designs. However, it is possible to analyze categorical data (binary, nominal, and ordinal) for a potentially unlimited range of cross-over designs with methods based on explicit models (see Kenward & Jones 1994). These methods are based on generalized linear models (McCullagh & Nelder 1989; Agresti, 1990; Dobson 1990; Crawley 1993). Generalized linear models are extensions of linear models that make it possible to analyze data in which a function —called the link function— of the mean response (but not the response itself) is linearly related to a set of predictors, and where the variance of the response might be a function of the mean response; generalized linear models have become the standard way of analyzing categorical data.

With categorical data (and also with other data, such as survival; see below) we need to distinguish between different types of models, the two most common being marginal or population averaged, and subject-specific or random-effects (see discussion in Kenward & Jones 1994; Albert 1999; Diggle et al. 1994, ch. 7; Lindsey 1993, ch. 2; Liang et al. 1992; Zeger et al. 1988). Briefly, marginal models model the marginal distribution of the response as a function of the explanatory variables; this modeling is done separately from the within-subject correlation across time (which is treated as a nuisance) and the estimated coefficients have a population interpretation (not an individual interpretation). In contrast, in subject-specific models a random effect for an individual is introduced (as was done in the linear mixed models), and the parameter estimates (say, for treatment effects) modify the probability of a specific subject giving one response instead of another. The distinction between marginal and subject-specific models is not important for linear models because we can formulate the two approaches so that the coefficients have the same interpretation; however, with categorical (and survival) data this is generally not the case for most link functions.

Generalized estimating equations (GEE) are marginal models and can be implemented (see Horton & Lipsitz 1999) using SAS (PROC GENMOD) and R and S-Plus (package gee; for S-Plus also library yags at http://www.biostat.harvard.edu/˜carey); GEE's should perform relatively well in experiments with at least 20 subjects; estimators (e.g., of treatment effects) are consistent even when the correlation structure is misspecified, and testing is done using a robust estimator of variance; Albert (1999) and Horton & Lipsitz (1999) present useful tutorials on GEE's. However, J. K. Lindsey, has pointed out —pers. comm.— that GEE's are not appropriate for cross-over designs, because GEE's treat dependence among observations as if treatments were between subjects, instead of within subjects; thus, the corrected standard errors from GEE's are inflated instead of reduced —the opposite of what one wants—, and therefore result in lower statistical power. Generalized linear mixed models are subject-specific models in which the random subject effects are assumed to follow some distribution; these models can be fitted with SAS (PROC NLMIXED and macro GLIMMIX —Littell et al. 1996), and R (package repeated, from J. Lindsey, available at http://www.luc.ac.be/˜jlindsey/rcode.html; see also Lindsey's packages gnlmm for generalized non-linear mixed models and library growth) but might not perform adequately with small sample sizes. Alternatively, generalized linear mixed models can be estimated using bayesian methods, for instance with the GLMMGibbs package for R (see Myles & Clayton 2001 in the documentation for GLMMGibbs) or using the BUGS and WinBUGS programs (Spiegelhalter et al. 1996, 2000). Conditional likelihood models are also subject-specific models (but here the subject effects are eliminated), and they can be fitted using software for log-linear models, such as SAS's PROC CATMOD (see Kenward & Jones 1991, for examples), and for some conditional models distribution-free and exact permutation tests are available (Agresti 1993; Kenward & Jones, 1994). Discussion and references of GEE's and generalized linear mixed effects models can be found in Albert (1999), Horton & Lipsitz (1999), Littell et al. (1996, ch. 11), Vonesh & Chinchilli (1997, ch. 8), Diggle et al. (1994, ch. 7-9), Kenward & Jones (1994), Lindsey (1993, ch. 2), Lipsitz et al. (1994), and SAS's manual (which includes a cross-over example). Recent examples of applications to cross-over trials are shown in Diggle et al. (1994; GEE's in pp. 154-159; conditional likelihood in pp. 175-181), Kenward & Jones (1994) and Lindsey (1993, pp. 201-204).

Chapter 3

Time to event data: censored observations

Many studies in animal behavior collect time to event data (also called failure time data or survival data) such as time until a certain behavior is displayed (e.g., time to re-emerge from a refuge following a predator's attack). Generally, animals are observed for a predetermined time, and the observer records when the event takes place. If the event takes place in every period for every subject, these are metric data (and can be analyzed with either parametric or nonparametric methods). However, for some subjects the event might not occur within the observation period, which results in censoring (i.e., all we know is that the time till the event occurs is larger than the observation time). Although a small number of censored observations probably does not preclude the use of the parametric and nonparametric methods above, censored observations make usual techniques for metric data, including non-parametric ones (see France et al. 1991; Ducroq, 1997), inappropriate. Censoring can violate several of the assumptions of both parametric and non-parametric tests and will result in tests insensitive to treatment effects and biased estimates of treatment effects. In particular, converting survival data into 0/1 data (for no-event and event respectively) is not only arbitrary (the coding depends on the time at which the categorization is made) but is also a very inefficient use of information. Moreover, 0/1 scores do not really facilitate the analysis with cross-over designs.

Censoring can be of several types (for details see, e.g., Klein & Moeschberger, 1997; Lee 1992). The most common in behavioral studies is Type I censoring, where the event is observed only if it occurs before some predetermined time. This censoring time is usually common for all individuals; with random censoring — censoring time a random variable— data can be analyzed with methods for Type I censoring, provided that censoring and survival times are independent (O'Brien & Fleming 1987; Heimann & Neuhaus 1998).

Analysis of censored data, generally referred to as survival analysis or reliability analysis, is well developed (e.g., Klein & Moeschberger 1997; Collett 1994; Lee 1992; Lawless 1982; Kalbfleisch & Prentice 1980), but techniques applicable to experiments where the same individual experiences the event repeatedly are not common. Some methods have been proposed to analyze paired censored data (e.g., Woolson

& O'Gorman 1992; O'Brien & Fleming 1987), but these methods cannot be applied to cross-over designs if there are period effects.

Two recent techniques available to analyze repeated time to event derive from the analysis of multivariate time to event data, but might not be appropriate with small sample sizes. The method developed by Lee et al. (1992; see also Lin 1994, 1993; Wei et al. 1989; also Hougaard 2001 and Therneau & Grambsch 2000) assumes a marginal proportional hazards model; it does not require that we specify the form of the joint distribution of the observations of each subject. Frailty models (e.g., Klein & Moeschberger 1997, ch. 13; Ducrocq 1997; Hougaard 2001, Therneau & Grambsch 2000) are subject-specific models in which all the observations from a subject share a common frailty (a common random effect that affects the hazard rates of all the observations of a subject); frailty models require that we assume a particular distribution for the frailty (generally a gamma). Both the marginal and frailty models are available in R and S-Plus(package survival5) and in SAS (PROC PHREG —Allison 1995, pp. 236-247).

