Some problems in paired comparisons and goodness of fit for logistic regression

by

Kenneth Alec Butler

B. Sc., University of Birmingham, 1985

M. Sc., Simon Fraser University, 1989

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in the Department

of

Mathematics & Statistics

SIMON FRASER UNIVERSITY

July 1997

APPROVAL

Name: Kenneth Alec Butler

Degree: Doctor of Philosophy

Title of thesis: Some problems in paired comparisons and goodness of fit for logistic regression

Examining Committee: Dr. C. Schwarz
Chair

Dr. R. Lockhart
Senior Supervisor

Dr. C. Dean

Dr. R. Routledge

Dr. M. A. Stephens

Dr. R. R. Davidson
External Examiner
Department of Mathematics and Statistics
University of Victoria

Date Approved: July 17, 1997

Abstract

In this thesis, we have two general aims: to learn more about design and analysis of paired-comparison experiments, and to develop some tests of fit for logistic regression models, of which some commonly-used models for paired-comparison data are a special case.

We begin with some "model-free" considerations for paired-comparison experiments. We review some constructions for round-robin designs, especially those with certain desirable properties such as the arrangement of home and away games and the equalization of "carry-over effects", where we improve on some previous work in the field. Also considered are non-parametric tests of equality of strength of the teams in a tournament, where we obtain some new asymptotic results in the presence of ties and order effects, and show by simulation that the asymptotic approximations are generally good.

We then review the Bradley-Terry model for paired comparisons, and its generalizations to handle ties and order effects. We review some algorithms for fitting the model, and illustrate with examples the kinds of data for which the algorithms perform well.

Next, we turn to the issue of optimal design for the Bradley-Terry model. We review the theory of continuous designs, and show that the special structure of the Bradley-Terry model enables the D-optimal continuous design to be found explicitly in many cases. We then discuss the implementation of a well-known algorithm for finding exact D-optimal designs and indicate by example that the algorithm usually works well enough. Since both of the preceding design types assume known parameters, we also investigate sequential designs in which each stage of the design is deduced from parameter estimates obtained from the previous stages of the design, and show that the obvious algorithm works reasonably well. We then carry out numerical efficiency comparisons of the D-optimal designs with the round-robin and Swiss designs introduced in Chapter 2.

The final chapter is an investigation of some goodness-of-fit tests for logistic regression

models. We arrange the (known or fitted) response success probabilities in ascending order, and then look at the process given by the cumulative difference between observed and expected successes. We obtain asymptotic distribution theory for a class of statistics based on the integral of this process, by central limit theorem arguments, and for a class of statistics based on the integrated squared process, which rests on the theory of weak convergence of quadratic forms in sequences of random variables. Finally, we show by examples that the asymptotic distributions usually form good approximations for finite samples, and so can be recommended for use in practice.

Acknowledgments

I am grateful to the Department of Mathematics and Statistics, and to NSERC through the generosity of Dr. M. A. Stephens and Dr. R. Lockhart, for the financial support that made this work possible.

I would also like to thank my examining committee for their time and their comments on this work.

Most of all, however, I would like to express my thanks to my Senior Supervisor, Dr. R. Lockhart, for the energy, the good ideas and at times the patience he devoted to me and to this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivations

Most of the work in this thesis was motivated by a desire to understand the design and analysis of paired-comparison experiments. There seem to be several problems worthy of consideration, and these are dealt with separately in the following chapters. As a result, the chapters may appear to have little to do with each other, but each, it is hoped, serves to shed light on a part of the problem.

In a paired-comparison experiment, some number t of objects are to be compared, but they can only be compared two at a time, and each comparison yields only the result that one of the two objects was "preferred" to the other (or, possibly, that the objects were equally preferable). In sporting parlance, this corresponds to a tournament containing t teams (or players); the comparisons correspond to games between the teams, and preferences to wins for one team or the other, with "equally preferable" corresponding to a tie. It is also possible to see an "order effect", whereby (say) the first object in each comparison has an advantage purely by being first; in sporting terms, this corresponds to a home field advantage.

The sporting terminology seems easier to follow, and will be used frequently in this thesis, but it should be borne in mind that the methods apply equally to other applications, such as taste-testing, where objects are compared by a judge.

From such an experiment, it is natural to want an overall ranking of the objects, or, better yet, an estimate of their "strengths". Furthermore, we will want to design the experiment in such a way as to achieve an accurate estimation of team strengths with the smallest possible experimental effort.

Finally, we will want to assess the goodness of fit of our proposed models. It turns out that our proposed testing procedure extends easily to the more general case of logistic regression, and so we have investigated the asymptotic theory behind our test statistics in some detail in this more general setting.

1.2 The chapters

The organization of the chapters in this thesis is intended to lead the reader from the simple towards the more complex, with Chapter 5 containing the most general and the most mathematical work. In particular, the different chapters aim to address the issues described below.

1.2.1 Chapter 2

This Chapter is concerned with the simplest case, where modelling is to be kept to a minimum. In particular, attention is focused on round-robin tournaments, which seem naturally to yield a "fair" ranking of the teams. There are numerous design issues here; we review some different constructions from the graph-theory literature, and show how tournaments can be (exactly or approximately) balanced for home and away games and for nuisance factors known as "carry-over effects". In particular, for the class of designs known as "general cyclic designs", we investigate the pattern of carry-over effects in some detail. We also review an algorithm for generating "random" round-robin designs, and give an algorithm of the same type for generating such a design with a specified pattern of home and away games for each team.

We then turn to tests of overall equality in round-robin tournaments. Such tests are an obvious first step in the analysis of paired comparison data. We review a test based on a natural idea, namely the variability of the "scores" (the number of wins for each team); clearly, the greater this variability, the less tenable a hypothesis of the teams being of equal strength. We extend these tests to enable an assessment of overall equality in the presence of ties and order effects. All the test statistics have asymptotic chi-squared distributions. We show, by simulation and by comparison with previous results, that the asymptotic distribution is a remarkably good approximation even when the number of teams and games is small. We also study the distribution of the range of scores, for which an asymptotic approximation based on the range of normal random variables is generally adequate, and

indicate how this may be used for a multiple-comparison procedure like that used in analysis of variance.

Finally, we introduce the Swiss tournament, which is played in rounds like a round-robin tournament, but the constitution of each round is determined from the results of previous rounds rather than being fixed in advance. Swiss tournaments, which are used in such games as chess and bridge, can be thought of as an alternative to round-robins when the number of participants is large. We give a detailed algorithm for use in the typical case where no prior information is available about the teams; this algorithm chooses between otherwise equally-preferable matchups at random. Non-random algorithms are available that use prior information about the team strengths as a seeding mechanism, but we do not pursue these ideas.

1.2.2 Chapter 3

In this chapter, we look at the Bradley-Terry model, which is the standard model used in paired-comparison experiments, and extensions of the model for estimating the effects of ties and home field advantage. The likelihood derivatives are obtained based on an additive version of the model, and, as has been previously shown, the maximum likelihood estimates are obtained by equating observed and expected wins or "points" in a manner reminiscent of contingency table analysis.

We next turn to estimation procedures, considering the obvious candidate of Newton's method along with a simple method due to Ford (in two guises) and Jacobi's method, which assumes, incorrectly, that the second derivative matrix of the log-likelihood is diagonal. We give a detailed investigation of the computational complexity of these algorithms, showing how many additions, multiplications and exponentiations one iteration of each algorithm requires, and then, by means of examples, we assess the actual number of operations required by each algorithm, allowing for the fact that simpler algorithms will tend to require a larger number of iterations. Except for the very largest data sets, Newton's method tends to come out best, despite the complexity of each iteration, because it requires very few iterations. The other methods are preferable when the number of teams is large; Jacobi's method works best when the tournament design is approximately balanced, and Ford's method is good otherwise.

Implicit in the work of Chapter 2 was the idea that ranking teams in a round-robin or Swiss tournament by the wins (or more generally points) they had obtained was a reasonable

thing to do. When the Bradley-Terry model holds, it is known that the ranking of teams in a round-robin by wins is identical to the ranking by Bradley-Terry strengths in the absence of ties and order effects. We show that, when one point is awarded for a win and a half point for a tie, this result continues to hold in the presence of ties and order effects, provided that the home and away games for each team are properly balanced. For Swiss tournaments, no such general results are available, however. We show by simulation that, when ranking teams by points and breaking ties by a standard quantity known as the Buchholz score, the agreement between this ranking and the one obtained from the Bradley-Terry strengths is F¡4However, most of the mis-rankings *v:hmM:Wxs kEquite high. Though there are few mis-rankings, those that do occur can be attributed to imbalance in the strength of opposition faced by the teams involved. We demonstrate that a composite score based on both the points and the Buchholz score produces a ranking that is consistently closer to that based on the Bradley-Terry model.

1.2.3 Chapter 4

In this chapter, we apply the theory of D-optimal designs to paired comparisons, and consider how this theory might be used in practice. We also compare the D-optimal designs with some more familiar types of tournament.

After a brief review of the ideas of optimal design, we show how D-optimality can be applied to paired comparisons when the Bradley-Terry model, in its simplest form, holds. We first need to consider how to deal with an information matrix that is singular; two methods are proposed for "fixing up" the matrix, and these are shown to be equivalent, allowing us to use whichever is more convenient.

Since optimal design for non-linear models requires knowledge of the true parameter values, at least if a complete design is to be generated, we are forced to assume that the parameter values (true team strengths) are indeed known. This offers a conditional approach to design issues: "if these are the team strengths, then the optimal design is this".

We first consider "continuous designs", which indicate the fraction of games in the tournament that should be played between each pair of teams, without regard to the practical issue of obtaining a design in which the numbers of games are integers. We review a result known as the General Equivalence Theorem which yields checkable conditions for the optimality or otherwise of a candidate design, and show how these conditions translate to our case. As we show, this leads to a method which can be used to obtain the D-optimal

continuous design explicitly, at least when games between every pair of teams feature in the optimal design, and we offer some ideas concerning the way to proceed if the method fails.

Practical designs require the numbers of games between each pair of teams to be integers. Such designs are called "exact". We review some known algorithms for obtaining exact designs, and consider the details of implementation of one of the simpler algorithms, which is based on the idea of adding games to and removing games from the design one at a time until no further improvement can be made. The structure of the Bradley-Terry model permits a straightforward assessment of which games to add or remove, as well as simple updates of the inverse and determinant of the information matrix. Some examples are given; the structure of the D-optimal designs shows up clearly, and the choice of a simple algorithm is justified by noting that it finds the best design on a fair proportion of its runs and never (in the examples given) finds a badly sub-optimal design.

We then drop the assumption of known team strengths and consider sequential designs, in which the design is constructed in stages based on the game results previously observed. An obvious algorithm is given, and an example is presented of the algorithm in action.

It is natural to wish to compare different designs; we do so by means of a well-known criterion called D-efficiency, which indicates how a design performs relative to a D-optimal (exact) design of the same size. We look at round-robin and Swiss designs for various numbers of teams and various spreads in team strength, comparing these with the same-size exact D-optimal design (which assumes that the true team strengths are known) and one of the class of sequential designs that we call "sequential-1" (which makes no such assumption). Because the Swiss and sequential-1 designs are generated sequentially, their behaviour can only be assessed by simulation. We find that both round-robin and Swiss designs decrease in efficiency relative to the D-optimal exact designs as the variability in team strengths increases, with the latter decrease being slower, results that correspond to intuition; relative to the sequential-1 designs, we see a slower decrease in efficiency for the round-robin designs and essentially no decrease in efficiency at all for the Swiss designs. This indicates that, relative to the best designs that can be realized in practice, the Swiss tournament performs very well, at least for the range of tournament sizes given.

1.2.4 Chapter 5

This chapter provides the mathematical culmination of the thesis. In considering the assessment of goodness of fit of the Bradley-Terry model introduced in Chapter 3, it became

clear that an obvious idea there, namely that of comparing observed and expected frequencies in a cumulative fashion, could be extended to general logistic regression (of which the Bradley-Terry model is a special case). We therefore consider, in detail, the asymptotic theory of our proposed statistics in this more general case, and, by simulations, assess the quality of the asymptotic distributions as approximations for finite-sample tests.

The idea of comparing observed and expected "successes" cumulatively leads naturally to an empirical process on which test statistics can be based. Standard work with such processes normally results in proofs of weak convergence to a Gaussian process; statistics based on the empirical process can then easily be shown to have asymptotic distributions based in the same way on a Gaussian process. The presence of estimated parameters makes such an approach too difficult in our case, and so we have studied our chosen statistics individually. We look at two families of statistics, an "area family", based on the average of the empirical process itself, and a "quadratic family", based on the average of the squared empirical process. In each family, we begin with the statistic based on all the parameters being known, and work towards the practically-useful statistic in which all parameters are estimated from the data. We are able to find asymptotic distributions, under general conditions, for all the statistics in the area family and (with some difficulty) for the statistic of the quadratic family that is based on known parameters; we also obtain asymptotic results for the entire quadratic family based on a rather more restrictive limiting process.

In order to assess the quality of the asymptotic distribution theory as an approximation for finite samples, we carry out some simulation studies of examples that are intended to be "typical". The studies show that, when parameters are known, convergence to the asymptotic distributions is fast, especially for the quadratic statistic. When parameters are estimated, convergence is (not surprisingly) slower. For the quadratic statistic, the correspondence between simulated and asymptotic distribution is generally good in the lower tail, which is the important tail for inference, but the tails of the simulated distribution of the area statistic are less well behaved.

Chapter 2

Model-free design and analysis

2.1 Introduction

The kinds of tournaments that are easiest to interpret are those in which the teams can be assessed by their wins and losses. For example, a knockout tournament is very commonly used; these tournaments are discussed further in David (1988). However, while these tournaments are very effective in eliciting the best team (Maurer, 1965), they are much less helpful in producing a ranking of all the teams, or an assessment of their strengths. In this thesis, we are concerned more with these last two issues and less with the problem of finding the best team, and so we do not consider knockout tournaments further.

The most intuitively satisfactory approach for paired comparisons is to make each possible comparison the same number of times. This design is called a **round-robin tournament**, and if each comparison is made r times, it is a **r-tuple round robin tournament**. In a round-robin tournament, it is "fair" to compare the objects by comparing their wins and losses, in a sense that can be made precise (see Theorem 3.4), and so the results are easy to interpret. In Section 2.2, we review some known methods of constructing round-robin tournaments, indicate how these methods may be extended or simplified in some cases, and consider some properties of these tournaments concerning home and away games and "carry-overs". We obtain some new results concerning carry-overs, and show that in general it is impossible to design a round-robin tournament with desirable properties for both home and away games and carry-overs.

We then turn to some basic non-parametric tests for use in round-robin tournaments, considering an overall test of equality and some ideas of multiple comparisons. We wish to

consider the effect on these tests of the possibility of ties or the presence of an "order effect" or "home field advantage"; we begin by reviewing known results in the case when ties are impossible and there is no order effect, and then develop tests for use when either or both of these effects are present.

Another design, often used in games such as chess and bridge, is the **Swiss tournament**, which occupies a position between the knockout and the round-robin, and in which it is at least reasonably fair to rank the teams according to their wins and losses. In Section 2.5, we describe how such a tournament may be constructed, and briefly discuss its properties.

2.2 Round-robin designs

2.2.1 Introduction

A general round-robin tournament can have each pair of objects compared any number of times, but in this chapter we consider designs for a single round-robin, where each comparison is only made once. An r-tuple round-robin tournament can be, and usually is, constructed by combining r single round-robins. In practice, single and double round-robins are most common.

It might seem a trivial task to design a round-robin tournament. However, in practice we often prefer to do more than simply choose a pair, compare them, and move on to another pair. We may wish to arrange the pairs so that the comparisons for each object are spread evenly through the design, rather than being concentrated at the beginning or at the end. Most commonly, this is achieved by partitioning the comparisons into "rounds" so that each object has precisely one comparison in each round; this is what we will attempt to do. Alternatively, if the pairs are simply compared one after the other, we may wish to ensure "maximal spacing": that is, once each object has appeared in a comparison, we perform as many comparisons as possible between other objects before this object is compared again. Ross (1939) gives a procedure for an odd number of objects.

Splitting a round-robin into rounds is definitely not a trivial task, as Rosa and Wallis (1982) observe. Consider the following tournament for $t = 6$ objects, with the first three rounds as given. There are $t - 1 = 5$ rounds altogether.

Round 1:	1 vs 4	2 vs 5	3 vs 6
Round 2:	1 vs 5	2 vs 6	3 vs 4
Round 3:	1 vs 6	2 vs 4	3 vs 5

The comparisons 1 vs 2, 1 vs 3 and 2 vs 3 are still to be made, but it is impossible to arrange them into only two rounds.

In this chapter, we will examine a number of round-robin designs and algorithms, and we will consider how well these are equipped to deal with practical issues such as the order within a pair (which object is presented first to a judge, which player or object is at home, etc.). We consider only even t; when t is odd, we can construct a round-robin design for $t + 1$ objects and ignore any comparisons containing object $t + 1$.

2.2.2 Relation to graph theory

A paired-comparison experiment of any type, not just a round-robin tournament, can be expressed as a graph whose vertices represent the objects being compared. If two objects are compared during the experiment, the graph has an edge between their two vertices.

In a single round-robin tournament, all pairs of objects are compared once, and so there is a single edge between each pair of vertices on the graph. This is the "complete graph" on the t vertices.

We wish to partition the tournament into rounds, such that each object appears precisely once in each round. On the graph, a round is represented as a disjoint set of edges with each vertex appearing on precisely one edge. Such a set of edges is called a **1-factor**, and the collection of rounds that make up the tournament is represented as a collection of 1-factors, known as a **1-factorization**.

Mendelson and Rosa (1985) give a detailed survey of known results about 1-factorizations; because of the equivalence just described, these results may also be applied to round-robin tournaments.

2.2.3 Carry-over effects

Suppose object i is paired with object j in one round, and object k in the next. Suppose also that object j is very strong; then object i will probably suffer a heavy defeat against j and will be "discouraged", and perform below par, against k. This happens in sports,

but also in comparisons by a judge: the judge may consciously or subconsciously remember that object i was much less preferable than object j, and may more easily regard i as less preferable than k also. Because j is strong, k benefits from being compared to i after j, and k is said to receive a **carry-over effect** from j.

If object k frequently receives carry-over effects from object j, the benefit to k could be substantial. We therefore prefer designs at least approximately "balanced" for carry-over effects: that is, each object receives carry-over effects from as many different objects as possible.

2.3 Designs

2.3.1 The design GK_t

This is probably the oldest and certainly the most commonly used design for round-robin tournaments. It was known long before 1-factorizations were studied in general, and appears for example in Kraitchik (1953), who gives a simple alternative derivation.

Construction

To obtain the design, number the t (t even) objects $1, 2, \ldots, t-1$ and *; the last object plays a special role. Two objects are paired in round k if their numbers sum to either k or $k+t-1$; one object will be left over in each round, and that object is paired with *.

For example, let $t = 8$; the first three rounds of GK_8 are then:

Round 1:	(Sum = 1 or 8)	1 vs 7	2 vs 6	3 vs 5	4 vs *
Round 2:	(Sum = 2 or 9)	2 vs 7	3 vs 6	4 vs 5	1 vs *
Round 3:	(Sum = 3 or 10)	1 vs 2	3 vs 7	4 vs 6	5 vs *

To ensure that this does indeed generate a round-robin tournament, we must establish the following:

1. The pairing is well-defined: that is, for each object, there is exactly one possible opponent in each round.

2. Each pair of objects does indeed occur precisely once.

Consider object i in round k. Its possible opponents are $k-i$ and $t-1+k-i$. If $k-i$ is a possible opponent, we must have $k-i \geq 1$, but then $t-1+k-i \geq n$. Conversely, if

$t-1+k-i$ is a possible opponent, $t-1+k-i \leq t-1 \Rightarrow k-i \leq 0$. Since the opponent must have a number between 1 and $t-1$ inclusive, only one of the two alternatives can yield an opponent for object i. However, it may happen that $k-i=i$, when $2i=k$ (k even), or $t-1+k-i=i$, when $2i=t-1+k$ (k odd). For any k, there is only one i for which this happens, and then i is paired with * in round k.

To show that each pair of objects does occur precisely once, note that any two objects i and j must have $1 \leq i+j \leq 2(t-1)$, since $1 \leq i,j \leq t-1$. If $i+j \leq t-1$, then objects i and j are paired in round $i+j$; otherwise, i and j are paired in round $i+j-t+1$. We already saw that the object * is paired with objects $1,2,\ldots,t/2-1$ in rounds $2,4,\ldots,t-2$, and with objects $t/2,t/2+1,\ldots,t-1$ in rounds $1,3,\ldots,t-1$. Thus object * is also paired with each of the other objects exactly once.

Home and away games

In practice, it may well make a difference which object of a pair is presented to a judge first, or, in sporting encounters, which player plays first (as in chess), or which object plays on its home field. In this situation, we list the pairings so that the object listed first is the one playing at home.

The design GK_t has two desirable properties concerning home and away games:

1. It is possible to arrange for each object to alternate home and away games with at most one exception. In fact, two of the objects alternate home and away games precisely, while of the remaining $t-2$ objects, half have two consecutive home games at some stage, while half have two consecutive away games.

2. The objects can be grouped into pairs so that whenever one object of the pair has a home game, the other has an away game, and *vice versa*. This is important if pairs of objects (teams) have the same home stadium.

Proofs of these results can be found in de Werra (1980). We illustrate with an example for $t=6$:

Round 1:	1 vs 5	2 vs 4	3 vs *
Round 2:	* vs 1	5 vs 2	4 vs 3
Round 3:	2 vs 1	3 vs 5	4 vs *
Round 4:	1 vs 3	* vs 2	5 vs 4
Round 5:	4 vs 1	3 vs 2	5 vs *

The procedure is as follows: in the first round, the lower-numbered object is at home (* is highest); after that, alternate home and away games wherever possible, but ensure that object * always alternates, even at the expense of one of the other objects.

We see that objects 3 and * alternate home and away games throughout, while object 1 has two consecutive away games in rounds 2 and 3, object 2 has two away in rounds 4 and 5, object 4 has two home games in rounds 2 and 3, and object 5 has two in rounds 4 and 5.

Furthermore, precisely one of objects 1 and 4 is at home in each round, and the same holds true for 2 and 5, and 3 and *.

These results are the best possible, as de Werra proves, but we will see in Section 2.3.3 that there exist designs different from GK_t for which these properties also hold.

A double round-robin is usually constructed from a single round-robin by repeating the $t-1$ rounds in the same order, but with allocation of home and away reversed. Unfortunately, using the allocation of home and away games described above, an object can then have *three* consecutive home or away games, as we see below with $t = 6$:

Round 4	* vs **2**	1 vs 3	*5* vs 4
Round 5	3 vs **2**	4 vs 1	*5* vs *
Round 6	4 vs **2**	* vs 3	*5* vs 1

where Round 6 is the same as Round 1, but with home and away allocations reversed; here, object 2 plays three consecutive away games and object 5 plays three consecutive home games. This problem is addressed in practice in a number of ways: some tournaments schedule a break at the halfway point, so that the presence of the consecutive games is "forgotten"; in some cases, * plays two consecutive home or away games earlier in the tournament in such a way that the last two rounds of the first half of the tournament have each team at home once and away once, thus avoiding the problem, and in some other cases, the order of rounds in the second half of the tournament is rearranged – for example, playing the rounds of the second half in reverse order ensures that the last round of the first half and the first round of the second half feature the same games, but with home and away reversed.

Unfortunately, the design GK_t is very unbalanced for carry-over effects. De Werra (1982) proves the following result:

Theorem 2.1 *In GK_t, objects $1, 2, \ldots, t-1$ each receive $t-3$ carry-over effects from one other object. This result assumes that carry-overs from round $t-1$ to round 1 are also*

counted.

2.3.2 The general cyclic design

A more general design for round-robin tournaments is the **general cyclic design**. Given a **strong starter**, which we define below, each round of the design is determined inductively from the previous round. It turns out that GK_t is, in a way, a special case of the general cyclic design.

Construction

As before, we label the t objects $1, 2, \ldots, t-1, *$. A **strong starter** (Dinitz and Stinson, 1981) from a set $S = \{1, 2, \ldots, t-1\}$ is defined to be a partitioning of the set into pairs $(a_i, b_i), a_i, b_i \in S, i = 1, 2, \ldots, t/2 - 1$ such that

1. Each element of S appears in no more than one pair.
2. The differences $a_i - b_i$ *and* the differences $b_i - a_i$, $i = 1, 2, \ldots, t/2 - 1$ are all distinct modulo $t - 1$.

Since there are $t-1$ elements of S and only $t/2-1$ pairs, one element c of S does not appear in any pair.

A strong starter always exists for even t: for example, take $a_i = i$ and $b_i = t - i$ for $i = 1, 2, \ldots, t/2 - 1$, with object $t/2$ being left over. Generally, there will be many other strong starters for given t; if a "random" strong starter is desired, the algorithm of Dinitz and Stinson (1981) can be used.

The first round of a general cyclic design for n objects is then determined by finding a strong starter from the set $S = \{1, 2, \ldots, t-1\}$. The pairings for the first round are the pairs of the strong starter, together with the pair $(c, *)$ where c is the element of S that does not appear in the strong starter.

The remaining rounds are determined by induction. If, in round k, objects i and j are paired, with $i, j \neq *$, objects $i+1$ and $j+1$ are paired in round $k+1$, with arithmetic being, as before, modulo $t-1$ with the result being taken from the set $\{1, 2, \ldots, t-1\}$. Round k contains one other pairing, that of * and some other object i; in round $k+1$, * and $i+1$ are paired.

As an example, suppose that $t = 6$ and we choose the strong starter $(1,2),(3,5)$. This is indeed a strong starter since, modulo 5, the differences $1-2 = 4, 3-5 = 3, 2-1 = 1, 5-3 = 2$ are all distinct. The first round is completed by pairing 4 and *, and the entire design follows:

Round 1:	1 vs 2	3 vs 5	4 vs *
Round 2:	2 vs 3	4 vs 1	5 vs *
Round 3:	3 vs 4	5 vs 2	1 vs *
Round 4:	4 vs 5	1 vs 3	2 vs *
Round 5:	5 vs 1	2 vs 4	3 vs *

It is evident that a sixth round generated in the same way would be identical to Round 1; this explains the name "cyclic".

The structure of the design is apparent with the example laid out in this form. The first column of pairings contains all the pairings i vs j for which $i - j \equiv 1$ or $4 \pmod{t-1}$, the second all those for which the difference is congruent to 2 or $3 \pmod{t-1}$, and the final column contains all the pairings involving *.

As we did with GK_t, we need to prove that this construction does indeed yield a design in which each object appears precisely once in each round, and that each possible pairing occurs in precisely one of the rounds.

Note first that there are $t/2 - 1$ pairs in the strong starter, and so $2(t/2-1) = t-2$ differences that must be distinct. But, working modulo $t-1$, there are only $t-2$ possible non-zero differences, so, given a difference d, there must exist some pair (a_i, b_i) in the starter such that either $a_i - b_i \equiv d \pmod{t-1}$ or $b_i - a_i \equiv d \pmod{t-1}$.

Now, list the design as above for $t = 6$, with each pairing listed below the pairing it was generated from. Each object appears exactly once in the first round, because that round is based on the strong starter. Now suppose that each object appears exactly once in some round k; then, in round $k+1$, object * will be listed below itself in round k, while any other object i will be listed below $i-1$, with object 1 listed below object $t-1$. Thus each object also appears exactly once in round $k+1$.

Finally, note that all pairings of * with any other object are found in the last column. For any other pairing i vs j, let $d = i - j \pmod{t-1}$, and find the pair of the strong starter, (k,l) say, for which $k - l \equiv d \pmod{t-1}$ or $l - k \equiv d \pmod{t-1}$. Such a pair must exist, as we observed above. If $k - l \equiv d \pmod{t-1}$, i and j are paired in round $r = i - k \pmod{t-1} + 1$; otherwise, they are paired in round $s = j - l \pmod{t-1} + 1$.

Since $1 \leq r, s \leq t-1$, each possible pairing i vs j must appear in exactly one of rounds $1, 2, \ldots, t-1$.

Relation to GK_t

Theorem 2.2 *The design GK_t is a special case of the general cyclic design, although with the rounds in a different order. In particular, the first round of GK_t, with the pairing containing * removed, is a strong starter, and rounds $1, 2, \ldots, t-1$ of the general cyclic design are rounds $1, 3, \ldots, t-1, 2, 4, \ldots, t-2$ of GK_t.*

Proof: Except for the pairing $t/2$ vs *, the other pairings in the first round of GK_t are of the form i vs. $t-i$, $i = 1, 2, \ldots, t/2-1$, with, by this definition, $i < t-i$. $i < t-i$. The differences $d_i = (t-i) - i = t - 2i$ are always positive and also less than t, so that d_i is already reduced modulo $t-1$; this implies that these d_i are all distinct and all even (since t and $2i$ are both even). Now let $d_i = i - (t-i) = 2i - t$; these are all negative but larger than $-t$. They can therefore be "reduced" modulo $t-1$ by adding $t-1$, so that $d_i = 2i - t + t - 1 = 2i - 1$; these are all distinct and odd. Thus the first round of GK_t, with $t/2$ vs * removed, is indeed a strong starter.

In the first round of both designs, the objects paired, except for *, sum to 1 or t. Thus, for any objects i and j that are paired in the first round, $i + j \equiv 1 \pmod{t-1}$. In round k of the general cyclic design, $i+k-1$ and $j+k-1$ are paired (both numbers being reduced modulo $t-1$ if necessary). In this round, therefore, $(i+k-1)+(j+k-1) = i+j+2k-2 = 2k-1$, meaning that objects meeting in round k of the general cyclic design meet in round $2k-1$, or $2k-1-(t-1) = 2k-t$ (reducing modulo $t-1$), of GK_t, and this proves the result.

Of course, by choosing a strong starter that does not correspond to a round of GK_t, we can obtain a design that is very different from GK_t.

Home and away games

There do not appear to be any useful results concerning allocation of home and away games in a general cyclic design. Even those general cyclic designs that correspond to GK_t cannot benefit from the results discussed in Section 2.3.1 on this subject, because the order of the rounds is rearranged.

The strongest results available seem to be those of de Werra (1980) and Wallis (1983), and these results apply to *any* round-robin design in which the pairings are arranged with

$t/2$ of them in each of $t-1$ rounds. The results are as follows:

Theorem 2.3 *Arrange, in any fashion, $t-2$ of the $t-1$ rounds into disjoint pairs of rounds. Then, within each pair of rounds, each object can be allocated one home game and one away game.*

The (constructive) proof depends on the fact that a pair of rounds can be represented as a graph which is the union of two 1-factors, and so cannot have cycles of odd length. This permits the allocation to be made.

In practice, the result is most useful if the pairs of rounds are adjacent. For example, the pairs might be of rounds $(1,2),(3,4),\ldots,(t-3,t-2)$. In this case, in the first k rounds (k even), each object has exactly $k/2$ home games and $k/2$ away games.

Theorem 2.4 *For any grouping of the objects into pairs, an allocation of home and away games can be found in which one of the objects in each pair plays at home and the other plays away.*

This is a general version of the result we had for GK_t. However, in general, the theorems of this section do not combine usefully: one can have pairs of rounds in which each object plays once at home and once away, or pairs of objects in which exactly one plays at home in each round, but not both in general. As an example, consider the following general cyclic design with $t = 10$:

Round 1:	1 vs 2	3 vs 5	4 vs 8	6 vs 9	7 vs 10
Round 2:	2 vs 3	4 vs 6	5 vs 9	7 vs 1	8 vs 10
Round 3:	3 vs 4	5 vs 7	6 vs 1	8 vs 2	9 vs 10
Round 4:	4 vs 5	6 vs 8	7 vs 2	9 vs 3	1 vs 10
Round 5:	5 vs 6	7 vs 9	8 vs 3	1 vs 4	2 vs 10
Round 6:	6 vs 7	8 vs 1	9 vs 4	2 vs 5	3 vs 10
Round 7:	7 vs 8	9 vs 2	1 vs 5	3 vs 6	4 vs 10
Round 8:	8 vs 9	1 vs 3	2 vs 6	4 vs 7	5 vs 10
Round 9:	9 vs 1	2 vs 4	3 vs 7	5 vs 8	6 vs 10

We pair the rounds (1 and 2, 3 and 4, etc.), and arrange the comparisons so that each object has one comparison at home and the other away in each pair of rounds, as in

Theorem 2.3. We find the following, where the first five objects listed for each round pair are at home in the first round and away in the second, or vice versa:

Round pair	Home in one	Home in other
1 and 2	1 3 4 9 10	2 5 6 7 8
3 and 4	1 4 7 8 9	2 3 5 6 10
5 and 6	1 2 3 6 9	4 5 7 8 10
7 and 8	1 6 7 9 10	2 3 4 5 8

The home and away games for round 9 can be assigned in any fashion. Consider object 2: during the course of the round-robin, this object is at home and away in the same round as every other object (for example, in rounds 5 and 6, but at no other time, object 2 plays at home at the same time as object 9). In other words, it is impossible to find, in the manner of Theorem 2.4, a pairing of the objects such that exactly one of each pair is at home in each round, without destroying the property constructed from Theorem 2.3.

In this example, a partial pairing of the teams is available, however. If object 3 is paired with object 7, object 4 with object 6 and object 5 with either 1 or 9, exactly one object of each pair is at home in each round. Note also that the nature of this example is not changed if we pair rounds 2 and 3, 4 and 5, etc.; object 3 is then left without a pair, and we obtain a partial pairing similar to the above, but with the objects relabelled (specifically, the partial pairing is 4 and 8, 5 and 7, 6 and 1 or 2). In practice, the existence of a partial pairing of the teams may be sufficient (for example, in a sporting context, if only some of the teams in a league are geographically close to another team). We do not have general results concerning the nature of partial pairings of the teams, although we suspect that the size of the partial pairing will generally become small relative to t as t increases, because each object is at home at the same time as a greater number of other objects for larger t. For example, with $t = 8$, there are three round pairs and four comparisons in each round, so that each team is at home at the same time as $3 \times 3 = 9$ other (non-distinct) objects, whereas with $t = 10$, each object is at home at the same time as $4 \times 4 = 16$ other objects. It is more likely, in the absence of any special structure, that the 16 objects will exhaust the nine potential team-pairs available when $t = 10$, as compared to the situation when $t = 8$, where there are seven potential team-pairs and only 9 objects at home at the same time.

Carry-over effects

In contrast to the paucity of useful results in the previous section, the carry-over effect situation is a good deal brighter for the general cyclic design than for GK_t.

Russell (1980) shows the following:

Theorem 2.5 *If $t = 2^k$ for some k, there exists a round-robin design which is completely balanced for carry-over effects.*

The converse of Russell's result does not hold: we will show that there is a general cyclic design for $t = 20$ that is completely balanced for carry-over effects. A more useful result in practice is that for any (even) t, there are designs which are "approximately" balanced for carry-over effects.

In order to investigate carry-over patterns in general cyclic designs, the following results are useful. We define the **carry-over pattern** for any object to be a list of the frequencies of carry-over effects received from the other objects, without regard to the identities of the other objects. For example, if object 4 receives 5 carry-over effects from object 3 and one each from object 2 and object *, the carry-over pattern is $[5 \times 1, 1 \times 2]$, read as "five carry-over effects from one object and one each from two other objects". A carry-over pattern of $[1 \times (t-1)]$ denotes a balanced set of carry-over effects for an object: one each received from the $t-1$ other objects.

Theorem 2.6 *In the general cyclic design, object * always receives one carry-over effect from each other object, and for objects $1, 2, \ldots, t-1$, the carry-over pattern is the same.*

Proof: Suppose that objects i and * are paired in round k, and that i is paired with j in round $k-1$. Then, in round k, * receives a carry-over effect from j. But $i+1$ and * are paired in round $k+1$, and $i+1$ and $j+1$ in round k, so that in round $k+1$, * receives a carry-over effect from $j+1$. (In all the arithmetic here, the result is reduced modulo $t-1$ if necessary.) In general, in round $k+d$, * receives a carry-over effect from $j+d$, and it is seen that object * must therefore receive exactly one carry-over effect from each other object.

Now suppose that object m receives a carry-over effect from object $j \leq t-1$ in round k. This means that some object i must have been paired with object j in round $k-1$ and object m in round k. Now, from the construction, it must also be true that object $i+1$ is paired with object $j+1$ in round k, and object $i+1$ is paired with object $m+1$ in round

$k+1$. Hence object $m+1$ receives a carry-over effect from object $j+1$ in round $k+1$. The same applies if $j = *$: if object m receives a carry-over effect from object t in round k, object $m+1$ receives a carry-over effect from object t in round $k+1$. Thus, for example, if object m has a carry-over pattern of $[3 \times 1, 1 \times (t-4)]$, the same will be true for all objects $1, 2, \ldots, t-1$.

As a result of Theorem 2.6, it makes sense to talk about "*the* carry-over pattern" of a general cyclic design, where the carry-over pattern applies to all the objects $1, 2, \ldots, t-1$, and the carry-over pattern for object * is always $[1 \times (t-1)]$.

Theorem 2.7 *The carry-over pattern for any object in a general cyclic design can be determined from the first- and second-round pairings.*

Proof: Consider an object $i \leq t-1$ whose first two opponents are not *, and call these opponents objects j and $j+d$, with difference $d \pmod{t-1}$ between their numbers. Then, starting in round k, object $i+k-1$ plays object $j+k-1$ followed by $j+d+k-1$, and the numbers of these objects also differ by d. If $j+d+k-1 = 1$, object 1 receives a carry-over effect from object $j+k-1 = 1-d = t-d \pmod{t-1}$. Thus, object 1 receives m carry-over effects from object $t-d$ if and only if there are m objects whose first two opponents differ in number by d, and are not *. In addition, object 1 receives one carry-over effect from *, and one from object $t-1$ (since objects $t-1$ and 1 are successive opponents of object t). The carry-over effect from object $t-1$ must be the only one, since if $d = 1$, the pairing of objects i and j first round implies the pairing of both i and $j+1$ (since $d = 1$) *and* $i+1$ and $j+1$ (by construction), and this is impossible.

Since the theorem has been proved for object 1, Theorem 2.6 implies that the result holds for objects $1, 2, \ldots, t-1$.

Since the second round is determined from the first, Theorem 2.7 can be extended:

Theorem 2.8 *In a general cyclic design, the carry-over pattern for all objects can be determined from the pairings for the first round.*

Proof: To prove this, let $\mathbf{v}$ be a $(t-1)$-vector with v_i being the number of the first-round opponent of object i. Then object i's first two opponents' numbers differ by d (and neither opponent is *) if and only if $v_{i-1} - v_i = d-1$ (following our usual arithmetic conventions, $v_{i-1} = v_{t-1}$ if $i = 1$). Consequently, object 1 receives m carry-over effects from object d if

and only if there are m i's, $1 \leq i \leq t-1$, such that $v_{i-1} - v_i \equiv d-1 \pmod{t-1}$, and $v_{i-1}, v_i \neq t$.

As an example, we investigate two different admissible first-round pairings for $t = 8$ objects. It is helpful to define $e_i = v_{i-1} - v_i$, provided $v_{i-1}, v_i \neq t$ (the value of e_i not being needed otherwise). Consistently with our arithmetic being modulo $t-1$, we define $e_1 = v_{t-1} - v_1$. We also define the $(t-1)$-vector n such that n_j is the number of e_i that equal j.

Consider the first-round pairing 1 vs 2, 3 vs 7, 4 vs 6, 5 vs 8. The 7-vector $\mathbf{v}$ is $(2,1,7,6,8,4,3)$. Then $e_1 = e_2 = e_3 = e_4 = e_7 = 1$, with e_5 and e_6 undefined. So object 1 receives 5 carry-over effects from object $t-1-1 = 6$, and, as mentioned before, one each from objects 8 and 7. Therefore the carry-over pattern for this design is $[5 \times 1, 1 \times 2]$, and $n = (5,0,0,0,0,0,0)$.

Compare now the first-round pairing 1 vs 2, 3 vs 5, 4 vs 7, 6 vs 8: the vector $\mathbf{v}$ is now $(2,1,5,7,3,8,4)$, and so $e_1 = 2, e_2 = 1, e_3 = 3, e_4 = 5, e_5 = 4$, and e_6 and e_7 are undefined. Since the e_i are all distinct, object 1 receives (and hence all objects receive) exactly one carry-over effect from each other object. This is a design completely balanced for carry-over effects, and clearly the carry-over pattern for the design is $[1 \times 7]$, with $n = (1,1,1,1,1,0,0)$.

In order to compare these carry-over patterns with those from other designs (in which the carry-over patterns may not be the same for all objects), we define a matrix M for any design with element m_{ij} being the number of carry-over effects received by object j from object i. We have seen that a design that is balanced for carry-over effects will have $m_{ij} = 1$ for $i \neq j$, whereas an unbalanced design will exhibit greater variability in the m_{ij}; it is thus natural to use the variance of the elements of M, or equivalently the quantity $S = \sum_{i,j} m_{ij}^2 = MM'$, the sum of squares of the elements of M, as a measure of a design's balance or otherwise for carry-over effects. For general cyclic designs, the last column of M contains $t-1$ ones, while the remaining $t-1$ columns each contain the same set of numbers (permuted), so that each column's sum of squares is related to the sum of squares of n. Specifically, noting that object t receives one carry-over effect from each of the $t-1$ other objects, and that, for $1 \leq i \leq t-1$, object i, receives an additional carry-over effect from each of two other teams aside from the ones summarized in n, we find that S can be written as

$$S = (t-1)\left(\sum_{i=1}^{t-1} n_i^2 + 3\right).$$

If we couple Theorem 2.8 with a method for generating all possible strong starters for any given t, we can find the general cyclic design which comes closest to complete balance, simply by an exhaustive search. Such a simple-minded approach is in fact practical, at least for $t, \leq 20$ or so. Compared with Russell (1980), this approach appears more capable of finding designs that achieve or approach balance.

An alternative approach is to use a modification of an algorithm given by Dinitz and Stinson (1981) that generates "random" strong starters; by repeatedly generating a strong starter in this way, and using the above results to calculate S for the general cyclic design that uses the strong starter as its first round, we may be able to find balanced or nearly-balanced designs with less effort than is required for a complete enumeration. Of course, by doing this, we cannot be sure that the most balanced design has been found unless some external check is available. Dinitz and Stinson derive their algorithm in a context that is more general than we require; we found that the following simplified algorithm works quickly and successfully:

1. Let the set of objects (for this algorithm) be $T = \{1, 2, \ldots, t-1\}$ and the set of "differences" be $D = \{1, 2, \ldots, t/2 - 1\}$.

2. Mark all objects and differences as "unused".

3. Choose an unused object k and an unused difference d, both choices being made at random with equal probability from the available possibilities.

4. Let $r_1 = k - d$ and $r_2 = k + d$, in both cases reducing the answer modulo $t - 1$ to the set T.

5. If objects r_1 and r_2 are both used or both unused, choose one of them at random, and denote that object r. Otherwise, let r be the unused object out of r_1 and r_2.

6. If object r is already used, there exists an object r_3 and a difference d_3 such that the comparison r vs. r_3, corresponding to the difference d_3, is currently a part of the strong starter. Remove r vs. r_3 from the strong starter, and mark object r_3 and difference d_3 unused.

7. Mark objects k and r and the difference d used, and add k vs. r to the strong starter, corresponding to the difference d.

8. If there are any unused differences, go back to step 3. Otherwise stop.

Table 2.1 shows the minimal values of S found, along with an example first round, for general cyclic designs of sizes up to $t = 32$. The values for $t \leq 20$ were found by enumeration, and are thus the smallest possible for general cyclic designs; the remaining values were found using our version of the Dinitz-Stinson algorithm, as given above, with 10,000 random strong starters generated for each t. Also shown are the values of S obtained by Russell (1980) for the designs he studied. Our result for $t = 32$ is surprising, since a design is known to exist that is completely balanced for carry-over effects. We do not know whether this design cannot be expressed as a general cyclic design, or whether it can but we were unable to find it.

The algorithm given above for random strong starters was also able to reproduce the values of S obtained in Table 2.1 for $t \leq 20$ with a smaller computational effort than was required to find the optimal values of S by enumeration. Though we generated 10,000 random strong starters for each value of t, a much smaller number would have been sufficient for smaller values of t. This is not surprising, considering that the number of different general cyclic designs increases rapidly with t; we would expect to require a more extensive search to locate a design with minimal S when t is larger. Nonetheless, for larger values of t, for which an exhaustive search is not feasible, this method provides a means for generating designs with near-minimal S in a reasonably short time.

2.3.3 Random round-robin designs

In this Section we consider two algorithms which generate round-robin designs that do not necessarily follow any particular pattern. As remarked at the beginning of this Chapter, it is not a trivial matter to generate such designs, and so a certain amount of care may be needed.

A construction

As well as giving an algorithm for generating strong starters, Dinitz and Stinson (1987) also give an algorithm for generating random round-robins, in which the pairings in one round are not necessarily determined by the pairings in another. The algorithm consists of two "heuristics" H_1 and H_2, which are used as described below. Object i is called **live** if there exists another object j such that the pairing (i, j) does not appear in any round; similarly,

	Minimum values of S			
t	Balanced	This thesis	Russell	Example first-round pairing
4	12	12	12	1 vs 2, 3 vs 4
6	30	60	60	1 vs 2, 3 vs 5, 4 vs 6
8	56	56	56	1 vs 2, 3 vs 5, 4 vs 7, 6 vs 8
10	90	108	138	1 vs 2, 3 vs 5, 4 vs 8, 6 vs 9, 7 vs 10
12	132	176	196	1 vs 2, 3 vs 5, 4 vs 10, 6 vs 9, 7 vs 11, 8 vs 12
14	182	234	260	1 vs 2, 3 vs 5, 4 vs 9, 6 vs 13, 7 vs 10, 8 vs 12, 11 vs 14
16	240	240	240	1 vs 2, 3 vs 8, 4 vs 6, 5 vs 11, 7 vs 15, 9 vs 12, 10 vs 14, 13 vs 16
18	306	340	428	1 vs 2, 3 vs 7, 4 vs 11, 5 vs 14, 6 vs 9, 8 vs 18, 10 vs 16, 12 vs 17, 13 vs 15
20	380	380	520	1 vs 2, 3 vs 9, 4 vs 12, 5 vs 7, 6 vs 11, 8 vs 15, 10 vs 19, 13 vs 16, 14 vs 18, 17 vs 20
22	462	546		1 vs 2, 3 vs 5, 4 vs 10, 6 vs 13, 7 vs 15, 8 vs 12, 9 vs 18, 11 vs 21, 14 vs 19, 16 vs 22, 17 vs 20
24	552	644		1 vs 7, 2 vs 21, 3 vs 15, 4 vs 20, 5 vs 24, 6 vs 14, 8 vs 10, 9 vs 18, 11 vs 12, 13 vs 23, 16 vs 19, 17 vs 22
26	650	750		1 vs 13, 2 vs 25, 3 vs 22, 4 vs 19, 5 vs 23, 6 vs 15, 7 vs 21, 8 vs 9, 10 vs 18, 11 vs 16, 12 vs 26, 14 vs 17, 20 vs 24
28	756	918		1 vs 10, 2 vs 24, 3 vs 5, 4 vs 14, 6 vs 25, 7 vs 21, 8 vs 9, 11 vs 22, 12 vs 27, 13 vs 28, 15 vs 18, 16 vs 20, 17 vs 23, 19 vs 26
30	870	1102		1 vs 23, 2 vs 19, 3 vs 8, 4 vs 5, 6 vs 17, 7 vs 15, 9 vs 18, 10 vs 20, 11 vs 30, 12 vs 16, 13 vs 26, 14 vs 29, 21 vs 24, 22 vs 28, 25 vs 27
32	992	1240		1 vs 27, 2 vs 15, 3 vs 12, 4 vs 14, 5 vs 6, 7 vs 30, 8 vs 20, 9 vs 25, 10 vs 16, 11 vs 28, 13 vs 17, 18 vs 29, 19 vs 22, 21 vs 23, 24 vs 31, 26 vs 32

Table 2.1: Values of S for various t

a round r is live if fewer than $t/2$ pairings appear in round r. At the beginning, therefore, all objects and rounds are live, while at the end, there are no live objects or rounds. All choices in the algorithm and heuristics are made at random with equal probability from the available choices.

The algorithm is simply:

1. While there exists a live object:
   (a) Choose either H_1 or H_2, and carry out the chosen heuristic.
   (b) End.
2. End.

Heuristic H_1 is defined as follows:

1. Choose a live object i.
2. Choose an object j for which the pairing (i, j) has not been assigned to a round.
3. Choose a round r in which object i does not appear.
4. Add the pairing (i, j) to round r.
5. If j is paired with another object k in round r, remove the pairing (j, k) from round r.
6. End.

Heuristic H_2 is:

1. Choose a live round r.
2. Choose two objects i and j that do not appear in round r.
3. Add the pairing (i, j) to round r.
4. If the pairing (i, j) appears in any other round s, remove it from round s.
5. End.

Table 2.2: Example run of Dinitz-Stinson round-robin algorithm for $t = 4$

Step	Heuristic	i	j	r	Round 1	Round 2	Round 3
1	2	2	1	1	1-2		
2	2	1	4	3	1-2		1-4
3	1	3	1	3	1-2		1-3
4	1	3	4	2	1-2	3-4	1-3
5	2	4	2	3	1-2	3-4	1-3, 2-4
6	1	1	4	2	1-2	1-4	1-3, 2-4
7	1	4	3	1	1-2, 3-4	1-4	1-3, 2-4
8	2	2	3	2	1-2, 3-4	1-4, 2-3	1-3, 2-4

The algorithm is "downhill" in the sense that the number of pairings currently assigned to rounds never decreases as the algorithm proceeds. There are only three possibilities at each step: a new pairing is added, one pairing is replaced by another, or a pairing is moved from one round to another. This means that, unless the algorithm reaches a point at which there is a live object but no new pairings can be added (no matter how many moves or replacements are made), it must eventually produce a design. In fact, the algorithm has never been known to fail, even though a proof of its certain success has not been found.

As an example, we show a run of the algorithm for $t = 4$. The details are given in Table 2.2, where it is seen that the algorithm required eight steps to complete a design with six games. (In fact, $t = 4$ is generally easy for the algorithm; our experience indicates that the ratio of steps to total games increases with t.) In this run, the algorithm makes what seems a pointless change at step 3, replacing 1 vs. 4 with 1 vs. 3. In general, however, the ability of the algorithm to make this kind of change can prevent it from becoming blocked. In step 4, the addition of 3 vs. 4 to round 2, instead of round 1 where it "belongs", seems to be a mistake, but in step 6, everything is sorted out: team 1 has one remaining game, against team 4, and team 1 has a game in every round but round 2, so that 1 vs. 4 must be placed in round 2, replacing the game involving team 4 that was already there. From this point, the algorithm has no trouble placing the two remaining games to complete the design.

Home and away games

The results of Section 2.3.2 apply here also. With the random nature of the designs produced by the algorithm, it is impossible to make more general statements. The same is true concerning carry-over effects; it is possible that there exist random designs that are more balanced for carry-over effects than any general cyclic design. The situation with home and away games can be improved, however, by use of the algorithm given below.

One reason for the popularity of the design GK_t is, as was discussed, the ease with which home and away games can be assigned to ensure that objects alternate home and away games as far as possible, with, at the same time, pairs of objects never both being at home in the same round. It is possible, however, to generate designs different from GK_t which nonetheless still have the same pattern of home and away games for each object.

Consider, for example, the pattern of home and away games generated from GK_8. An "H" in row i and column j in the table below indicates that object i plays at home in round j; an "A" in that position indicates an away game.

Object	1	2	3	4	5	6	7
1	H	A	A	H	A	H	A
2	H	A	H	A	A	H	A
3	H	A	H	A	H	A	A
4	H	A	H	A	H	A	H
5	A	H	H	A	H	A	H
6	A	H	A	H	H	A	H
7	A	H	A	H	A	H	H
*	A	H	A	H	A	H	A

In this table, we see that one of objects 1 and 5 plays at home in each round, while the other plays away. The same applies for objects 2 and 6, 3 and 7, 4 and *. A pair of objects can be paired in a round only if one of them is at home in that round and the other is away. This means, for example, that there is only one round, namely round 7, in which 3 and 4 can be paired. The same applies for objects 7 and *. In contrast, objects 1 and 5 can be paired in any round. An algorithm seeking to be successful in assigning pairings to rounds should, intuitively, deal first with those pairings for which there is a small number of "live" rounds. The following simple procedure seems to work, at least for the home and away pattern generated from GK_t:

1. List the pairings which either:
   (a) can only appear in one round
   (b) can be added to a round that currently contains $t/2-1$ pairings (ie. all but one).
2. If this list is empty, instead list the pairings which either:
   (a) can only appear in two rounds
   (b) can be added to a round that currently contains $t/2-2$ pairings.
3. Choose a pairing at random (with equal probability) from the list.
4. Choose a round at random for the pairing to be assigned to, and assign it. (There may, of course, be only one possible round.)
5. If there still exists a live round, go back to 1. Otherwise end.

Since the carry-over effects in GK_t are so unbalanced, this method provides the possibility of balancing out the carry-over effects somewhat while still maintaining a desirable pattern of home and away games for each object.

A proof that this method will always work is as elusive as one for Dinitz and Stinson's original method, and for the same reason. Our limited experience suggests that when a round-robin schedule can be generated at all, the algorithm will probably find one, though not necessarily quickly; the practical solution is to stop it and try again if it seems to be taking too long.

2.4 Tests of overall equality and multiple comparisons

2.4.1 Introduction

Suppose that t objects are compared in a possibly replicated round-robin tournament. Later, we wish to investigate order effects and ties, so that in this Section a single round-robin consists of all $t(t-1)/2$ pairs of objects being compared *twice*, once in each order. Thus the number r of replications is half the n of David (1988). In the presence of ties, we also consider the "score" a_i for each object to include a half-point for each tie as well as a point for each time the object was preferred in a comparison.

2.4.2 Overall test of equality

We will usually first wish to test whether the objects are all equally preferable. To do this, we assume that each comparison is independent, and that (in the most general case) our null hypothesis states that the probabilities of the first object being preferred, the second object being preferred or a tie being declared do not depend on which objects are being compared (they are the same for all comparisons). Natural statistics for a test of equality are measures of the spread of the a_i, such as the variance or the range. It turns out that the variance of the a_i, or equivalently $\sum_{i=1}^{t} a_i^2$, is also the score test for overall equality if and only if the Bradley-Terry model holds (for which see Chapter 3 and Bühlmann and Huber, 1963). We consider this variance test first.

David (1988) shows that, in the absence of order effects and ties, the suitably scaled sum of squared scores has an asymptotic χ^2_{t-1} distribution (for large r), and provides tables for small experiments. Starks (1958) shows that use of the asymptotic distribution is accurate for moderately-sized experiments. Gillot and Caussinus (1966) derive some exact results for the joint distribution of the a_i in the presence of ties, but do not consider order effects. When order effects or ties are present, one would expect the variability in scores to be smaller, and so use of David's results in such situations would lead to (possibly very) conservative tests. We show that this is indeed the case.

We begin by considering what happens in the absence of ties, but with order effects. Under the null hypothesis that all objects are equally preferable, the only factor affecting the preference probabilities is the order effect; specifically, H_0 assumes that for all pairs of objects, the probability that the one presented first is preferred is p, where $p = 0.5$ in David's results.

Under this hypothesis, the total score for any object i is the sum of two binomial random variables: the number of preferences in the $r(t-1)$ cases where object i is presented first, and the number in the $r(t-1)$ cases where object i is presented second. Thus, for each i,

$$\begin{aligned} E(a_i) &= r(t-1)\{p + (1-p)\} = r(t-1) \\ \text{var}(a_i) &= 2r(t-1)p(1-p) = \sigma^2, \text{ say.} \end{aligned}$$

Defining

$$d_i = \frac{a_i - r(t-1)}{\sqrt{2rtp(1-p)}},$$

it follows that $E(d_i) = 0$, $\text{var}(d_i) = (t-1)/t$. The d_i are clearly correlated, and, under

H_0, equally correlated, but we can find the correlation by the same method as David: $\sum_{i=1}^{t} a_i = rt(t-1)$, since each of the $rt(t-1)$ comparisons yield a total of one point, and so $\sum_{i=1}^{t} d_i = 0$. Since these sums are fixed, their variances are zero, and so, letting ρ be the common correlation, we have

$$\begin{aligned} 0 = \mathrm{var}\left(\sum_i d_i\right) &= \sum_i \mathrm{var}(d_i) + \sum_{i \neq j} \mathrm{cov}(d_i, d_j) \\ &= \frac{t(t-1)}{t} + \frac{t(t-1)(t-1)}{t}\rho, \end{aligned}$$

and thus $\rho = -1/(t-1)$ and the common covariance is $-1/t$.

Thinking of $\mathbf{d} = \{d_i\}$ as a random vector, the covariance matrix of $\mathbf{d}$ has diagonal entries $(t-1)/t$ and off-diagonal entries $-1/t$ in a $t \times t$ matrix. Such a matrix has one zero eigenvalue and $t-1$ unit eigenvalues; as a result, $\sum_{i=1}^{t} d_i^2 = \mathbf{d}'\mathbf{d}$ has an asymptotic χ^2 distribution with $t-1$ d. f., just as when order effects are absent. Writing

$$\sum_{i=1}^{t} d_i^2 = \frac{\sum_{i=1}^{t}\{a_i - r(t-1)\}^2}{2rtp(1-p)},$$

we see that the tendency for values of p far from $\frac{1}{2}$ to decrease the variability in the a_i is balanced by the presence of a $p(1-p)$ term in the denominator of the test statistic.

When there are ties as well, the score for any single object is no longer binomial, but the same sort of argument goes through. Let p_1 denote the probability of the first object being preferred in any paired comparison (since the objects are equally preferable under H_0, the probability is the same for all comparisons), let p_2 be the probability of the second object being preferred, and let p_0 denote the probability of a tie. Counting one point when object i is preferred and half a point for a tie, the expected number of points for object i in a single comparison is $p_1 + p_0/2$ when object i is presented first, and $p_2 + p_0/2$ when object i is presented second. The variance of the number of points in a single comparison is $p_1(1-p_1) + p_0(1-p_0)/4 - p_0p_1$ in the first case and $p_2(1-p_2) + p_0(1-p_0)/4 - p_0p_2$ in the second. After a little algebra, we find that

$$\begin{aligned} E(a_i) &= r(t-1), \\ \mathrm{var}(a_i) &= r(t-1)\{p_1(1-p_1) + p_2(1-p_2) - p_0(1-p_0)/2\}. \end{aligned}$$

Defining, then,

$$d_i = \frac{a_i - r(t-1)}{\sqrt{rt\{p_1(1-p_1) + p_2(1-p_2) - p_0(1-p_0)/2\}}}, \tag{2.1}$$

enables us to show in exactly the same way as before that $\sum_{i=1}^{t} d_i^2$ has an asymptotic χ^2_{t-1} distribution.

We do not present any theory concerning the range of the a_i or d_i; David (1988, p. 35) indicates, in the absence of ties and order effects, that an approximation based on the distribution of the range of normal random variables will work well.

It is also possible to simulate the behaviour of our statistics, for comparison with these asymptotic results and David's small-sample tables. For each combination of values t, r, p_1, p_2, p_0, we simulated 10,000 values of the sum of squares and of the range. Table 2.3 shows the results of our simulations for $\sum_i a_i^2$ and Table 2.4 shows the corresponding results for the range of the a_i. It is clear that the effect both of increasing the order effect and increasing the tie probability is to decrease the critical values of the test statistics, in some cases considerably. The agreement between the simulations and the exact results of David, where they overlap, is good (the case $p_1 = p_2 = 0.5$ in Tables 2.3 and 2.4, and Tables 1 and 3 of David, for the sum of squares and the range respectively). Further (see Table 2.3), the agreement between the simulated distribution of the sum of squares and the critical values obtained from the chi-squared asymptotic distribution is extremely good — even for $t = 3, r = 1$, where $\sum_i a_i^2 \leq 20$ and some percentage points do not exist, the ones that do exist are very well approximated using the asymptotic distribution. For the range (Table 2.4), the general picture is that the points obtained from the distribution of the range are somewhat anti-conservative; for moderate sample sizes, adding 1 to the approximated critical values seems to improve matters somewhat.

One could also consider using the range of the scores as an overall test of equality. It seems, however, that the test based on the variance of the scores will be more powerful because it uses *all* the scores, not just the most extreme ones. The range of scores, however, is a natural candidate for use in multiple comparisons, as we discuss below.

The above results consider the distribution of the sum of squared scores and the range of scores conditional on the probability of the first object being preferred and on the probability of a tie. In practice, though, these probabilities will not be known, and will have to be estimated by the observed proportions of first-object preferences and of ties. The effect of this estimation procedure on the true level of the test is not clear.

2.4.3 Multiple comparisons

As in analysis of variance, having shown that some difference exists between the objects, we then wish to decide which objects differ from the others. David (1988) discusses tests which are analogous to those used in analysis of variance: a least significant differences method, a multiple range test paralleling Tukey's, and a method like Scheffé's for judging contrasts. These tests, like the corresponding tests in analysis of variance, attempt to control error rates when a null hypothesis of equality holds; in particular, when the objects are equally preferable, these tests, run at level α, will have probability $1 - \alpha$ of declaring *none* of the objects to be different. When this null hypothesis is false, the behaviour of the tests is less clear, but in that case the experimenter may prefer to fit a model, such as the Bradley-Terry model described in Chapter 3, in which the relative strengths of the objects are estimated directly.

The range of scores is a natural statistic for use in multiple-comparison tests. When there are no order effects or ties, Table 3 of David (1988) can be used for small experiments. Asymptotically, David states that the distribution of the range of the d_i is that of the range W_t of t independent normal random variables with variance $\sigma^2(1 - \rho) = 1$. When there are order effects, the d_i are still asymptotically normal with the same variance and covariances (because of the presence of the factor $p(1 - p)$ in the denominator of the d_i), so this result still holds. The discussion above also shows that the same asymptotic distribution holds for suitably-defined d_i in the presence of ties.

The distribution of the range can then be used for a Tukey-like multiple range test, by declaring significantly different any objects whose scores differ by more than the critical range for that value of t. Alternatively, one can carry out a Student-Newman-Keuls-type procedure, using the distribution of the range of a smaller number of normal random variables within a group of scores where differences have been shown to exist.

2.4.4 Example

In the 1995-96 season, the Scottish soccer league had 10 teams, who played each other twice each at home and away. Since the home team is listed first, the order effect here indicates a home field advantage. In the notation above, $t = 10$ and $r = 2$. Of the 180 games, 45% were won by the home team, 33% by the away team, and 22% were drawn (tied). The scores for each team are shown in Table 2.5, with the d_i calculated using (2.1).

For the overall test of equality, $\sum_i a_i^2 = 3679$. Comparing this to the line of Table 2.3 for which $t = 10$, $r = 2$, $p_1 = 0.5$, $p_2 = 0.25$, $p_0 = 0.25$, whose p_i-values are fairly close to the ones observed, the test statistic is easily significant at the 1% level, as it would be for any of the p_i combinations shown. For comparison, $\sum_i d_i^2 = 57.3$, which, when compared with the χ_9^2 distribution, is significant even beyond the 0.0005 level. There is clearly a difference between the teams in this league.

To decide which teams differ from which, we turn to the corresponding line of Table 2.4, to find that two teams are significantly different at the 5% level if their scores differ by 13 or more. Thus the top two teams are significantly stronger than the bottom six, but no other differences are revealed (the top two teams are not quite significantly stronger than the 3rd- and 4th-placed teams). For comparison, using the asymptotic distribution, the upper 5% point of the range of 10 independent standard normal random variables is 4.47 (using the "∞ d. f." line of a table of the Studentized Range), so that any teams whose d_i differ by more than this are significantly different. This yields the same result as the previous procedure.

2.5 The Swiss tournament

2.5.1 Introduction and construction

Often, the number of participating teams in a tournament is too large for a round-robin, but not so large that a knockout is the only possible alternative. Various solutions are possible, such as running a double or triple knockout (in which a team is not eliminated until it has lost two or three games), or dividing the teams into groups small enough for a round-robin to be feasible in each group, and then playing a further round-robin or knockout between the top teams in each group. These solutions do not permit easy comparison between teams in different groups or different parts of the knockout. A Swiss tournament, however, provides an immediate ranking for all the teams, by virtue of not splitting them into groups.

In a Swiss tournament, the idea is to play most of the games between teams that are evenly matched, as far as the games previously played allow this to be judged. This is achieved by the following procedure. If the number of teams is odd, introduce a fictitious team called "Bye", where any team drawn against "Bye" does not play in that round, but is awarded a "free" win. In this way, the effective number of teams in the tournament t is even.

Table 2.3: Sum of squares: simulated and approximate points

					5%		1%	
t	r	p_1	p_2	p_0	Sim.	Asymp.	Sim.	Asymp.
3	1	50	50	0	21	21.0	21	25.8
3	1	75	25	0	19	18.7	21	22.4
3	1	50	25	25	19	18.2	21	21.5
3	1	20	20	60	16	15.6	17	17.5
3	1	30	10	60	16	15.2	17	17.0
3	3	50	50	0	135	135.0	147	149.4
3	3	75	25	0	127	128.2	135	139.1
3	3	50	25	25	127	126.5	135	136.5
3	3	20	20	60	119	118.8	124	124.6
3	3	30	10	60	118	117.7	123	122.9
3	5	50	50	0	343	344.9	363	369.1
3	5	75	25	0	333	333.7	351	351.8
3	5	50	25	25	331	330.9	346	347.5
3	5	20	20	60	319	318.0	327	327.6
3	5	30	10	60	316	316.2	325	324.9
4	2	50	50	0	175	175.3	185	189.4
4	2	75	25	0	169	167.4	177	178.0
4	2	50	25	25	166	165.5	174	175.2
4	2	20	20	60	157	156.5	162	162.2
4	2	30	10	60	156	155.3	160	160.3
4	3	50	50	0	371	370.9	391	392.1
4	3	75	25	0	361	359.2	375	375.1
4	3	50	25	25	356	356.2	370	370.8
4	3	20	20	60	343	342.8	352	351.2
4	3	30	10	60	342	340.9	349	348.5
6	1	50	50	0	183	183.2	191	195.3
6	1	75	25	0	175	174.9	183	183.9
6	1	50	25	25	173	172.8	180	181.1
6	1	20	20	60	164	163.3	168	168.1
6	1	30	10	60	162	162.0	166	166.3
10	1	50	50	0	895	894.6	913	918.3
10	1	75	25	0	873	873.4	889	891.2
10	1	50	25	25	868	868.2	882	884.5
10	1	20	20	60	844	843.8	854	853.3
10	1	30	10	60	841	840.5	849	849.0
10	2	50	50	0	3409	3409.2	3451	3456.7
10	2	75	25	0	3367	3366.9	3407	3402.5
10	2	50	25	25	3357	3356.3	3387	3389.0
10	2	20	20	60	3308	3307.7	3326	3326.7
10	2	30	10	60	3301	3300.9	3318	3318.0
14	1	50	50	0	2521	2522.5	2553	2559.8
14	1	75	25	0	2483	2483.4	2511	2511.4
14	1	50	25	25	2472	2473.6	2496	2499.2
14	1	20	20	60	2429	2428.6	2443	2443.5
14	1	30	10	60	2422	2422.4	2435	2435.8

Table 2.4: Range: simulated and approximate points

					5%		1%	
t	r	p_1	p_2	p_0	Sim.	Asymp.	Sim.	Asymp.
3	1	50	50	0	5	4.1	5	5.4
3	1	75	25	0	4	3.5	5	4.7
3	1	50	25	25	4	3.4	5	4.5
3	1	20	20	60	4	2.6	4	3.4
3	1	30	10	60	4	2.4	4	3.2
3	3	50	50	0	8	7.0	9	9.4
3	3	75	25	0	7	6.1	8	8.1
3	3	50	25	25	7	5.8	8	7.8
3	3	20	20	60	6	4.4	7	5.9
3	3	30	10	60	5	4.2	6	5.6
3	5	50	50	0	10	9.1	12	12.1
3	5	75	25	0	9	7.9	11	10.5
3	5	50	25	25	9	7.5	11	10.0
3	5	20	20	60	7	5.7	8	7.7
3	5	30	10	60	7	5.4	8	7.3
4	2	50	50	0	8	7.3	9	9.4
4	2	75	25	0	7	6.3	8	8.1
4	2	50	25	25	7	6.0	8	7.8
4	2	20	20	60	6	4.6	7	5.9
4	2	30	10	60	6	4.4	6	5.6
4	3	50	50	0	10	8.9	11	11.5
4	3	75	25	0	9	7.7	10	9.9
4	3	50	25	25	9	7.4	10	9.5
4	3	20	20	60	7	5.6	8	7.3
4	3	30	10	60	7	5.3	8	6.9
6	1	50	50	0	8	7.0	9	8.7
6	1	75	25	0	7	6.0	8	7.5
6	1	50	25	25	7	5.8	8	7.2
6	1	20	20	60	6	4.4	6	5.5
6	1	30	10	60	5	4.2	6	5.2
10	1	50	50	0	11	10.0	12	12.1
10	1	75	25	0	10	8.7	11	10.5
10	1	50	25	25	9	8.3	11	10.0
10	1	20	20	60	8	6.3	8	7.7
10	1	30	10	60	7	6.0	8	7.3
10	2	50	50	0	15	14.1	17	17.1
10	2	75	25	0	13	12.2	15	14.8
10	2	50	25	25	13	11.7	15	14.2
10	2	20	20	60	10	8.9	11	10.8
10	2	30	10	60	10	8.5	11	10.3
14	1	50	50	0	13	12.5	15	14.3
14	1	75	25	0	12	10.9	13	12.4
14	1	50	25	25	12	10.4	13	11.8
14	1	20	20	60	9	7.9	10	9.0
14	1	30	10	60	9	7.5	10	8.6

Table 2.5: Scores for the Example

Team	a_i	d_i
Rangers	30.0	4.34
Celtic	29.5	4.16
Aberdeen	19.5	0.54
Hearts	19.5	0.54
Hibernian	16.0	-0.72
Raith Rovers	15.5	-0.90
Kilmarnock	15.0	-1.08
Motherwell	15.0	-1.08
Partick	11.0	-2.53
Falkirk	9.0	-3.25

1. Start:

   (a) If this is the first round, arrange the teams in random order.

   (b) Otherwise, rank the teams by points, breaking ties randomly.

2. Arrange the teams in "pairing order": for a team whose rank is $j \leq t/2$, its pairing order is $2j - 1$, and for a team with rank $k > t/2$, its pairing order is $2(t + 1 - k)$. When this is done, the team that is first in pairing order is the top-ranked team, the second in pairing order is the lowest-ranked team, and so on, alternating high- and low-ranked teams.

3. For $i = 1, 2, \ldots, t$:

   (a) Let r be the team with pairing order i.

   (b) If team r has already been paired, proceed to the next value of i. Otherwise, continue.

   (c) Make a list of the available opponents for team r (where "available" opponents are those which have not yet played team r and are not yet paired in the current round). If team r has rank no larger than $t/2$, arrange the available opponents with the highest-ranked listed first; otherwise arrange them with the lowest-ranked listed first. (This list is of the opponents for team r in order of desirability.)

   (d) If it is not known to cause a blockage (see below), pair team r with the first team on its available opponent list; otherwise, proceed down the list until an opponent

is found that is not known to cause a blockage, in which case team r is paired with that opponent, or until the list is exhausted.