Lindsey et al. (1996) present a method specific for cross-over designs based on log-linear models, which has the advantage that it works with relatively small sample sizes and can be fitted with software that handles generalized linear models such as R, S-Plus, SAS, GLIM. The R package event (available at http://www.luc.ac.be/˜jlindsey/rcode.html; the syntax for model building with this library is somewhat different from other R statistical models) will fit these (see function ehr) and other models for repeated censored data. Segal & Neuhaus (1993) present a related marginal method that combines Poisson regression with GEE and can be implemented with SAS, S-Plus, or R. Two advantages of all these four methods are: a) they can accommodate covariates and factorial designs that mix within- and among-subject treatments —although not necessarily nested designs; b) they can be used to analyze experiments where we have measured more than one response variable. Many modeling strategies for these methods are common with linear models (see above).

Feingold and Gillespie (1996) suggested two nonparametric-like approaches for two-treatment designs. Their second method is tailored to the 2x2 design but is difficult to extend to other designs. Their first method has wider generality; one first ranks (see below) all the observations, and then applies the procedures for complete data to these ranks (i.e., one applies within-individual contrasts to the ranks, and later compares the within-subject contrasts between the sequences; note that with Koch's (1972) method, however, one first computes within-individual contrasts and then ranks them). There are several ways of ranking the observations in the context of survival analysis; Feingold & Gillespie (1996) employ Gehan's (1965a & b) scores; log-rank scores (see explanation in, e.g., Lawless 1982, p. 420; Lee 1992, p.109-112) might be preferable (Prentice & Marek 1979; O'Brien & Fleming 1987; Kalbfleisch & Prentice 1980; Lee 1992; Lawless, 1982). The p-value for this test could be obtained with a t-test, a Mann-Whitney test, or a randomization test. This method is easy to apply, and it can be used with multiple strata or trials composed of dual designs, e.g., by using the extended Mantel-Haenszel test with the log-ranked data (e.g., Tudor & Koch 1994) or using randomization tests where randomization is constrained within strata. An example of the application of this method to a behavioral experiment is given in Díaz-Uriarte (1999). An alternative to Feingold & Gillespie's (1996) approach is to apply the methods in "Ordinal responses" to log-ranks of the data (see Tudor & Koch 1994, p. 365).

Chapter 4

Multivariate responses

4.1 Multivariate repeated measures: parametric methods

Behavioral ecology experiments frequently collect more than one response variable (e.g., in an anti-predator experiment in each period we might measure distance from the predator and time to re-emerge from the refuge, so we would have measured q=2 different response variables). This is somewhat similar to making repeated measurements (of the same response variable) within each time period (e.g., in each one of p periods, we might record the preferred perch height at 5 min intervals during 1 h; thus we have q "sub-periods" —here q=12— or different measurement occasions within each period). In both cases these are called "doubly multivariate" or "multivariate repeated measures". Multiple univariate tests of each one of the response variables (or at each one of the repeated observation times) can result in inferential problems as they ignore possible dependencies between observations (e.g., Krzanowski, 1990, p. 235 & ff.; Johnson & Wichern 1998). Sometimes there is a large increase in Type I error rate (i.e., the true experiment-wise alpha level is larger than the nominal alpha level); other times fully multivariate approaches can attain larger power by using the information from the correlation among variables. With multiple responses it is frequently advised (e.g., Johnson & Wichern, 1998) that one should initially use a multivariate test and only if it reveals significant differences employ univariate tests on each response variable.

For metric data, Jones & Kenward (1989) devote a chapter (ch. 6) to **repeated observations of the same variable**. First, we could summarize the repeated data for each individual into one or a few statistics, such as area under the curve, slope and intercept, etc.; this is the simplest approach. However, this approach is problematic when the data are incomplete, and when covariates take different values during the observation session. Moreover, use of this approach requires obtaining a scientifically meaningful data summary, and thus assuming that all the information in the data that is not reflected by the summary statistic(s) is scientifically uninteresting (see also Crowder & Hand 1990, ch. 1; Diggle et al. 1994, ch. 6 for discussion).

With **two-sequences designs**, a second approach (see Jones & Kenward 1989, ch. 6) is to obtain within-individual contrasts (see 1.1.1) for each sub-period q; thus, we reduce the data from a total of q*p to q derived measurements, and can analyze these q derived measurements with appropriate repeated measures techniques (e.g., MANOVA). For instance, Patel & Hearne (1980; see also Rodríguez-Carvajal & Freeman 1999, p. 399) use a multivariate linear model and obtain, for each subject, a new transformed variable which is a linear combination of the original responses over the q sub-periods, and then use a two-sample Student's t-test on the transformed variable. This procedure tests the hypothesis that the sum of treatment effects over all periods is the same for the two sequences (and thus would not be appropriate with multiple responses —different variates).

A third approach, more satisfactory and flexible (and a necessity with more complicated designs) is to fit all the data in a single model (i.e., avoid reducing the data to q derived measurements). We can use a split-plot in time repeated measures ANOVA where we have three strata: between-subjects, within-subjects-among-periods, and within-period (i.e., the "sub-period" level). These analyses, like other split-plot-in-time repeated measures, make assumptions about the covariance structure which might not be appropriate; moreover, they are cumbersome if the spacing between successive measurements is unequal or if there are missing data. Thus we can also employ linear mixed effects models by specifying the corresponding random effects and covariance structures (see an example in Littell et al. 1996, pp. 388 and ff.). In addition, Galecki (1994) discusses some covariance structures which can be used with mixed models and allow flexibility for modeling the correlation structures for each repeated factor. These structures can be fitted using SAS's PROC MIXED; with the nlme library for R and S-Plus these structures can be fitted by defining the appropriate correlation structure.

With multiple response variables, application of Galecki's (1994) structures might not be appropriate (as they require that the marginal covariance structure associated with time be the same for every response variable). Thus, mixed models with more complex covariance structures (and a larger number of parameters) need to be fitted (e.g., Amemiya 1994; Vonesh & Chinchilli 1997). These models could be fitted, for instance, using a completely unstructured (positive-definite) variance-covariance matrix (but in this case we would probably be estimating too many parameters). Alternatively, in S-Plus or R it might be possible to define special covariance structures tailored to our specific situation (e.g., unstructured except for blocks along the diagonal with particular structures for the within-variate covariance structure).

With categorical data, both GEE and generalized linear mixed models can accommodate multiple responses, although the latter requires that we specify the covariance structure. With time-to-event data, multiple responses can be easily analyzed with the marginal approach of Lee et al (1992; we only need to obtain the quadratic form for the multivariate tests as in pp. 1066 and 1070 in Wei et al. 1989; see also documentation of library survival5) and the log-linear models of Lindsey et al. (1996; see pp. 531 and ff. for a worked example).

For some cross-over designs with multivariate normal responses, some simple approaches have been worked out. Rodríguez-Carvajal & Freeman (1999) show how to carry out a multivariate analysis in the 2x2 case using Hotelling's T^2 (a common statistic for multivariate comparisons of two groups; e.g., Morrison, 1990; Krzanowski 1990). Grender & Johnson (1993; pp. 71-74 and 84) had proposed a similar but more general approach that can be extended to some higher-order designs, and it is applicable to both repeated measures and multiple responses, and to multiple responses with repeated measures for each response. The tests of Rodríguez-Carvajal & Freeman (1999) and Grender & Johnson (1993) for the multiple response situation is a simultaneous (multivariate) test of the hypothesis that the treatment effect vectors are the same in both sequences (which is appropriate when variates are not measured in the same scale), and differs from the test of Patel & Hearne (1980) explained above.