(e) If the list of opponents for team r is empty, or if it has been exhausted by the previous step, then a blockage has occurred: with the pairings already made, it is not possible to pair the remaining teams without causing some of them to meet an opponent for the second time. In this case, find the pairing s_1 vs. s_2 that was the last to be made, mark this as being known to cause a blockage, and reset i to the pairing order of team s_1. (In plainer terms: go back and find a new opponent for team s_1, until one is found for which all teams can be paired.)

4. In each game, award the winning team 1 point, and give $\frac{1}{2}$ point to each team for a tie (draw). (Any linear transformation of this scale will give the same result.)

5. If the desired number of rounds has not yet been played, go back to Step 1.

6. The teams are ranked by points, with ties broken by the "Buchholz score". For each team, the Buchholz score is the total number of points obtained by that team's opponents, and is therefore a measure of the quality of opposition faced by that team. Using the Buchholz score as a tiebreaker is desirable, since each team has faced different opponents.

The number of rounds played is typically somewhat larger than the number of rounds contained in a knockout for the same number of teams, so that the teams have the opportunity to play most of the other teams of similar strength. The restriction to a single meeting between each pair of teams, while making the pairing procedure more difficult, is an attempt to even out the randomness present in the assignment of opponents, as well as to ensure that each team does indeed play a collection of similar-strength opponents, rather than the same opponent repeatedly.

The algorithm used in chess tournaments differs slightly from the above in that chess players have "ratings" which offer prior information about the relative strengths of the players in the tournament, and these can be used as a seeding mechanism. In a chess tournament, players with the same number of points are sorted by their ratings, and the highest-rated players play against the lowest-rated players in the next round. While this information typically permits a ranking of the players in fewer rounds than would be necessary under the

procedure given above, we have preferred to assume that, in general, this prior information will not be available.

2.5.2 Discussion

Since each team faces different opposition, it is not immediately obvious that it is fair to compare teams by the number of points they have obtained. However, a Swiss tournament is designed, roughly speaking, to allow each team to "find its own level", in that a team with an incorrectly high ranking will face difficult opposition in the next round and will tend to lose, and a team with an incorrectly low ranking will face easy opposition in the next round, and should move towards their correct position. This ameliorates (although does not always completely overcome) the effect of the randomness in pairing each round.

We further discuss the issue of ranking teams from a Swiss tournament in Section 3.6.2, after we have introduced the Bradley-Terry model.

Chapter 3

The Bradley-Terry model

3.1 Introduction

The Bradley-Terry model (Bradley and Terry, 1952) is commonly fitted to paired-comparison data. It offers a means for modelling the probabilities of winning or tying in terms of parameters which can be interpreted on an odds or log-odds scale (depending on the parameterization chosen) as the "strengths" of the teams participating in the tournament. The original model has been extended by Davidson (1970) to accommodate ties, and by Davidson and Beaver (1977) to permit the estimation of an order effect, which has a natural interpretation of "home field advantage" in a sporting context. While, as pointed out in Davidson and Beaver (1977), it is straightforward to extend the models further to allow the order effect and tie parameters to be team-dependent, in practical situations there is rarely sufficient data to be able to demonstrate that such a model is appropriate. Below, therefore, we consider a version of the Bradley-Terry model with a single tie parameter and order effect, common to all comparisons.

An extensive bibliography of the Bradley-Terry and related models is given by Davidson and Farquhar (1976). A good reference to paired comparisons in general is David (1988).

There are two equivalent parameterizations of the Bradley-Terry model. Davidson (1977) uses parameters which combine multiplicatively; here, we use an additive parameterization that we find more convenient for estimation.

Specifically, suppose there are t teams in the tournament, and let $\beta_1, \ldots, \beta_t$ represent their "strengths". Let d be the tie parameter, and let h denote the home advantage (in

contrast to Davidson, we assume that the first team in each game has the advantage, following the European sports tradition, although there is no problem if the second team has the advantage, for h will simply be negative). For a comparison between teams i and j in that order, let p_{ij1} denote the probability that team i wins, p_{ij2} denote the probability of team j winning, and let p_{ij0} denote the probability of a tie, with $\sum_{k=0}^{2} p_{ijk} = 1$. Then the Bradley-Terry model asserts that the following relations hold:

$$\left.\begin{array}{rcl} p_{ij1} & \propto & \exp(h + \beta_i) \\ p_{ij2} & \propto & \exp(\beta_j) \\ p_{ij0} & \propto & \exp\{d + (h + \beta_i + \beta_j)/2\}, \end{array}\right\} \tag{3.1}$$

with the constant of proportionality, denoted D_{ij}, being chosen so that these probabilities sum to 1. The equivalence to the Davidson (1977) model is seen by expressing our parameters β_i, h, d in terms of his parameters π_i, γ, ν as $\beta_i = \log \pi_i$, $h = -\log \gamma$ and $d = \log \nu - (\log \gamma)/2$.

The first two of these probabilities are apparently reasonable: the probability of a team winning increases with that team's strength, while team i, playing at home, has its "effective" strength increased by h. The last, p_{ij0}, has behaviour which is less clear; its properties are given in the following two Lemmas.

Lemma 3.1 *The probabilities are unchanged if any constant is added to both β_i and β_j.*

Proof: Let c be the constant added; then $p_{ij1} \propto \exp(h + \beta_i + c)$, $p_{ij2} \propto \exp(\beta_j + c)$, and $p_{ij0} \propto \exp\{d + (h + \beta_i + \beta_j)/2 + c\}$. We see that there is a common value e^c multiplying each probability, and thus also the constant of proportionality, so that the probabilities when properly scaled do not depend on c.

This result shows that there are only $t - 1$ freely varying β_i, since, by the above result, any one β_i can be set equal to zero. As we see in Section 3.3, this requires us to make some adjustments to the estimation process.

Lemma 3.2 *For fixed $d > -\infty$, the tie probability p_{ij0} is maximum when $|h + \beta_i - \beta_j| = 0$, and decreases monotonically with $|h + \beta_i - \beta_j|$.*

Proof: By Lemma 3.1, we can, without affecting the probabilities, replace $h + \beta_i$ by $h + \beta_i - (h + \beta_i) = 0$ and β_j by $\beta_j - (h + \beta_i) = \epsilon$, say, so that $|\epsilon| = |h + \beta_i - \beta_j|$. With this parameterization, $p_{ij1} \propto 1$, $p_{ij2} \propto \exp(\epsilon)$ and $p_{ij0} \propto \exp(d + \epsilon/2)$, so that

$$p_{ij0} = \frac{\exp(d + \epsilon/2)}{\{1 + \exp(\epsilon) + \exp(d + \epsilon/2)\}}$$

$$= \frac{ax}{1 + ax + x^2},$$

where $x = \exp(\epsilon/2)$ and $a = \exp(d)$. This, as a function of x, is positive for $x > 0$ (since $a > 0$), and is zero for $x = 0$ and as $x \to \infty$. It has one stationary point, at $x = 1$, which must therefore be a maximum, indicating that p_{ij0} has a maximum at $\epsilon = 2 \log 1 = 0$ and decreases with $|\epsilon|$, as required.

This result says that, after allowing for home field advantage, the probability of a tie is greatest when the teams are evenly matched, ie. $\beta_i + h = \beta_j$, which is a property we are entitled to expect.

The quantities being exponentiated on the right-hand sides of (3.1), being linear in the parameters, have the air of a linear predictor in a generalized linear model. To that end, let $\eta_{ij1} = h + \beta_i$, $\eta_{ij2} = \beta_j$, $\eta_{ij0} = d + (h + \beta_i + \beta_j)/2$. When there are no ties, the third relation of (3.1) is discarded, and the correspondence is exact: the Bradley-Terry model is a special case of logistic regression. The design matrix has a special form: for a meeting between teams i and j, the row of the design matrix has a 1 in position i, a -1 in position j and zeroes elsewhere, with the intercept corresponding to the home field advantage (order effect). In the more general case, the link with generalized linear modelling is less useful, though the η_{ijk} notation continues to be helpful. Specifically, note that $p_{ijk} = \exp \eta_{ijk}/D_{ij}$ for $k = 0, 1, 2$.

3.2 Likelihood and derivatives

Let y_{ij1} denote the observed frequency of wins for team i against team j when the former is playing at home, and let y_{ij2}, y_{ij0} denote the observed frequencies of wins for team j and of ties in these games. Further, let $y_{ij+} = y_{ij1} + y_{ij2} + y_{ij0}$ denote the total number of games played between teams i and j with team i at home.

The likelihood is simply

$$L = \prod_{i,j,k} p_{ijk}^{y_{ijk}},$$

and the log-likelihood

$$l = \sum_{i,j,k} y_{ijk} \log p_{ijk},$$

where the product and sum extend over $1 \le i, j \le t$ with $i \ne j$ and $k = 0, 1, 2$, and the dependence on the parameters is contained within p_{ijk}.

It is most convenient to develop the likelihood derivatives in stages. First, since the η_{ijk} are linear in the parameters,

$$\frac{\partial \eta_{ij1}}{\partial \beta i} = \frac{\partial \eta_{ij1}}{\partial h} = \frac{\partial \eta_{ij2}}{\partial \beta j} = \frac{\partial \eta_{ij0}}{\partial d} = 1,$$

and

$$\frac{\partial \eta_{ij0}}{\partial \beta i} = \frac{\partial \eta_{ij0}}{\partial \beta j} = \frac{\partial \eta_{ij0}}{\partial h} = \frac{1}{2},$$

with the remaining derivatives being zero.

Next, noting that $D_{ij} = \sum_{k=0}^{2} \exp \eta_{ijk}$, and therefore that

$$\frac{\partial D_{ij}}{\partial \theta} = \sum_{k} \exp \eta_{ijk} \frac{\partial \eta_{ijk}}{\partial \theta},$$

where θ denotes a "generic" parameter, it follows that

$$\begin{aligned} \frac{\partial D_{ij}}{\partial \beta i} = \frac{\partial D_{ij}}{\partial h} &= \exp \eta_{ij1} + \frac{1}{2} \exp \eta_{ij0}, \\ \frac{\partial D_{ij}}{\partial \beta j} &= \exp \eta_{ij2} + \frac{1}{2} \exp \eta_{ij0}, \\ \frac{\partial D_{ij}}{\partial d} &= \exp \eta_{ij0}. \end{aligned}$$

Dividing through by D_{ij}, we therefore obtain

$$\begin{aligned} \frac{1}{D_{ij}} \frac{\partial D_{ij}}{\partial \beta_i} = \frac{1}{D_{ij}} \frac{\partial D_{ij}}{\partial h} &= p_{ij1} + \frac{1}{2} p_{ij0}, \\ \frac{1}{D_{ij}} \frac{\partial D_{ij}}{\partial \beta_j} &= p_{ij2} + \frac{1}{2} p_{ij0}, \\ \frac{1}{D_{ij}} \frac{\partial D_{ij}}{\partial d} &= p_{ij0}. \end{aligned}$$

Now, once again letting θ denote any of the parameters, and noting that

$$p_{ijk} = \exp \eta_{ijk} / D_{ij},$$

we find that

$$\frac{\partial p_{ijk}}{\partial \theta} = p_{ijk} \left(\frac{\partial \eta_{ijk}}{\partial \theta} - \frac{1}{D_{ij}} \frac{\partial D_{ij}}{\partial \theta} \right).$$

As a result, we can write down the derivatives of p_{ijk} with respect to the parameters, for all k, as follows:

$$\frac{\partial p_{ij1}}{\partial \beta_i} = \frac{\partial p_{ij1}}{\partial h} = p_{ij1} \left(1 - p_{ij1} - \frac{1}{2} p_{ij0} \right)$$

$$
\begin{aligned}
\frac{\partial p_{ij2}}{\partial \beta_i} = \frac{\partial p_{ij2}}{\partial h} &= p_{ij2}\left(0 - p_{ij1} - \frac{1}{2}p_{ij0}\right) \\
\frac{\partial p_{ij0}}{\partial \beta_i} = \frac{\partial p_{ij0}}{\partial h} &= p_{ij0}\left(\frac{1}{2} - p_{ij1} - \frac{1}{2}p_{ij0}\right) \\
\frac{\partial p_{ij1}}{\partial \beta_j} &= p_{ij1}\left(0 - p_{ij2} - \frac{1}{2}p_{ij0}\right) \\
\frac{\partial p_{ij2}}{\partial \beta_j} &= p_{ij2}\left(1 - p_{ij2} - \frac{1}{2}p_{ij0}\right) \\
\frac{\partial p_{ij0}}{\partial \beta_j} &= p_{ij0}\left(\frac{1}{2} - p_{ij2} - \frac{1}{2}p_{ij0}\right) \\
\frac{\partial p_{ij1}}{\partial d} &= p_{ij1}\left(0 - p_{ij0}\right) \\
\frac{\partial p_{ij2}}{\partial d} &= p_{ij2}\left(0 - p_{ij0}\right) \\
\frac{\partial p_{ij0}}{\partial d} &= p_{ij0}\left(1 - p_{ij0}\right),
\end{aligned}
$$

where the zeroes, while not necessary, illustrate the pattern.

Since

$$\frac{\partial l}{\partial \theta} = \sum_{i,j,k} \frac{y_{ijk}}{p_{ijk}} \frac{\partial p_{ijk}}{\partial \theta},$$

the derivatives of the log-likelihood with respect to the parameters can now be found. With respect to β_r, the derivative must be summed over all opponents for team r and all games, home and away; with respect to h and d, the sum is simply over all i and j. After some algebra, we find that the likelihood derivatives are readily interpretable:

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_r} &= \sum_i \left\{ y_{ir2} + \frac{1}{2}y_{ir0} - y_{ir+}\left(p_{il2} + \frac{1}{2}p_{il0}\right)\right\} \\
&\quad + \sum_j \left\{ y_{rj1} + \frac{1}{2}y_{rj0} - y_{rj+}\left(p_{rj1} + \frac{1}{2}p_{rj0}\right)\right\}.
\end{aligned}
$$

Summing over all home and away games played by team r, this is the difference between the observed "points" and expected "points" obtained by team r, where a win is worth one point and a tie a half point. (Any linear transformation of this point scale would also work, such as two points for a win and one for a tie.)

Moving on, we find that

$$\frac{\partial l}{\partial h} = \sum_{i,j} \left\{ y_{ij1} + \frac{1}{2}y_{ij0} - y_{ij+}\left(p_{ij1} + \frac{1}{2}p_{ij0}\right)\right\},$$

which is the difference between observed and expected points obtained by the home teams in all the games. Likewise,

$$\frac{\partial l}{\partial d} = \sum_{i,j} y_{ij0} - y_{ij+}p_{ij0},$$

which is the difference between observed and expected ties summed over all the games.

In summary, we see that the maximum likelihood estimates of the parameters are such that the sufficient statistics (numbers of points for each team, total points scored by home teams, total number of draws) are equal to their expectations, and that a score statistic would be based on the degree to which these differ under some hypothesis. These results were anticipated by Fienberg (1979), who showed that the Bradley-Terry model given here, when suitably parameterized, could be considered as an incomplete contingency table, and therefore proved that the maximum likelihood estimates of the parameters were obtained by equating observed and expected frequencies. Fienberg's results are valid in greater or lesser generality: to fit team-specific home field or tie effects, we can match the observed and expected home wins or ties for each team; if we do not wish to fit a home field advantage, we set $h = 0$ and do not match the observed and expected frequencies, and if we do not wish to fit ties, we set $d = -\infty$, and do not match the observed and expected frequencies of ties.

The foregoing calculations also enable us to find the second derivatives of the log-likelihood without great difficulty. They are, after some algebra:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_r^2} &= -\sum_i y_{ir+}\{p_{ir1}p_{ir2} + \frac{1}{4}p_{ir0}(1 - p_{ir0})\} \\
&\quad -\sum_j y_{rj+}\{p_{rj1}p_{rj2} + \frac{1}{4}p_{rj0}(1 - p_{rj0})\} \\
\frac{\partial^2 l}{\partial \beta_r\, \partial \beta_s} &= y_{sr+}\{p_{sr1}p_{sr2} + \frac{1}{4}p_{sr0}(1 - p_{sr0})\} \\
&\quad + y_{rs+}\{p_{rs1}p_{rs2} + \frac{1}{4}p_{rs0}(1 - p_{rs0})\} \\
\frac{\partial^2 l}{\partial \beta_r\, \partial h} &= \sum_i y_{ir+}\{p_{ir1}p_{ir2} + \frac{1}{4}p_{rj0}(1 - p_{rj0})\} \\
&\quad -\sum_j y_{rj+}\{p_{rj1}p_{rj2} + \frac{1}{4}p_{ir0}(1 - p_{ir0}) \\
\frac{\partial^2 l}{\partial \beta_r\, \partial d} &= \frac{1}{2}\sum_i y_{ir+}p_{ir0}(p_{ir2} - p_{ir1}) + \frac{1}{2}\sum_j y_{rj+}p_{rj0}(p_{rj1} - p_{rj2})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial h^2} &= -\left[\sum_{i,j} y_{ij+}\{p_{ij1}p_{ij2} + \frac{1}{4}p_{ij0}(1-p_{ij0})\}\right] \\
\frac{\partial^2 l}{\partial h\,\partial d} &= \frac{1}{2}\sum_{i,j} y_{ij+}p_{ij0}(p_{ij1}-p_{ij2}) \\
\frac{\partial^2 l}{\partial d^2} &= -\sum_{i,j} y_{ij+}p_{ij0}(1-p_{ij0}).
\end{aligned}$$

Alternatively, one can note that each game makes a contribution to the (observed) information matrix. Letting l_{rs} denote the contribution to the log-likelihood from a game between teams r and s in that order, we find that

$$\frac{\partial^2 l_{rs}}{\partial \beta_r^2} = \frac{\partial^2 l_{rs}}{\partial \beta_s^2} = \frac{\partial^2 l_{rs}}{\partial h^2} = \frac{\partial^2 l_{rs}}{\partial \beta_r\,\partial h} = -\frac{\partial^2 l_{rs}}{\partial \beta_r\,\partial \beta_s} = -\frac{\partial^2 l_{rs}}{\partial \beta_s\,\partial h} = -p_{rs1}p_{rs2} + \frac{1}{4}p_{rs0}(1-p_{rs0}),$$

$$\frac{\partial^2 l_{rs}}{\partial \beta_r\,\partial d} = \frac{\partial^2 l_{rs}}{\partial h\,\partial d} = -\frac{\partial^2 l_{rs}}{\partial \beta_s\,\partial d} = -p_{rs0}(p_{rs1}-p_{rs2}),$$

and

$$\frac{\partial^2 l_{rs}}{\partial d^2} = -p_{rs0}(1-p_{rs0}).$$

3.3 Solving the likelihood equations

There are a couple of practical problems with the results we have obtained so far. First, since the probabilities are obtained from *differences* between the β_i, and in the light of Lemma 3.1, it is impossible to obtain unique estimates for the parameters unless one of the β_i is fixed. Consequently, the information matrix cannot be inverted, since it is rank-deficient by one (or more). Second, it is possible in practice for the parameter estimates to be infinite. This usually happens when there are teams that defeat every other team they play, or are defeated in all their games. From the likelihood equations, we see that such a team r must have $p_{rj1} = 1$ or 0, and then $\hat{\beta}_i = \infty$ or $\hat{\beta}_i = -\infty$. This is a curious result, given that it is perfectly possible to observe 100% success or failure with β_i that are finite, but it is an instance of a phenomenon known as "separation" which can afflict any binary-response model. Albert and Anderson (1984) and Santner and Duffy (1986) have carried out detailed studies of separation and quasi-separation, showing that infinite parameter estimates will exist if there is a plane in X-space such that all the observations on one side of the plane are successes and all those on the other side are failures. In our

application, as mentioned, this usually takes the form of one team winning or losing all its games, but it is also possible to observe an interdependence with the home field effect that renders finite estimation impossible.

A simple approach that fixes both these problems is to introduce a "fictitious" team $t+1$ against which all other teams are given a neutral-field tie, say. If we then fix the rating of this $(t+1)$-th team at 0, or some other convenient figure, we can then obtain unique estimates for the other ratings. Furthermore, simply from its comparisons with team $t+1$, it is impossible for any team to attain 100% success or failure, and hence all teams will have finite strengths. If a home field or tie effect is being fitted, a fictitious game can be included in which was "observed" a third of a home win, a third of an away win, and a third of a tie (or, if a tie effect is not being fitted, half a home win and half an away win). This guarantees that finite estimates will exist for all the parameters no matter what the arrangement of games played, and also ensures that the information matrix is strictly positive definite rather than merely positive semi-definite. (Since the "likelihood equation" for the fictitious team is never actually used, being a consequence of the other equations, the effect on the information matrix is purely to add positive quantities to the diagonal, a standard way of making a matrix positive definite.)

Such a procedure is necessarily *ad hoc*, but is not without precedent; consider, for example, the practice of adding $\frac{1}{2}$ to the entries in a contingency table to deal with zeroes. Though we approach our estimation from a classical rather than a Bayesian viewpoint (for which see Davidson and Solomon, 1973), it is interesting to note that this procedure can be viewed as a hypothesis of "prior equality" from which we proceed, in a Bayesian fashion, towards a posterior distribution for the parameters, by combining this prior distribution with the data, with the maximum likelihood estimate corresponding to the mode of the joint posterior distribution.

We now turn to the problem of obtaining the maximum likelihood estimates. In the remainder of our discussion of this subject, we assume that the "fictitious team" has been introduced, and we are thereby assured that the estimates are finite. (Strictly, once the fictitious team has been introduced, we are no longer maximizing a likelihood, but solving a set of equations for the parameters that happens to bear a close resemblance to the likelihood equations.)

While the likelihood equations are linear in the p_{ijk}, they are certainly not linear in the β_i, h or d, and so an iterative procedure will be necessary. Below, we consider some

candidate methods and discuss their advantages and disadvantages. Detailed theoretical comparisons (in terms of operation counts) are given in Section 3.4, and some examples are considered in Section 3.5.

Any iterative algorithm needs a place to start; often the choice of initial estimate can determine whether or not the algorithm converges. In this problem, the log-likelihood tends to be well approximated by a quadratic in most cases, and is unimodal due to general results concerning exponential families. The second derivative is also positive definite for any finite values of the parameters. Therefore, any choice of initial values that is "not too unreasonable" should suffice. We set the strength of the fictitious team equal to 0, and then set the initial values for $\beta_1, \beta_2, \ldots, \beta_t, h, d$ equal to zero also.

3.3.1 Newton's method

Since all the second derivatives are readily available, and since the likelihood is approximately quadratic, an obvious choice for this problem is the multivariate Newton's method (Dennis and Schnabel, 1989). The first derivative vector and second derivative matrix are evaluated at the current parameter estimate and used to find a quadratic approximation to the log-likelihood; the maximum of this quadratic is the next parameter estimate.

Let θ denote the entire parameter vector $\beta_1, \ldots, \beta_t, h, d$; if a home field effect or tie parameter is not being fitted, it is omitted from the parameter vector and its likelihood derivatives are omitted from the estimation procedure. Let g denote the vector of first derivatives of the log-likelihood l, and let H denote the matrix of second derivatives, so that $g_i = dl/d\theta_i$ and $h_{ij} = d^2l/d\theta_i d\theta_j$. Further, let l_c, g_c, H_c denote l, g and H evaluated at $\theta = \theta_c$. Then a quadratic approximation to the log-likelihood is

$$Q_c(\theta) = l_c + g_c^T(\theta - \theta_c) + \frac{1}{2}(\theta - \theta_c)^T H_c(\theta - \theta_c)$$

and, provided H_c is negative definite, this quadratic will have a single maximum at the value θ_+ which satisfies

$$H_c(\theta_+ - \theta_c) = -g_c.$$

Newton's method for our problem consists of repeating the following steps:

1. Evaluate g_c and H_c at the current value $\theta = \theta_c$.

2. Solve $H_c(\theta_+ - \theta_c) = -g_c$ for θ_+.

3. Set $\theta_c = \theta_+$.

until convergence is attained.

In this problem, H_c is always negative definite. Thus, at each iteration, Newton's method is finding the maximum of the quadratic approximation. In principle, we also need to check that each step actually increases the likelihood (for example, a step might be in the right direction but too long, passing over the maximum to a point where the likelihood is smaller), though in practice for this problem, the likelihood is sufficiently well-behaved for such a check to be unnecessary.

The principal advantage of Newton's method is its local quadratic convergence: once θ_c is close to the maximizing θ, the number of correct significant figures approximately doubles on each iteration. As a result, the total number of iterations required is usually small. On the other hand, each iteration of Newton's method requires the solution of a system of equations, which is considerable numerical work if t is large. Other methods which can avoid solving this system may therefore converge using a smaller number of arithmetic operations, even if the number of iterations is larger.

3.3.2 Ford's algorithm

Ford (1954) proposed a model for paired comparisons which is equivalent to the Bradley-Terry model, and gives an iterative method, and a proof of its convergence, for estimating the parameters of the model. We describe the method, as did Ford, for the simple model lacking home field advantage and draw parameter.

Let $\pi_i = \exp(\beta_i)$, and let $\pi_{i,r}$ be the value of π_i after the r-th iteration. The algorithm begins by setting $\pi_{i,1} = 1$ (say) for all i, then, for $i = 1, 2, \ldots, t$, cyclically, setting

$$\begin{aligned} \pi_{i,r+1} &= W_i / \textstyle\sum_j A_{ij}/(\pi_{i,r} + \pi_{j,r}) \\ \pi_{j,r+1} &= \pi_{j,r} \qquad\qquad\qquad\qquad \text{if } j \neq i, \end{aligned}$$

where W_i is the observed number of wins for team i and $A_{ij} = y_{ij} + y_{ji}$ is the total number of games played between teams i and j.

As with Jacobi's method, described in Section 3.3.3 below, Ford's algorithm ignores any correlation between the ratings, and may therefore be expected to converge slowly if any correlations are appreciable. (Strictly, Ford's method is a kind of Gauss-Seidel algorithm, since the most recent values of π_j are used to update π_i.) In practice, convergence generally requires a large number of iterations (David, 1988, p. 62), but is guaranteed.

The algorithm as described above seems to have been intended for hand calculation, where the W_i and A_{ij} will be either available or easily calculated at the beginning. A perhaps more natural implementation for machine use consists of passing through the data set one game at a time (for further discussion of this issue see Section 3.4.5), for which we define one iteration to consist of the following steps:

1. For $i = 1, 2, \ldots, t$:
   (a) Set $S_i = 0$.
   (b) For each game in the data set:
      i. If team i played in this game, against team j say, let $S_i = S_i + 1/(\pi_i + \pi_j)$.
   (c) Set $\pi_i = W_i/S_i$.

This implementation of Ford's algorithm, which we call "Ford-1", is analyzed in Section 3.4 and used in the Examples of Section 3.5. Viewed in this light, however, it seems that an unnecessary number of passes is made through the data set (in searching for games involving team i). This suggests a modification whereby the changes to all the π_i are computed from one pass through the data, as follows:

1. Set $S_i = 0$ for $i = 1, 2, \ldots, t$.
2. For each game in the data set:
   (a) Let $\delta = 1/(\pi_i + \pi_j)$.
   (b) Add δ to both S_i and S_j.
3. Set $\pi_i = W_i/S_i$ for $i = 1, 2, \ldots, t$.

We call this algorithm "Ford-2". One might suspect, since the previous values of π_i are being used rather than the most recent ones, that this algorithm will require a higher number of iterations; on the other hand, this loss may be offset by the smaller amount of arithmetic at each iteration.

3.3.3 Jacobi's method

A simple method that can work reasonably well is to pretend that the matrix H of second derivatives of the log-likelihood is diagonal; each iteration is then reduced to a collection of

one-variable equations to solve. This is known as Jacobi's method (Dennis and Schnabel, 1984). This method does not appear to have been proposed as a solution to this problem, despite its promise in dealing with tournaments with large numbers of teams.

In practice, convergence will be quickest if the off-diagonal elements of H are much smaller than the diagonal elements. When fitting a Bradley-Terry model, this will be true if (a) the parameter values are not too widely dispersed (so that the p_{ijk} are not too different from each other) and (b) if each team has been compared with many of the other teams (rather than being compared many times with a small number of other team).

Each iteration of the method is, in general, as follows. As with Newton's method, we let g denote the vector of likelihood derivatives, θ denote the entire parameter vector including home field effect and tie parameter if included in the model, and subscripts c and $+$ denote the current and updated versions of the quantities of which they are subscripts.

1. Calculate g_c and the diagonal elements h_{iic} of H_c.
2. Let $\theta_{i+} = \theta_{ic} + g_{ic}/h_{iic}$.
3. Replace θ_{ic} by θ_{i+}.

Relative to Newton's method, the number of iterations required for Jacobi's method is usually large. However, each Jacobi iteration is quick to complete, and so even if many iterations are required, the total amount of computation is still small.

3.4 Computational complexity of the algorithms

3.4.1 Introduction

A simple way of comparing the algorithms given above is by the number of iterations they require on "typical" problems. This, however, ignores the fact that some algorithms have iterations that are much simpler than others. For example, Newton's method requires the solution of a $t \times t$ (or bigger) system of linear equations at each iteration, whereas an iteration of Ford's method requires only the calculation of some simple quantities on one pass through the data set. A fair comparison of the algorithms needs to take this into account; one approach is to measure the time taken by each algorithm on a problem, but here we prefer to count the number of floating-point arithmetic operations performed by each

method. This is most easily done by calculating the number of additions, multiplications and exponentiations required for one iteration of each method, as functions of the number of teams t and the number of games n, and then multiplying by the number of iterations required for each problem.

We restrict our analysis of floating-point operations to the simplest version of the Bradley-Terry model, with neither home field advantage nor ties included in the estimation. (Some of the data sets used in Section 3.5 contain ties, but these are counted as half a win and half a loss if a tie parameter is not present.) We count the number of additions (and subtractions), the number of multiplications (and divisions) and the number of exponentiations required in each case.

In counting operations, we have assumed a data layout in which one row of the design matrix represents one game; the operation counts will therefore contain a term proportional to the number of games n. With some precomputation, the data can be aggregated so that one row of the design matrix represents *all* the games between a particular pair of teams. We return to this issue in Section 3.4.5.

3.4.2 Ford's methods

In counting the number of arithmetic operations, we assume that the observed numbers of wins W_i have already been calculated for each team; this can be done in one pass through the data using only integer arithmetic.

The fundamental operation in Ford's methods is the calculation of $1/(\pi_i + \pi_j)$, which requires one addition and one multiplication (which is actually a division).

In Ford-1, this calculation is made twice for each game, once each on the two separate passes for the two teams involved in each game. On each occasion, a value of S_i is updated, requiring another addition. Finally, for each team, the calculation of W_i/S_i takes another multiplication. There are thus $4n$ additions and $2n + t$ multiplications in each iteration of Ford-1.

In Ford-2, the value of $1/(\pi_i + \pi_j)$ is used for both teams, saving an addition and a multiplication for each game. Apart from the smaller number of passes through the data (we ignore any saving of time due to this), the algorithm has otherwise the same number of operations. Thus Ford-2 has $3n$ additions and $n + t$ multiplications per iteration.

3.4.3 Jacobi's method

An iteration of Jacobi's method consists of two parts: the calculation of the derivatives g_i and h_{ii}, and the use of these derivatives to update the β_i.

For a game between teams i and j, the derivative calculation consists of the calculation of the win probability $p = \{1 + \exp(\beta_j - \beta_i)\}$, which requires two additions, a multiplication and an exponentiation, the calculation of $y - p$, the increment to the first derivative and its addition to g_i and subtraction from g_j (three additions), and the calculation of $p(1-p)$ and its addition to h_{ii} and h_{jj} (three additions and a multiplication). Since there are n games, each of these operation counts must be multiplied by n.

The update to β_i consists of adding g_i/h_{ii}, and so contributes an extra addition and multiplication for each of the t teams. One iteration of Jacobi's method therefore requires $8n + t$ additions, $2n + t$ multiplications and n exponentiations. (We note that it may be possible to save some exponentiations by storing and updating $\exp(\beta_i)$ instead of β_i; an exponentiation is then necessary for each of the t updates, but one is saved for each of the n probability calculations.)

3.4.4 Newton's method

One iteration of Newton's method is considerably more complicated than for any of the other methods. Specifically, one iteration consists of

1. Calculation of the derivative vector g and second derivative matrix H.
2. Decomposition of H into a suitable form for solving $H\delta = g$ for δ, where δ is the update to the vector of β_i. We use the square-root-free Cholesky decomposition $H = LDL'$, where L is unit lower triangular and D is diagonal.
3. Solution of $H\delta = g$ using the decomposition.
4. Updating of the β_i.

Step 1 takes a similar form to that for Jacobi's method. The operations for calculating a probability for each game, calculating the updates to g and H and carrying out the update to g are identical, but two extra additions occur because *four* elements of H must be updated, two off-diagonal elements as well as the two diagonal ones. Over all n games, there are

therefore $10n$ additions, $2n$ multiplications and n exponentiations (some of which can be avoided with sufficient care; see the discussion of this issue under Jacobi's method).

Step 2 simply requires a count of the operations involved in the standard algorithm for this decomposition. There are $t(t-1)(t+1)/6$ additions and $t(t-1)(2t+5)/6$ multiplications (a result which seems to be correct, though which differs from Table 3.2.2 of Dennis and Schnabel (1983, p. 51), where the coefficient of the cubic term is asserted to be $\frac{1}{6}$ for both the number of additions and the number of multiplications).

Step 3 requires the sum of the operation counts for a solution of a diagonal system of equations (t multiplications only) and two solutions of triangular systems, one for L and one for L' ($t(t-1)/2$ additions and multiplications each). The total number of additions in this step is then $t(t-1)$ and of multiplications is $t(t-1)-t = t^2$.

Step 4 simply consists of the addition of the solution of the set of equations calculated in Step 3 to the current values of β_i, which requires t additions.

Combining these results, we find that Newton's method requires $10n + \frac{1}{6}t(t^2+6t-1)$ additions, $2n + \frac{1}{6}t(t+5)(2t-1)$ multiplications and n exponentiations on each iteration.

3.4.5 Summary and additional notes

In the application of these methods to real data, we also include "fictitious games", one for each team, to ensure that the estimated β_i are all finite. A data set with n "real" games thus actually contains a total of $n+t$ games as far as the algorithms are concerned. To get accurate operation counts, we therefore need to replace n in the formulas derived above by $n+t$. When this is done, and the resulting quantities simplified, we obtain the operation counts shown in Table 3.1.

As was remarked earlier, our data layout, in which one row of the design matrix represents one game, leads to components of the operation counts that are proportional to n. By doing some precomputation, all the games between each pair of teams can be aggregated into one row of the design matrix; the effect of this, for all the algorithms, is to reduce n to the number n_1 of distinct games. In tournaments where each pair of teams meets several times, the difference between n and n_1 will be considerable, and the gains achieved by aggregation will be large.

As one would expect, the operation counts are linear in t for all the methods except Newton, where the number of additions and multiplications is cubic in t because of the necessity of solving a system of linear equations at each iteration. Of the other methods,

Table 3.1: Operation counts for the algorithms

Algorithm	Adds	Multiplies	Exp'ns
Newton	$10n + \frac{1}{6}t(t^2 + 6t + 59)$	$2n + \frac{1}{6}t(2t^2 + 9t + 7)$	$n + t$
Jacobi	$8n + 9t$	$2n + 3t$	$n + t$
Ford-1	$4n + 4t$	$2n + 3t$	0
Ford-2	$3n + 3t$	$n + 2t$	0

Ford-2 has the smallest coefficients attached to both n and t, and so this will be the method whose individual iterations are quickest.

In deciding which algorithm to apply to a particular problem, we will also need to consider the typical numbers of iterations required for convergence of the algorithms, so that the total amount of arithmetic for each can be assessed. In the next Section, we will see how this works out for some real data sets.

3.5 Examples

3.5.1 Introduction

We consider four real data sets here, from the sports of soccer, ice hockey and basketball. These are intended to illustrate how different aspects of the data set influence the convergence properties, and hence the desirability, of the algorithms we have considered. Two of the data sets are from round-robin tournaments, while the other two are less balanced and therefore provide more of a test for the algorithms.

In each of our examples, the algorithms were run until the largest change in any of the β_i (and h and d, if fitted) was less than 10^{-4}, an attempt (not always successful) to obtain four decimals of accuracy. It is unlikely that greater accuracy would be required in practice; indeed, three- or even two-decimal accuracy might be sufficient to obtain fitted probabilities to the accuracy desired. With our choice of convergence criterion, it was generally true that the estimates obtained from Newton and Jacobi agreed to four decimals, but those obtained from Ford's methods tended to differ in the fourth place. This seems to be a consequence of the local convergence rates of Ford's methods being slower than their competitors', a property that also suggests that these methods will be more competitive if only moderate accuracy is desired.

Table 3.2: Scores for the Iceland example

Team	Score
Akranes	13.5
KR	13.0
Leiftur	10.5
IBV	8.5
Valur	8.5
Stjarnan	8.5
Keflavik	7.5
Grindavik	7.0
Fylkir	6.5
Breidablik	6.5
Home team	52.0
Ties	20

Table 3.3: Iterations and operation counts for the Iceland example

Algorithm	Iterations	Adds	Multiplies	Exp'ns
Newton	4	5,060	2,700	400
Jacobi	30	24,300	6,300	3,000
Ford-1	122	48,800	25,620	0
Ford-2	133	39,900	14,630	0

3.5.2 A small round robin

Weather conditions in Iceland ensure that the soccer season has to run during the (short) summer. In 1996, the $t = 10$ teams in the top division played a double round-robin tournament for a total of 18 games per team and $n = 90$ in total. The scores for the teams (one point for a win, half a point for a tie) are shown in Table 3.2, along with the total number of points obtained by the home teams and the total number of ties.

Fitting the simple Bradley-Terry model using the four different algorithms required the numbers of iterations shown in Table 3.3. Of greater interest is the total number of arithmetic operations required in each case; these are shown in the remaining columns of the Table, and were calculated by substituting $t = 10, n = 90$ into the formulas of Table 3.1 and multiplying by the number of iterations.

With this small number of teams, Newton's method clearly comes out best, the additional

complexity of each iteration more than offset by the much smaller number of iterations required. Ford's methods are not competitive, but it is worth noting that the operation count of Ford-2 is noticeably smaller than that of Ford-1, despite requiring a greater number of iterations. Ford's methods compare in this way in all of our examples, suggesting that this behaviour persists in some generality.

In this round-robin, each pair of teams plays twice, so that even though $n = 90$, there are only $n_1 = 45$ different games played. Aggregating the data produces the most dramatic reduction in operations for our two implementations of Ford's method, but with the small t of this problem, Newton's method is still preferred.

Newton's and Jacobi's methods can also be used to fit models containing a tie parameter, a home field effect, or both. We have not carried out an analysis of operation counts for these models, but it is instructive to compare the iteration counts — Newton's method required four iterations no matter which model was fitted, but Jacobi required a greater number of iterations when a tie parameter was included: 44 for the tie parameter only, and 42 when a home field effect was fitted as well.

3.5.3 A larger round robin

The soccer season in England traditionally runs from late summer to spring. In the 1996/97 season, the top ("Premier") division contained $t = 20$ teams who played 38 games each for a total of $n = 380$. The scores for the teams, giving one point for a win and half a point for a tie, the total points for the home teams, and the total number of ties, are shown in Table 3.4.

The numbers of iterations required and the operation counts for each method are shown in Table 3.5. Ford's methods are again uncompetitive, but this time, with the increased number of teams, Jacobi's method is superior to Newton, indeed almost matching Newton's number of iterations (this is probably due to the balance in the data).

As in the previous Example, each pair of teams meets twice, so that $n_1 = n/2 = 190$. Recalculating the numbers of operations, however, does not lead to any change in the relative merits of the algorithms: Jacobi's method is still preferable to Newton, with the operation counts for Ford-1 and Ford-2 being higher.

When a home field effect and a tie parameter are fitted, Newton's method continues to require only four iterations for convergence, but Jacobi requires one extra iteration with a home field effect and an additional 9 iterations with a tie parameter. (The numbers of

Table 3.4: Scores for the England example

Team	Score
Manchester Utd	27.0
Newcastle	24.5
Arsenal	24.5
Liverpool	24.5
Aston Villa	22.0
Chelsea	21.5
Sheffield Wed	21.5
Wimbledon	20.5
Leicester	17.5
Leeds	17.5
Derby	17.5
Tottenham	16.5
Blackburn	16.5
West Ham	16.0
Everton	16.0
Coventry	16.0
Middlesbrough	16.0
Southampton	15.5
Sunderland	15.0
Nottingham F	14.0
Home team	221.5
Ties	119

Table 3.5: Iterations and operation counts for the England example

Algorithm	Iterations	Adds	Multiplies	Exp'ns
Newton	4	22,920	16,200	1,600
Jacobi	5	16,100	4,100	2,000
Ford-1	140	224,000	114,800	0
Ford-2	159	190,800	66,780	0

Table 3.6: Iterations and operation counts for college hockey

Algorithm	Iterations	Adds	Multiplies	Exp'ns
Newton	5	121,630	164,510	4,100
Jacobi	119	785,876	200,396	97,580
Ford-1	289	947,920	486,676	0
Ford-2	321	789,660	277,344	0

iterations happen to behave additively.) In the presence of ties, therefore, the advantage of Jacobi seems to have been lost.

3.5.4 College ice hockey

In the US, 44 college teams compete in the top level, "Division 1", of ice hockey organized by the National Collegiate Athletic Association. These teams are mostly based in the northeastern and midwestern US, and play anywhere between fewer than 20 and more than 40 games in a season that stretches from October to March. In the 1996/97 season, a total of $n = 776$ games were played between Division 1 teams; a fair number of games were played between Division 1 teams and those at lower levels, but these are not included here. The range in strength of teams is quite wide, though "upsets" do happen. The teams are arranged in conferences, and play most of their games against teams in the same conference; there are also some "independent" teams, not affiliated with any conference, who tend to play a smaller number of games than average. This lends a fair imbalance to the data, although the four conferences were, in this season, of similar strength, so that teams in different conferences will have a similar calibre of opposition. In terms of wins and losses, as well as according to the Bradley-Terry model, the strongest team was Michigan, with 34 wins, 4 losses and 4 ties, while the weakest on both counts was Air Force, who had 2 wins, 16 losses and 1 tie.

The iteration and operation counts are shown in Table 3.6. Despite the large number of teams, Newton's method is clearly the best here; it maintains its record of convergence in a small number of iterations while the imbalance in the data adversely affects the other methods. Indeed, there is little to choose between Jacobi and Ford-2 here.

Because of the arrangement of the teams into conferences, some pairs of teams play many games against each other and many pairs of teams do not meet at all. As a result, the

Table 3.7: Iterations and operation counts for college basketball

Algorithm	Iterations	Adds	Multiplies	Exp'ns
Newton	6	29,484,846	58,200,570	27,072
Jacobi	112	4,077,024	1,044,960	505,344
Ford-1	360	6,497,280	3,358,800	0
Ford-2	393	5,319,648	1,893,474	0

value of n_1 is 317, less than half of $n = 776$. Running the algorithms on the aggregated data therefore saves a considerable number of operations, especially with Ford-1 and Ford-2, but the pattern remains the same, with Newton preferred and little to choose between Jacobi and Ford-2.

Ties are possible, though rare, in college hockey, and the advantage of home ice seems to be small, though the assessment of home ice advantage is clouded by some of the games being played on neutral ice (which was not accounted for in the fitting process). The models including these parameters were fitted anyway, using Newton and Jacobi. Newton's method contined to require five iterations for convergence, while Jacobi's method was essentially unaffected by the extra parameters, requiring from 117 iterations with one extra parameter to 120 when both were included.

3.5.5 College basketball

For our final example, we have an extremely large data set. Basketball is played at an enormous number of colleges and universities in the US; our data, which is taken only from the top level ("Division 1") of the National Collegiate Athletic Association's structure, consists of 4,206 games played by 306 different teams during the period from October 1996 to March 1997. As with the previous example, the teams are arranged in conferences and play most of their games against teams in their own conference, although there is a greater amount of inter-conference play in basketball due to a large number of small tournaments played before the turn of the year and the 64-team knockout competition that ends the season. Not only do the teams vary widely in strength, but the conferences do as well; a team's won-lost record is generally a poor indicator of the team's strength.

With the large number of teams and the unbalancedness of the data, estimation is challenging for any algorithm. Table 3.7 shows how our four algorithms fared. With this

many teams, Newton's method, even though convergent in six iterations, simply cannot compete; Jacobi appears to be the method of choice, though Ford-2 is not far behind, especially when the cost of an exponentiation is considered.

The conference structure of this data set means that many pairs of teams never meet, but any particular pair of teams plays no more than twice. The value of n_1 is 2861, which is close to $\frac{2}{3}n$, a larger ratio than for the other Examples. The effect of aggregating the data is therefore not large: the operation counts are reduced by roughly a third for Jacobi, Ford-1 and Ford-2, while the operation counts for Newton, which are dominated by t, are almost unchanged.

Basketball games cannot end in a tie, and I have not attempted to distinguish in the data set between home and away teams or games played on a neutral court, so we do not assess the effect of estimating home advantage or a tie parameter here.

3.5.6 Discussion

These examples have shown that Newton's method, though the most complicated to program, requires the smallest number of operations for convergence for small and moderate values of t. Ford-2 always seems to outperform Ford-1, and tends to perform similarly to Jacobi's method in large or unbalanced data sets. All these methods are far superior to Newton when t is large. Jacobi's method, on the other hand, appears to perform well when the data set is balanced.

Newton's method appears to be the most reliable, converging always in a small number of iterations - this is unusual behaviour, since Newton's method generally requires some sort of safeguard to prevent occasional large steps being taken. The reason for this may be that the log-likelihood is close to quadratic for typical Bradley-Terry problems. When the number of teams is small, Newton's method can be recommended without question, but as the number of teams increases, the number of operations per iteration becomes insupportably large, at which point Jacobi's method or Ford-2 are to be preferred.

3.6 Comparisons with round-robin and Swiss tournaments

3.6.1 Round-robins

It is known (see David, 1988, p. 104; Bühlmann and Huber, 1963) that, in the simplest version of the Bradley-Terry model in which there is no home field advantage (order effect) or possibility of ties, the ranking of the teams given by the Bradley-Terry model is identical to that given by the number of wins. Further, two teams with the same number of wins will have the same strengths as given by the Bradley-Terry model. Bühlmann and Huber also show that the converse holds: only under the Bradley-Terry model are the rankings guaranteed to be identical.

The identity of rankings extends to tournaments with ties (Davidson, 1970): the teams are now ranked by points, with 1 point given for a win and $\frac{1}{2}$ point for a tie. As noted in the discussion of the likelihood equations in Section 3.2, a linear transformation of this scale has no effect on the estimation procedure, and since such a transformation does not affect the ranking of the teams by points either, the result continues to hold. However, there is no guarantee of equality of rankings when some other point system is used, such as awarding 3 points for a win and 1 for a tie (which is the world standard for soccer).

We show in Theorem 3.4 that equality of rankings holds in the presence of a home field advantage, provided that the round-robin is properly "balanced". As a preliminary, we require a Lemma that enables us to assert that the expected number of points for a team from a game increases with the difference in strengths:

Lemma 3.3 *The expected number of points $p_{ij1} + p_{ij0}/2$ is an increasing function of $h + \beta_i - \beta_j$.*

Proof: Let $x^2 = \exp(h + \beta_i)$, $y^2 = \exp(\beta_j)$ and $\delta = \exp(d)$. Then $p_{ij1} = x^2/(x^2 + y^2 + \delta xy)$ and $p_{ij0} = \delta xy/(x^2 + y^2 + \delta xy)$. Thus the expected number of points is

$$s_{ij} = \frac{x^2 + \delta xy/2}{x^2 + y^2 + \delta xy}.$$

For fixed y, this is seen to be a function of x that is zero when $x = 0$, and increases to 1 with x, and thus, since x is an increasing function of $h + \beta_i$ for β_j fixed, the result is proved.

Theorem 3.4 *In a round-robin where, for each i and j play against each other r times at the field of team i and r times at the field of team j, the teams are ranked identically by the*

Bradley-Terry model and by points, with one point awarded for a win and a half point for a tie (or any linear transformation of this scale).

Proof: It suffices to show that all pairs of teams (i, j) are correctly ranked; in particular, we show that the observed numbers of points are correctly greater, equal or smaller as $\hat{\beta}_i > \hat{\beta}_j$, $\hat{\beta}_i = \hat{\beta}_j$, $\hat{\beta}_i < \hat{\beta}_j$. Furthermore, since the Bradley-Terry likelihood equations require $\hat{\beta}$ to be found such that the observed and expected points are equal, we need only show that the expected numbers of points are ordered correctly. Let $s_{mn} = p_{mn1} + p_{mn0}/2$ be the expected number of points for team m in a game at home against team n.

Consider a pair of teams i and j for which $\hat{\beta}_i > \hat{\beta}_j$. For $k \neq i, j$, both teams play team k at home r times and away r times. Now, $s_{ik} > s_{jk}$ because of Lemma 3.3 and $s_{ki} < s_{kj}$ by the same Lemma, so that, for each opponent k, team i has a greater number of expected points than team j, both home and away. Also, $\hat{h} + \hat{\beta}_i - \hat{\beta}_j > \hat{h} + \hat{\beta}_j - \hat{\beta}_i$, so that team i is expected to gain more points than team j in the totality of games between team. Thus the result holds if $\hat{\beta}_i > \hat{\beta}_j$.

The other cases follow similarly; if $\hat{\beta}_i < \hat{\beta}_j$, the preceding argument can be used with i and j exchanged, whereas if $\hat{\beta}_i = \hat{\beta}_j$, $s_{ik} = s_{jk}$ and $s_{ki} = s_{kj}$ for all k and $s_{ij} = s_{ji}$. Thus the proof is complete.

It is worth noting that the Theorem does not in general hold if the balance condition is not satisfied. For example, if the home field advantage is large, then a team that plays most of its games at home will be expected to gain a larger number of points than a team that plays the same opposition predominantly away from home. Even if the *number* of home and away games is balanced for each team, there is still an advantage in expected points for the team that plays the very strongest and very weakest opposition away from home, where the probabilities are affected least, and the other teams at home, where the probabilities are affected most by the home field advantage.

3.6.2 Swiss tournaments

In Section 2.5, we introduced Swiss tournaments as an alternative to round-robins when the number of teams t is too large for a round-robin to be feasible.

It should be noted first that, because the design produced by a Swiss tournament is sequential in nature, standard likelihood results cannot be applied blindly. In particular, the probability of a particular game ending in a win, loss or tie depends on the teams involved in

that game, and this in turn depends on the results of previous games, so that the likelihood is a product not of independent probabilities, but of conditional ones that express this dependence. This has an impact both on the information matrix and on repeated-sampling inference: up to now, we have not had to worry about distinguishing between observed and expected information, since the second derivative of the log-likelihood for independent data does not contain the data, and therefore the observed and expected information matrices are equal. However, calculation of the expected information for a sequential design requires us to include the conditioning, and this will in general be very difficult. Similarly, repeated-sampling inference from a sequential design requires us not to conceive of replications of results from the design that was observed, but to consider that the design itself will vary in repeated sampling. (Adherents to the "pure likelihood" principle of inference face no such problem, since for them the only relevant issue is what was actually observed, rather than any rules determining the design of the experiment.)

This issue also arises in the sequential construction of optimal designs, as we shall see in Chapter 4. Silvey (1980, p. 63) gives a discussion in this context. While acknowledging that these problems exist, however, we continue to estimate parameters as if the design had been fixed in advance; this approach seems to yield sensible results with considerably less attention to detail than a more careful approach would require.

The lack of balance inherent in a Swiss tournament, as well as the sequential nature of the tournament design discussed above, precludes any easy theoretical results concerning the equality of rankings obtained from the tournament itself and from Bradley-Terry estimation. We turn, therefore, to simulation studies to investigate the nature and strength of agreement between the two rankings.

In order to do this, we need to recall that, in a Swiss tournament, the teams are first ranked by points, denoted here a_i, and then, if tied, by Buchholz score (the sum of points obtained by all the opponents of each team, a measure of "strength of schedule"), which is here denoted b_i. Then we need a measure of the agreement between the two rankings. We have chosen Kendall's τ for our measure of rank correlation, since it is based on the number of "discordances", that is, the number of pairs of teams (i, j) such that team i is ranked above team j on one ranking but below on the other; the number of discordances seems a natural way to quantify disagreement between rankings.

Simulations were carried out for various combinations of number of teams, disparity of team strengths and tendency for ties to happen. Specifically, we assumed that the true β_i for

the teams are equally spaced and in descending order, with the difference $\beta_i - \beta_{i+1}$ chosen so that the probability of team i defeating team j (given a non-tie, in the cases where ties are possible) is p_W, for $p_W = 0.5, 0.6, 0.7$. The tie parameter in the Bradley-Terry model can be related to the probability of a tie occurring between two teams of equal strength; in our simulations, the parameter is chosen so that this probability is $p_T = 0, 0.1, 0.25, 0.4$. Finally, the numbers of teams t investigated were $t = 10, 20, 30, 45$, the number of rounds in the tournaments were respectively $r = 6, 9, 10, 12$ (approximately $2 \log_2 t$), and 100 simulations were carried out for each combination.

The results of the simulations are shown in Table 3.8. The column marked τ_0 shows the mean "raw" Kendall correlations between the rankings obtained from the Swiss tournaments and the Bradley-Terry β_i. The correlations are generally high, but tend to decrease as t increases and increase as p_W increases (the relationship with p_T seems unclear). As p_W increases for fixed t and p_T, the true ranking becomes easier to discern, and so it seems likely that both rankings are approaching the true ranking. As t increases for fixed p_W and p_T, the number of rounds in the tournament becomes a smaller fraction of $t - 1$, the number of rounds in a single round-robin, so that it becomes increasingly difficult to rank close-together teams correctly.

A more detailed investigation was conducted of the situations in which a pair of teams would be misranked in the Swiss tournament (relative to the Bradley-Terry β_i). In the vast majority of cases (typically 70%-80%), especially for $p_W > 0.5$, a pair of teams (i, j) would be misranked because team i had more points but a distinctly smaller Buchholz score than team j. This means that team i faced considerably easier opposition than team j through the tournament, and so, even though $a_i > a_j$, is actually assessed as weaker ($\beta_i < \beta_j$) by the Bradley-Terry model. This suggests that a better approach might be to calculate a Swiss tournament score by $a_i + \alpha b_i$, for some value of α, and rank the teams in this fashion. Note that the standard procedure consists of taking α to be a slightly larger than zero.

For each of the simulated tournaments, the value α_{max} of α was found such that the mean value of τ was maximized for that combination of t, p_W, p_T; in other words, α was chosen to make the rankings based on this modified score correspond as closely as possible on average to the rankings from the β_i. Because a small change in α may not change the ranking, and hence τ has a plateau as a function of α, the maximizing value α_{max} is not uniquely defined, and in particular cannot be pinned down to more than two or three decimals. In Table 3.8, the column labelled τ_{max} shows the maximum value of τ (which *is* precisely defined), and

Table 3.8: Kendall rank correlations for simulated Swiss tournaments

t	p_W	p_T	τ_0	τ_1	τ_{max}	α_{max}
10	0.50	0.00	0.9098	0.9522	0.9531	0.18
10	0.50	0.10	0.9298	0.9571	0.9573	0.16
10	0.50	0.25	0.9253	0.9627	0.9642	0.16
10	0.50	0.40	0.9369	0.9649	0.9649	0.17
10	0.60	0.00	0.9287	0.9424	0.9438	0.19
10	0.60	0.10	0.9382	0.9498	0.9518	0.13
10	0.60	0.25	0.9311	0.9578	0.9593	0.17
10	0.60	0.40	0.9447	0.9656	0.9660	0.14
10	0.70	0.00	0.9533	0.9609	0.9629	0.21
10	0.70	0.10	0.9493	0.9627	0.9636	0.16
10	0.70	0.25	0.9376	0.9507	0.9511	0.23
10	0.70	0.40	0.9329	0.9522	0.9547	0.19
20	0.50	0.00	0.9157	0.9473	0.9478	0.18
20	0.50	0.10	0.8854	0.9445	0.9459	0.16
20	0.50	0.25	0.8987	0.9523	0.9528	0.16
20	0.50	0.40	0.9088	0.9548	0.9559	0.16
20	0.60	0.00	0.9209	0.9376	0.9377	0.16
20	0.60	0.10	0.9000	0.9376	0.9388	0.15
20	0.60	0.25	0.8915	0.9386	0.9393	0.14
20	0.60	0.40	0.9015	0.9421	0.9427	0.14
20	0.70	0.00	0.9249	0.9413	0.9415	0.14
20	0.70	0.10	0.9086	0.9421	0.9424	0.16
20	0.70	0.25	0.8921	0.9378	0.9384	0.15
20	0.70	0.40	0.8937	0.9385	0.9393	0.14
30	0.50	0.00	0.8977	0.9385	0.9389	0.18
30	0.50	0.10	0.8709	0.9417	0.9419	0.17
30	0.50	0.25	0.8630	0.9414	0.9415	0.16
30	0.50	0.40	0.8835	0.9474	0.9477	0.16
30	0.60	0.00	0.9109	0.9264	0.9270	0.20
30	0.60	0.10	0.8967	0.9292	0.9295	0.19
30	0.60	0.25	0.8884	0.9346	0.9353	0.19
30	0.60	0.40	0.9001	0.9383	0.9392	0.19
30	0.70	0.00	0.9189	0.9323	0.9324	0.18
30	0.70	0.10	0.9040	0.9302	0.9303	0.16
30	0.70	0.25	0.8896	0.9282	0.9283	0.14
30	0.70	0.40	0.8883	0.9308	0.9309	0.16
45	0.50	0.00	0.8466	0.8948	0.8951	0.16
45	0.50	0.10	0.8036	0.8868	0.8878	0.14
45	0.50	0.25	0.8037	0.8870	0.8871	0.17
45	0.50	0.40	0.8146	0.8912	0.8924	0.15
45	0.60	0.00	0.9080	0.9195	0.9197	0.18
45	0.60	0.10	0.8899	0.9209	0.9209	0.18
45	0.60	0.25	0.8825	0.9210	0.9210	0.17
45	0.60	0.40	0.8881	0.9238	0.9242	0.17
45	0.70	0.00	0.9131	0.9253	0.9256	0.18
45	0.70	0.10	0.8984	0.9232	0.9233	0.18
45	0.70	0.25	0.8886	0.9242	0.9247	0.21
45	0.70	0.40	0.8878	0.9237	0.9238	0.18

the column labelled α_{max} shows the (unique to two decimals) maximizing α.

The values of τ_{max} are substantially greater than those of τ_0 in almost all cases; furthermore, the variability in the values of α_{max} is small over the entire range of values of t, p_W, p_T in the study. In particular, the value $\alpha = \frac{1}{6}$ seems to be "typical" of the α_{max} values, and is also a readily memorable number. We can therefore propose ranking teams in a Swiss tournament, instead of by points with the Buchholz score as a tiebreaker, by the "scores" $s_i = a_i + b_i/6$. This was done for each simulated tournament, and the values τ_1 of the mean Kendall correlation for each combination of t, p_W, p_T are shown in the Table. It is seen that the values of τ_1 are very close to τ_{max}, even when α_{max} is not especially close to $\frac{1}{6}$, so that, as with τ_{max}, the rank correlations are substantially improved in almost all cases.

The simulations were carried out by awarding two points for a win and one for a tie, so as to enable integer arithmetic to be used, but it should be noted that any linear transformation of the point scale has the same effect on the Buchholz scores as it does on the points, so that the same value $\alpha = \frac{1}{6}$ can be used for any such point scale.

In a Swiss tournament with r rounds, the points scored by each team are proportional to r, while the Buchholz scores are the sum of r quantities each proportional to r, and so themselves are proportional to r^2. With this in mind, one might suspect that a score of the form $a_i + \alpha'\sqrt{b_i}$ would give better results. This was tried with the results of the simulation. It turned out that the values of τ_{max} were very similar (sometimes slightly larger, sometimes slightly smaller) to those obtained with α_{max}, but, rather less conveniently, the values of α'_{max} were very much more variable with no obvious pattern. So our recommendation stands: namely, to rank teams by $s_i = a_i + b_i/6$.

It should be borne in mind that these conclusions are only valid in the range of the simulations performed. In particular, the gains in using the s_i rather than ranking by points may be negligible for $\alpha = \frac{1}{6}$ for larger values of t or p_T. The evidence of our simulation study, however, suggests that the variability in values of α_{max} is small for large t, and that the biggest gains $\tau_1 - \tau_0$ are obtained for large values of p_T.

Chapter 4

Optimal designs and comparisons

4.1 Introduction

In Chapter 2, we considered designs for paired-comparison experiments in a general setting, and in Chapter 3, we looked at the Bradley-Terry model. It is natural now to consider the design problem in cases where the Bradley-Terry model is tenable; in particular, we can use the theory of optimal design to develop designs which should permit accurate estimation of the parameters in a Bradley-Terry model, and we can investigate the efficiency relative to these "optimal" designs of round-robin and Swiss tournaments.

Throughout this chapter, our objective is to estimate accurately the strengths of *all* teams in a tournament. There are other possible objectives, such as maximizing the probability of detecting the best team, for which the design goals are different. (For example, a knockout tournament is, relative to the number of games that need to be played, an effective way of finding the best team but a very ineffective way of ranking all the teams; see David, 1988, Section 6.4.)

We begin with a short review of the ideas of optimal experimental design. Atkinson and Donev (1992) and Silvey (1980) provide a more general background. We then investigate the nature of optimal designs and their determination when the Bradley-Terry model holds. The problem is interesting because of the design space: rather than being able to choose from an infinite set of x-values in some closed region, our design space consists of the $t(t-1)/2$ different games that could be played between the teams, and is a matter of choosing which games to play, and how frequently. Like Atkinson and Donev, we consider "continuous", "exact" and "sequential" designs, and offer some ideas about the application of optimal

design in practice. Finally, we present some efficiency comparisons between round-robins, Swiss tournaments and sequential designs with optimal exact designs of the same size.

4.2 Optimal design

The meaning of a "best" design depends on the objectives of the experiment. One might, for example, be interested in estimating the parameters of a model accurately, or be interested in predicting the response variable with high accuracy in some region. There are thus numerous optimality criteria in use – see, for example, Atkinson and Donev (1992, ch. 10). However, since accuracy of estimation and prediction are both governed by the information matrix, most of the commonly-used criteria are functions of the information matrix. There are, in fact, some equivalence results, given in (4.2) and Theorem 4.2.

For our notation, let $\mathbf{u}$ denote a design, that is, a collection of values $\{x_i, w_i\}$ where the x_i are design points and the w_i indicate the amount of experimental effort at that point. We distinguish between "continuous" designs, in which $\sum_i w_i = 1$ and the w_i indicate the fraction of observations to be taken at design point x_i, and "exact" designs, where the w_i are integers with $\sum w_i = n$ for some chosen value of n. Let $M(\mathbf{u}, \theta)$ denote the $p \times p$ information matrix based on a design $\mathbf{u}$, evaluated at parameter value θ. We will tend to have a particular value of θ in mind, in which case the dependence on θ will be suppressed. Let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of $M(\mathbf{u}, \theta)$; note that the eigenvalues of $M^{-1}(\mathbf{u}, \theta)$, the asymptotic covariance matrix of the parameters, are $1/\lambda_1, \ldots, 1/\lambda_p$.

Continuous designs are generally easier to obtain or verify by means of the theory, but in practice, exact designs are required. For large samples, exact designs can be obtained by the obvious process of multiplying the w_i in a continuous design by n and then rounding, but for small samples, and commonly in the designs required for the Bradley-Terry model, the loss of accuracy due to rounding can be too great, and so we will consider methods for generating exact designs directly.

Some commonly-used optimization criteria are expressed in terms of the eigenvalues as follows:

A-optimality Minimize $\sum_i (1/\lambda_i)$, the sum (or equivalently average) of the variances of the parameter estimates.

D-optimality Minimize $1/\prod_i \lambda_i$, which is the inverse determinant of the information matrix (the "generalized variance" of the parameter estimates").

E-optimality Minimize $\max_i(1/\lambda_i)$, the variance of the least well-estimated contrast $a'\theta$ where $a'a = 1$.

In linear models, designs can typically be found which are optimal under the chosen criterion for all parameter values, and therefore there is no doubt about the optimality of the design. For non-linear models, however, of which the Bradley-Terry model is an example, the optimal design depends on the values of the parameters. In theory, this presents no problem, since optimal designs can be found for any given values of the parameters, but in practice the dependence on the true parameter values presents a difficulty. Two possible approaches suggest themselves: to design based on *a priori* values for the parameters, as we do in Sections 4.3.2 and 4.3.3, or to adopt a sequential approach, making a few observations, estimating the parameters, constructing a small design based on the current information about the parameters, and repeating as necessary. We investigate this approach in Section 4.3.4.

We concentrate on D-optimality, which turns out to have some attractive properties for the Bradley-Terry model.

4.3 D-optimal design for the Bradley-Terry model

4.3.1 Introduction

As discussed above, we will need to consider a number of different design problems. Assuming the parameters to be known, we wish to find continuous and, if possible, exact designs. Then we discuss sequential designs. The algorithms used in each case are well-known (see Atkinson and Donev, 1992); it is the nature of the designs produced by these algorithms that is of particular interest here.

In Chapter 3, we found the likelihood and derivatives for a general form of the Bradley-Terry model. In this thesis, we consider only the simplest form of the model, in which there is no home field (order) effect and no tie parameter. In this case, the design space consists of the set of $t(t-1)/2$ different games between the t teams, and a design consists of this list of games together with a value $w_{ij}, i, j = 1, \ldots, t, i \neq j$, that represents its frequency or relative frequency in the design. There is no longer any need to distinguish between, say,

i vs. j and j vs. i, so we adopt the convention that $w_{ji} = w_{ij}$. The notation of Chapter 3 simplifies: since $p_{ij0} = 0$ for all i, j, $p_{ij2} = 1 - p_{ij1}$ and $p_{ji1} = p_{ij2} = 1 - p_{ij1}$. Noting also that the design implies $w_{ij} = w_{ji} = y_{ij+} = y_{ji+}$, we find that the off-diagonal elements of the information matrix $M(\mathbf{u}, \beta)$ become

$$m_{ij} = -w_{ij}p_{ij1}(1 - p_{ij1}), \quad i \neq j,$$

and the diagonal elements become

$$m_{ii} = \sum_{j \neq i} w_{ij}p_{ij1}(1 - p_{ij1}).$$

The diagonal elements are such that the row sums of this matrix are all zero, so that M is rank-deficient by one (or more, if too many of the w_{ij} are zero). We cannot therefore immediately apply D-optimality, $\det M$ being zero. Two possible remedies are:

- Remove one row and column of M, producing a $(t-1) \times (t-1)$ matrix of full rank for all reasonable designs.
- Add a constant δ to the diagonal of M, which will increase all the eigenvalues by δ, including any of them that were zero, and thus render M positive definite.

Either course can be helpful in certain circumstances. In dealing with continuous designs, we know or suspect that the information matrix will be rank-deficient by exactly one, and so the first course will be easier to follow. However, with exact and sequential designs, it is generally easier to take the second course (choosing δ large enough to ensure that M is numerically positive definite but small enough not to interfere with the determination of the design), because in the early stages of design construction, we wish to be able to proceed sensibly even if the information matrix has multiple zero eigenvalues. The two approaches give answers that are "equivalent" in the following sense:

Theorem 4.1 *Let an $n \times n$ matrix A have eigenvalues $\lambda_1, \ldots, \lambda_n$ with $\lambda_i \geq 0$ for all i. Suppose that k eigenvalues of A are exactly zero. Then the determinant of the matrix obtained from A by deleting k linearly dependent rows and columns is approximately $\delta^{-k} \det(A + \delta I)$ for δ small.*

Proof: The matrix $A + \delta I$ has eigenvalues $\lambda_1 + \delta, \ldots, \lambda_n + \delta$, and these eigenvalues are all positive if $\delta > 0$. The determinant of $A + \delta I$ is $\prod_{i=1}^{n}(\lambda_i + \delta)$, which can be written as the

sum of $\prod_{i=1}^{n} \lambda_i$, δ times the sum of all products containing $n-1$ λ_i, ..., δ^k times the sum of all products containing $n-k$ λ_i, ..., and δ^n times the sum of the λ_i. Thus if A contains k zero eigenvalues, $\det(A+\delta I) = \delta^k \prod_{i:\lambda_i \neq 0} \lambda_i$ plus terms containing higher powers of δ. Neglecting these terms, we see that this determinant contains the product of the non-zero λ_i values.

Now, the matrix A is of rank $n-k$, and thus has k rows that can be expressed as linear combinations of other rows of A. Deleting these rows and their corresponding columns yields a non-singular matrix A' with non-zero eigenvalues the same as those of A. (This can be shown by first applying a linear transformation to set the elements of the rows and columns concerned to zero, then noting that zero rows and columns can be deleted to yield a matrix with the same non-zero eigenvalues.) Thus, provided higher powers of δ can be neglected, $\det A' = \delta^{-k} \det(A+\delta I)$, as we wished to prove.

4.3.2 Continuous designs when parameters known

The w_{ij} of a continuous design are the proportion of all games that are played between teams i and j, based on the assumption that all β_i and p_{ij1} are known. Since the w_{ij} of a D-optimal continuous design depend continuously on the β_i and p_{ij1}, it is possible to state general results which allow one to check the optimality of particular designs, and to use methods of continuous optimization to find optimal (continuous) designs. These issues are considered in Chapter 9 of Atkinson and Donev (1992).

For small t, the continuous D-optimal design can sometimes be found analytically. For example, consider a tournament with $t=3$ teams, for which $p_{121} = \frac{3}{4}$ and $p_{231} = \frac{2}{3}$. Since the Bradley-Terry model is based on additivity of log-odds, it follows that $p_{131} = \frac{6}{7}$. For a general design $\mathbf{u}$, it follows that $M(\mathbf{u})$ is

$$\begin{bmatrix} \frac{3}{16}w_{12} + \frac{6}{49}w_{13} & -\frac{3}{16}w_{12} & -\frac{6}{49}w_{13} \\ -\frac{3}{16}w_{12} & \frac{3}{16}w_{12} + \frac{2}{9}w_{23} & -\frac{2}{9}w_{23} \\ -\frac{6}{49}w_{13} & -\frac{2}{9}w_{23} & \frac{6}{49}w_{13} + \frac{2}{9}w_{23} \end{bmatrix}.$$

The determinant is zero, because the row and column sums are zero, but the determinant of the upper left 2×2 submatrix is $\frac{9}{392}w_{12}w_{13} + \frac{1}{24}w_{12}w_{23} + \frac{4}{147}w_{13}w_{23}$.

Maximizing the determinant subject to the constraint $w_{12} + w_{13} + w_{23} = 1$ yields the solution $w_{12} = 0.420, w_{13} = 0.146, w_{23} = 0.434$. Thus D-optimality tells us to play fewest

games between the teams 1 and 3 that are most different in strength, and most games between the teams 2 and 3 that are most evenly matched.

This is typical of D-optimal designs for Bradley-Terry models – the "information" in a game is $p_{ij1}(1 - p_{ij1})$, which is maximum when $p_{ij1} = \frac{1}{2}$, so that the optimal design tends to contain as many of these "informative" games as is consistent with being able to estimate all the team strengths accurately. One other noteworthy feature in general is that the teams do not play the same number of games: in this example, team 2 features in over 85% of the games, while team 1 appears in fewer than 57%. This is in contrast with the other tournament types we have considered; in both round-robins and Swiss tournaments, all teams play the same number of games.

If the teams are all assumed to be of equal strength, there should be no reason for the D-optimal design to call for any one pair of teams to meet more often than any other pair. We will show that this is indeed the case, and we will also derive a method for finding D-optimal designs in general for the Bradley-Terry model.