4.2 Nonparametric and randomization tests

A different approach is to use nonparametric, rank-based, and randomization multivariate tests. Analogous to robust and nonparametric tests of section 1.1.1, the first step is to reduce the p*q measurements of each individual to a set of q variates by applying within-individual contrasts separately to each variate. We will refer to these as w-q. (With survival data a possibility is to apply the methods of Feingold & Gillespie (1996) by obtaining the w-q from the log-ranks or Gehan's scores of the data —not the original, censored, data; however, it is unknown how well this approach works). This first step of obtaining the w-q variates will be common to all the remaining multivariate tests. The next step is to compare, with the appropriate multivariate test, the w-q variates among sequences. Therefore, we can apply any multivariate test provided that we can set the hypothesis test as a comparison among sequences of within-individual contrasts. This will be possible (see Jones & Kenward 1989, pp. 171 & ff.; Senn 1993a, pp. 144-152; "Metric responses: nonparametric and robust methods" section) with two-treatment designs composed of pairs of dual sequences and with designs for more than two treatments that have the special structure in Table B.2 b, but it might not be possible otherwise; this emphasizes again the need to consider design and analysis before conducting the experiment. As was done before, we might want to start with within-individual contrasts that include carry-over effects, and later recompute the w-q from contrasts without carry-over if multivariate and univariate tests show no evidence of carry-over effects in any variable.

A very simple approach is to use the test in O'Brien (1984); first, each w-q is ranked separately; next, for each individual i we compute S_i as the sum of the ranks of all of the w-q. We test the null hypothesis of no overall difference between treatments by comparing the S_i's between sequences, using a two-sample t-test, a rank-sum test, or a randomization test. This method can be extended to accommodate individual-level covariates (by using, e.g., a linear model with S_i as the response and sequence and covariate as independent variables) and blocking (see section 1.1.2). This application of O'Brien's test is very similar to Patel & Hearne's (1980) method, except that we use a linear combination of the ranks instead of the original variables (which is what makes it possible to apply the test to variables measured in different scales). A drawback of O'Brien's test is that it is appropriate only for some limited alternative hypotheses (see below).

After ranking each w-q separately, an alternative to O'Brien's (1984) test is to use the multivariate extension of the Kruskal-Wallis test (Puri & Sen 1971, p. 184 & ff.), which is equivalent to applying a MANOVA on the separately ranked w-q variates (Zwick 1985; note that with only two groups a MANOVA is the same as Hotelling's T^2). This is the test discussed in Johnson & Grender (1993; however they compute the test statistic using N*Pillai-Bartlett's trace, instead of (N-1)*Pillai-Bartlett's trace, as in Zwick 1985; this is inconsequential if a randomization test is used, but not if the chi-square approximation is used).

A different test is obtained by applying the procedures of Mielke and collaborators (Mielke et al. 1976, 1981 a & b) to the w-q variates (without ranking), as explained in Johnson & Mercante (1996). This method does not assume any particular distribution for the data or homoscedasticity. We compute the average distance among the individuals of the two sequences in the q-dimensional space defined by the w-q variates, using an appropriate distance metric (e.g., Euclidean distance —but distance metric can affect power; Díaz-Uriarte & Nordheim, in prep.). Under the null hypothesis, permuting individuals randomly between the two sequences should have no effect on the average within-sequence distance, but under the alternative hypothesis permuting individuals should increase the average within-sequence distance. (P-values can be obtained from randomization tests, or using an approximation; see Mielke et al. 1976, 1981b; Berry & Mielke 1983). When different response variables are measured in different scales, we will probably want to give equal weights to all variables; equal weights can be achieved by scaling the data (e.g., to a mean of zero and variance of one) before computing the within-subject comparisons or by applying the test to the ranks of the w-q variates —where each w-q is ranked separately—; (see Johnson & Mercante 1996). An example of the application of this method to a behavioral study is given in Díaz-Uriarte (1999).

The tests discussed so far have been previously used with cross-over designs. Besides them, other randomization (e.g., Manly 1997, ch.12; Edgington 1995, ch. 8) and rank-based (e.g., Puri & Sen 1971, 1985; Thompson 1991; Choi & Marden 1997; Hettmansperger et al. 1998) multivariate tests could potentially be applied, either to the w-q variates or their ranks (with ranks computed for each variate separately or all together, depending on the test).

4.3 A quick discussion of the previous methods

In summary, we can apply a fully multivariate approach to the original responses; this requires modeling the variance-covariance matrix in linear mixed models but not necessarily with GEE's or marginal survival models. When this is not feasible, multivariate and repeated measures tests can be applied to the w-q variates/responses. The latter, although more robust than, say, a fully multivariate linear mixed model, can also be considerably less powerful as we lose degrees of freedom when we reduce the data to w-q contrasts The appropriate statistic will depend on the null and alternative hypotheses and the structure of the data (and should not be decided based upon the results of the tests). For example, O'Brien's (1984) test is not designed to detect treatment effects that occur in only a few variates, or when the responses in different variates are not consistent (e.g., if there are negative correlations among variates). On the other hand, Hotelling's T^2 is not the most powerful test against restricted alternatives. Moreover, among nonparametric and rank-based multivariate tests, performance can be strongly affected by the shape of the distributions. Finally, different multivariate tests make different assumptions (normality, homoscedasticity, symmetry of distributions, etc.). Discussion can be found in Smith (1998), Choi & Marden (1997), Manly (1997, ch. 12), Edgington (1995, ch. 8), Westfall & Young (1993, ch. 6), Lachin (1992), Bernstein et al. (1988), and O'Brien (1984).

4.4 Multiple univariate tests

A different approach is to adjust the p-values to control for the increase in Type-I error rate from multiple univariate tests (e.g., Wright 1992 and references therein; two articles in biological journals are Rice 1989 and Chandler, 1995). These adjustments are better suited for situations (such as data snooping or the so-called "fishing expeditions") where we are testing many individual hypotheses and want to control overall Type I error rates (e.g., we want to examine in which of five response variables a treatment has some effect), but are probably not the best approach when we conduct our experiment with the objective of testing a particular multivariate hypothesis (specified before the experiment was conducted); this approach is also useful when it is not possible to combine the different tests into a single multivariate test. Most of the most recent methods (e.g., Hochberg's and Holm's sequential Bonferroni methods) provide much higher power than the traditional Bonferroni method (without increasing experiment-wise error rates), and some of them increase this power further by taking into account possible covariation among variables (e.g., Westfall & Young 1993). For instance, the resampling-based methods in Westfall & Young (1993; see also SAS Institute 1996, documentation for PROC MULTTEST) could be applied to the between sequence comparison of the w-q variates. Alternatively, we can employ the usual methods for cross-over trials with each variable independently, and later make an overall statement about the effect of a treatment by using, for example, Holm's multiple comparisons method.

Even in the absence of rigorous statistical methods for dealing with multiple response variables, some of the inferential problems arising from multiple responses can be minimized with careful experimental design and analysis. For instance, what hypotheses will be tested, and with what variables, can be specified a priori; also, different variables can be used to test different (biological) hypotheses, so that even if the data are not statistically independent, they at least refer to very different biological phenomena. This is not to suggest that other variables should not be examined for treatment effects, but just that testing of pre-specified hypotheses should be differentiated from hypotheses generation, for which data snooping might be well suited (see also discussion in Stewart-Oaten 1995). Paraphrasing Rice (1989, p. 225), adjustment for multiple testing is necessary because, otherwise, as authors we will be spending many pages discussing spurious results, and as readers we will be wasting our time reading about relationships that can be explained just by chance.