To do this, we review a result, known as the General Equivalence Theorem, that yields checkable conditions for D-optimality. We present the result in greater generality first.

Let $\Psi\{M(\mathbf{u})\}$ denote a convex function of the information matrix which is to be minimized. All of the optimality criteria above can be written in this way; for example, the D-optimality criterion can be taken as $-\log\det\{M(\mathbf{u})\}$. Let $\bar{\mathbf{u}}$ denote the design measure that puts unit mass at u, and let $\phi(x, \mathbf{u})$ denote the directional derivative of $\Psi\{M(\mathbf{u})\}$ in the direction $\bar{\mathbf{u}}$. The General Equivalence Theorem (as given in Atkinson and Donev, 1992) is then:

Theorem 4.2 *The following three statements are equivalent:*

1. *The design* $\mathbf{u}^*$ *minimizes* $\Psi\{M(\mathbf{u})\}$.

2. $\min \phi(x, \mathbf{u}^*) \geq 0$.

3. $\phi(x, \mathbf{u}^*)$ *achieves its minimum at the points of the design.*

Proof: See Silvey (1980, p. 22).

The nature of the theorem can be seen more easily in the case of D-optimality. Let p be the number of parameters being estimated. When

$$\Psi\{M(\mathbf{u})\} = -\log\det\{M(\mathbf{u})\},$$

it follows that $\phi(x, \mathbf{u}) = p - d(x, \mathbf{u})$, where $d(x, \mathbf{u})$ is the standardized variance of the predicted response at x (see Atkinson and Donev, 1992, p. 95). The three equivalent conditions can therefore be expressed:

1. $\mathbf{u}^*$ is D-optimal.
2. $\max d(x, \mathbf{u}^*) \leq p$.
3. $d(x, \mathbf{u}^*)$ achieves its maximum at the points of the design.

It is worth noting that, by this result, D-optimality is the same as a criterion, known as G-optimality, which tries to minimize the maximum prediction variance over the design space. This explains the name of the Theorem. The same result, however, does not hold for exact designs; see Atkinson and Donev (1992, p. 43–44) for an example.

These results can be used to check a candidate continuous design to see whether it is D-optimal or not. Typically, the maximum of the second condition is equal to p. It should be noted that this maximum is over the *entire* design space; it is not enough to check the standardized variance only at the points in the design.

For the Bradley-Terry model, with information matrix $M(\mathbf{u})$ for a continuous design $\mathbf{u}$ (evaluated at some fixed parameter value β), the standardized variance takes the simple form

$$d(i, j, \mathbf{u}) = p_{ij1}(1 - p_{ij1})(m^{ii} + m^{jj} - 2m^{ij}),$$

where m^{ij} denotes the (i, j)-th element of $\{M(\mathbf{u})\}^{-1}$, for a game between teams i and j.

The General Equivalence Theorem then asserts for the D-optimal design that

$$d(i, j, \mathbf{u}) = p_{ij1}(1 - p_{ij1})(m^{ii} + m^{jj} - 2m^{ij}) \leq t - 1$$

for all i and j. We can therefore check a candidate design for D-optimality by checking this condition. The following Theorem is a useful example.

Theorem 4.3 *When $p_{ij1} = p_{ij2} = \frac{1}{2}$ for all i and j, the D-optimal design sets $w_{ij} = 2/\{t(t-1)\}$ for all i and j.*

Proof: Since $p_{ij1}(1 - p_{ij1}) = \frac{1}{4}$ for all i, j, each off-diagonal element of the information matrix for the postulated design is $m_{ij} = -1/\{2t(t-1)\}$. The diagonal elements are the negative sums of the $t - 1$ other entries in each row (or column), and are therefore

$m_{tt} = 1/(2t)$. Applying the tactic of removing row and column t to produce a matrix M' that is non-singular, we can apply a well-known result to deduce the inverse of M', and thus obtain an "inverse" for M by appending a row and column of zeroes to M'^{-1}. This "inverse" has, for $1 \leq i, j \leq t-1$, $m^{ii} = 4(t-1)$ and $m^{ij} = 2(t-1)$ for $i \neq j$. Thus, for such i and j, $d(i,j,\mathbf{u}) = \frac{1}{4}\{4(t-1) + 4(t-1) - 2 \cdot 2(t-1)\} = t-1$, and also $d(i,t,\mathbf{u}) = \frac{1}{4}\{4(t-1)\} = t-1 = d(t,i,\mathbf{u})$. In other words, for each pair of teams i, j with $1 \leq i, j \leq t$, the standardized variance is exactly equal to $p = t-1$. Since the design space consists of all such pairs i, j, the General Equivalence Theorem can be applied, and we conclude that the hypothesized design is indeed D-optimal, completing the proof.

An interpretation of this result is that "equally matched teams should meet an equal number of times". In other words, if the w_{ij} are scaled up to an integer, a round-robin tournament is D-optimal when the teams are evenly matched. This adds a mathematical justification to the heuristic notion of round-robin tournaments being "sensible".

It also turns out to be possible, at least in certain circumstances, to obtain the D-optimal continuous design when the teams are not evenly matched. This is done by using the General Equivalence Theorem to show that the inverse of the information matrix for the D-optimal design must take a certain form, and showing how to extract the w_{ij} for a design from the information matrix. The details are given in the next two Theorems.

Theorem 4.4 *For any information matrix $M(\mathbf{u})$ that comes from a Bradley-Terry model, the continuous design that produced $M(\mathbf{u})$ has w_{ij} given by $w_{ij} = -m_{ij}/\{p_{ij1}(1-p_{ij1})\}$ for $i \neq j$.*

Proof: The proof is straightforward, since the only contribution to m_{ij} comes from games between teams i and j. In particular, if a continuous design calls for a fraction w_{ij} of the sampling effort to be applied to i vs. j, $m_{ij} = -w_{ij}p_{ij1}(1-p_{ij1})$. The result follows.

Theorem 4.5 *If the D-optimal design is such that all games i vs. j have weight $w_{ij} > 0$, then the upper left $(t-1) \times (t-1)$ submatrix of $M(\mathbf{u})$ for the D-optimal design has an inverse whose elements are*

$$m^{ii} = \frac{t-1}{p_{it1}(1-p_{it1})}$$

and

$$m^{ij} = \frac{m^{ii} + m^{jj} - (t-1)/\{p_{ij1}(1-p_{ij1})\}}{2}$$

for $i \neq j$.

Proof: The proof consists simply of showing that $d(i,j,\mathbf{u}) = t-1$ for all i and j based on this matrix $M(\mathbf{u})$. First, for $i,j \leq t-1$,

$$\begin{aligned} d(i,j,\mathbf{u}) &= p_{ij1}(1-p_{ij1})(m^{ii}+m^{jj}-2m^{ij}) \\ &= p_{ij1}(1-p_{ij1})\left\{m^{ii}+m^{jj}-m^{ii}-m^{jj}+\frac{(t-1)}{p_{ij1}(1-p_{ij1})}\right\} \\ &= t-1. \end{aligned}$$

Then

$$d(i,t,\mathbf{u}) = p_{it1}(1-p_{it1})m^{ii} = t-1.$$

The same argument holds for $d(t,i,\mathbf{u})$. Thus, without assuming that $w_{ij} > 0$ for all $i \neq j$, we have obtained the form of the inverse. However, Example 2 below illustrates that if $w_{ij} = 0$ for some i,j with $i \neq j$, the above procedure does not yield a sensible $M(\mathbf{u})$. Thus the condition is necessary, and the proof is complete.

We look at two examples to illustrate the above theory. For Example 1, we take $t = 6$ teams whose strengths are equally spaced on the logistic scale; specifically, suppose that $\beta_i - \beta_{i+1} = 0.2$ for $1 \leq i \leq 5$. This means that each team has probability approximately 0.55 of defeating its "immediate neighbour", and, by the Bradley-Terry model, team 1 will defeat team 6 with probability approximately 0.73. These are six reasonably well-matched teams, so we might expect that the theory above will apply and will yield a design with w_{ij} positive and not too dispersed. Table 4.1 shows the results. The theory has indeed worked correctly, but the optimal design shows that games between teams 1 and 2 (and teams 5 and 6) should occur almost three times as frequently as games between 1 and 6. In general, the design calls for games between close teams to be most frequent. Note, however, that games between teams 3 and 4 are relatively infrequent, because these two teams are already often compared indirectly, and therefore direct comparisons between them are not especially profitable.

For Example 2, we return to $t = 3$ teams, but space their strengths more widely; specifically, we let $\beta_1 = 3, \beta_2 = 1, \beta_3 = 0$. This means that $p_{121} = 0.88$ and $p_{231} = 0.73$ approximately. Applying the theory in this case, under the assumption that $w_{ij} > 0$ for all $i \neq j$, yields the absurd result $w_{12} = 0.9855, w_{23} = 1.2290, w_{13} = -1.2145$. While it is still true that the w_{ij} sum to 1, the value of w_{13} can only point to the fact that there are no games between teams 1 and 3 in the D-optimal design. With this and the symmetry of the problem in mind, it seems likely that the optimal design in fact places equal weight of 0.5

Table 4.1: Relative frequencies of games in Example 1

Game(s)	Rel. frequency
1-2, 5-6	.0909
1-3, 4-6	.0781
2-3, 4-5	.0717
1-4, 3-6	.0654
3-4	.0653
2-4, 3-5	.0646
2-5	.0581
1-5, 2-6	.0512
1-6	.0333

on the other two games. This design produces (to the accuracy shown)

$$M^{-1} = \begin{bmatrix} 26.827 & 13.4145 \\ 13.4145 & 13.4145 \end{bmatrix},$$

from which, by straightforward calculation, it follows that for this design $d(1,2,\mathbf{u}) = 2 = t-1 = d(2,3,\mathbf{u})$ and $d(1,3,\mathbf{u}) = 1.2126$. The General Equivalence Theorem then shows that this is indeed the D-optimal design, since the maximum value, 2, of $d(i,j,\mathbf{u})$ occurs at the games featured in the design; $d(1,3,\mathbf{u}) < 2$, but this does not matter since the design contains no games between teams 1 and 3.

For $t = 3$, assuming without loss of generality that $\beta_1 \geq \beta_2 \geq \beta_3$, we can show that either Theorems 4.5 and 4.4 yield a D-optimal design with all three possible games occurring, or (when application of these results yields $w_{13} \leq 0$), that the D-optimal design sets $w_{12} = w_{23} = 0.5, w_{13} = 0$. The latter case is easily demonstrated to be D-optimal when $w_{13} = 0$ by directly maximizing the determinant of the upper left 2×2 submatrix of M subject to $w_{13} = 0$ and $w_{12} + w_{23} = 1$. For larger values of t, however, it does not seem to be possible to make general statements about D-optimality when some of the possible games are missing from the optimal design.

For general t we can note that for games between i and j not in the design, $d(i,j,\mathbf{u}) < t-1$, so that there exists a value of $d(i,j,\mathbf{u})$ for which, if we calculate M^{-1} and M as described above, we will find that $m_{ij} = 0$ and hence that $w_{ij} = 0$ as required. If values $d(i,j,\mathbf{u})$ can be found for all i,j pairs for which i vs. j is not in the design in such a way that each $m_{ij} = 0$, then the General Equivalence Theorem tells us that the D-optimal design has

been found. For designs where the number n_e of games excluded is small, this leads to a practical method - a grid search of trial $d(i,j,\mathbf{u})$ values will reveal the values for which m_{ij} is approximately zero. For instance, in Example 2, reducing $d(1,3,\mathbf{u})$ to 1.6 while leaving $d(1,2,\mathbf{u}) = d(2,3,\mathbf{u}) = 2$ produces $m_{13} = 0.0002$, $w_{12} = 0.5017, w_{23} = 0.5008, w_{13} = -0.0025$, which strongly suggests that the design with $w_{12} = w_{23} = 0.5, w_{13} = 0$ is the D-optimal design, as we showed to be the case.

When n_e is larger, however, grid searches become impractical, and greater insight into the interdependence of $d(i,j,\mathbf{u})$ and the elements of $M(\mathbf{u})$ is needed.

4.3.3 D-optimal exact designs when parameters known

A D-optimal exact design $\mathbf{u}$ is one for which $\det M(\mathbf{u})$ achieves its maximum over the set of integer w_{ij}, where now $\sum_{i,j} w_{ij} = n$, the desired number of games in the tournament.

If n in a tournament is large, the most straightforward way to design the tournament is to find the D-optimal continuous design, using the methods of Section 4.3.2, and then to multiply the fractional w_{ij} thus obtained by n and round off to integers. For large n, the distortion induced by the rounding process will be slight, and we may confidently assert that this exact design is D-optimal or very close to it.

On the other hand, when n is small, the rounding process may induce sufficient distortion for there to exist a different exact design with noticeably larger $\det M$. In this case, we will wish to consider algorithms for producing exact designs directly, so as to avoid the rounding issue.

There are numerous algorithms available (Atkinson and Donev, 1992, ch. 15). They share a philosophy of maintaining a "current" design and adjusting it by adding or deleting design points, continuing until the design is of the right size and cannot be improved by exchanging one design point for another. Such algorithms are well suited for the Bradley-Terry model, where the set of candidate design points is discrete and finite; in our discussion below, we focus on application of these algorithms to the Bradley-Terry problem. For D-optimality, there are two other helpful facts. First, the greatest increase in $\det M(\mathbf{u})$ is obtained by adding the point where $d(i,j,\mathbf{u})$ is largest (and, correspondingly, the smallest decrease is obtained by removing the point where $d(i,j,\mathbf{u})$ is smallest). This is shown by (4.2) below. Second, formulas exist (eg. Atkinson and Donev, 1992, p. 170; Dennis and Schnabel, 1983, p. 188; Thisted, 1988, p. 117) to enable M, $\det M$ and M^{-1} to be updated without having to be recalculated from scratch each time a design point is added, removed

or exchanged with another point.

In dealing with exact optimality, it is usually easiest to work with a "regularized" version of M obtained by adding a small positive constant δ to each diagonal entry. As noted earlier, this does not, for small δ, affect the optimality or otherwise of designs, but in deriving exact designs, it is useful to have an information matrix that is guaranteed to be non-singular even when the number of points in the current design is very small. In the remainder of this section, we assume that the information matrix has been so regularized.

The simplest algorithm is known as the "forwards procedure". It starts by choosing one game at random, and thereafter adding a game i vs. j for which $d(i, j, \mathbf{u})$ is maximized for the current design (breaking ties at random), until the design contains the desired number n of games. It is known (Wynn, 1972) that this procedure produces designs which, as $n \to \infty$, converge to the D-optimal continuous design, so there is some hope that stopping when the design contains n games will yield a reasonably good, if perhaps not optimal, design.

Complementary to the forwards procedure is the "backwards procedure". A design containing $n_0 > n$ games is chosen (perhaps randomly), and the game in the current design for which $d(i, j, \mathbf{u})$ is minimum is removed. This removal process is continued until the design contains the desired n games.

In the same way that explanatory variables in a regression can be chosen "automatically" by forward selection, backward selection, or a stepwise procedure, the natural extension of these design algorithms is to allow games to be both added to and deleted from the design as the algorithm progresses. This is desirable in the design problem for the same reason as in the regression variable-selection problem: whether or not a game should be added to or removed from a design typically depends on which other games are also in the design. There are various ways in which this addition and removal can be handled, for example:

- Add the game for which d is largest for the current design, then remove the game for which d is smallest in the resulting design (Mitchell and Miller, 1970; Wynn, 1970), or do the removal first (Van Schalkwyk, 1971).

- Add, and then remove, more than one game at each iteration (Mitchell, 1974).

- Combine the addition and removal processes by considering all possible games i vs. j that could be added to the design and games k vs. l that could be removed (because they are currently in the design), and exchange the pair of games for which the increase in the determinant is largest (Fedorov, 1972, p. 164).

- Instead of finding the largest increase in $\det M$, choose pairs of games at random and exchange *any* games for which $\det M$ is increased (Cook and Nachtsheim, 1980).
- Speed up the Fedorov algorithm by considering only the K games with the largest values of d for addition and the L games with smallest values of d for deletion (Atkinson and Donev, 1992, p. 173).

In the same way that stepwise regression does not necessarily detect the best set of explanatory variables in a regression, these algorithms are not guaranteed to find a design with the largest possible $\det M$, since this is a combinatorial optimization problem (unlike the finding of a continuous D-optimal design, which is a continuous optimization problem), and the algorithms may find a local rather than a global maximum. Generally speaking, the above algorithms trade intelligence in adding and removing games for speed; the more sophisticated algorithms will tend to produce better designs, but will take longer to do so. Of course, the user is at liberty to run a faster algorithm repeatedly from different starting designs, and to take the best design generated by any of the runs; whether this is preferable to running a slower algorithm once will be problem-dependent.

Since the Bradley-Terry design problem seems to be a reasonably co-operative one, we do not carry out a comparison of algorithms here; the Mitchell-Miller-Wynn algorithm, re-run a number of times and the best design chosen from these runs, seems to work well at a reasonable speed and without undue computational complexity. Our implementation of the algorithm follows these steps to produce a design with n games:

1. Set $M^{-1} = I/\delta$ for some small δ, and set $\det M = \delta^t$. (This corresponds to an initial $M = \delta I$, so that the matrix M is the Fisher information with δI added.)
2. Select n games at random (with equal probability and with replacement) from the set of possible games, and add them to the design, keeping track of M^{-1} and $\det M$ as each game is added.
3. Save the current value of $\det M(\mathbf{u})$.
4. Find a game i vs. j for which $d(i, j, \mathbf{u})$ is maximized (in the case of a tie, choose one game at random or arbitrarily) and add it to the design, updating M^{-1} and $\det M$.
5. Find a game k vs. l currently appearing at least once in the current design for which $d(k, l, \mathbf{u})$ is minimum, breaking ties as in the previous step. Remove (one instance of)

this game from the design, and "downdate" M^{-1} and $\det M$.

6. If the current value of $\det M$ is no bigger than the saved value, then stop. (In practice, one compares the increase in the logarithm of the determinant against a small tolerance such as 10^{-4}.) Typically, the algorithm will stop with the same game being added and then removed. Otherwise, if $\det M$ is still increasing, go back to step 3.

There are ways to select the initial n-game design other than that given in Step 2. Our choice, it is hoped, will allow the algorithm to explore different parts of the design space on different runs, and therefore should have an improved chance of finding a D-optimal design. Other possibilities are to use either the forwards algorithm or the scaled-up continuous D-optimal design to generate the initial n-game design, which gives a better initial design at the expense of allowing the algorithm to explore less of the design space, or a compromise version of this in which $n_0 < n$ games are generated in this way with the remaining $n - n_0$ games of the initial design generated at random.

As noted earlier, there are also different ways to implement the instruction to update M^{-1} and $\det M$. Note that the effect on M of adding a game i vs. j to the design is to add the quantity $p_{ij1}(1 - p_{ij1})$ to m_{ii} and m_{jj} and to subtract the same quantity from m_{ij} and m_{ji}. If x_{ij} denotes the vector with 1 in the i-th position, -1 in the j-th, and zeroes elsewhere, this means that M becomes $M + p_{ij1}(1 - p_{ij1})x_{ij}x'_{ij}$, or, writing $v = \sqrt{p_{ij1}(1 - p_{ij1})}x_{ij}$, M becomes $M + vv'$. This is a rank-one update to M, as is the case generally when adding points to a regression; so is the update connected with removing a game from the design, since M becomes $M - vv'$.

We have chosen to use the Sherman-Morrison-Woodbury formula to M^{-1} without recalculating the inverse from scratch. There is also a corresponding formula for the update of the determinant. They are given in Dennis and Schnabel (1983, p. 188) and Thisted (1988, p. 117), and are:

$$\begin{aligned}(M \pm vv')^{-1} &= M^{-1} \mp \frac{(M^{-1}v)(M^{-1}v)'}{1 \pm v'M^{-1}v}, \\ \det(M \pm vv') &= (1 \pm v'M^{-1}v)\det M.\end{aligned}$$

In our case, some further simplifications are possible. Let $t = M^{-1}x_{ij}$; the nature of x_{ij} implies that $t = m^i - m^j$, where m^l denotes column l of M^{-1}. The calculation of $v'M^{-1}v$ then has only four nonzero terms based on $m^{ii}, m^{jj}, m^{ij}, m^{ji}$; in fact, $v'M^{-1}v = d(i, j, \mathbf{u})$.

Thus, letting $\sigma = 1 \pm d(i, j, \mathbf{u})$, the updates become

$$(M \pm vv')^{-1} = M^{-1} \mp \frac{p_{ij1}(1 - p_{ij1})}{\sigma} tt', \tag{4.1}$$

$$\det(M \pm vv') = \sigma \det M. \tag{4.2}$$

We have chosen to implement this update directly because of its simplicity. Dennis and Schnabel (1983) note that this is not the most numerically stable way to proceed, especially if M^{-1} becomes nearly singular during the course of the computation, which can happen, especially when removing games from the design (Thisted, 1988). By only ever removing points from our designs when they reach size $n + 1$, we hope to avoid the worst of these problems, although we have coded the algorithm in double precision as a precaution. A more numerically stable alternative is to update a factorization (such as Cholesky or QR) of either M or M^{-1}; Thisted (1988, p. 118) gives an example and some references.

One way to check the numerical quality of the algorithm using the Sherman-Morrison-Woodbury formulas is to take the purported M^{-1} matrix of a final design, invert it accurately, and compare with the M that would have been calculated directly from the design. As an example, consider an experiment with $t = 3$, $\beta_1 = \beta_2 = \beta_3 = 0$, $n = 10$. The D-optimal design consists of two of the three games played three times and the other game played four times (by symmetry, it does not matter which). In our case, it happened that 1 vs. 3 was played four times. The correct information matrix is therefore, using $\delta = 0.01$ and displaying the upper triangle only:

$$M = \begin{bmatrix} 1.76 & -0.75 & -1.00 \\ & 1.51 & -0.75 \\ & & 1.76 \end{bmatrix}$$

while the inverted M^{-1} from the algorithm, implemented on an IBM 386SX with numeric coprocessor in Turbo Pascal, came out to be

$$\begin{bmatrix} 1.77101 & -0.75424 & -1.00677 \\ & 1.51848 & -0.75424 \\ & & 1.77101 \end{bmatrix}.$$

There is clearly some degradation of accuracy, though not, in this case, nearly enough to affect the optimality or otherwise of designs. This, however, was a reasonably co-operative case; when the D-optimal design calls for some matches to be played many more times than

Example no.	No. of games	Strength vector β
1	24	$(0,0,0,0,0,0)$
2	30	$(1.0,0.8,0.6,0.4,0.2,0.0)$
3	30	$(2.5,2.0,1.5,1.0,0.5,0.0)$
4	30	$(2.5,1.5,1.3,0.9,0.2,0.1)$

Table 4.2: Examples for exact design algorithm

others, it may take a large number of additions to and deletions from the initial random design in order to obtain an optimal design. In such cases, one might expect more serious numerical difficulties, although even then, the likelihood of being unable to find a D-optimal design as a result seems slight.

In Section 4.3.2, we noted some properties of D-optimal continuous designs, and saw in general that more games need to be played between teams of similar strength than between those of dissimilar strength, subject to the demands of an overall level of comparison between the teams. Likewise, we found that if the teams were all of equal strength, then they should play each other an equal number of times. The picture is similar for exact designs, though the effect of requiring the w_{ij} to be integral is serious for small n.

We now consider some examples. While these examples all feature $t = 6$, for ease of comparison between them, what is observed is, in our experience, similar for all numbers of teams. Table 4.2 shows the true team strengths, arranged in a vector β, as well as the number of games the design should contain. The best designs found by the algorithm are shown in tables 4.3–4.6. In each case, the algorithm was run 20 times and the best design (ranked by calculated $\det M$) was chosen. The designs are shown as grids where, for example, the number in the first row and fourth column indicates the number of times teams 1 and 4 will meet. The grids are symmetric, but giving the whole grid makes it easier to judge the opponents that will be faced by a particular team, as well as to count up the total number of games played by each team in a particular exact design.

As a prelude to Example 1, it is worth noting that when all the teams are of equal strength and the desired number of games is a multiple of $t(t-1)/2$, the algorithm will produce a round-robin as the optimal design without any trouble; indeed, in re-running Example 1 with 30 games instead of 24, the algorithm produced a double round-robin design on each of its 20 runs.

For Example 1, therefore, it is of interest to see what kinds of design are optimal when

–	1	1	2	2	2
1	–	1	2	2	2
1	1	–	2	2	2
2	2	2	–	1	1
2	2	2	1	–	1
2	2	2	1	1	–

Table 4.3: Optimal design for Example 1

–	3	2	2	2	1
3	–	2	2	1	2
2	2	–	2	2	2
2	2	2	–	2	2
2	1	2	2	–	3
1	2	2	2	3	–

Table 4.4: Optimal design for Example 2

–	4	3	1	0	0
4	–	3	3	0	0
3	3	–	2	3	1
1	3	2	–	3	3
0	0	3	3	–	4
0	0	1	3	4	–

Table 4.5: Optimal design for Example 3

–	4	3	1	0	0
4	–	2	3	1	1
3	2	–	2	2	2
1	3	2	–	3	2
0	1	2	3	–	4
0	1	2	2	4	–

Table 4.6: Optimal design for Example 4

the number of games is not a multiple of $t(t-1)/2$. Simply allocating the nine leftover games from a single round-robin at random is not good enough; the optimal designs have structure. This structure varies according to the number of teams and the number of games, of course, but takes a particularly interesting form for $t = 6, n = 24$ as shown in Table 4.3. The teams are split into two groups (in this case the first three teams and the last three), and the additional games after the single round-robin are all those featuring a team from one group against a team from the other. This means that, amongst the additional games, the teams within a group do not play against each other, but they have three common opponents, namely the teams in the other group, and so within-group comparisons can still be made with high precision.

For the remainder of the examples, n was taken equal to 30, to facilitate comparison between the designs produced and the double round-robin design with the same number of games. In Example 2, the teams are somewhat closely matched: a difference of 0.2 on the β scale corresponds to a probability of close to 0.55. Even so, the probability of the best team defeating the worst in this example is still over 0.73. Nonetheless, Table 4.4 indicates that the optimal design is close to a round-robin, with only one extra game between the two best and the two worst teams, and one fewer instance of two of the less evenly-matched games. This indicates, as we investigate further in Section 4.4.2, that round-robin designs have a certain amount of robustness to unequal strengths of the teams involved; however, the next Example shows that this robustness extends only so far.

In Example 3, the gap in probability terms between neighbouring teams is about 0.62, and between best and worst is about 0.92. Now, Table 4.5 shows that the optimal design is anything but balanced, with a large majority of games being between teams that are close together in strength. The relationship is not monotonic, however – for example, teams 3 and 4 play only twice – but this is illustrative of the need for the design to be balanced enough to provide good estimation of the relative strengths of all the teams. Here, teams 3 and 4 are compared well by the games between them and teams 2 and 5, so that additional games between the two teams are not necessary.

The last example had a pattern to it which seemed to be the result of happenstance with the number of games and the equal spacing of the team strengths. In Example 4, we space the team strengths irregularly, with one clearly strongest team and two weak teams that are close together. Table 4.6 shows that the resulting optimal design has no particular pattern, other that previously noted of generally calling for more games between teams closer in

Table 4.7: Summary of designs found by the algorithm in the Examples

Ex.	Runs	Times found Opt. design	Designs	log det M High	Low
1	20	7	2	0.7852	0.7821
2	20	20	1	1.7943	1.7943
3	20	7	2	0.7039	0.7022
4	20	6	6	0.7275	0.7228

strength.

This Example and the previous one also show that, when the teams vary widely in strength, there is no restriction on the teams each playing the same number of games; in these Examples, the total numbers of games vary from 8 to 12, with the middle-strength teams playing the most games and the strongest and weakest teams playing the fewest. This phenomenon occurs generally with D-optimal designs (it could also be observed in the continuous designs of Section 4.3.2), in contrast to round-robins and Swiss tournaments where each team plays the same number of games.

Another issue of interest is the variety of supposed "optimal" designs found by different runs of the same algorithm on the same design problem. Table 4.7 shows the results for our four examples. In no case was a seriously sub-optimal design found, and on these examples, the algorithm found the best design frequently enough to offer convincing evidence that it indeed *is* the best design. The algorithm is not always as convincing: runs with the parameters of Example 2, but with $n = 15$ instead of 30 yielded the (apparently) optimal design only about once every 50 runs, amid a large variety of other designs. These experiences indicate that the likelihood of finding the optimal design and the number of different designs generated by the algorithm are very problem-dependent. In practice, the only advice that can be given is to run the algorithm a few times, then look at the values of $\det M$ and decide whether a maximum seems to have been attained. Sometimes, as when most or all of the values are the same, this decision is easy; otherwise, a decision has to be made about whether it is worth running the algorithm again. Some optimal designs are simply harder to find than others, although it is some comfort to know that the algorithm is unlikely ever to find a seriously sub-optimal design.

4.3.4 *D*-optimal sequential designs

All of the foregoing has assumed that the team strength parameters β_i are known, an assumption that is highly unrealistic in practice. We have also seen that the designs generated from known β_i are quite strongly dependent on the values of the β_i; in other words, the optimal design for one set of values β_i can be far from optimal for a different set of β_i. Designing an entire tournament based on possibly bad guesses of the β_i is therefore hazardous, and it is natural to proceed sequentially: design a small subtournament based on current knowledge about the β_i, run this subtournament, use the results to improve knowledge about the β_i, and then repeat as desired. The final tournament design is then obtained by combining the subtournaments.

Let s denote the number of "stages", that is, subtournaments, in the sequential tournament, and let n_j denote the number of games in stage j, with $n = \sum_{i=j}^{s} n_i$. For simplicity, we will assume that s and the n_j are known before the tournament begins. Typically, one might take all the n_j equal. Choice of the value or values of the n_j might be made from cost considerations, balancing the cost of stopping the tournament to re-estimate the β_i and to design the next stage with the benefit of possessing the most accurate estimates of the β_i at all times. An obvious way to proceed is then as follows:

1. Set $j = 1$.
2. Assuming that all β_i are equal (or, perhaps, obtaining values for the β_i from prior knowledge), use the algorithm of Section 4.3.3 to obtain a D-optimal design with n_1 games.
3. Play the games in stage j.
4. Estimate the β_i based on the game results from the current design, by maximum likelihood, using the methods of Chapter 3.
5. If $j = s$, stop; otherwise, continue.
6. Add 1 to j.
7. Use a modified version of the algorithm of Section 4.3.3 to find the n_j games which, when added to the games in the current design, produce a D-optimal $(\sum_{k=1}^{j} n_k)$-game design. (The modification required is that the algorithm does not delete games that have already been played.)

8. Go back to step 3.

There is nothing new here. Indeed, the above is merely a slight generalization of an algorithm given in Silvey (1980, p. 62). Interest therefore centres mostly in the performance of the algorithm in practical cases. However, two issues arise immediately, one technical and one practical.

The technical issue is one that afflicts all sequential experimentation, where the choice of experimental conditions for one data point is dependent on previous observations. The likelihood itself is unaffected by the kind of experimentation performed, but repeated-sampling inferences based on the likelihood are certainly affected, because a repeated sample would consist of observations at different data points, not merely different observations at the same data points, as would happen if the design were fixed. In our context, two sequentially-designed tournaments with an identical set of teams will consist of two different sets of games, so that inference about the team strengths has to consider not only the variability in results of particular games, but also the variability in designs in repeated samples. It is natural to hope that inferences which ignore the sequential nature of the design will be approximately correct, but this hope should be supported by simulated repetitions of the entire sequentially-designed structure. We do not pursue this idea here, however.

Given all the above, it is natural to ask what it is that a D-optimal sequential design is optimizing. The answer seems to be that, at each stage, it is producing the D-optimal design conditional on the current estimates of the β_i being correct. This does not sound especially compelling; on the other hand, the algorithm given above does seem sensible on practical grounds, and therefore its effectiveness in practice is worth investigating.

When we were looking at the designs generated in Sections 4.3.2 and 4.3.3, we found that some of the designs resembled a round-robin tournament (in particular, those where the β_i did not vary widely). In Chapter 2, we also looked at the Swiss tournament, which is constructed sequentially by pairing teams of similar apparent strength, subject to the constraint that two teams may not meet more than once. We will therefore investigate similarities between the structure of sequential D-optimal designs and Swiss tournaments.

In our Example below, and also in Section 4.4, we concentrate on what might be called a "sequential-1" design, where only one game is selected and played at each stage. In our notation, this sets $s = g$, the total number of games, and $n_j = 1$ for all j. By re-estimating the team strengths β_i as often as possible, we should be making the best use of

the information collected throughout the tournament, and therefore should be constructing the best possible sequential D-optimal design. Sequential-1 designs thus give us an idea of the potential of sequential designs in general. Note that the algorithm above simplifies when $n_j = 1$, since step 7 consists only of finding the game i vs j for which $d(i, j, \mathbf{u})$ is greatest, based on the current design, and then adding this game to the design.

Let us examine an example of a sequential-1 design. In this example, $t = 6$, and the true team strengths are all equal, although of course this affects only the simulated game results and not the design procedure. The estimation of the team strengths β_i is carried out using the device of a "fictitious team", as described in Chapter 3, so that the estimated team strengths are always finite. We begin the procedure by setting $M(\mathbf{u}) = \delta I$ for some small δ, and, whenever a tie exists for the best game to add, we choose one of the games involved at random.

Not surprisingly, the first three games ensure that all of the six teams appear in the design. Our algorithm produced the games 1 vs 2, 3 vs 4, 5 vs 6, with teams 1, 4 and 6 winning. (By symmetry, any other pairing involving all six teams is equally good.) This yields estimated strengths of 0.76 for the winning teams and -0.76 for the losers.

As in a Swiss tournament, the next two games have two of the winners and two of the losers play each other. In our case, the games were 1 vs 4 and 2 vs 5, with 1 and 5 winning. Now we see our first divergence from a Swiss tournament: there, teams 3 and 6 would meet to complete the second round. Here, however, there is less information to be gained from playing a game between these two teams than there is from playing 1 vs 6, the two undefeated teams, and 2 vs 3, the two winless teams. In these games, teams 6 and 3 were the winners.

At this point, the design is

$$\begin{array}{cccccc} - & 1 & 0 & 1 & 0 & 1 \\ 1 & - & 1 & 0 & 1 & 0 \\ 0 & 1 & - & 1 & 0 & 0 \\ 1 & 0 & 1 & - & 0 & 0 \\ 0 & 1 & 0 & 0 & - & 1 \\ 1 & 0 & 0 & 0 & 1 & - \end{array}$$

with the estimated team strengths $\hat{\beta}$ being

$$(0.69, -2.01, -0.61, 0.03, -0.03, 1.87).$$

Note that team 2 has played the next two weakest teams, according to this estimation, and team 6 has played the second and fourth strongest teams. Despite team 6 having played only twice, this team does not feature in any of the next five games, and then team 6 plays team 1 again, losing this time. After the next game, 4 vs 5, with 5 winning, the design is the following:

-	1	0	1	1	2
1	-	2	1	1	0
0	2	-	1	1	0
1	1	1	-	2	0
1	1	1	2	-	1
2	0	0	0	1	-

with $\hat{\beta}$ being

$$(1.67, -1.70, -1.68, -0.51, 0.60, 1.56).$$

At this point, the teams seem to have split themselves into two groups, even though in truth they are all of equal strength. The next six games are all "within-group", but eventually 1 plays 3 twice consecutively, losing both, and $\hat{\beta}$ becomes

$$(0.09, -1.55, -0.02, -0.50, 0.55, 1.34),$$

based on the design

-	1	2	1	2	3
1	-	3	2	1	0
2	3	-	2	1	0
1	2	2	-	2	0
2	1	1	2	-	2
3	0	0	0	2	-

which is far from balanced. Indeed, continuing until 100 games have been played does not yield a design that looks noticeably more balanced:

-	4	4	9	9	8
4	-	14	5	3	6
4	14	-	5	3	4
9	5	5	-	9	8
9	3	3	9	-	9
8	6	4	8	9	-

with $\hat{\beta}$ of

$$(0.42, -0.92, -1.08, 0.55, 0.62, 0.30).$$

We see that a sequential-1 design can be noticeably sub-optimal, in that too few games are played between teams that are actually equal in strength but seem to be far apart (such as 2 vs 5, 3 vs 5). It is also worth noting that the sequential-1 strategy of playing games early in the tournament between teams with similar numbers of wins and losses guarantees that, after these games, there will be one team that looks very strong and another that looks very weak, even if the truth is otherwise.

This example suggests that sequential-1 designs will not be optimal when the teams are close together in strength. On the other hand, when the teams vary widely in strength, the teams that win their first few games in a sequential-1 tournament are likely to be the strongest, so that later games will concentrate on teams of similar (true) strength. In this case, we might expect a sequential-1 design to be nearly D-optimal, especially as the number of games increases.

4.4 Efficiency comparisons of designs

4.4.1 Introduction

In Chapter 2, we examined round-robin and Swiss tournaments, and saw that they behaved in a reasonably intuitive fashion. In this Chapter, we have seen that a round-robin tournament is D-optimal if the teams are of equal strength, and we have also seen that the guiding principle of the Swiss tournament, namely to pair teams of similar strength provided that they have not met too many times before, is similar to the construction of a D-optimal design. It is natural then to define a measure of "efficiency" and to see how the tournament designs compare.

Atkinson and Donev (1992) define the D-efficiency of a design $\mathbf{u}_1$ relative to $\mathbf{u}_2$ as

$$D_{\text{eff}} = \left(\frac{\det M(\mathbf{u}_1)}{\det M(\mathbf{u}_2)} \right)^{1/(t-1)}, \tag{4.3}$$

where $t-1$ is the number of estimable parameters in the Bradley-Terry model. To see that extracting the $(t-1)$-th root is the appropriate scaling, consider two designs $\mathbf{u}_1$ and $\mathbf{u}_2$ for which the D-efficiency of $\mathbf{u}_1$ relative to $\mathbf{u}_2$ is 0.5 for the same number of games. Replicating the design $\mathbf{u}_1$ has the effect of doubling $M(\mathbf{u}_1)$, which means that its eigenvalues are also

doubled. This increases $\det M(\mathbf{u}_1)$ by a factor of 2^{t-1}, since in general $t-1$ of the eigenvalues are non-zero; therefore, from (4.3), the D-efficiency of the replicated $\mathbf{u}_1$ relative to $\mathbf{u}_2$ is 1. Like other measures of efficiency, the D-efficiency can therefore be interpreted as the ratio of sample sizes necessary to estimate the parameters with equal precision from the two designs.

The definition (4.3) can be applied to any designs, though a typical application has $\mathbf{u}_2$ as a D-optimal design (continuous, exact, sequential) and $\mathbf{u}_1$ as some other design.

We now look at the different designs we have seen, and see how efficient they are as the number of teams and variability in strength of the teams changes. A natural measure of the variability is the standard deviation σ of the β_i; we use this as our parameterization of the variability of team strength. In our calculations, we assume that $\beta_i = F^{-1}\{i/(t+1)\}$ where $F(x)$ is the cumulative distribution function of a logistic distribution with mean 0 and scale parameter d. This assumption was made because it mimics the pattern of team strengths often found in practice, with many teams of similar strengths and a few teams noticeably stronger or weaker than the rest. The scale parameter d can be chosen to provide greater or lesser overall variability in the team strengths. For any fixed t, there is a simple relationship between σ and d, namely

$$\sigma^2 = \frac{d^2}{t}\sum_{i=1}^{t}\left\{\log\left(\frac{t+1-i}{i}\right)\right\}^2.$$

In our calculations, we choose d such that $\sigma = 0, 0.2, 0.4, 0.6, 1, 2$, values again intended to be illustrative of what occurs in practice.

Other methods of measuring team strength variability are possible. A rather more intuitive measure is the expected fraction w of wins by the strongest team when playing the remaining teams once each; the larger this "expected winning percentage", the greater the variability in team strengths. It turns out that w, for fixed σ, is only weakly dependent on t (it has a limit as $t \to \infty$), and so the categorization of variability in strength can be viewed in terms of w as well as in terms of σ, as shown in Table 4.8.

4.4.2 Efficiency of round-robin tournaments

The Examples of Section 4.3.3 showed that when the teams are of equal strength, the round-robin design was D-optimal, but the D-optimal designs in other cases bore little resemblance to round-robins. We would therefore expect round-robin tournaments to be fully efficient when $\sigma = 0$, but for the D-efficiency to drop off fairly quickly as σ increases.

		σ					
		0	0.2	0.4	0.6	1	2
	6	0.500	0.590	0.671	0.740	0.836	0.941
	10	0.500	0.594	0.678	0.748	0.843	0.943
	14	0.500	0.598	0.685	0.756	0.852	0.948
t	20	0.500	0.603	0.694	0.766	0.862	0.954
	30	0.500	0.610	0.705	0.780	0.876	0.962
	40	0.500	0.615	0.713	0.790	0.885	0.967
	50	0.500	0.619	0.720	0.798	0.893	0.971

Table 4.8: Values of w in terms of t and σ

It is straightforward to calculate $\det M$ directly for round-robin designs, and these values are then compared with the best design containing the same number of games as found by the Mitchell-Miller-Wynn algorithm. This was done for values of t between 6 and 20, with the results shown in Table 4.9.

We see that our expectations were justified; the second-to-last column of the Table shows that once the teams differ too much in strength, round-robin tournaments are very inefficient. This is because, in a round-robin, too much experimental effort is devoted to games which are "foregone conclusions", in that p_{ij1} is very close to 0 or 1 and the result of the game gives very little information.

The results shown in Table 4.9 are for single round-robins, in which the teams play each other only once, but r-tuple round-robins tell exactly the same story. Indeed, the D-efficiencies are almost exactly the same, because M for an r-tuple round-robin is exactly r times that for a single round-robin, and the D-optimal exact design for the larger tournament has w_{ij} close to r times those for the smaller (though not exactly equal because of requiring an exact design with integer w_{ij}).

4.4.3 Efficiency of Swiss tournaments

One might expect Swiss tournaments to have higher D-efficiency when the teams are more different in ability, since the structure of the tournament means that teams of approximately equal strength will tend to play each other, at least in rounds after the first.

To assess this, Swiss tournaments were simulated for $t = 6, 10, 14, 20, 30, 40, 50$. The number of rounds in each tournament was $3, 6, 7, 9, 10, 11, 12$ respectively; these are approximately $2\log_2 t$, and might be expected to be typical numbers of rounds for Swiss

Table 4.9: D-efficiency of round-robin designs

			log det M			D-efficiency	
Teams	Games	σ	RR	Seq.-1	D-opt.	RR	Seq.-1
6	15	0.0	-2.545	-2.828	-2.545	1.000	0.945
6	15	0.2	-2.661	-2.954	-2.661	1.000	0.943
6	15	0.4	-2.993	-3.243	-2.989	0.999	0.950
6	15	0.6	-3.486	-3.681	-3.349	0.973	0.936
6	15	1.0	-4.750	-4.635	-4.126	0.883	0.903
6	15	2.0	-8.240	-7.304	-5.947	0.632	0.762
10	45	0.0	3.677	3.155	3.677	1.000	0.944
10	45	0.2	3.482	2.973	3.482	1.000	0.945
10	45	0.4	2.936	2.543	3.046	0.988	0.946
10	45	0.6	2.131	1.892	2.547	0.955	0.930
10	45	1.0	0.105	0.550	1.528	0.854	0.897
10	45	2.0	-5.259	-2.548	-0.861	0.613	0.829
14	91	0.0	11.718	11.025	11.718	1.000	0.948
14	91	0.2	11.444	10.757	11.450	1.000	0.948
14	91	0.4	10.682	10.159	10.912	0.982	0.944
14	91	0.6	9.566	9.368	10.309	0.944	0.930
14	91	1.0	6.792	7.720	9.118	0.836	0.898
14	91	2.0	-0.440	4.231	6.326	0.594	0.851
20	190	0.0	26.012	25.061	26.012	1.000	0.951
20	190	0.2	25.620	24.767	25.649	0.998	0.955
20	190	0.4	24.532	23.941	24.970	0.977	0.947
20	190	0.6	22.960	23.036	24.245	0.935	0.938
20	190	1.0	19.054	21.202	22.825	0.820	0.918
20	190	2.0	9.037	16.895	19.547	0.575	0.870

tournaments for these numbers of teams. For each combination of t and σ, 100 tournaments were simulated, and $\log\det M$ calculated for each based on the true β_i. The mean log-determinant is shown in Table 4.10, along with the "exact" log-determinant for the D-optimal exact design with the same number of games. The standard deviation of values of $\log\det M$ from the simulated Swiss tournaments (not shown) increased dramatically with σ. This means that the relative efficiencies are not accurately determined for larger values of σ.

For $\sigma = 0$, we see from the second-to-last column of Table 4.10 that the Swiss tournaments are essentially fully efficient relative to the exact D-optimal design. As with round-robin tournaments, the D-efficiency drops off rapidly as σ increases; however, the initial rate of decrease is smaller, and in none of the cases shown in the Table is the D-efficiency truly small. The inaccuracy in estimation of the "true" D-efficiencies for larger values of σ does not have a serious effect, since it is clear that the D-efficiency in these cases is decreasing with σ.

4.4.4 Efficiency of sequential-1 designs

As with the Swiss tournaments, we can only assess the efficiency of the sequential-1 designs by simulation. For ease of comparison, we investigated designs with the same numbers of games as for the round-robin and Swiss tournaments considered in the previous two Sections, although of course any number of games is possible. In each case, 100 simulations were run and the average log-determinant calculated. (It is worth noting that an alternative approach, for the simulated Swiss tournaments as well as here, would be to calculate the "average information matrix" over all the simulations, since this is then an estimate of the observed information matrix even though the tournaments are designed sequentially and therefore the information matrix is not constructed from independent observations.)

The results of the simulations appear in Tables 4.9 and 4.10, for comparison with single round-robin and Swiss tournaments respectively. The last column in each table shows the D-efficiency of the sequential-1 designs relative to the exact D-optimal design for the same number of games. In both cases, we see that the D-efficiency for the sequential-1 designs, as for the round-robin and Swiss tournaments, is close to 1 when $\sigma = 0$ and drops off slowly as σ increases. The reason appears to be that when σ is large, some games are played at the beginning of the sequential-1 tournament that later turn out to be between teams of widely different strength, so that in retrospect it would have been more informative to play

Table 4.10: D-efficiency for Swiss tournaments

			log det M			D-efficiency	
Teams	Games	σ	Swiss	Seq.-1	D-opt.	Swiss	Seq.-1
6	9	0.0	-5.350	-5.343	-5.291	0.988	0.990
6	9	0.2	-5.459	-5.453	-5.389	0.986	0.987
6	9	0.4	-5.760	-5.758	-5.670	0.982	0.983
6	9	0.6	-6.267	-6.273	-6.004	0.949	0.948
6	9	1.0	-7.526	-7.530	-6.646	0.839	0.838
6	9	2.0	-10.930	-10.912	-8.582	0.625	0.628
10	30	0.0	-0.184	-0.185	-0.167	0.998	0.998
10	30	0.2	-0.379	-0.377	-0.322	0.994	0.994
10	30	0.4	-0.886	-0.911	-0.683	0.978	0.975
10	30	0.6	-1.593	-1.674	-1.149	0.952	0.943
10	30	1.0	-3.494	-3.418	-2.109	0.857	0.865
10	30	2.0	-8.660	-8.497	-4.507	0.630	0.642
14	49	0.0	3.285	3.271	3.352	0.995	0.994
14	49	0.2	2.997	2.995	3.122	0.990	0.990
14	49	0.4	2.271	2.286	2.683	0.969	0.970
14	49	0.6	1.250	1.306	2.142	0.934	0.938
14	49	1.0	-1.249	-1.093	1.068	0.837	0.847
14	49	2.0	-7.305	-7.131	-1.711	0.650	0.659
20	90	0.0	11.283	11.260	11.381	0.995	0.994
20	90	0.2	10.885	10.889	11.087	0.989	0.990
20	90	0.4	9.889	9.856	10.531	0.967	0.965
20	90	0.6	8.557	8.689	9.881	0.933	0.939
20	90	1.0	5.259	5.526	8.586	0.839	0.851
20	90	2.0	-1.648	-1.839	5.356	0.692	0.685
30	150	0.0	22.033	22.043	22.194	0.994	0.995
30	150	0.2	21.478	21.460	21.799	0.989	0.988
30	150	0.4	19.983	19.944	21.136	0.961	0.960
30	150	0.6	17.949	17.970	20.359	0.920	0.921
30	150	1.0	13.744	13.962	18.804	0.840	0.846
30	150	2.0	4.134	4.343	15.128	0.684	0.689
40	220	0.0	34.589	34.594	34.760	0.996	0.996
40	220	0.2	33.823	33.809	34.280	0.988	0.988
40	220	0.4	31.882	31.899	33.502	0.959	0.960
40	220	0.6	29.468	29.550	32.620	0.922	0.924
40	220	1.0	24.060	24.374	30.878	0.840	0.846
40	220	2.0	12.053	12.350	26.789	0.685	0.691
50	300	0.0	48.698	48.682	48.879	0.996	0.996
50	300	0.2	47.726	47.739	48.334	0.988	0.988
50	300	0.4	45.420	45.423	47.443	0.960	0.960
50	300	0.6	42.597	42.457	46.465	0.924	0.921
50	300	1.0	36.038	36.165	44.551	0.841	0.843
50	300	2.0	22.245	22.389	40.075	0.695	0.697

another game instead. However, of course, these games need to be played first in order to learn which teams are strong or weak.

It is of interest to see how the round-robin and Swiss tournaments compare in efficiency with the sequential-1 tournaments, since the sequential-1 tournaments are, one might say, the best that the D-optimality theory can provide in the absence of prior knowledge.

Comparison of the last two columns of Table 4.9 shows that the round-robin design becomes clearly less efficient as σ increases. This is expected: eventually, the sequential-1 procedure will recognize that the teams vary in strength, and will begin to design accordingly, producing designs closer to those of Section 4.3.3. It should be noted that the sequential-1 efficiency figures given in Table 4.9 are only estimates; nonetheless, the pattern seems clear. Further simulation with double round-robins (not given here) shows that the pattern of relative efficiency is the same there, with a larger rate of decrease, corresponding to the idea that a larger round-robin is "further away" from the exact D-optimal design or an approximation to it.

The last column of Table 4.10 shows the relative efficiency of the sequential-1 designs relative to the D-optimal exact design with the same number of observations. The last two columns of this Table therefore indicate how the Swiss and sequential-1 tournaments compare in terms of D-efficiency. Although these figures are subject to appreciable variability, since the log-determinants for both the Swiss and sequential-1 tournaments are estimated from simulations, it seems clear that there is little to choose between these two tournament types. For values of σ in the range shown here, therefore, Swiss tournaments perform well, and can be recommended.

Chapter 5

Goodness of fit for logistic regression models

5.1 Introduction

When the response variable in a model is binary or binomial, it is natural to consider assessing the fit of the model by comparing the observed successes with the expected successes (that is, the exact or estimated success probabilities) in some way.

Two obvious approaches suggest themselves: to partition the observations, and then to carry out the usual chi-squared test based on the observed and expected successes within each group, or to compare observed and expected successes in a cumulative fashion. Within each approach, there is also the choice of using the x-variables to guide the partitioning or cumulation, or of using the fitted probabilities of success. In the partitioning camp, Tsiatis (1980) uses the x-variables, while Hosmer and Lemeshow (1980) use the fitted probabilities. Su and Wei (1991) define a supremum statistic based on cumulation by the x-variables; they are able to obtain the asymptotic distribution of their statistic by showing weak convergence of an process to a Gaussian process, which is the usual technique for statistics based on cumulation. We will investigate the other possibility on the cumulation side, namely to cumulate according to the fitted probabilities. We will not be able to use the same kinds of weak convergence arguments as Su and Wei, however, because of the additional presence of estimated quantities; we will therefore derive the asymptotic distributions of our statistics directly. Although, by doing this, we give up the ability to prove general results about

statistics based on the empirical process, we are able to obtain, without excessive difficulty, useful results about the statistics we do study.

A third approach is possible if there are repeated observations at each x-value (compare Section 5.3.1), and that is to cumulate, not by the p_i, but by the numbers of successes in each covariate group. This approach was explored by Spinelli (1994), and leads, for each k, to a comparison between the observed and expected numbers of groups for which the number of successes is less than or equal to k. Such an approach focuses on whether the responses in each covariate group are truly binomial with a success probability that depends only on x.

We consider that tests based on cumulation by the p_i are most suitable for our purposes here. In particular, a logistic regression model may fit badly because some other binary-response model is more appropriate or because the relationship between the explanatory variables and the fitted probabilities is mis-specified. In these cases, departures from the hypothesized logistic regression will tend to be smooth. Tests of the chi-squared type are sensitive to *all* departures from the hypothesized model, including many non-smooth departures which are not of interest in this context. It therefore seems better to base tests on the cumulative difference between observed and expected successes, since such tests will be more sensitive to the smooth departures that are of interest.

Let us suppose that there are n observations in total, and let $y_i, i = 1, 2, \ldots, n$, denote the responses. In most of the following work, the responses are assumed to be independently Bernoulli, with $P(y_i = 1)$ denoted by p_i (and thus $P(y_i = 0) = 1 - p_i$), so that $y_i = 1$ denotes "success". At one point when dealing with the quadratic statistics defined below, the responses are instead assumed to be binomial with index parameters n_i; this occasion is noted when it occurs.

The situation in which the p_i are known and we desire to test fit based on these known p_i is rarely of interest in practice, but the theory provides a useful stepping-stone to the more practical cases, and is therefore detailed below. It is perhaps worth noting that no model is involved in the known-p_i case (except for the untested assertion that the y_i really are independently Bernoulli).

It is more useful to address the case when the success probabilities are estimated. The most common model, and the one addressed here, is the logistic model, in which the explanatory variables, arranged in a design matrix X, and the "slope" parameters β, arranged

in a vector, are connected to the p_i as follows:

$$\text{logit } p_i = \log\{p_i/(1-p_i)\} = \eta_i, \tag{5.1}$$

where η_i is the so-called "linear predictor", and

$$\eta_i = x_i'\beta, \tag{5.2}$$

where x_i' denotes the i-th row of X, and, in general, primes denote vector and matrix transposes. We shall need the score vector and information matrix, as functions of the parameter vector β: these are most compactly written as

$$s_n(\beta) = \sum_{i=1}^{n}(y_i - p_i)x_i \tag{5.3}$$

and

$$F_n(\beta) = \sum_{i=1}^{n} p_i(1-p_i)x_i x_i', \tag{5.4}$$

where s_n denotes the score vector and F_n the information matrix, and are functions of β implicitly, since each p_i is a function of β. The maximum likelihood estimate $\hat{\beta}$ is found in the usual way, by solving $s_n(\hat{\beta}) = 0$ (which must be done numerically in general).

Our strategy for testing fit was outlined above; specifically, we define a process $X_n(p)$ by

$$X_n(p) = n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)I(p_i \le p), \tag{5.5}$$

where the indicator restricts the comparison to those observations where p_i is no bigger than p, so that $X_n(p)$ compares the observed and expected successes cumulatively. When, as is usually the case, the probabilities have been estimated, p_i is replaced by $\hat{p}_i$. We then construct test statistics by averaging (integrating) some function of this process over p, so as to get a single-number summary of the discrepancy between observed and expected successes. The particular functions we investigate are the process itself and its square.

This strategy parallels that of the standard tests of fit based on the empirical process $W_n(x) = n^{-1/2}\{\sum_{i=1}^{n} I(x_i \le x) - nx\}$, which are studied by Stephens (1986). In Section 5.2 we study statistics based on the integral of $X_n(p)$, which correspond to the goodness-of-fit statistic $n^{1/2}(\bar{x} - \frac{1}{2})$, and in Section 5.3 we study statistics based on the integral of $\{X_n(p)\}^2$, which correspond to the Cramér-von Mises statistic.

5.2 Asymptotic theory for the area family of statistics

5.2.1 The area statistics

These statistics are based on the integral of $X_n(p)$ itself, on the basis that if $X_n(p)$ is usually close to zero, so is its integral. The area statistics take a simple form, being linear in the y_i, and yield relatively straightforward asymptotic distributions because of this. On the other hand, these statistics are also close to zero if $X_n(p)$ oscillates around zero without always being small, and so one might expect a loss of power in comparison with the quadratic statistics considered in the next Section.

Specifically:

$$A_0 = n^{-1/2} \int_0^1 \sum_{i=1}^{n} (y_i - p_i) I(p_i \leq p)\, dp,$$

and the corresponding statistic, which we call A_2, in which the p_i are replaced by $\hat{p}_i$, their maximum likelihood estimates under a logistic model. In addition, we look at a statistic A_1 formed by replacing only the first occurrence of p_i by $\hat{p}_i$; this is of no practical purpose, but eases the theoretical development.

Interchanging the order of integration and summation and noting that

$$\int_0^1 I(p_i \leq p)\, dp = 1 - p_i,$$

we find that:

$$A_0 = n^{-1/2} \sum_{i=1}^{n} (y_i - p_i)(1 - p_i), \tag{5.6}$$

$$A_1 = n^{-1/2} \sum_{i=1}^{n} (y_i - \hat{p}_i)(1 - p_i), \tag{5.7}$$

$$A_2 = n^{-1/2} \sum_{i=1}^{n} (y_i - \hat{p}_i)(1 - \hat{p}_i). \tag{5.8}$$

The linear appearance of these statistics suggests that we might reasonably expect them to have asymptotic normal distributions; we devote this section to showing that this is indeed the case, under certain reasonable assumptions about the limiting behaviour of the p_i and, in the case of A_1 and A_2, about the x-variables in the logistic regression.

5.2.2 Two invariance results

One might imagine that some arbitrary decisions have been made in the definition of the area statistics. Specifically, it is of interest to know what happens if the values of $y_i - p_i$

are cumulated downwards, with decreasing p_i rather than increasing p_i, or when successes and failures are interchanged. The following two results show that tests based on estimated parameters are unaffected, at least when there is an intercept in the model.

Theorem 5.1 *If the model has an intercept, the statistic A_2' obtained by cumulating the values of $y_i - \hat{p}_i$ downwards is the negative of the statistic A_2. An analogous result holds for A_1.*

Proof: The definition of A_2' is

$$A_2' = n^{-1/2} \int_0^1 \sum_i (y_i - \hat{p}_i)\, I(\hat{p}_i \geq p)\, dp.$$

since the indicator selects only those values $\hat{p}_i$ that exceed p. The sum is over $1 \leq i \leq n$. Thus:

$$\begin{aligned} A_2' &= n^{-1/2} \sum_i (y_i - \hat{p}_i) \int_0^1 I(\hat{p}_i \geq p) \\ &= n^{-1/2} \sum_i (y_i - \hat{p}_i)\hat{p}_i \\ &= n^{-1/2} \sum_i (y_i - \hat{p}_i)\{1 - (1 - \hat{p}_i)\} \\ &= n^{-1/2} \sum_i (y_i - \hat{p}_i) - A_2. \end{aligned}$$

When the model has an intercept, one of the likelihood equations is $\sum_i y_i - \hat{p}_i = 0$, and so the result is proved for A_2. The same algebra, with p_i replacing $\hat{p}_i$ in the indicator function, shows that the corresponding result also holds for A_1.

The second result concerns the interchange of successes and failures:

Theorem 5.2 *If the model has an intercept, the statistic A_2'' obtained by cumulating the observed and expected numbers of failures is the negative of A_2. The corresponding result holds for A_1.*

Proof: Let $w_i = 1 - y_i$ be the observed numbers of failures and $\hat{q}_i = 1 - \hat{p}_i$ be the estimated probabilities of failure. Then

$$n^{-1/2} A_2'' = \int_0^1 \sum_i (w_i - \hat{q}_i)\, I(\hat{q}_i \leq q)\, dq.$$

Making the change of variable $p = 1 - q$, and writing w_i and $\hat{q}_i$ in terms of y_i and $\hat{p}_i$, we find that

$$A_2'' = n^{-1/2} \int_0^1 \sum_i (y_i - \hat{p}_i)\, I(\hat{p}_i \geq p)\, dp = A_2' = -A_2,$$

since two minus signs cancel from the change of variable. The proof for A_2 is complete: the effect of interchanging success and failure is the same as that of cumulating downwards rather than upwards. And, as before, the proof for A_1 is carried out with exactly the same algebra, replacing $\hat{p}_i$ in the indicator with p_i.

For A_0, and for models without an intercept, we have the results

$$A_0 + A_0' = A_0 + A_0'' = n^{-1/2} \sum_i (y_i - p_i),$$

and

$$A_j + A_j' = A_j + A_j'' = n^{-1/2} \sum_i (y_i - \hat{p}_i),$$

for $j = 1, 2$. One can argue in this case that if the right-hand side differs too much from zero, this in itself is evidence of a lack of fit, since the observed number of successes was very much larger or smaller than the probabilities would lead one to expect. Thus, in the more interesting cases where the lack of fit is due to the pattern of the y_i rather than the number of them that are 1, we can expect the right-hand side to be small, and thus the test statistics obtained by downward cumulation or exchange of successes and failures to be approximately the negatives of the original statistics.

As a final remark, it should be noted that tests based on the area statistics will be two-tailed in general, rejecting for large $|A_j|$, so the appearance of minus signs in the preceding Theorems has no effect on the P-values of tests based on these statistics.

5.2.3 Statistic A_0

When the p_i are known, statistic A_0 of (5.6) can be used. The only random quantities it contains are the y_i, which are Bernoulli with success probability p_i. The appearance of the statistic, a linear function of the y_i, suggests that its asymptotic normality may be demonstrated using a version of the Central Limit Theorem such as that of Lyapunov which applies to a sum of random variables with non-identical variances. This is in fact exactly how it works.

We begin by proving a more general result, one which will be used again later in our work with A_1.

Theorem 5.3 *Let* $\{u_{in}\}$, $1 \leq i \leq n$, *be a triangular array of numbers, and let* $m_n = \max_{1 \leq i \leq n} |u_{in}|$. *Let* $s_n^2 = n^{-1} \sum p_i(1-p_i)u_i^2$; *if* $\lim_{n\to\infty} s_n^2$ *exists and is equal to* s^2, *say, with* $s^2 > 0$, *and if* $\lim_{n\to\infty} n^{-1/2} m_n = 0$, *then*

$$n^{-1/2}\left\{\sum_{i=1}^{n}(y_i - p_i)u_{in}\right\}/s_n \xrightarrow{D} N(0,1).$$

Proof: Let $T_i = n^{-1/2}(y_i - p_i)u_{in}$, so that it is $\sum_{i=1}^n T_i$ whose limiting distribution we seek. Since $s_n^2 \to s^2$ (s_n^2 is not random), it suffices to show that $\sum_{i=1}^n T_i/s \xrightarrow{D} N(0,1)$. Now:

$$\begin{aligned}
&E(y_i - p_i) = 0 \Rightarrow E(T_i) = 0;\\
&\text{var}(y_i - p_i) = p_i(1-p_i) \Rightarrow \text{var}(T_i) = p_i(1-p_i)u_{in}^2/n;\\
&E(|y_i - p_i|^3) = (1-p_i)^3 p_i + p_i^3(1-p_i) \Rightarrow\\
&E(|T_i|^3) = n^{-3/2} p_i(1-p_i)\{p_i^2 + (1-p_i)^2\}|u_{in}|^3.
\end{aligned}$$

Noting that $p_i^2 + (1-p_i)^2 \leq 1$ for each i, consider

$$\begin{aligned}
\frac{\sum_{i=1}^n E(|T_i|^3)}{\{\sum_{i=1}^n \text{var}(T_i)\}^{3/2}} &= \frac{\sum_{i=1}^n p_i(1-p_i)\{p_i^2 + (1-p_i)^2\}|u_{in}|^3}{\{\sum_{i=1}^n p_i(1-p_i)u_{in}^2\}^{3/2}}\\
&\leq \frac{m_n \sum_{i=1}^n p_i(1-p_i)u_{in}^2}{\{\sum_{i=1}^n p_i(1-p_i)u_{in}^2\}^{3/2}}, \quad \text{since } |u_{in}| \leq m_n,\\
&= n^{-1/2} m_n \left\{n^{-1}\sum_{i=1}^{n} p_i(1-p_i)u_i^2\right\}^{-1/2}\\
&\to 0 \cdot (s^2)^{-1/2} = 0 \quad \text{since } s^2 > 0.
\end{aligned}$$

Therefore, by Lyapunov's version of the Central Limit Theorem (as in, for example, Grimmett and Stirzaker, 1982, p. 110), the random variable $\sum_{i=1}^n T_i$ divided by its standard deviation converges in distribution to standard normal, and the result is proved.

The specialization of this result to A_0 is immediate:

Theorem 5.4 *Let* $s_n^2 = n^{-1}\sum_{i=1}^n p_i(1-p_i)^3$. *If* $s^2 = \lim_{n\to\infty} s_n^2$ *exists with* $s^2 > 0$, *then* $A_0/s_n \xrightarrow{D} N(0,1)$.

Proof: Take $u_{in} = 1 - p_i$ in Theorem 5.3. Since $m_n \leq 1$ for all n, the result holds.

We remark that the condition on the p_i for the theorem is an eminently reasonable one, since it will fail only if the p_i are tending towards 0 or 1. In this case the "information" in the data is not increasing quickly enough, so that other problems such as inconsistency of estimators will also occur.

5.2.4 Some additional results

Before moving on to the other two statistics, we give some general results that will be used in the sequel. The first gives a Taylor series expansion in terms of $\hat{\beta} - \beta$, and the second gives conditions for the asymptotic normality of $\hat{\beta}$ in the logistic regression case.

Expanding $\hat{p}_i = p_i(\hat{\beta})$ in a Taylor series about $p_i = p_i(\beta)$ gives

$$\hat{p}_i - p_i = p_i(1 - p_i)x_i'(\hat{\beta} - \beta) + \frac{1}{2}\tilde{p}_i(1 - \tilde{p}_i)(1 - 2\tilde{p}_i)\left\{x_i'(\hat{\beta} - \beta)\right\}^2,$$

where $\tilde{p}_i = p_i(\tilde{\beta}_i)$, and each $\tilde{\beta}_i$ lies on the line joining β and $\hat{\beta}$.

Written this way, the quantity $\tilde{\beta}_i$ in the remainder term depends on i, but we will use the series expansion in sums of the form $\sum_{i=1}^n f(p_i)(\hat{p}_i - p_i)$, for which one quantity $\tilde{\beta}$ suffices, since the series is for the whole sum:

$$\begin{aligned}\sum_{i=1}^n f(p_i)(\hat{p}_i - p_i) &= \sum_{i=1}^n \Big[f(p_i)p_i(1 - p_i)x_i'(\hat{\beta} - \beta) \\ &\quad + \frac{1}{2}f(p_i)\tilde{p}_i(1 - \tilde{p}_i)(1 - 2\tilde{p}_i)\left\{x_i'(\hat{\beta} - \beta)\right\}^2\Big], \qquad (5.9)\end{aligned}$$

where now $\tilde{p}_i = p_i(\tilde{\beta})$, and $\tilde{\beta}$ lies on the line joining β and $\hat{\beta}$.

Fahrmeir and Kaufmann (1985) give conditions for the asymptotic existence, (weak and strong) consistency and asymptotic normality of maximum likelihood estimators in generalized linear models. Their Corollary 2 concerns the case where the response variable is bounded, as is the case here, since $0 \leq y_i \leq 1$ for all i.

Theorem 5.5 (Fahrmeir and Kaufmann) *If* $\max_{1 \leq i \leq n} x_i' F_n^{-1} x_i \to 0$, *then* $\hat{\beta}$ *is weakly consistent for* β *and*

$$L_n'(\hat{\beta}_n - \beta) \xrightarrow{D} N(0, I),$$

where L_n' *is such that the information matrix* $F_n = L_n L_n'$.

We will use this result frequently. The result is true for any type of "square root" matrix L_n that depends continuously on β, provided that the same type is used for all n. In our calculations we use the Cholesky square root. For a matrix A, this is the unique lower triangular matrix R with positive diagonal elements for which $RR' = A$. To the condition of Fahrmeir and Kaufmann we will, in our work, add the assumption that $\lim_{n\to\infty}(F_n/n)$ exists and is positive definite; we will call the limit matrix G.

Fahrmeir and Kaufmann (1985) show that, in models like logistic regression which have a canonical link function and a bounded response, the condition of Theorem 5.5 implies their condition (N); combining this condition with our assumption of convergence of F_n/n to G, and using some elementary properties of matrix norms, we see that

$$\sup_{\{\beta^*:\|L_n(\beta^*-\beta)\|\leq C\}} \|n^{-1}\{F_n(\beta^*) - F_n(\beta)\}\| \xrightarrow{P} 0,$$

for all finite C, and equivalently

$$\sup_{\{\beta^*:\|n^{1/2}(\beta^*-\beta)\|\leq C\}} \|n^{-1}\{F_n(\beta^*) - F_n(\beta)\}\| \xrightarrow{P} 0.$$

Now we obtain asymptotic results for linear and quadratic functions of $\hat{\beta}$:

Theorem 5.6 *Let $S_n = n^{-1/2}\sum_{i=1}^n g(p_i)x_i'(\hat{\beta}-\beta)$. If the condition of Theorem 5.5 holds, together with*

$$\lim_{n\to\infty} F_n/n = G \tag{5.10}$$

with G positive definite, and

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^n g(p_i)x_i' = v', \tag{5.11}$$

then

$$S_n \xrightarrow{D} N(0, v'G^{-1}v).$$

Proof: Write

$$S_n = n^{-1/2}\left\{\sum_{i=1}^n g(p_i)x_i'\right\}(L_n')^{-1}L_n'(\hat{\beta}-\beta).$$

Noting that $\lim_{n\to\infty} n^{-1/2}L_n = M$ for some matrix M with $MM' = G$, as a result of (5.10), and using (5.11), we find that

$$S_n \xrightarrow{D} n^{1/2}v'n^{-1/2}(M')^{-1}Z = v'(M')^{-1}Z,$$

where Z has a standard multivariate normal distribution, using Theorem 5.5. Thus it follows that

$$S_n \xrightarrow{D} N(0, v'(M')^{-1}M^{-1}v) = N(0, v'G^{-1}v),$$

completing the proof.

Once again, the conditions imposed seem reasonable. Condition (5.10) requires the information per observation to tend to a limit, a standard requirement in inference, and

condition (5.11) is roughly a requirement that the "average x_i" does not get too large. In fact, since $g(p_i)$ is typically a function like $p_i(1-p_i)$, (5.11) is not as strong as this, because for such $g(p_i)$, large x_i can be, and often are, counterbalanced by the corresponding p_i being close to 0 or 1.

Considering now quadratic forms in $\hat{\beta}-\beta$, we have the following result:

Theorem 5.7 *Let* $T_n = n^{-1/2}\sum_{i=1}^n h(p_i)\left\{x_i'(\hat{\beta}-\beta)\right\}^2$, $A_n = \sum_{i=1}^n h(p_i)x_ix_i'$. *Under the conditions of Theorem 5.5, and assuming also that* $\lim_{n\to\infty} F_n/n = G, \lim_{n\to\infty} A_n/n = A$ *for some matrices* A *and* G *with* G *positive definite, we have* $T_n \xrightarrow{P} 0$.