4.5 PCA in lieu of MANOVA?

A potential mistake in the analysis of multiple responses is to try to use Principal Components Analysis (e.g., Morrison 1990; Krzanowski 1990; Bernstein et al., 1988) to reduce the dimensionality of the response space, and then analyze the principal components scores as if they were independent response variables. This procedure is inappropriate for two reasons. First, if we want to reduce the dimensionality of the problem in the context of considering differences between groups, we should use canonical variates, which are different from principal components; canonical variates are closely related to MANOVA, canonical correlation, and discriminant analysis (see Krzanowski 1990, p. 291-300 and 370-385; Bernstein et al. 1988, ch. 10; Digby & Kempton 1987, pp. 75-77). Second, when using PCA we would be mixing within and among-individual covariation in the response variables. However, it should be possible to use canonical variate analysis on the w-q variates (including randomization-based canonical variate analysis —Manly 1997, p. 274).

Chapter 5

Plotting in cross-over designs

Plotting is a key tool in statistical analysis and can help us spot patterns and problems in the original data and the fitted models. We can plot the original data, plot some linear functions of the data, or make plots that are specific for the types of analyses carried out (particularly helpful to examine violations of model assumptions, such as residual plots). I will briefly review the first two here.

Initial plots of the data will help detect errors in the transcription or recording of data, and will give an idea of the results that could be expected. Jones & Kenward (1989 p. 20) refer to **subject profile** plots where, for each sequence, the response of each subject is plotted over the different treatment periods, and the responses of each subject are connected with a line. These plots help identify period and treatment effects, potential outliers, and variation within and among sequences. For designs with more than two treatments, it is convenient to add treatment labels in the x-axis. In **treatment by treatment scatter-plots** (Senn 1993a, p. 188), we plot each patient's values using each treatment response as a dimension.

The **response by patient scatterplot** (Senn 1993a, p. 125 and 187) depicts the response variable (y-axis) by the sequence, using the same symbol across sequences to identify treatments; all the responses of a subject are shown in the same vertical line (x-axis position). This plot conveys a lot of information: variation within-subjects, variation among sequences, magnitude of differences between treatments, and possible differences in treatment effects across sequences (e.g., treatment*period interaction), as well as potential outliers (either a whole subject or observations within an otherwise non-outlying subject). This plot and the subject profiles plot complement each other, as they convey similar information in different ways. In these plots, covariates or other factors can be added by using symbols. Plots for time to event data are based on the survival function and are shown in Feingold & Gillespie (1996). Non-metric data are generally difficult to plot conveniently, and tables are probably more useful (but see Senn 1993a, p. 188-190).

The second type of plots are those that depict some function of the data, such as the linear contrasts. These plots are very useful at the initial and intermediate stages of formal analyses. For the 2x2 design, Jones & Kenward (1989, p. 28-30) discuss a plot that helps understand the **role of carry-over and treatment effects**. In a scatterplot, each individual's sum over the two periods is shown in the x-axis and each individual's difference between the first and second periods in the y-axis; individuals from each of the two sequences are plotted with different symbols, and the outermost points of each sequence are joined (i.e., we draw the convex hull of each sequence group). If there are only strong treatment effects, we will see two non-overlapping curves that are separated in the vertical direction; if there are carry-over effects, the separation will be along the horizontal axis. This plot also gives visual information on the variability in each sequence (for parametric analyses, variance should be the same in each group). The **groups-by-period plot** (Jones & Kenward 1989, p. 20) shows the group by period means for each sequence, connected by a segment. These are very similar to the usual interaction plots in linear models. Plotting the linear contrast by a covariate can be particularly helpful to understand the role of continuous covariates. Miller (1999) has proposed two types of plots that help identify outliers and indicate whether representing differences between samples by a single statistic (such as the mean) is appropriate; these plots allow us to examine subject by treatment interactions and changes in carry-over effect over time.

Summary plots of results should avoid two potential **pitfalls**. First, if analyses have been nonparametric it is misleading to use plots that represent a mean and its standard error, as these have no relationship with the actual analyses conducted (and could suggest that the mean and s.e. are adequate characterizations of the data distribution, which they are not). Second, in cross-over trials the estimator is based on within-individual differences, and the relevant source of variance is the within-individual variability, not the among-individual variability. Thus, a plot of the overall mean of treatments A and B, each with an standard error, would be of little use as the analyses were conducted using within individual differences; moreover, this plot can suggest no effect even when there is a strong one. Instead, it is preferable to plot the estimated treatment difference with its standard error (with no treatment differences, the confidence interval should cover 0). If we need to present the estimates of the actual responses with with some measure of variability, it is best if those treatment means are adjusted treatment means (as obtained from, e.g., linear models after correcting for effects of period and other fixed effects), and if a cautionary statement is added to the figure legend indicating that those means and s.e. cannot be used to conduct a visual test of the hypothesis.

Chapter 6

Sample size and missing data

Discussion of sample size and power is provided in Senn (1993a, p. 211-219), Hills & Armitage (1979), and Ezzet & Whitehead (1992). Sample size calculations can be extremely complicated except for the simplest designs, and when planning trials we would need information on variances, which is not always available before the trial starts.

The consequences of missing data can be particularly serious for the 2x2 design; the simplest strategy is to use only subjects without missing data, but other strategies are possible (Jones & Kenward 1989, p. 76-80). For other designs, the consequences of missing data are not necessarily that serious, and probably all the available data from every subject should be used (see Senn 1993a, p. 219-221; see also Low et al. 1999 for discussion of robustness of cross-over designs to dropouts).

It is important to understand what is the missing data mechanism (e.g., Diggle et al. 1994, ch. 11; Albert 1999). A common classification is based on Littell & Rubin (1987). Data are missing completely at random (MCAR) if the missing mechanisms is independent of both the observed and actual missing value; they are missing at random (MAR) when the missing mechanism is independent of the actual missing value but depend on observed data (e.g., if it depends on previously observed values); and they are missing non-randomly (= informative missing mechanism or non-ignorable missingness) when the missing mechanisms depends on the values of the missed observations. For instance, suppose we are measuring fight duration in an experiment where each subject is scheduled to be observed five times per day, but occasionally we can not obtain complete records for each individual. If there is a constant probability that we cannot find the subject for the scheduled observation we have a MCAR mechanism. If, however, long previous fights make it more unlikely that we will able to find the subject for the following trial (e.g., following a long fight an animal is more likely to move somewhere else), then we have a MAR mechanism. We will have non-ignorable missingness if the probability that we observe a short fight is smaller than that of observing a long fight (i.e., the probability of recording a fight increases with fight duration, the variable we are measuring).

The statistical methods discussed above can accommodate MCAR data; some of them (e.g., linear mixed models, but not GEE) also accommodate MAR data; but most methods will be biased with informative missing values (e.g., experiments where the probability of having missing data depend on the treatment applied). Application of multivariate/repeated measures within periods techniques can be much more complicated in the presence of missing values or incomplete observations (see, e.g., Davis 1991; Lachin 1992; Palesch & Lachin 1994).