Proof:

$$\begin{aligned} n^{1/2}T_n &= (\hat{\beta}-\beta)'\left\{\sum_{i=1}^n h(p_i)x_ix_i'\right\}(\hat{\beta}-\beta) \\ &= (\hat{\beta}-\beta)'A_n(\hat{\beta}-\beta). \end{aligned}$$

Now, from Theorem 5.5, $L_n'(\hat{\beta}-\beta) \xrightarrow{D} N(0,I)$, so write $n^{1/2}T_n$ as

$$n^{1/2}T_n = (\hat{\beta}-\beta)'L_nL_n^{-1}A_n(L_n')^{-1}L_n'(\hat{\beta}-\beta).$$

The matrices $n^{-1/2}L_n$ and A_n/n both have limits by hypothesis; denote $\lim_{n\to\infty} n^{-1/2}L_n$ by M. Then

$$n^{1/2}T_n \xrightarrow{D} Z'(n^{-1/2}M^{-1})(nA)\{n^{-1/2}(M')^{-1}\}Z = Z'QZ,$$

where $Q = M^{-1}A(M')^{-1}$ and Z has a standard multivariate normal distribution. The distribution of $Z'QZ$ is well known to be that of $\sum_{i=1}^p \lambda_iY_i^2$, where p is the dimension of β, the Y_i are independent standard normal random variables, and the λ_i are the eigenvalues of Q. Since this limiting distribution of T_n itself is this multiplied by $n^{-1/2}$, it follows that $T_n \xrightarrow{P} 0$.

In dealing with remainder terms, we will also be faced with quadratic forms containing possibly random quantities $\tilde{p}_i$ as well as p_i. An extension of Theorem 5.7 enables us to deal with these as well, provided that we now allow the function h to depend on both p_i and $\tilde{p}_i$:

Theorem 5.8 *Let*

$$\tilde{T}_n = n^{-1/2}\sum_{i=1}^n h(p_i,\tilde{p}_i)\{x_i(\hat{\beta}-\beta)\}^2,$$

and assume that there exists a constant C *such that*

$$h(p_i,\tilde{p}_i) \le C\{p_i(1-p_i)+\tilde{p}_i(1-\tilde{p}_i)\}.$$

Then $\tilde{T}_n \xrightarrow{P} 0$.

Proof: By hypothesis,

$$\begin{aligned} n^{1/2}\tilde{T}_n &\leq C(\hat{\beta}-\beta)'\left\{\sum_{i=1}^{n}\tilde{p}_i(1-\tilde{p}_i)x_i x_i'\right\}(\hat{\beta}-\beta) \\ &\quad + C(\hat{\beta}-\beta)'\left\{\sum_{i=1}^{n}p_i(1-p_i)x_i x_i'\right\}(\hat{\beta}-\beta) \\ &= C(\hat{\beta}-\beta)'F_n(\tilde{\beta})(\hat{\beta}-\beta) + C(\hat{\beta}-\beta)'F_n(\hat{\beta}-\beta) \end{aligned}$$

Now, $n^{1/2}(\hat{\beta}-\beta)$ has a limiting normal distribution, and, as observed in the discussion following Theorem 5.5, $n^{-1}F_n(\tilde{\beta})$ and $n^{-1}F_n$ both converge to G, the former in probability since the sequence is of random variables. Thus $n^{1/2}\tilde{T}_n$ is bounded in probability, and hence $\tilde{T}_n$ itself converges in probability to zero as claimed.

Notice that in fact the conditions of Theorems 5.7 and 5.8 lead to limiting distributions for $n^{1/2}T_n$ and $n^{1/2}\tilde{T}_n$.

With these results in hand, we can now move on to proofs of the asymptotic distributions of A_1 and A_2.

5.2.5 Statistic A_1

The asymptotic distribution of this statistic is not important in practice, but it turns out to be a useful stepping stone to that of A_2. We give the theory for A_1 in two stages: first we show that its limiting distribution is that of the sum of two random variables, one of which is A_0, and then we show that the distribution of this sum is in fact normal.

Theorem 5.9 *If the conditions of Theorems 5.4, 5.5, 5.6 and 5.8 hold, with* $h(p_i,\tilde{p}_i) = (1-p_i)\tilde{p}_i(1-\tilde{p}_i)(1-2\tilde{p}_i)$, *then*

$$A_1 = A_0 - n^{-1/2}\sum_{i=1}^{n}p_i(1-p_i)^2x_i'(\hat{\beta}-\beta) + o_p(1).$$

Proof:

$$\begin{aligned} A_1 &= n^{-1/2}\sum_{i=1}^{n}(y_i-\hat{p}_i)(1-p_i) \\ &= n^{-1/2}\sum_{i=1}^{n}\{(y_i-p_i)-(\hat{p}_i-p_i)\}(1-p_i) \\ &= A_0 - n^{-1/2}\sum_{i=1}^{n}(1-p_i)(\hat{p}_i-p_i). \end{aligned}$$

Using (5.9) with $f(p_i) = 1 - p_i$,

$$\begin{aligned} A_1 &= A_0 - n^{-1/2}\sum_{i=1}^{n} p_i(1-p_i)^2 x_i'(\hat{\beta}-\beta) \\ &\quad - \frac{1}{2}n^{-1/2}\sum_{i=1}^{n}(1-p_i)\tilde{p}_i(1-\tilde{p}_i)(1-2\tilde{p}_i)\{x_i'(\hat{\beta}-\beta)\}^2. \end{aligned}$$

Since

$$h(p_i, \tilde{p}_i) = (1-p_i)\tilde{p}_i(1-\tilde{p}_i)(1-2\tilde{p}_i) \leq 3\{p_i(1-p_i) + \tilde{p}_i(1-\tilde{p}_i)\},$$

the condition of Theorem 5.8 is satisfied, and hence the last term converges in probability to zero (in other words, is $o_p(1)$). The result is therefore proved.

It remains to discover what the limiting distribution of A_1 is. The only random quantities in A_0 are the $y_i - p_i$, so we can attempt to write the second term as a function of $y_i - p_i$ also. The next theorem shows how this may be done:

Theorem 5.10 *Assume that the conditions of Theorem 5.9 hold. Let $G = \lim_{n\to\infty} F_n/n$ (the limit being assumed to exist and be positive definite), let $w_{in} = \sum_{j=1}^{n} p_j(1-p_j)^2 x_j' F_n^{-1} x_i$, and let $s_n^2 = n^{-1}\sum_{i=1}^{n} p_i(1-p_i)(1-p_i-w_{in})^2$. Provided that $s^2 = \lim_{n\to\infty} s_n^2$ exists and is positive, it follows that $A_1/s_n \xrightarrow{D} N(0,1)$.*

Proof:

$$\begin{aligned} A_1 &= A_0 - n^{-1/2}\sum_{i=1}^{n} p_i(1-p_i)^2 x_i'(\hat{\beta}-\beta) - R_n \\ &= n^{-1/2}\sum_{i=1}^{n}(y_i-p_i)(1-p_i) \\ &\quad - n^{-1/2}\sum_{i=1}^{n} p_i(1-p_i)^2 x_i'(L_n')^{-1}L_n'(\hat{\beta}-\beta) - R_n, \end{aligned} \tag{5.12}$$

where R_n is a quantity that converges in probability to zero. Under the conditions that we have imposed here, Fahrmeir and Kaufmann (1985) show (in the proof of their Theorem 3), by using the mean-value theorem for vector-valued functions, that

$$L_n^{-1}\sum_{i=1}^{n}(y_i-p_i)x_i = \left[\int_0^1 V_n\{\beta + t(\hat{\beta}-\beta)\}\,dt\right] L_n'(\hat{\beta}-\beta),$$

where $V_n(\beta^*) = L_n^{-1}F_n(\beta^*)(L_n')^{-1}$, and the integral of this matrix-valued function is understood elementwise. Letting $\hat{U}_n$ denote this integral, Fahrmeir and Kaufmann show (later on

in the same proof) that $\hat{U}_n \xrightarrow{P} I$, so that for all n sufficiently large, $\hat{U}_n$ is invertible with a probability that can be made as close to 1 as we please. For such n,

$$L_n'(\hat{\beta} - \beta) = \hat{U}_n^{-1} L_n^{-1} \sum_{i=1}^{n} (y_i - p_i) x_i. \tag{5.13}$$

Thus, changing the index of summation in the second term of (5.12) to j, using (5.13), and neglecting the term R_n that converges in probability to zero,

$$\begin{aligned}
A_1 &= n^{-1/2} \Big\{ \sum_{i=1}^{n} (y_i - p_i)(1 - p_i) \\
&\quad - \sum_{j=1}^{n} p_j (1 - p_j)^2 x_j' (L_n')^{-1} \hat{U}_n^{-1} L_n^{-1} \sum_{i=1}^{n} (y_i - p_i) x_i \Big\} \\
&= n^{-1/2} \sum_{i=1}^{n} (y_i - p_i) \left\{ 1 - p_i - \sum_{j=1}^{n} p_j (1 - p_j)^2 x_j' (L_n')^{-1} \hat{U}_n^{-1} L_n^{-1} x_i \right\} \\
&= n^{-1/2} \sum_{i=1}^{n} (y_i - p_i)(1 - p_i - W_{in}) \qquad (5.14) \\
&= n^{-1/2} \sum_{i=1}^{n} (y_i - p_i)(1 - p_i - w_{in}) \\
&\quad - n^{-1/2} \sum_{i=1}^{n} (y_i - p_i)(W_{in} - w_{in}), \qquad (5.15)
\end{aligned}$$

where (5.14) defines the random variable W_{in}.

We next show that the last term on the right-hand side of (5.15) converges in probability to zero:

$$\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} (y_i - p_i)(W_{in} - w_{in}) \\
&= n^{-1/2} \sum_{i=1}^{n} (y_i - p_i) \left[\sum_{j=1}^{n} p_j (1 - p_j)^2 x_j' \{ (L_n \hat{U}_n L_n')^{-1} - F_n^{-1} \} \right] x_i \\
&= n^{-1} \sum_{j=1}^{n} p_j (1 - p_j)^2 x_j' \{ n (L_n \hat{U}_n L_n')^{-1} - (F_n/n)^{-1} \} \\
&\quad \times n^{-1/2} L_n L_n^{-1} \sum_{i=1}^{n} (y_i - p_i) x_i.
\end{aligned}$$

By the hypotheses of previous theorems, these quantities all have limits:

$$\begin{aligned} n^{-1}\textstyle\sum_{j=1}^{n} p_j(1-p_j)^2 x_j' &\to v' \quad \text{as in Theorem 5.6,} \\ n(L_n \hat{U}_n L_n')^{-1} - (F_n/n)^{-1} &\xrightarrow{P} 0 \quad \text{since } \hat{U}_n \xrightarrow{P} I, \\ n^{-1/2} L_n &\to M \quad \text{since } L_n L_n' = F_n, \\ L_n^{-1}\textstyle\sum_{i=1}^{n}(y_i - p_i)x_i &\xrightarrow{D} N(0, I). \end{aligned}$$

Thus $n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)(W_{in} - w_{in}) \xrightarrow{P} 0$ as claimed. It follows that A_1 converges to the same distribution as $n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)(1 - p_i - w_{in})$.

To complete the proof, we wish to take $u_{in} = 1 - p_i - w_{in}$ in Theorem 5.3; this requires us to show that $n^{-1/2}\max_{1\le i\le n}(1 - p_i - w_{in})$ tends to zero. Since $1 - p_i$ is bounded, it remains to show that $n^{-1/2}\max|w_{in}|$ tends to zero. We do this by showing that, in fact, the convergence holds for *every* i.

Let $b_n = \sum_{j=1}^{n} p_j(1-p_j)^2 x_j$; then

$$\begin{aligned} w_{in} &= b_n' F_n^{-1} x_i \\ &= (n^{-1} b_n')(n^{-1} F_n)^{-1} x_i \\ &= \{(n^{-1/2} L_n^{-1})(n^{-1} b_n)\}'\{(n^{-1/2} L_n^{-1}) x_i\}. \end{aligned}$$

Thus

$$|w_{in}| \le \|(n^{-1/2} L_n^{-1})(n^{-1} b_n)\| \|(n^{-1/2} L_n^{-1}) x_i\|.$$

As $n \to \infty$, the first norm has a limit, namely $\|M^{-1} v\|$. The second norm can be written as

$$\sqrt{x_i'(n^{-1}F_n)^{-1}x_i} = n^{1/2}\sqrt{x_i' F_n^{-1} x_i}.$$

Since the radicand tends to zero by hypothesis, it follows that $n^{-1/2}|w_{in}|$ also does for each i, and hence that $n^{-1/2}\max|w_{in}|$ converges to zero as well.

Finally, appeal to Theorem 5.3 shows that $A_1/s_n \xrightarrow{D} N(0,1)$, completing the proof.

5.2.6 Statistic A_2

Our strategy with A_2 is to show that it is the sum of A_1 and some other quantities that converge in probability to zero, which would show that A_2 is also asymptotically normal.

First, however, we have to note that in practice, the statistic will be normalized using a quantity $\hat{s}_n$ which contains estimated parameters, so that we must also show that $\hat{s}_n^2$ and s_n^2 converge in probability to the same non-zero limit.

Specifically, define

$$\hat{s}_n^2 = n^{-1} \sum_{i=1}^{n} \hat{p}_i(1-\hat{p}_i)(1-\hat{p}_i-\hat{w}_{in})^2, \tag{5.16}$$

where $\hat{w}_{in} = \sum_{j=1}^{n} \hat{p}_j(1-\hat{p}_j)^2 x_j' F_n^{-1}(\hat{\beta}) x_i$, analogously to the definition of s_n^2. Also let $b_n = \sum_{j=1}^{n} p_j(1-p_j)^2 x_j$, and define $\hat{b}_n$ analogously.

We proceed to establish that $\hat{s}_n^2 - s_n^2 \xrightarrow{P} 0$ by means of a series of Lemmas.

Lemma 5.11 *The quantity s_n^2 can be written as*

$$s_n^2 = n^{-1} \sum_{i=1}^{n} p_i(1-p_i)^3 - n^{-1} \sum_{i=1}^{n} p_i(1-p_i)^2 w_{in}. \tag{5.17}$$

Proof: Expand out s_n^2 to obtain

$$s_n^2 = n^{-1} \sum_{i=1}^{n} p_i(1-p_i)^3 - 2n^{-1} \sum_{i=1}^{n} p_i(1-p_i)^2 w_{in} + n^{-1} \sum_{i=1}^{n} p_i(1-p_i) w_{in}^2. \tag{5.18}$$

Writing b_n for $\sum_{j=1}^{n} p_j(1-p_j)^2 x_j$, we find that

$$\begin{aligned} \sum_{i=1}^{n} p_i(1-p_i) w_{in}^2 &= \sum_{i=1}^{n} b_n' F_n^{-1} p_i(1-p_i) x_i x_i' F_n^{-1} b_n \\ &= b_n' F_n^{-1} b_n \\ &= \sum_{j=1}^{n} p_j(1-p_j)^2 b_n' F_n^{-1} x_j \\ &= \sum_{j=1}^{n} p_j(1-p_j)^2 w_{in}. \end{aligned}$$

In other words, the third term on the right-hand side of (5.18) can be subsumed into the second, and

$$s_n^2 = n^{-1} \sum_{i=1}^{n} p_i(1-p_i)^3 - n^{-1} \sum_{i=1}^{n} p_i(1-p_i)^2 w_{in},$$

as we wished to prove.

Note that the same algebra yields the corresponding result

$$\hat{s}_n^2 = n^{-1} \sum_{i=1}^{n} \hat{p}_i(1-\hat{p}_i)^3 - n^{-1} \sum_{i=1}^{n} \hat{p}(1-\hat{p})^2 \hat{w}_{in}. \tag{5.19}$$

Our convergence proof therefore rests on the difference between the right-hand sides of (5.17) and (5.19). Lemmas 5.12 and 5.16 will show that the two constituents of this difference converge in probability to zero.

Lemma 5.12 *Let* $R_n = n^{-1}\sum_{i=1}^{n}\{\hat{p}_i(1-\hat{p}_i)^3 - p_i(1-p_i)^3\}$. *Then* $R_n \xrightarrow{P} 0$.

Proof: R_n contains a polynomial function of the p_i, evaluated at $\hat{p}_i$ and p_i. This polynomial is seen to have derivative with respect to a particular p_i of $(1-p_i)^2(1-4p_i)$. Thus, thinking of R_n as a function of β, R_n can be written in a Taylor series about β as

$$R_n = n^{-1}\sum_{i=1}^{n}\tilde{p}_i(1-\tilde{p}_i)^3(1-4\tilde{p}_i)x_i'(\hat{\beta}-\beta),$$

for $\tilde{p}_i$ between p_i and $\hat{p}_i$ for all i. (The process of differentiating p_i with repect to β produces a term $\tilde{p}_i(1-\tilde{p}_i)$.) Using the Cauchy-Schwarz inequality with a judicious splitting of the terms in the sum, we find that

$$\begin{aligned} R_n &\leq n^{-1}\sqrt{\sum_{i=1}^{n}\tilde{p}_i(1-\tilde{p}_i)^5(1-4\tilde{p}_i)^2}\sqrt{(\hat{\beta}-\beta)'\sum_{i=1}^{n}\tilde{p}_i(1-\tilde{p}_i)x_ix_i'(\hat{\beta}-\beta)} \\ &\leq n^{-1}\sqrt{9n}\sqrt{(\hat{\beta}-\beta)'F_n(\tilde{\beta})(\hat{\beta}-\beta)}, \end{aligned}$$

since each term in the sum of the first radicand is bounded by 9. The second radicand can be written as

$$(\hat{\beta}-\beta)'L_nL_n^{-1}F_n(\tilde{\beta})(L_n')^{-1}L_n'(\hat{\beta}-\beta),$$

where $L_n^{-1}F_n(\tilde{\beta})(L_n')^{-1} \xrightarrow{P} I$ for all $\tilde{\beta}$ in a suitable neighbourhood of β and $L_n'(\hat{\beta}-\beta) \xrightarrow{D} N(0,I)$ (see Fahrmeir and Kaufmann, 1985, and the discussion following Theorem 5.5). The second radicand is therefore bounded in probability, and the remaining powers of n are sufficient to ensure that $R_n \xrightarrow{P} 0$, as we wished to prove.

There now follow three small lemmas asserting some other useful convergence results.

Lemma 5.13 *Let* $L_n^{-1}(\hat{\beta})$ *denote* L_n^{-1} *evaluated at* $\hat{\beta}$. *Then* $n^{1/2}L_n^{-1}(\hat{\beta}) \xrightarrow{P} M^{-1}$.

Proof:

$$n^{1/2}L_n^{-1}(\hat{\beta}) = n^{1/2}L_n^{-1}(\hat{\beta})L_nL_n^{-1}.$$

By assumption, $n^{1/2}L_n^{-1} \to M^{-1}$; condition (Q) of Fahrmeir and Kaufmann (1985) shows that $L_n^{-1}(\hat{\beta})L_n \xrightarrow{P} I$.

Lemma 5.14 $n^{-1}b_n$ *is bounded as* $n \to \infty$.

Proof: We show that each element b_{nj} of b_n is, when divided by n, bounded in absolute value by a quantity that has a limit. Specifically,

$$\begin{aligned} |b_{nj}|/n &= n^{-1}\sum_{i=1}^{n} p_i(1-p_i)^2|x_{ij}| \\ &\leq n^{-1}\sqrt{\sum_{i=1}^{n} p_i(1-p_i)^3}\sqrt{\sum_{i=1}^{n} p_i(1-p_i)x_{ij}^2} \\ &\leq n^{-1/2}\sqrt{(F_n)_{jj}}. \end{aligned}$$

$n^{-1}(F_n)_{jj}$ has a limit as $n \to \infty$, namely G_{jj}, the j-th diagonal element of the matrix G.

Lemma 5.15 $n^{-1}(\hat{b}_n - b_n) \xrightarrow{P} 0$.

Proof:

$$\begin{aligned} n^{-1}(\hat{b}_n - b_n) &= n^{-1}\sum_{i=1}^{n}\{\hat{p}_i(1-\hat{p}_i)^2 - p_i(1-p_i)^2\} \\ &= n^{-1}\sum_{i=1}^{n}\tilde{p}_i(1-\tilde{p}_i)^2(1-3\tilde{p}_i)x_i x_i'(\hat{\beta}-\beta) \\ &\leq 3n^{-1}\sum_{i=1}^{n}\tilde{p}_i(1-\tilde{p}_i)x_i x_i'(\hat{\beta}-\beta) \\ &= 3n^{-1/2}\{F_n(\tilde{\beta})/n\}\{n^{1/2}(\hat{\beta}-\beta)\}. \end{aligned}$$

Since $F_n(\tilde{\beta})/n$ converges in probability to G, by our assumptions and the discussion after Theorem 5.5, and $n^{1/2}(\hat{\beta}-\beta)$ converges in distribution to normal, it follows that $n^{-1}(\hat{b}_n - b_n)$ converges in probability to zero.

Lemma 5.16 *Let* $T_n = n^{-1}\sum_{i=1}^{n}\{\hat{p}_i(1-\hat{p}_i)^2\hat{w}_{in} - p_i(1-p_i)^2 w_{in}\}$. *Then* $T_n \xrightarrow{P} 0$.

Proof: Using the definitions of w_{in} and $\hat{w}_{in}$,

$$\begin{aligned} T_n &= n^{-1}\sum_{i=1}^{n}\{\hat{p}_i(1-\hat{p}_i)^2\hat{b}_n' F_n(\hat{\beta})^{-1}x_i - p_i(1-p_i)^2 b_n' F_n^{-1}x_i\} \\ &= n^{-1}\{\hat{b}_n' F_n(\hat{\beta})^{-1}\hat{b}_n - b_n' F_n^{-1}b_n\} \\ &= n^{-1}[\{L_n^{-1}(\hat{\beta})\hat{b}_n\}'\{L_n^{-1}(\hat{\beta})\hat{b}_n\} - \{L_n^{-1}b_n\}'\{L_n^{-1}b_n\}] \\ &= n^{-1/2}\{L_n^{-1}(\hat{\beta})\hat{b}_n - L_n^{-1}b_n\}' n^{-1/2}\{L_n^{-1}(\hat{\beta})\hat{b}_n + L_n^{-1}b_n\}. \end{aligned}$$

The last line expresses T_n as the inner product of two vectors. Our aim is to show that the first vector converges in probability to zero, while the second is bounded in probability.

The first vector is

$$\begin{aligned} & n^{-1/2}\{L_n^{-1}(\hat{\beta})\hat{b}_n - L_n^{-1}b_n\} \\ &= \{n^{1/2}L_n^{-1}(\hat{\beta})\}\{n^{-1}(\hat{b}_n - b_n)\} + [n^{1/2}\{L_n^{-1}(\hat{\beta}) - L_n^{-1}\}](n^{-1}b_n). \end{aligned}$$

Lemmas 5.13 and 5.15 show that the first term converges in probability to zero, while Lemmas 5.14 and 5.15 show that the second term also does, both being the product of a factor converging in probability to zero and a factor with a limit in probability. The second vector is attacked in the same way:

$$\begin{aligned} & n^{-1/2}\{L_n^{-1}(\hat{\beta})\hat{b}_n + L_n^{-1}b_n\} \\ &= \{n^{1/2}L_n^{-1}(\hat{\beta})\}\{n^{-1}(\hat{b}_n - b_n)\} + [n^{1/2}\{L_n^{-1}(\hat{\beta}) + L_n^{-1}\}](n^{-1}b_n). \end{aligned}$$

This time, Lemmas 5.13 and 5.15 show that the first term converges in probability to zero, while the second term has a limit in probability because both of its factors do. Putting these results together, we see that $T_n \xrightarrow{P} 0$ as claimed.

We now have the tools to prove the following:

Theorem 5.17 $\hat{s}_n^2 - s_n^2 \xrightarrow{P} 0$.

Proof: By Lemma 5.11, we have

$$\begin{aligned} \hat{s}_n^2 - s_n^2 &= n^{-1}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)^3 - n^{-1}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)^2\hat{w}_{in} \\ &\quad - n^{-1}\sum_{i=1}^{n}p_i(1-p_i)^3 + n^{-1}\sum_{i=1}^{n}p_i(1-p_i)^2w_{in} \\ &= R_n - T_n. \end{aligned}$$

Since both R_n and T_n converge in probability to zero, by Lemmas 5.12 and 5.16 respectively, so does their difference, and the result is proved.

Finally, we turn to the main result, the convergence in distribution of A_2:

Theorem 5.18 *Under the conditions of Theorem 5.10, A_2 has the same limiting distribution as A_1, that is, defining $\hat{s}_n^2$ analogously to the s_n^2 of Theorem 5.10, replacing the p_i and w_{in} by estimates, $A_2/\hat{s}_n \xrightarrow{D} N(0,1)$.*

Proof: First, Theorem 5.17 shows that $\hat{s}_n^2$ and s_n^2 have the same limit, assumed nonzero by Theorem 5.10.

$$\begin{aligned} A_2 &= n^{-1/2}\sum_{i=1}^{n}(y_i - \hat{p}_i)(1 - \hat{p}_i) \\ &= n^{-1/2}\sum_{i=1}^{n}(y_i - \hat{p}_i)\{(1 - p_i) - (\hat{p}_i - p_i)\} \\ &= A_1 - n^{-1/2}\sum_{i=1}^{n}(y_i - \hat{p}_i)(\hat{p}_i - p_i) \\ &= A_1 - n^{-1/2}\sum_{i=1}^{n}\{(y_i - p_i) - (\hat{p}_i - p_i)\}(\hat{p}_i - p_i) \\ &= A_1 - n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)(\hat{p}_i - p_i) + n^{-1/2}\sum_{i=1}^{n}(\hat{p}_i - p_i)^2. \end{aligned} \tag{5.20}$$

The first term of the right-hand side of (5.20), A_1, is already known to be asymptotically normal, so we turn to the second, using the expansion of (5.9):

$$n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)(\hat{p}_i - p_i) \tag{5.21}$$

$$\begin{aligned} &= n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)p_i(1 - p_i)x_i'(\hat{\beta} - \beta) \\ &\quad + \frac{1}{2}n^{-1/2}\sum_{i=1}^{n}(y_i - p_i)(1 - p_i)\tilde{p}_i(1 - \tilde{p}_i)(1 - 2\tilde{p}_i)\{x_i'(\hat{\beta} - \beta)\}^2, \end{aligned} \tag{5.22}$$

where $\tilde{\beta}$ lies on the line joining β and $\hat{\beta}$. Looking at the second term, we note that, with probability 1, $|y_i - p_i| \leq 1$, so in absolute value, this term is no bigger than

$$\frac{1}{2}n^{-1/2}\left|\sum_{i=1}^{n}h(p_i, \tilde{p}_i)\{x_i'(\hat{\beta} - \beta)\}^2\right|,$$

with $h(p_i, \tilde{p}_i) = (1 - p_i)\tilde{p}_i(1 - \tilde{p}_i)(1 - 2\tilde{p}_i)$. Thus, by Theorem 5.8, this term converges in probability to zero.

Turning to the first term of (5.22), and taking its absolute value, we find

$$\begin{aligned} & n^{-1/2}\left|\sum_{i=1}^{n}(y_i - p_i)p_i(1 - p_i)x_i'(\hat{\beta} - \beta)\right| \\ &= n^{-1/2}\left|\left\{\sum_{i=1}^{n}(y_i - p_i)p_i(1 - p_i)x_i'\right\}(L_n')^{-1}L_n'(\hat{\beta} - \beta)\right| \end{aligned}$$

$$\begin{aligned} &\leq n^{-1/2}\left|\left\{\sum_{i=1}^{n}(y_i-p_i)x_i'\right\}(L_n')^{-1}L_n'(\hat{\beta}-\beta)\right| \\ &= n^{-1/2}\left|(L_n^{-1}s_n)'L_n'(\hat{\beta}-\beta)\right|. \end{aligned}$$

The two random variables here are both converging in distribution to the standard multivariate normal distribution, and so each, and therefore their product, is bounded in probability. It follows that the term as a whole, being multiplied by $n^{-1/2}$, is converging in probability to zero.

Finally, we expand the last term of (5.20) in a Taylor series:

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^{n}(\hat{p}_i-p_i)^2 \\ &= 2n^{-1/2}\sum_{i=1}^{n}\{\tilde{p}_i^2(1-\tilde{p}_i)^2+(\tilde{p}_i-p_i)\tilde{p}_i(1-\tilde{p}_i)(1-2\tilde{p}_i)\}[x_i'(\hat{\beta}-\beta)]^2. \end{aligned}$$

Letting $h(p_i,\tilde{p}_i)$ denote the quantity enclosed in braces, Theorem 5.8 shows that the whole term converges in probability to zero. Thus we have shown that the statistic A_2 can be written as the sum of A_1 and some "correction" terms that converge in probability to zero, and therefore that A_1 and A_2 have the same asymptotic distribution.

5.3 Asymptotic theory for the quadratic family of statistics

5.3.1 Introduction

Starting once again from our "empirical process"

$$X_n(p)=n^{-1/2}\sum_{i=1}^{n}(y_i-p_i)I(p_i\leq p),$$

or the corresponding versions with one or both of the p_i replaced by estimates, we define the quadratic family of statistics by squaring the process and integrating. This can be expected to yield test statistics that are sensitive to any departures of $X_n(p)$ away from zero.

For example, Q_0 is found as follows:

$$\begin{aligned} Q_0 &= \int_0^1\{X_n(p)\}^2\,dp \\ &= n^{-1}\int_0^1\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i-p_i)(y_j-p_j)I(p_i\leq p)I(p_j\leq p)\,dp \end{aligned}$$

$$
\begin{aligned}
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - p_i)(y_j - p_j)\int_0^1 I\{\max(p_i, p_j) \le p\}\, dp \\
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - p_i)(y_j - p_j)\{1 - \max(p_i, p_j)\} \\
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}(y_i - p_i)(y_j - p_j) \qquad (5.23)
\end{aligned}
$$

where $q_{ij} = 1 - \max(p_i, p_j)$.

Unfortunately, the theory for this general situation is difficult for the statistics Q_1 and Q_2 (especially the latter). For Q_0, a general result can be given, as shown in Section 5.3.6. First, however, we obtain results for the simpler case described below.

Suppose that, instead of n distinct success probabilities, we have p_i that take on only a fixed number d of different values. Suppose there are n_i observations at the design point corresponding to p_i, and assume that $\nu_i = \lim_{n\to\infty}(n_i/n)$ exists and lies strictly between 0 and 1, where now $n = \sum_{i=1}^{d} n_i$. Let $y_{ij}, i = 1, 2, \ldots, d, j = 1, 2, \ldots, n_i$ denote the result of the j-th trial at the design point with success probability p_i, and let $y_{i+} = \sum_{j=1}^{n_i} y_{ij}$ be the total number of successes observed for this value p_i. Under our assumptions, y_{i+} has a binomial distribution with index n_i and probability p_i.

It is also convenient to express these quantities in vector-matrix terms. Let X denote the design matrix, let y denote the d-vector with i-th element y_{i+}, and let p denote the d-vector whose i-th element is p_i. Further, let N_n denote the diagonal matrix with (i, i)-th element n_i. (The rather cumbersome notation is necessary because the n_i depend on n.)

Some additional notation is desirable: as we shall see, the statistics depend on $q_{ij} = 1 - \max(p_i, p_j)$; let Q be the matrix with (i, j)-th element q_{ij}. In addition, let $V_n(\beta)$ be the diagonal matrix with (i, i)-th entry $n_i p_i(1 - p_i)$, where p_i is really $p_i(\beta)$; $V_n(\beta)$ is the covariance matrix of the y_{i+}, diagonal because the y_{i+} are independent. Let $V(\beta)$ be the diagonal matrix with (i, i)-th entry $\nu_i p_i(1 - p_i)$; by assumption, $\lim_{n\to\infty} V_n(\beta)/n = V(\beta)$. The matrices V_n and V without argument are evaluated at the true β.

When dealing with estimated parameters, we also use the obvious notation $\hat{p}$ for the vector of $\hat{p}_i$ and $\hat{Q}_n$ for the matrix of $\hat{q}_{ij} = 1 - \max(\hat{p}_i, \hat{p}_j)$ based on a sample of size n.

With the new notation, our empirical process becomes

$$
X_0(p) = n^{-1/2}\sum_{i=1}^{d}\sum_{k=1}^{n_i}(y_{ik} - p_i)I(p_i \le p)
$$

$$= n^{-1/2}\sum_{i=1}^{d}(y_{i+} - n_i p_i)I(p_i \le p). \tag{5.24}$$

Clearly, it follows that the statistic Q_0, and the corresponding statistics Q_1 and Q_2 obtained by replacing some or all of the parameters with estimates, can be written as follows:

$$\begin{aligned} Q_0 &= n^{-1}\sum_{i=1}^{d}\sum_{j=1}^{d} q_{ij}(y_{i+} - n_i p_i)(y_{j+} - n_j p_j) \\ &= n^{-1}(y - N_n p)'Q(y - N_n p) \qquad (5.25) \\ Q_1 &= n^{-1}\sum_{i=1}^{d}\sum_{j=1}^{d} q_{ij}(y_{i+} - n_i \hat{p}_i)(y_{j+} - n_j \hat{p}_j) \\ &= n^{-1}(y - N_n \hat{p})'Q(y - N_n \hat{p}) \qquad (5.26) \\ Q_2 &= n^{-1}\sum_{i=1}^{d}\sum_{j=1}^{d} \hat{q}_{ij}(y_{i+} - n_i \hat{p}_i)(y_{j+} - n_j \hat{p}_j), \\ &= n^{-1}(y - N_n \hat{p})'\hat{Q}(y - N_n \hat{p}). \qquad (5.27) \end{aligned}$$

Our strategy is similar to that with the area family: we show that Q_0 and Q_1 have different asymptotic distributions from the same family, while Q_1 and Q_2 have the same asymptotic distribution. In the proofs, we can take advantage of the fact that the number of the terms in the sums, and hence the dimension of the vectors y and p, remains finite, whereas within the sums, we have binomial random variables with increasing n_i, for which convergence to normality applies.

In the proofs, we shall need the forms of the score vector and information matrix under this new arrangement of the data. For general β, these are seen to be

$$\begin{aligned} s_n(\beta) &= \sum_{i=1}^{d}(y_i - n_i p_i)x_i, \\ &= X'(y - N_n p) \qquad (5.28) \\ F_n(\beta) &= \sum_{i=1}^{d} n_i p_i(1 - p_i)x_i x_i'. \\ &= X'V_n(\beta)X \qquad (5.29) \end{aligned}$$

The design matrix X, with i-th row x_i', now has d rows rather than n. As before, s_n and F_n without argument denote the score vector and information matrix evaluated at the true parameter value β.

5.3.2 Two invariance results

As with the area family of statistics, two invariance results are available for the quadratic family, showing that, whenever an intercept is estimated, the statistics Q_1 and Q_2 are unchanged under downward cumulation and exchange of successes and failures. These results hold under the most general conditions, since they do not depend on asymptotics.

Theorem 5.19 *When the model includes an intercept, the statistic Q_2' obtained by cumulating $(y_i - \hat{p}_i)(y_j - \hat{p}_j)$ downwards instead of upwards is equal to Q_2 calculated from the same data. The corresponding result holds for Q_1.*

Proof: Beginning from the integral defining the quadratic statistics, and modifying to cumulate downwards, we find

$$\begin{aligned} Q_2' &= n^{-1} \int_0^1 \sum_i \sum_j (y_i - \hat{p}_i)(y_j - \hat{p}_j)\, I(p_i \geq p)\, I(p_j \geq p)\, dp \\ &= n^{-1} \sum_i \sum_j (y_i - \hat{p}_i)(y_j - \hat{p}_j) \min(p_i, p_j) \\ &= n^{-1} \sum_i \sum_j (y_i - \hat{p}_i)(y_j - \hat{p}_j)\{p_i + p_j - 1 + 1 - \max(p_i, p_j)\} \\ &= n^{-1} \sum_i \sum_j (y_i - \hat{p}_i)(y_j - \hat{p}_j)(p_i + p_j - 1) + Q_2. \end{aligned} \tag{5.30}$$

Since the model contains an intercept, one of the likelihood equations is $\sum_k (y_k - \hat{p}_k) = 0$. It follows that, for any γ_i, γ_j,

$$\sum_i \sum_j \gamma_i (y_i - \hat{p}_i)(y_j - \hat{p}_j) = \sum_i \sum_j \gamma_j (y_i - \hat{p}_i)(y_j - \hat{p}_j) = 0,$$

by carrying out the summations in the right order, and therefore that the double sum on the right-hand side of (5.30) is also zero. Thus $Q_2' = Q_2$. As with the area family, the same argument shows that $Q_1' = Q_1$, and so the proof is complete.