Chapter 7

Conclusions

Cross-over designs can be very useful in many behavioral experiments (see Díaz-Uriarte 2002); however, their analyses are more complicated than those of parallel trials. When planning a cross-over trial we should consider both the design and analysis, as the type of response variable can affect the choice of design. Cross-over trials will be much easier to analyze if we can keep the design simple, minimizing nesting and crossing of among-subject treatments (but if the setup does include these factors, they should be incorporated in the analyses).

Analysis of categorical data (specially ordered responses) can be complicated with cross-over designs, and generally requires at least moderate sample sizes ($\geq$10 individuals per sequence group); even with moderate sample size, power might be too low to detect small, but biologically relevant, differences between treatments. Analysis of time to event data can also be unsatisfactory, but is easier if censoring time is common for all individuals. More complex designs, such as those that include blocks and covariates, can make analysis of categorical and time to event data very complicated. Modifying the experimental protocol might ameliorate some of these problems; for example, to avoid censored data we might make observation periods longer, and to eliminate categories such as ”low perch”, ”medium height”, ”high perch” we might be able to actually measure perch height. In particular, it is best to always obtain data at as high a level as possible in the measurement hierarchy (i.e., as close to interval as possible), and to remember that degrading data into categories such as orderings or 0/1 will make analyses more complicated. Experiments with three or more treatments are inherently more complicated to design and analyze, in particular if nonparametric and robust methods will be used. Experiments that measure multiple responses should use multivariate techniques. Finally, how carry-over and period effects are dealt with should be made explicit.

Appendix A

Terminology

Direct treatment effect The effect of a treatment at the time of its application. Often times abbreviated to treatment effect (when there is no possibility of confusion).

Period effect A period is each one of the occasions in which a treatment is applied, and thus period effects refer to those changes in the value of the response that are due to the response variable being measured at, say, time t instead of $t+1$ or $t-1$.

Carry-over effects The effects of a treatment that persist after the end of the treatment period. Carry-over effects appear when the response to a current treatment is affected by what treatment was applied in a previous period.

Sequence The order in which the within-individual treatments are applied. Designs are often referred to using sequences, such as ABB,BAA, which means that animals assigned to the sequence ABB are first given treatment A (1st period), then B (2nd period), then B (3rd period), and animals assigned to the BAA sequence are first given B, then A, then A (1st, 2nd, and 3rd periods, respectively).

Sequence effects or Group main effect Any effect related to a particular sequence of treatments, such as an overall difference in the responses to the treatments in animals of sequence AB compared to those of sequence BA. A sequence effect can result if animals assigned to one sequence are different to animals assigned to the other sequence, but under a randomized design it is reasonable to assume that there are no sequence effects (Crowder & Hand, 1990). In many designs, however, a sequence effect can be confounded with other effects (see Crowder & Hand, 1990, Jones & Kenward 1989, Ratkowsky et al. 1993).

Appendix B

Design of cross-over trials

Here I quickly review the main designs that could be useful in behavioral studies; more details are provided in Jones & Kenward (1989), Senn (1993 a), and Ratkowsky et al. (1993). I will only examine designs that consider period effects plausible. To maximize power, subjects should be allocated to treatments so that there are equal numbers of subjects for each sequence (and this restriction should be reflected when using randomization tests).

During the design phase, it is essential to understand how the data will be analyzed. For example, some nonparametric methods for more than two treatments require that the designs be of a specific kind or that allocation of subjects be done in a particular way; some other methods only work with large sample sizes. These requirements might prompt one to either change the design, to try to allocate more subjects or allocate subjects in different ways, or to measure different response variables.

B.1 Designs for two-treatment trials.

The most common cross-over design is the AB,BA design. This design is problematic in the absence of information about carry-over effects (see review in Díaz-Uriarte 2002). Even when carry-over effects are not present designs with more than two periods can be preferable as they lead to estimators of treatment effects with smaller variance, and therefore increase power (e.g., the ABB,BAA design has a variance for the estimate of treatment effects which is 19% of that from AB,BA –provided we use the same number of subjects, allocated in equal numbers to each sequence).

Table B.1 shows three two-treatment designs, and some of their basic properties which affect the degree of aliasing (aliasing refers to the presence, in the design matrix, of covariates which are linear combinations of other covariates; technically, it refers to the amount of overlap between the subspaces defined by the covariates; McCullagh & Nelder 1989, pp. 61-68). The consequence of aliasing is that we cannot obtain separate estimates of each parameter. Aliasing is a common problem in cross-over designs; the correlation between parameters is an indication of aliasing, and is listed for many designs in Jones & Kenward (1989) (and can also be easily obtained by matrix operations from the design matrix; see, e.g., Ratkowsky et al. 1993). For instance, in the design ABB,BAA the correlation between the estimate of treatment and carry-over effect is zero, and thus the estimate of treatment effects is the same in a model with or without carry-over effects, which is a good quality if the statistical model includes carry-over effects.

We can classify two-treatment designs by the number of sequences and the number of periods. Designs differ in the variance of estimated treatment effects (tabulated in Jones & Kenward 1989 for many designs). In general, the more periods the smaller the variance, but when sequences with many periods are used it is more likely that there will be missing data for later periods; thus, designs with more than 3 or 4 periods are not very advisable. Also, some designs are less affected by having to end a trial before it was expected: if one uses a design such as ABBA,BAAB and cannot collect data from the last period one is left with ABB,BAA which is a good design (in contrast with eliminating the last period from AAAB,BBBA). When only two periods can be used the AA,BB,AB,BA design (Balaam's design for two treatments) can minimize problems from carry-over effects; however, this design might be a worse choice than simply ensuring a long enough wash-out period and using AB,BA.

Designs composed of many sequences will be more complicated to use, in particular with limited sample sizes, as one will need sample sizes which are integer multiples of the number of sequences (to have the largest power). This is more problematic when one uses blocking or between-subject treatments (as one usually will want to use the complete design –i.e., all the sequences– in each block or between-subject treatment). Dual designs, or designs composed of dual sequences, (i.e., pairs of sequences where the second sequence is obtained by interchanging the treatment labels A and B of the first sequence) allow one to use simple and robust analysis based on within-individual comparisons (see section 1.1.1). The designs in Table B.1 are composed of dual sequences and are among the most useful for estimating treatment effects and also perform well under different within-individual correlation structures (Jones & Kenward 1989; Matthews 1990).

Table B.1: Some cross-over designs for two treatments(see Jones & Kenward 1989; definitions from Vonesh & Chinchilli 1997 are slightly different from those in Laska et al. 1983 and Jones & Kenward 1989).

Design	Uniform within sequences[1]	Uniform within periods[2]	Balanced[3]	Strongly balanced[4]	Variance of the estimator of treatment effects[5]	Variance of the estimator of treatment effects when carry-over effects are present[5]
ABB,BAA	No	Yes	Yes	Yes	0.375	0.375
ABBA,BAAB	Yes	Yes	Yes	No	0.250	0.275
ABBA,BAAB, AABB,BBAA	Yes	Yes	Yes	Yes	0.250	0.250

[1] A design is uniform within sequences if each treatment appears the same number of times within each sequence; sequence effects are not aliased with treatment effects.

[2] A design is uniform within periods if each treatment appears the same number of times within each period; period effects are then not aliased with treatment effects.

[3] A design is balanced if each treatment precedes each other treatment the same number of times; in this case, carry-over effects are aliased with treatment effects. A balanced design, as defined in Jones & Kenward (1989), is one that is balanced (as in this table), uniform within sequences –actually, each subject receives each treatment only once– and uniform within periods, and with equal number of subjects per sequence.

[4] A design is strongly balanced (or completely balanced) if each treatment precedes each other treatment, including itself, the same number of times; in this case, carry-over effects are not aliased with treatment effects.

[5] Expressed in multiples of (σ^2/Total number of subjects), assuming equal numbers of subjects allocated to each sequence.

B.2 Designs for more than two treatments

With more than two treatments we can distinguish between variance balanced (all pairwise differences between treatments are estimated with the same precision) and partially balanced designs (the variance of the comparison between two treatments depends on which two treatments are compared). Partially balanced designs might be the best choice when there are several experimental treatments and one control and we are most interested in minimizing the variance of contrasts between each experimental treatment and the control. We can also differentiate between complete and incomplete block designs (e.g., Senn 1993, p. 163 and ff.; Jones & Kenward 1989, p. 199 and ff.); in the latter the number of treatments is larger than the number of periods (so each individual is not subject to all the treatments). Incomplete block designs are particularly useful with large numbers of treatments; however, these are much more difficult to design and analyze, and thus are of limited interest in animal behavior studies.

If period can have an effect (as we generally assume), designs should be uniform within periods (see Table B.1, for explanation). Designs uniform within periods can be based on Latin squares (briefly, suppose we arrange our design as a square, with n rows and n columns; then, in a Latin square we can apply n treatments, and ensure that each treatment is applied once, and only once, in each row and column; for cross-over designs, the rows represent sequences and the columns represent periods). Williams designs or Williams squares (e.g., Table B.2a) are also balanced (with respect to carry-over; see Table B.1). Under certain assumptions, we can minimize problems from carry-over effects by using extra-period designs. For example, we can use a Williams designs to which we add a period so that the last treatment is equal to the previous one (e.g., in Table B.2, the first sequence would be ADBCC), and we obtain a strongly balanced design (see Table B.1). However Williams and strongly-balanced designs might not be particularly useful if carry-over is not an issue. Other designs based on Latin squares (e.g., Table B.2 b) have the property that, for every pair of treatments two sequences can be found where the treatments appear in interchanged periods (Senn 1993, p. 122 and 123); this property allows us to use some nonparametric and multivariate analyses (see section 1.1.3 and 4.2). Discussion of designs for three or more treatments can be found in Senn (1993, ch. 5, 9, 10), Jones & Kenward (1989, ch. 5), and Ratkowsky et al. (1993, ch. 5 and 6). In general, designs for more than four treatments will require sample sizes larger than those available in most behavioral studies.

The assignment of subjects to sequences (including blocking), and the election of the number of squares, are discussed in Senn (1993 p. 123 & 209-210) and Jones & Kenward (1989, p. 196-197; 198-199). In a three treatment trial, we can either use one or the two Latin squares (if carry-over effects are included in the model, we will use the two sets of Latin squares). For four treatments, either several squares or a single one can be used; the latter is generally simpler and will be less affected by loss of subjects.

Finally, the optimality of the designs discussed above depends on assumptions that might be inappropriate in some cases (e.g., when we expects treatment*carry-

Table B.2: Examples of cross-over designs for four treatments; a) Williams design; b) for every pair of treatments two sequences can be found where the treatments appear in interchanged periods (e.g., in sequence 1, A is in the 1st period and D in the 2nd period, whereas in sequence four the positions of A and D are reversed.

a)

Sequence	Period d			
	1	2	3	4
1	A	D	B	C
2	B	A	C	D
3	C	B	D	A
4	D	C	A	B

b)

Sequence	Period			
	1	2	3	4
1	A	D	B	C
2	B	C	A	D
3	C	B	D	A
4	D	A	C	B

over interaction). It is possible to construct optimal cross-over designs tailored to the particular assumptions of our model (see Donev 1998; Jones & Donev, 1996), and also use a sequential approach to trial design, so that assumptions can be incorporated as information becomes available.

B.3 Between-subjects designs and baseline data

Cross-over designs can be used in experiments that also include between-subject treatments (e.g., comparing the effect of female presence/absence in a cross-over trial, in which different individuals have been assigned to different hormonal manipulation treatments). Inclusion of these between-subject factors in the analyses is mentioned several times in these review (see index).

The use of baseline data (data collected before treatment(s) is(are) applied) can be found in Jones & Kenward (1989) and Senn (1993 see also Tsai & Patel (1996) for non-parametric analysis of a 2x2 design). Baseline data can increase the sensitivity of tests for treatment*period interactions and between-subject treatments; however, baseline data do not increase sensitivity of tests of direct treatment effects, and thus are unlikely to be useful in most behavioral studies.

References

Agresti, A. 1990. *Categorical Data Analysis.* New York: John Wiley & Sons.

Agresti, A. 1993. Distribution-free fitting of logit models with random effects for repeated categorical responses. *Statistics in Medicine*, 12, 1969-1987.

Albert, P. S. 1999. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine*, 18, 1707-1732.

Allison, P. D. 1995. *Survival Analysis Using the SAS System: a Practical Guide.* Cary, NC: SAS Institute Inc.

Amemiya, Y. 1994. On multivariate mixed model analysis. In: *Multivariate analysis and its applications.* IMS lecture notes, vol. 24. (Ed. by T. W. Anderson, K. T. Fang, and I. Olkin), pp. 83-95. Hayward, CA: Institute of Mathematical Statistics

Aragaki, D. L. R. and Meffert, L. M. 1998. A test of how well the repeatability of courtship predicts its heritability. *Animal Behaviour*, 55, 1141-1150.

Becker, M. P. and Balagtas, C. C. 1993. Marginal modeling of binary cross-over data. *Biometrics*, 49, 997-1009.

Bellavance, F. and Tardif, S. 1995. A nonparametric approach to the analysis of three-treatment three-period crossover designs. *Biometrika*, 82, 865-875.

Bennington, C. C. and Thayne, W. V. 1994. Use and misuse of mixed model analysis of variance in ecological studies. *Ecology*, 75, 717-722.

Bernstein, I. H., Garbin, C. P., and Teng, G. K. 1988. *Applied Multivariate Analysis.* New York: Springer-Verlag.

Berry, K. J. and Mielke, P. W. Jr. 1983. Computation of finite population parameters and approximate probability values for multi-response permutation procedures (MRPP). *Commun. Statist.-Simula. Computa.*, 12, 83-107.

Brown, H. K. and Kempton, R. A. 1994. The application of REML in clinical trials. *Statistics in Medicine*, 13, 1601-1617.

Brunner, E. and Neumann, N. 1987. Non-parametric methods for the 2-period cross-over design under weak model assumptions. *Biometrical Journal*, 29, 907-920.

Chandler, C. R. 1995. Practical considerations in the use of simulatenous inference for multiple tests. *Animal Behaviour*, 49, 524-527.

Choi, K. and Marden, J. 1997. An approach to multivariate rank tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 92, 1581-1590.

Collett, D. 1994. *Modelling Survival Data in Medical Research.* London: Chapman & Hall.

Conover, W. J. 1980. *Practical Nonparametric Statistics.* New York: John Wiley & Sons.

Crawley, M. J. 1993. *GLIM for Ecologists.* Oxford: Blackwell.

Crowder, M. J. and Hand, D. J. 1990. *Analysis of Repeated Measures.* New York: Chapman & Hall.

Crowley, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*, 23, 405-447.

Davidian, M. and Giltinan, D. M. 1995. *Nonlinear Models for Repeated Measurement Data.* London: Chapman & Hall.

Davis, C. S. 1991. Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine*, 10, 1959-1980.

DeWitt, T. J., Sih, A., and Hucko, J. A. 1999. Trait compensation and cospecialization in a freshwater snail: size, shape and antipredator behaviour. *Animal Behaviour*, 58, 397-407.

Díaz-Uriarte, R. 1999. Anti-predator behaviour changes following an aggressive encounter in the lizard *Tropidurus hispidus*. *Proceedings Royal Society of London, Series B*, 266, 2457-2464.

Díaz-Uriarte, R. 2001. Territorial intrusion risk and antipredator behaviour: a mathematical model. *Proceedings Royal Society of London, Series B*, 268, 1165-1173.

Díaz-Uriarte, R. 2002. Incorrect analysis of cross-over trials in animal behaviour research. *Animal Behaviour*, in press.

Digby, P. G. N. and Kempton, R. A. 1987. *Multivariate Analysis of Ecological Communities.* London: Chapman and Hall.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. 1994. *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Dobson, A. J. 1990. *An Introduction to Generalized Linear Models.* London : Chapman & Hall.

Donev, A. N. 1998. Crossover designs with correlated observations. *Journal of Biopharmaceutical Statistics*, 8, 249-62.

Ducrocq, V. 1997. Survival analysis, a statistical tool for longevity data. *48th Annual Meeting European Association for Animal Production*, 25-28 August, 1997, Vienna, Austria.

Edgington, E. S. 1995. *Randomization Tests*, 3rd Ed. New York: Marcel Dekker.

Elswick, R. K. and Uthoff, V. A. 1989. A nonparametric approach to the analysis of the 2-treatment, 2-period, 4-sequence crossover model. *Biometrics*, 45, 663-667.

Ezzet, F. and Whitehead, J. 1991. A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine*, 10, 901-6.

Ezzet, F. and Whitehead, J. 1992. A random effects model for binary data from crossover clinical-trials. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 41, 117-126.

Ezzet, F. and Whitehead, J. 1993. A random effects model for ordinal responses from a crossover trial - Reply. *Statistics in Medicine*, 12, 2150-2151.

Feingold, M. and Gillespie, B. W. 1996. Cross-over trials with censored data. *Statistics in Medicine*, 15, 953-967.

Fidler, V. 1984. Change-over clinical trial with binary data: mixed-model-based comparison of tests. *Biometrics*, 40, 1063-1070.

France, L. A., Lewis, J. A., and Kay, R. 1991. The analysis of failure time data in crossover studies. *Statistics in Medicine*, 10, 1099-1113.

Galecki, A. T. 1994. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Commun. Statist.-Theory Meth.*, 23, 3105-3119.

Gehan, E. A. 1965. A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, 52, 650-653.

Gehan, E. A. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52, 203-223.

Good, P. 1994. *Permutation Tests.* New York: Springer-Verlag.

Grender, J. M. and Johnson, W. D. 1993. Analysis of crossover designs with multivariate response. *Statistics in Medicine*, 12, 69-89.

Hafner, K. B., Koch, G. G., and Canada, A. T. 1988. Some analysis strategies for three-period changeover designs with two treatments. *Statistics in Medicine*, 7, 471-481.

Heimann, G. and Neuhaus, G. 1998. Permutational distribution of the log-rank statistic under random censorship with applications to carcinogenicity assays. *Biometrics*, 54, 168-184.

Hettmansperger, T. P., Mottonen, J., and Oja, H. 1998. Affine invariant multivariate rank tests for several samples. *Statistica Sinica*, 8, 785-800.

Hills, M. and Armitage, P. 1979. The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology*, 8, 7-20.

Horton, N. J. and Lipsitz, S. R. 1999. Review of software to fit Generalized Estimating Equation regression models. The American Statistician, 53, 160-169.

Hougaard, P. 2001. *Analysis of multivariate survival data.* New York: Springer-Verlag

Johnson, D. H. 1995. Statistical sirens: the allure of nonparametrics. *Ecology*, 76: 1998-2000.

Johnson, W. D. and Grender, J. M. 1993. Multivariate nonparametric analysis for the two-period crossover design with application in clinical trials. *Journal of Biopharmaceutical Statistics*, 3, 1-12.

Johnson, W. D. and Mercante, D. E. 1996. Analyzing multivariate data in crossover designs using permutation tests. *Journal of Biopharmaceutical Statistics*, 6, 327-42.

Johnson, R. A. and Wichern, D. W. 1998. *Applied Multivariate Statistical Analysis*, 4th ed. N.J., Prentice Hall.

Jones, B. and Kenward, M. G. 1989. *Design and Analysis of Cross-Over Trials.* New York: Chapman & Hall.

Jung, J. W. and Koch, G. G. 1999. Multivariate non-parametric methods for Mann-Whitney statistics to analyse cross-over studies with two treatment sequences. *Statistics in Medicine*, 18, 989-1017.

Kalbfleisch, J. D. and Prentice, R. L. 1980. *The Statistical Analysis of Failure Time Data.* New York: John Wiley & Sons.

Kenward, M. G. and Jones, B. 1991. The analysis of categorical-data from cross-over trials using a latent variable model. *Statistics in Medicine*, 10, 1607-1619.

Kenward, M. G. and Jones, B. 1994. The analysis of binary and categorical data from crossover trials. *Statistical Methods in Medical Research*, 3, 325-44.

Klein, J. P. and Moeschberger, M. L. 1997. *Survival Analysis.* New York: Springer-Verlag.

Koch, G. G. 1972. The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics*, 28, 577-584.

Koch, G. G. and Edwards, S. 1988. Clinical efficiency trials with categorical data. In: *Biopharmaceutical statistics for drug development.* (Ed. by K. E. Peace), pp. 403-457. New York: Marcel Dekker

Koch, G. G., Gitomer, S. L., and Skalland, L. 1983. Some non-parametric and categorical data analyses for a change-over design study and discussion of apparent carry-over effects. *Statistics in Medicine*, 2, 397-412.

Krzanowski, W. J. 1990. *Principles of Multivariate Analysis.* New York: Oxford University Press.

Lachin, J. M. 1992. Some large-sample distribution-free estimators and tests for multivariate partially incomplete populations. *Statistics in Medicine*, 11, 1151-1170.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data.* New York: John Wiley & Sons.

Lee, E. T. 1992. *Statistical Methods for Survival Data Analysis.* New York: John Wiley & Sons, Inc.

Lee, E. W., Wei, L. J., and Amato, D. A. 1992. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: *Survival analysis: state of the art.* (Ed. by J. P. Klein and P. K. Goel), pp. 237-247. The Netherlands: Kluwer Academic

Liang, K.-Y., Zeger, S. L., and Qaqish, B. 1992. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society B* , 54, 3-40.

Lin, D. Y. 1993. MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data. *Comp. Meth. Progr. Biom.*, 40, 279-293.

Lin, D. Y. 1994. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, 13, 2233-2247.

Lindsey, J. K. 1993. *Models for Repeated Measurements.* Oxford: Clarendon Press.

Lindsey, J. K., Jones, B., and Lewis, J. A. 1996. Analysis of cross-over trials for duration data. *Statistics in Medicine*, 15, 527-35.

Lipsitz, S. R., Kim., K., and Zhao, L. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, 13, 1149-1163.

Littell, R. C., Freund, R. J., and Spector, P. C. 1991. *SAS System for Linear Models*, 3rd Ed. Cary, NC: SAS Institute.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. 1996. SAS System for Mixed Models. Cary, NC: SAS Institute.

Littell, R. J. A. and Rubin, D. B. 1987. *Statistical Analysis With Missing Data.* New York: John Wiley & Sons.

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

Ludbrook, J. and Dudley, H. 1998. Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52, 127-132.

Manly, B. F. J. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, 2nd Ed. London: Chapman & Hall.

Maritz, J. S. 1995. *Distribution-Free Statistical Methods*, 2nd Ed. London: Chapman & Hall.

McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models*, 2nd Ed. New York: Chapman & Hall.

Mielke, P. W. Jr., Berry, K. J., and Johnson, E. S. 1976. Multi-response permutation procedures for a priori classifications. *Commun. Statist.-Theor. Meth.*, A5, 1409-1424.

Mielke, P. W. Jr., Berry, K. J., and Brier, G. W. 1981a. Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. *Monthly Weather Review*, 109, 120-126.

Mielke, P. W., Berry, K. J., Brockwell, P. J., and Williams, J. S. 1981b. A class of nonparametric tests based on multiresponse permutation procedures. *Biometrika*, 68, 720-724.

Myles, J., and Clayton, D. 2001. GLMMGibbs: An R package for estimating bayesian generalised linear mixed models by Gibbs sampling. Documentation for the R pckage GLMMGibbs, included with the package.

Miller, W. E. 1999. A visualization of cross-over data using linear functions. *Statistics in Medicine*, 18, 975-987.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, 3r Ed. New York: McGraw-Hill.

Noreen, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses: an Introduction.* New York: John Wiley & Sons.

O'Brien, P. C. 1984. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40, 1079-1087.

O'Brien, P. C. and Fleming, T. R. 1987. A paired Prentice-Wilcoxon test for censored paired data. *Biometrics*, 43, 169-180.

Ohrvik, J. 1998. Nonparametric methods in crossover trials. *Biometrical Journal*, 40, 771-789.

Palesch, Y. Y. and Lachin, J. M. 1994. Asymptotically distribution-free multivariate rank-tests for multiple samples with partially incomplete observations. *Statistica Sinica*, 4, 373-387.

Patel, H. I. and Hearne III, E. M. 1980. Multivariate analysis for the two-period repeated measures crossover design with application to clinical trials. *Commu. Statist.-Theor. Meth.*, A9, 1919-1929.

Peace, K. E. and Koch, G. G. 1993. Statistical methods for a three-period crossover design in which high dose cannot be used first. *Journal of Biopharmaceutical Statistics*, 3, 103-116.

Pinheiro, J. C & Bates, D. M. 2000. *Mixed-Effects Models in S.* Springer-Verlag, in press.

Prentice, R. L. and Marek, P. 1979. A qualitative discrepancy between censored data rank tests. *Biometrics*, 35, 861-867.

Puri, M. L. and Sen, P. K. 1971. *Nonparametric Methods in Multivariate Analysis.* New York: John Wiley & Sons.

Puri, M. L. and Sen, P. K. 1985. *Nonparametric Methods in General Linear Models.* New York: John Wiley & Sons.

Ratkowsky, D. A., Evans, M. A., and Alldredge, J. R. 1993. *Cross-Over Experiments: Design, Analysis, and Application.* New York: Marcel Dekker.

Rice, W. R. 1989. Analyzing tables of statistical tests. *Evolution*, 43, 223-225.

Rodriguez-Carvajal, L. A. and Freeman, G. H. 1999. Multivariate AB-BA crossover design. *Journal of Applied Statistics*, 26, 393-403.

SAS Institute Inc. 1996. *SAS/STAT Software: Changes and Enhancements Through Release 6.11.* Cary, NC: SAS Institute.

Seaman, J. W. and Jaeger, R. G. 1990. Statisticae dogmaticae: a critical essay on statistical practice in ecology. *Herpetologica*, 46, 337-346.

Segal, M. R. and Neuhaus, J. M. 1993. Robust inference for multivariate survival-data. *Statistics in Medicine*, 12, 1019-1031.

Senn, S. 1993a. *Cross-Over Trials in Clinical Research.* New York: John Wiley & Sons.

Senn, S. 1993b. A random effects model for ordinal responses from a crossover trial [letter; comment]. *Statistics in Medicine*, 12, 2147-51.

Shen, C. D. and Quade, D. 1983. A randomization test for a three-period three-treatment crossover experiment. *Commun. Statist.-Simula. Computa.*, 12, 183-199.

Smith, E. P. 1998. Randomization methods and the analysis of multivariate ecological data. *Environmetrics*, 9, 37-51.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. 1996. BUGS: Bayesian inference using gibbs sampling, verion 0.50. Technical report, Medical Research Conuncil Biostatistics Unit, Cambridge. Available from http://www.mrc-bsu.cam.ac.uk/bugs/)

Spiegelhalter, D. J., Thomas, A., Best, N. G. 2000. WinBUGS Version 1.3 user manual. Available from http://www.mrc-bsu.cam.ac.uk/bugs/)

Stewart-Oaten, A. 1995. Rules and judgments in statistics: three examples. *Ecology*, 76, 2001-2009.

Therneau, T. and Grambsch, P. 2000. *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag, New York, in press.

Thompson, G. L. 1991. A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, 86, 410-419.

Tudor, G. and Koch, G. G. 1994. Review of nonparametric methods for the analysis of crossover studies. *Statistical Methods in Medical Research*, 3, 345-81.

Verbeke, G. and Molenberghs, G (eds).1997. *Linear mixed models in practice. A SAS-Oriented approach.* New York: Springer-Verlag

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

Wei, L. J., Lin, D. Y., and Weissfeld, L. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

Westfall, P. H. and Young, S. S. 1993. *Resampling-Based Multiple Testing.* New York: John Wiley & Sons.

Woolson, R. F. and O'Gorman, T. W. 1992. A comparison of several tests for censored paired data. *Statistics in Medicine*, 11, 193-208.

Wrigth, S. P. 1992. Adjusted P-values for simultaneous inference. *Biometrics*, 48, 1005-1013.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

Zwick, R. 1985. Nonparametric one-way multivariate analysis of variance: a computational approach based on the Pillai-Bartlett trace. *Psychological Bulletin*, 97, 148-52.

Subject index