Theorem 5.20 *When the model contains an intercept, the statistic Q_2'' obtained by exchanging successes and failures is equal to Q_2 calculated on the same data. The corresponding result holds for Q_1.*

Proof: Let $w_i = 1 - y_i$ be the observed numbers of failures, and let $q_i = 1 - p_i$ be the failure probabilities. The definition of Q_2 gives

$$Q_2'' = n^{-1} \int_0^1 \sum_i \sum_j (w_i - \hat{q}_i)(w_j - \hat{q}_j)\, I(q_i \leq q)\, I(q_j \leq q)\, dq.$$

Making the change of variable $p = 1 - q$, and replacing w_i by $1 - y_i$ and q_i by $1 - p_i$, we obtain

$$Q_2'' = \int_0^1 \sum_i \sum_j (y_i - \hat{p}_i)(y_j - \hat{p}_j)\, I(p_i \geq p)\, I(p_j \geq p)\, dp.$$

This is the same integral as that defining Q_2'. It therefore follows that $Q_2'' = Q_2' = Q_2$. The corresponding algebra shows that $Q_1'' = Q_1' = Q_1$, completing the proof.

5.3.3 Statistic Q_0

The asymptotic distribution for Q_0 is easily obtained.

Theorem 5.21 *Let* $m_{ij} = q_{ij}\sqrt{\nu_i p_i(1-p_i)\nu_j p_j(1-p_j)}$, *and let* M *be the matrix with* (i,j)*-th element* m_{ij}. *Provided that* $0 < p_i, \nu_i < 1$, $Q_0 \xrightarrow{D} \sum_{i=1}^{d} \lambda_i z_i^2$, *where the* λ_i *are the eigenvalues of* M, *and the* z_i *are independent standard normal random variables.*

Proof: Each $y_{i+} - n_i p_i$, when suitably scaled, converges independently in distribution to normal, because of the normal approximation to the binomial distribution. In particular, let the vector $w_n = n^{-1/2}V^{-1/2}(y - N_n p)$; $E(w_n) = 0$, while $\mathrm{var}(w_n) = n^{-1}V^{-1}V_n \to I$ as $n \to \infty$. Thus $n^{-1/2}V^{-1/2}(y - N_n p) \xrightarrow{D} N(0, I)$. (Note that V_n and V are diagonal, so that raising these matrices to the power $-\frac{1}{2}$ is done by raising each diagonal element to the same power. The assumptions $0 < p_i, \nu_i < 1$ for all i prevent any of these elements being zero whenever n is sufficiently large.)

Starting from (5.25), Q_0 can be written as

$$\begin{aligned} Q_0 &= \{n^{-1/2}V^{-1/2}(y - N_n p)\}' V^{1/2} Q V^{1/2} \{n^{-1/2}V^{-1/2}(y - N_n p)\} \\ &= w_n' M_n w_n, \end{aligned}$$

where the matrix $M_n = V^{1/2}QV^{1/2}$ has (i,j)-th element

$$n^{-1} q_{ij} \sqrt{n_i p_i (1-p_i) n_j p_j (1-p_j)},$$

and w_n is a d-vector with i-th element $(y_{i+} - n_i p_i)/\sqrt{n_i p_i (1-p_i)}$. Now, as $n \to \infty$, $n^{-1}\sqrt{n_i n_j} \to \sqrt{\nu_i \nu_j}$ for each i and j, so that the (non-random) matrix $M_n \to M$. Thus, by a well-known result for quadratic forms of normal random variables, $Q_0 \xrightarrow{D} \sum_{i=1}^{d} \lambda_i z_i^2$, where the z_i are independent standard normal random variables and the λ_i are the eigenvalues of the matrix of the quadratic form, in this case M.

5.3.4 Statistic Q_1

In the light of our experiences with A_0 and A_1, we might expect Q_1 also to have an asymptotic distribution which is a sum of squares of normal random variables, but with different weights. This turns out to be the case, with an interesting parallel to weighted linear regression.

Note that the condition of Fahrmeir and Kaufmann (1985) now requires only that the maximum of $x_i' F_n^{-1} x_i$ tend to zero over the *finite* set of x_i, the dependence on n arising only through F_n.

Theorem 5.22 *Assume that* $\max_{1 \le i \le d} x_i' F_n^{-1} x_i \to 0$ *as* $n \to \infty$. *Assume also that* $G = \lim_{n\to\infty} F_n/n$ *exists and is positive definite, and that* $\lim_{n\to\infty} n_i/n = \nu_i$ *exists for all* i *with* $0 < \nu_i < 1$. *Then*

$$Q_1 \xrightarrow{D} n^{-1}(y - N_n p)'(I - H)'Q(I - H)(y - N_n p),$$

where $H = VX(X'VX)^{-1}X'$.

Proof: We begin by writing out Q_1 as follows:

$$\begin{aligned} Q_1 &= n^{-1}(y - N_n\hat{p})'Q(y - N_n\hat{p}) \\ &= n^{-1}\{(y - N_n p) - N_n(\hat{p} - p)\}'Q\{(y - N_n p) - N_n(\hat{p} - p)\} \\ &= n^{-1}(y - N_n p)'Q(y - N_n p) - n^{-1}(y - N_n p)'Q N_n(\hat{p} - p) \\ &\quad - n^{-1}(\hat{p} - p)'N_n'Q(y - N_n p) + n^{-1}(\hat{p} - p)'N_n'Q N_n(\hat{p} - p). \end{aligned} \tag{5.31}$$

The first term of (5.31) is just Q_0, and is already in the desired form. For the other terms, we require a link between $\hat{p} - p$ and $y - N_n p$, as we did in dealing with A_1. The link takes a somewhat different form here, since we are dealing with $\hat{p} - p$ as a vector; Dennis and Schnabel (1983, p. 74) give a suitable mean value theorem, from which it follows that

$$N_n(\hat{p} - p) = \left[\int_0^1 V_n\{\beta + t(\hat{\beta} - \beta)\}\, dt\right] X(\hat{\beta} - \beta),$$

since the (matrix) derivative of p with respect to β is $V_n(\beta)X$. The integral is to be interpreted elementwise. Let $\hat{V}_n$ denote the integral, and note that, from Fahrmeir and Kaufmann, $\hat{\beta} \xrightarrow{P} \beta$, so that $n^{-1}\hat{V}_n \xrightarrow{P} n^{-1}\int_0^1 V_n\, dt = V$. The link between $\hat{\beta} - \beta$ and $y - N_n p$ is the same as before, adjusted for the vector-matrix notation; it is

$$L_n'(\hat{\beta} - \beta) = \hat{U}_n L_n^{-1} X'(y - N_n p),$$

where $\hat{U}_n$ is a matrix converging in probability to the identity. Putting these two results together, we have

$$N_n(\hat{p} - p) = \hat{V}_n X(L_n \hat{U}_n L_n')^{-1} X'(y - N_n p). \tag{5.32}$$

Looking now at the second term of (5.31), and applying (5.32), we obtain

$$n^{-1}(y - N_n p)' Q N_n(\hat{p} - p) = n^{-1}(y - N_n p)' Q \hat{V}_n X(L_n \hat{U}_n L_n')^{-1} X'(y - N_n p).$$

The quantity $n^{-1/2}(y - N_n p)$ is bounded in probability (it has mean zero and variance which tends to V), so the convergence in probability of the second term of (5.31) rests on the convergence of the terms depending on n. $\hat{V}_n/n$ converges in probability to V; meanwhile, $\hat{U}_n$ converges in probability to the identity matrix, so that $L_n \hat{U}_n L_n'/n$ converges in probability to the same limit as F_n/n, which is here $X'VX$. The terms depending on n thus converge in probability to $VX(X'VX)^{-1}X'$, the n's cancelling. Letting H denote this limit, the quantity as a whole converges to

$$n^{-1}(y - N_n p)' Q H (y - N_n p). \tag{5.33}$$

The third term of (5.31) is the transpose of the second, so its limit must be

$$n^{-1}(y - N_n p)' H' Q (y - N_n p). \tag{5.34}$$

The fourth and final term of (5.31) also contributes to the asymptotic distribution of Q_1:

$$\begin{aligned} & n^{-1}(\hat{p} - p)' N_n' Q N_n(\hat{p} - p) \\ &= n^{-1}(y - N_n p)' X(L_n \hat{U}_n L_n')^{-1} X' \hat{V}_n Q \hat{V}_n X(L_n \hat{U}_n L_n')^{-1} X'(y - N_n p) \\ &\to n^{-1}(y - N_n p)' X(X'VX)^{-1} X' V Q V X(X'VX)^{-1} X'(y - N_n p) \\ &= n^{-1}(y - N_n p)' H' Q H (y - N_n p). \end{aligned} \tag{5.35}$$

Combining the first term of (5.31) with (5.33), (5.34) and (5.35), we find that Q_1 converges to the same distribution as

$$\begin{aligned} Q_1^* &= n^{-1}(y - N_n p)'(Q - QH - H'Q - H'QH)(y - N_n p) \\ &= n^{-1}(y - N_n p)'(I - H)'Q(I - H)(y - N_n p), \end{aligned}$$

which was to be proved.

It follows that the asymptotic distribution of Q_1 is $\sum_{i=1}^{J} \lambda_i z_i^2$, where the z_i are independent standard normal random variables, and now the λ_i are the eigenvalues of $V^{1/2}(I - H)'Q(I - H)V^{1/2}$.

It is also worth noting that the matrix H is idempotent, and indeed resembles the "hat matrix" in weighted least squares. This comes about because estimation in generalized linear models can be thought of as an iterated weighted least squares problem, and for large n, iterations after the first make almost no difference — "in the limit, everything is linear".

5.3.5 Statistic Q_2

As with the area family of statistics, we now show that Q_2 has the same asymptotic distribution as was just found for Q_1. The result rests on the continuity of $\hat{p}$ as a function of β.

Theorem 5.23 *Under the conditions of Theorem 5.22, the asymptotic distribution of the statistic Q_2 is the same as that of Q_1.*

Proof: Q_2 can be written as

$$\begin{aligned} Q_2 &= n^{-1}(y - N_n\hat{p})'\hat{Q}_n(y - N_n\hat{p}) \\ &= n^{-1}(y - N_n\hat{p})'Q(y - N_n\hat{p}) + n^{-1}(y - N_n\hat{p})'(\hat{Q}_n - Q)(y - N_n\hat{p}) \\ &= Q_1 + n^{-1}(y - N_n\hat{p})'(\hat{Q}_n - Q)(y - N_n\hat{p}). \end{aligned}$$

As was shown in Theorem 5.22, $n^{-1/2}(y - N_n\hat{p})$ converges in distribution to a normal distribution with mean zero and a variance that does not depend on n, and so is bounded in probability. It thus suffices to show that $\hat{Q}_n - Q$ converges in probability to zero. The (i, j)-th element of this matrix is

$$\begin{aligned} &\{1 - \max(\hat{p}_i, \hat{p}_j)\} - \{1 - \max(p_i, p_j)\} \\ &= \max(p_i, p_j) - \max(\hat{p}_i, \hat{p}_j). \end{aligned} \tag{5.36}$$

Now, the condition of Fahrmeir and Kaufmann (1985), namely that

$$\lim_{n\to\infty} \max_{1\le i\le d} x_i' F_n^{-1} x_i = 0,$$

ensures that $\hat{\beta} - \beta \xrightarrow{P} 0$. But there is a chain of functions linking $\hat{\beta} - \beta$ to (5.36), all of which are continuous: $\eta = X\beta$, the linear predictor, is linear and therefore continuous; $p : R^p \to R^d$ is a continuous though nonlinear function of β; $\max(p_i, p_j)$ is a continuous (though not differentiable) function of p for all i and j. Thus $\hat{Q}_n - Q$ is a continuous function of $\hat{\beta} - \beta$. Finally, since any continuous function of a convergent (in probability) sequence of random variables converges in probability to the corresponding limit, we have shown that $\hat{Q}_n - Q \xrightarrow{P} 0$, and thus that the statistics Q_2 and Q_1 have the same asymptotic distribution.

5.3.6 A more general result for statistic Q_0

The theory of Guttorp and Lockhart (1988) provides the means for a more general attack on Q_0. As was seen above, the restriction to a finite number of design points enables us to assert the asymptotic normality of each suitably standardized y_i, and thence to complete relatively straightforward convergence proofs. However, when each y_i is a Bernoulli trial with a possibly different success probability, the asymptotic distribution of a statistic must come from a central-limit-like argument whereby the statistic is expressed as a large number of independent quantities which are each almost certainly small, but not so small as to be converging to zero. Guttorp and Lockhart provide the theory to make such an argument rigorous in our case.

Equation (5.23) gives the general form of Q_0. It will, however, be more convenient for the Theorem given below if we write

$$Q_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} w_i w_j,$$

where

$$\begin{aligned} w_i &= \frac{y_i - p_i}{\sqrt{p_i(1-p_j)}}, \\ m_{ij} &= \frac{\{1 - \max(p_i, p_j)\}\sqrt{p_i(1-p_i)p_j(1-p_j)}}{n}. \end{aligned}$$

The m_{ij} can be thought of as elements of an $n \times n$ matrix M.

Theorem 5.24 *Let $B_\epsilon = \{i : \epsilon \le p_i \le 1 - \epsilon\}$, and let $Q_0^* = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} z_i z_j$ where the z_i are independent standard normal random variables. If, for some $\epsilon > 0$, $|B_\epsilon|/n \ge \eta > 0$*

for all n, then

$$\sup_{-\infty<u<\infty} |P(Q_0 \le u) - P(Q_0^* \le u)| \to 0$$

as $n \to \infty$.

Proof: Since $E(w_i^2) = E(z_i^2) = 1$, $E(Q_0) = E(Q_0^*) = \sum_{i=1}^n m_{ii} = \mu$, say. Let $\sigma^2 = 2\,\mathrm{tr}M^2$. Then, provided that $\sigma > 0$, we can write

$$\begin{aligned}&\sup_{-\infty<u<\infty} |P(Q_0 \le u) - P(Q_0^* \le u)| \\ &= \sup_{-\infty<u<\infty} \left| P\left(\frac{Q_0 - \mu}{\sigma} \le u\right) - P\left(\frac{Q_0^* - \mu}{\sigma} \le u\right)\right|\end{aligned}$$

because Q_0 and Q_0^* have undergone the same linear transformation, and therefore the value u' of u at which the maximum occurs on the left-hand side becomes $(u' - \mu)/\sigma$ on the right-hand side, and the difference in probabilities is the same. It suffices, therefore, to demonstrate that the right-hand side tends to zero as $n \to \infty$.

We now partition the set of pairs (i, j) as follows: let

$$\begin{aligned}A_\epsilon &= \{(i,j): i \ne j, i \in B_\epsilon, j \in B_\epsilon\}, \\ \bar{A}_\epsilon &= \{(i,j): i \ne j, (i,j) \notin A_\epsilon\}, \\ J_\epsilon &= \{(i,j): i = j, i \in B_\epsilon\}, \\ \bar{J}_\epsilon &= \{(i,j): i = j, i \notin B_\epsilon\}.\end{aligned}$$

Note that $|J_\epsilon| = |B_\epsilon|$ and $|A_\epsilon \cup J_\epsilon| = |B_\epsilon|^2$. Furthermore,

$$|A_\epsilon| = |B_\epsilon|^2 - |B_\epsilon| = |B_\epsilon|^2(1 - 1/|B_\epsilon|) \ge |B_\epsilon|^2/2 \ge n^2\eta^2/2$$

for all sufficiently large n, since $|B_\epsilon| \to \infty$ with n.

With these definitions, we can bound σ^2 above and below:

$$\sigma^2 = 2\,\mathrm{tr}M^2 = \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \{1 - \max(p_i, p_j)\}^2 p_i(1-p_i)p_j(1-p_j) \le 2 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{8};$$

and

$$\begin{aligned}\sigma^2 &\ge \frac{2}{n^2} \sum_{(i,j)\in A_\epsilon \cup J_\epsilon} \{1 - \max(p_i, p_j)\}^2 p_i(1-p_i)p_j(1-p_j) \\ &\ge 2\epsilon^4(1-\epsilon)^2|A_\epsilon \cup J_\epsilon|/n^2 \\ &\ge 2\epsilon^4(1-\epsilon)^2\eta^2 = \eta_1, \text{ say},\end{aligned}$$

with $\eta_1 > 0$.

We now focus our attention on a fixed value ϵ' that is no larger than the ϵ stated in the Theorem, and decompose Q_0 and Q_0^*. Note first that

$$Q_0 - \mu = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}w_i w_j - \sum_{i=1}^{n} m_{ii} = \sum_{i=1}^{n} m_{ii}(w_i^2 - 1) + \sum_{i\neq j} m_{ij}w_i w_j.$$

Thus

$$\begin{aligned}\frac{Q_0 - \mu}{\sigma} &= \frac{1}{\sigma}\sum_{i=1}^{n} m_{ii}(w_i^2 - 1) + \frac{1}{\sigma}\sum_{(i,j)\in A_{\epsilon'}} m_{ij}w_i w_j + \frac{1}{\sigma}\sum_{(i,j)\in A_{\epsilon'}} m_{ij}w_i w_j \\ &= T_1 + T_2 + T_3, \text{ say.}\end{aligned}$$

Letting T_k^* be T_k with w_i replaced by z_i, we obtain the same decomposition for $(Q_0^* - \mu)/\sigma$. We will show later that the terms with $k = 3$ dominate the asymptotic behaviour in both cases, and so we concentrate on T_3 and T_3^* for the moment.

Define $\tilde{M}$ to be the result of setting the diagonal elements of M to zero and deleting each row and column i for which $i \notin B_{\epsilon'}$. The matrix $\tilde{M}$ is therefore a $|B_{\epsilon'}| \times |B_{\epsilon'}|$ matrix with diagonal elements zero, and contains $|A_{\epsilon'}|$ nonzero elements. The subscripts i and j will be used equally for elements M and $\tilde{M}$, and quantities w_i, w_j, z_i, z_j, even though the matrices are of different sizes. No confusion should arise, however.

Let $\tilde{\sigma}^2 = 2\,\mathrm{tr}\tilde{M}^2$. Clearly $\tilde{\sigma}^2 \leq \sigma^2$, since $\tilde{M}$ was obtained from M by deleting some elements. Also, $\tilde{m}_{ij}$ is, for all $i \neq j$, based only on probabilities that lie between ϵ' and $1-\epsilon'$, so that such $\tilde{m}_{ij} \geq \epsilon'^2(1-\epsilon')/n$ (because $p_k(1-p_k) \geq \epsilon'(1-\epsilon')$ and $\max(p_k, p_l) \leq 1-\epsilon'$). Therefore

$$\begin{aligned}\tilde{\sigma}^2 = 2\,\mathrm{tr}\tilde{M}^2 &\geq 2\epsilon'^4(1-\epsilon')^2|A_{\epsilon'}|/n^2 \\ &\geq \epsilon'^4(1-\epsilon')^2\eta^2 = \eta_2 > 0,\end{aligned}$$

since for $\epsilon' \leq \epsilon$, $|B_{\epsilon'}| \geq |B_\epsilon|$, and where the above defines η_2.

Having shown that $\tilde{\sigma}$ is bounded away from zero and possesses an upper bound, it makes sense to consider the closeness of the distributions of $\tilde{T}_3 = \sum_i\sum_j(\tilde{m}_{ij}/\tilde{\sigma})w_i w_j$ and $\tilde{T}_3^* = \sum_i\sum_j(\tilde{m}_{ij}/\tilde{\sigma})z_i z_j$. In particular, we would like to apply Corollary 1 of Guttorp and Lockhart (1988).

The w_i are uniformly square integrable because they are bounded, for $\epsilon' \leq \frac{1}{2}$, by $\sqrt{(1-\epsilon')/\epsilon'}$. Also,

$$\max_i \sum_j \frac{\tilde{m}_{ij}^2}{\tilde{\sigma}^2} \leq \frac{1}{\tilde{\sigma}^2}\max_i \sum_j m_{ij}^2$$

$$
\begin{aligned}
&= \frac{1}{\tilde{\sigma}^2}\max_i \frac{1}{n^2}\sum_j \{1-\max(p_i,p_j)\}^2 p_i(1-p_i)p_j(1-p_j) \\
&\leq \frac{1}{\tilde{\sigma}^2}\frac{1}{n^2}\frac{n}{16} = \frac{1}{16n\tilde{\sigma}^2}.
\end{aligned}
$$

Since $\tilde{\sigma}^2 \geq \eta_2 > 0$, this quantity tends to zero as $n \to \infty$. The corollary applies, therefore, and so, as $n \to \infty$,

$$\sup_u |P(\tilde{T}_3 \leq u) - P(\tilde{T}_3^* \leq u)| \to 0.$$

Now,

$$\tilde{T}_3 = \frac{1}{\tilde{\sigma}}\sum_i\sum_j \tilde{m}_{ij}w_i w_j = \frac{\sigma}{\tilde{\sigma}}\frac{1}{\sigma}\sum_{(i,j)\in A_{\epsilon'}} m_{ij}w_i w_j = \frac{\sigma}{\tilde{\sigma}}T_3.$$

The same applies to $\tilde{T}_3^*$ with z_i, z_j replacing w_i, w_j. Thus $P(T_3 \leq u) - P(T_3^* \leq u) = P(\tilde{T}_3 \leq u\sigma/\tilde{\sigma}) - P(\tilde{T}_3^* \leq u\sigma/\tilde{\sigma})$. Since the distribution functions of $\tilde{T}_3$ and $\tilde{T}_3^*$ are being compared at the same place, and the supremum of this difference is, in absolute value, tending to zero, it follows that

$$\sup_u |P(T_3 \leq u) - P(T_3^* \leq u)| \to 0$$

as $n \to \infty$ as well.

The foregoing applies to a fixed $\epsilon' \leq \epsilon$. However, by using a result from Chung (1974), we can construct a sequence $\{\epsilon'_n\} \to 0$ for which the result still holds. Let $m = 1/\epsilon'$, and let $U(m,n) = \sup_u |P(T_3 \leq u) - P(T_3^* \leq u)|$, where T_3 and T_3^* depend on both m and n. Since $\lim_{n\to\infty} U(m,n) = 0$, Lemma 1 of Chung (1974, p. 206) shows that there exists a sequence $\{m_n\} \to \infty$ such that $\lim_{n\to\infty} U(m_n,n) = 0$. Then take $\epsilon'_n = 1/m_n$ to obtain the desired sequence.

In the remainder of the proof, we use the sequence $\{\epsilon'_n\}$ just constructed. Next, we show that $T_1, T_2 \xrightarrow{P} 0$.

Since $T_1 = \sigma^{-1}\sum_{i=1}^n m_{ii}(w_i^2 - 1)$ and $E(w_i^2) = 1$, $E(T_1) = 0$. Also,

$$
\begin{aligned}
\text{var}(T_1) &= \frac{1}{\sigma^2}\sum_{i=1}^n m_{ii}^2 \text{var}(w_i^2) \\
&= \frac{1}{n^2\sigma^2}\sum_{i=1}^n p_i^2(1-p_i)^4 \text{var}\left\{\frac{(y_i-p_i)^2}{p_i(1-p_i)}\right\} \\
&= \frac{1}{n^2\sigma^2}\sum_{i=1}^n (1-p_i)^2 \text{var}\{(y_i-p_i)^2\} \\
&\leq \frac{1}{n\sigma^2},
\end{aligned}
$$

since $(y_i - p_i)^2$, its variance, and $(1 - p_i)^2$ are all positive and less than 1. Since $\sigma^2 \geq \eta_1$, $\lim_{n\to\infty} \text{var}(T_1) = 0$. It follows that $T_1 \xrightarrow{P} 0$. Because $\text{var}(z_i^2)$ is also bounded, a similar argument shows that $T_1^* \xrightarrow{P} 0$.

For $T_2 = \sigma^{-1} \sum_{(i,j)\in A_{\epsilon_n'}} m_{ij} w_i w_j$, we find that $E(T_2) = 0$ since $E(w_i) = E(w_j) = 0$. Since the same is true for z_i, z_j, $E(T_2^*) = 0$ as well. The two quantities also have the same variance:

$$\text{var}(T_2) = \text{var}(T_2^*) = \frac{1}{\sigma^2} \sum_{(i,j)\in A_{\epsilon_n'}} m_{ij}^2.$$

Since one of $p_i, 1 - p_i, p_j, 1 - p_j$ is less than ϵ_n' and the others, as well as $1 - \max(p_i, p_j)$, are bounded above by 1, $m_{ij}^2 \leq \epsilon_n'/n^2$ for each term in the sum. There are no more than n^2 terms (indeed, the number of terms is typically only a small fraction of n^2), so that

$$\text{var}(T_2) = \text{var}(T_2^*) \leq \frac{\epsilon_n'}{\sigma^2},$$

which tends to zero since $\lim_{n\to\infty} \epsilon_n' = 0$ and $\sigma^2 \geq \eta_1 > 0$. Thus $T_2, T_2^* \xrightarrow{P} 0$.

Having disposed of T_1 and T_2 and their starred counterparts, we now look at the mean and variance of T_3 and T_3^*:

$$E(T_3) = \frac{1}{\sigma} \sum_{(i,j)\in A_{\epsilon_n'}} m_{ij} E(w_i w_j) = 0$$

since $i \neq j$ and $E(w_i) = 0$. The same applies for $E(T_3^*)$. Meanwhile,

$$\text{var}(T_3) = \text{var}(T_3^*) = \frac{1}{\sigma^2} \sum_{(i,j)\in A_{\epsilon_n'}} m_{ij}^2 = \frac{\tilde{\sigma}^2}{\sigma^2} \leq 1.$$

Since T_3 and T_3^* have constant means and bounded variances for all n, they are both bounded in probability; the families of distributions of T_3 and of T_3^* are "tight".

Now suppose that the theorem were false. This would mean that there exist sequences $\{n_k\}$ and $\{u_{n_k}\}$ and a δ such that

$$\left| P\left(\frac{Q_0 - \mu}{\sigma} \leq u_{n_k}\right) - P\left(\frac{Q_0^* - \mu}{\sigma} \leq u_{n_k}\right) \right| \geq \delta \quad \text{for all } k. \tag{5.37}$$

In other words, there has to exist a subsequence of points u_{n_k} for which the distribution functions of the standardized Q_0 and Q_0^* do not get close together. Passing to a subsequence is necessary because there may be some values of n for which the distribution functions are arbitrarily close together. Consider the (sub)sequence of distribution functions $\{H_{n_k}^*\}$

of T_3^*, where the sequence $\{n_k\}$ is that of the counterexample described above. By the Helly Selection Theorem (Theorem 25.10 of Billingsley, 1995, p. 336), there exists a further subsequence $\{H^*_{n_{k(j)}}\}$ and a bounded function H^*_∞ such that $H^*_\infty(u) = \lim_{j\to\infty} H^*_{n_{k(j)}}(u)$ exists for all u where $H^*_\infty(u)$ is continuous. Furthermore, since the family $\{H^*_n\}$ is tight, and therefore any subfamily is tight, this limit is a distribution function.

The same considerations apply to T_3, since its family of distribution functions is also tight. Thus $H_\infty(u) = \lim_{j\to\infty} H_{n_{k(j)}}(u)$ exists for all u such that $H_\infty(u)$ is continuous. But since $\sup_u |H_{n_k}(u) - H^*_{n_k}(u)| \to 0$ and $H_{n_k}(u) \to H_\infty(u)$, it must be that $H_\infty(u) = H^*_\infty(u)$.

At this point, we would like to be able to assert that H_∞ is continuous. This is in fact the case, but we defer its non-trivial proof to a following Lemma.

Since $(Q_0 - \mu)/\sigma$ is equal to the sum of T_3 and some quantities that converge in probability to zero, $(Q_0 - \mu)/\sigma \xrightarrow{D} Y$, where Y has distribution function H_∞. The same is true of $(Q_0^* - \mu)/\sigma$, since it is the sum of T_3^* and some quantities that converge in probability to zero. Since $H_\infty(u)$ is continuous, it follows that

$$P\left(\frac{Q_0 - \mu}{\sigma} \le u\right) \to H_\infty(u)$$
$$\text{and } P\left(\frac{Q_0^* - \mu}{\sigma} \le u\right) \to H_\infty(u)$$

for *all* u. While this demonstrates the pointwise convergence of the probabilities, it is not quite enough to establish the uniform convergence that we need. However, we can use a Lemma of Chung (1974, p. 133); taking the Q of the Lemma to be, for example, the set of rational numbers and the J of the Lemma to be empty since H_∞ is continuous, we can conclude that $P\{(Q_0 - \mu)/\sigma \le u\}$ converges to $H_\infty(u)$ uniformly, and so does $P\{(Q_0^* - \mu)/\sigma \le u\}$. Thus

$$\sup_u \left| P\left(\frac{Q_0 - \mu}{\sigma} \le u\right) - H_\infty(u)\right| \to 0,$$

and equally for Q_0^*. As a result,

$$\sup_u \left| P\left(\frac{Q_0 - \mu}{\sigma} \le u\right) - P\left(\frac{Q_0^* - \mu}{\sigma} \le u\right)\right|$$
$$\le \sup_u \left| P\left(\frac{Q_0 - \mu}{\sigma} \le u\right) - H_\infty(u)\right| + \sup_u \left| P\left(\frac{Q_0^* - \mu}{\sigma} \le u\right) - H_\infty(u)\right|$$
$$\to 0$$

since both terms do. As observed at the beginning, this suffices to show that

$$\sup_u |P(Q_0 \le u) - P(Q_0^* \le u)| \to 0$$

and hence the proof is complete.

This proof rested on a Lemma, which we now prove:

Lemma 5.25 *The function H_∞ defined in the proof of the preceding Theorem is continuous.*

Proof: Since $T_3^* \xrightarrow{D} Y$, where $P(Y \leq u) = H_\infty(u)$, every subsequence of the sequence of T_3^* also converges in distribution to Y. In the preceding proof, we showed that $0 < \eta_2 \leq \tilde{\sigma}^2 \leq \sigma^2 \leq \frac{1}{8}$, so that σ^2 and $\tilde{\sigma}^2$ are both bounded above and bounded away from zero. Thus, from the counterexample subsequence we can extract a subsequence for which both $\tilde{\sigma}^2 \to \tilde{\sigma}^2_\infty$ and $\sigma^2 \to \sigma^2_\infty$, with $\eta_2 \leq \sigma^2_\infty, \tilde{\sigma}^2_\infty \leq \frac{1}{8}$. (For example, the required subsequence can be found by first extracting a subsequence for which the sequence of $\tilde{\sigma}^2$ converges, and then extracting a further subsequence for which σ^2 converges.) For such a subsequence,

$$P(\tilde{T}_3^* \leq u) = P\left(T_3^* \leq \frac{u\tilde{\sigma}}{\sigma}\right) \to H_\infty\left(\frac{u\tilde{\sigma}_\infty}{\sigma_\infty}\right).$$

Now, from the subsequence just constructed, we can follow the proof of Corollary 1 of Guttorp and Lockhart (1988) to extract a further subsequence for which

$$\tilde{T}_3^* \xrightarrow{D} \lambda_0 z_0 + \sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1),$$

for scalars $\lambda_0, \lambda_1, \ldots$ with $2\sum_{i=1}^{\infty} \lambda_i^2 + \lambda_0^2 = 1$. In other words,

$$H_\infty\left(\frac{u\tilde{\sigma}_\infty}{\sigma_\infty}\right) = P\left\{\lambda_0 z_0 + \sum_{i=1}^{\infty} \lambda_i (z_i^2 - 1) \leq u\right\}.$$

Since, by Feller (1968, vol. 2, p. 144, Theorem 1), the right-hand side is continuous (if $\lambda_0 \neq 0$, it is the convolution of a normal random variable, which is continuous, with some other random variables, and if $\lambda_0 = 0$, it is the sum of at least one chi-squared random variable, also continuous), H_∞ is also continuous, as we wished to prove.

5.4 Finite samples

When carrying out a test of fit in practice, one would typically use the asymptotic distribution of the test statistic, hoping that this distribution is a reasonable approximation to the exact distribution. We assess the validity of this approach by simulations on three examples; in each case, three designs of the same type are chosen with n approximately equal to 20, 50 and 100, so that the effect of increasing n can be seen.

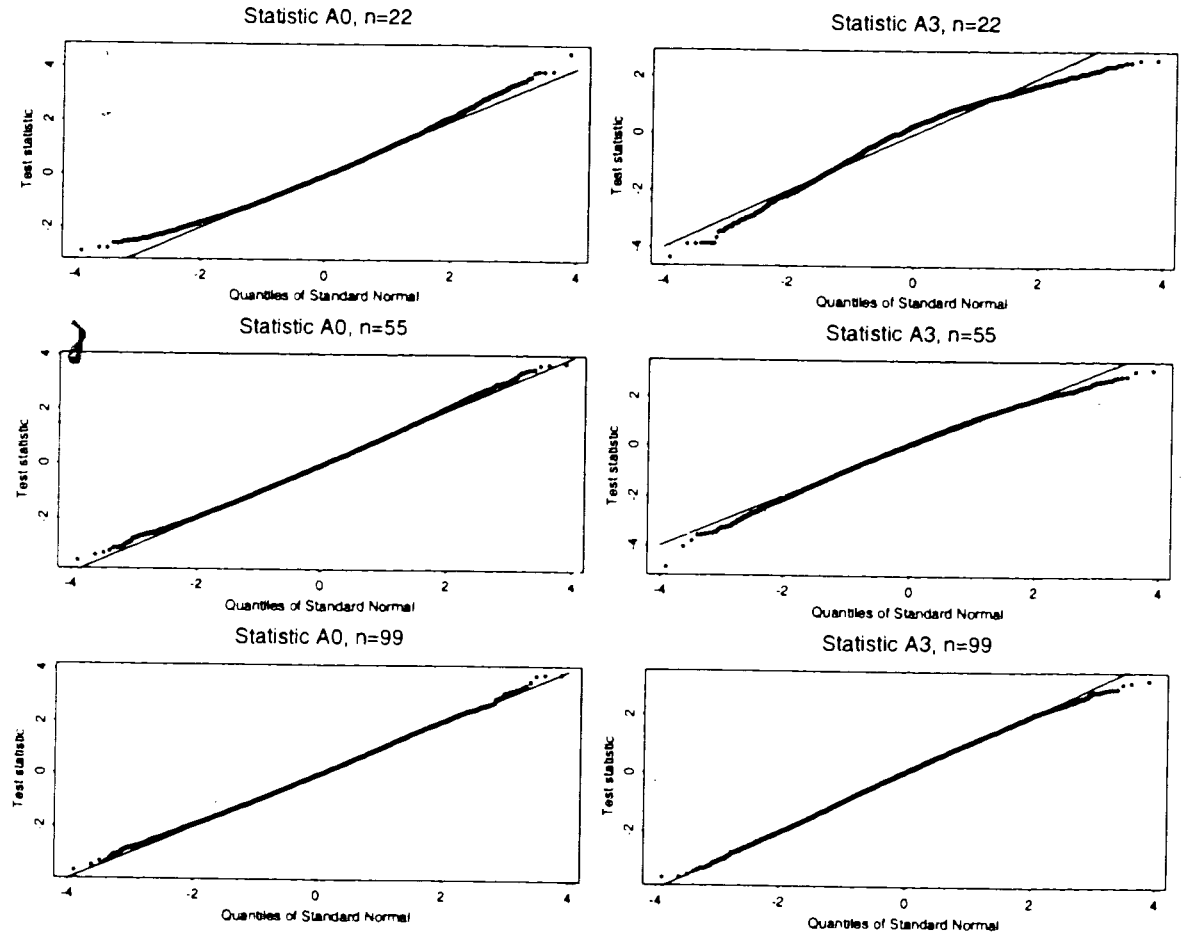

Figure 5.1: Normal Q-Q plots for statistics A_0 and A_3, example 1

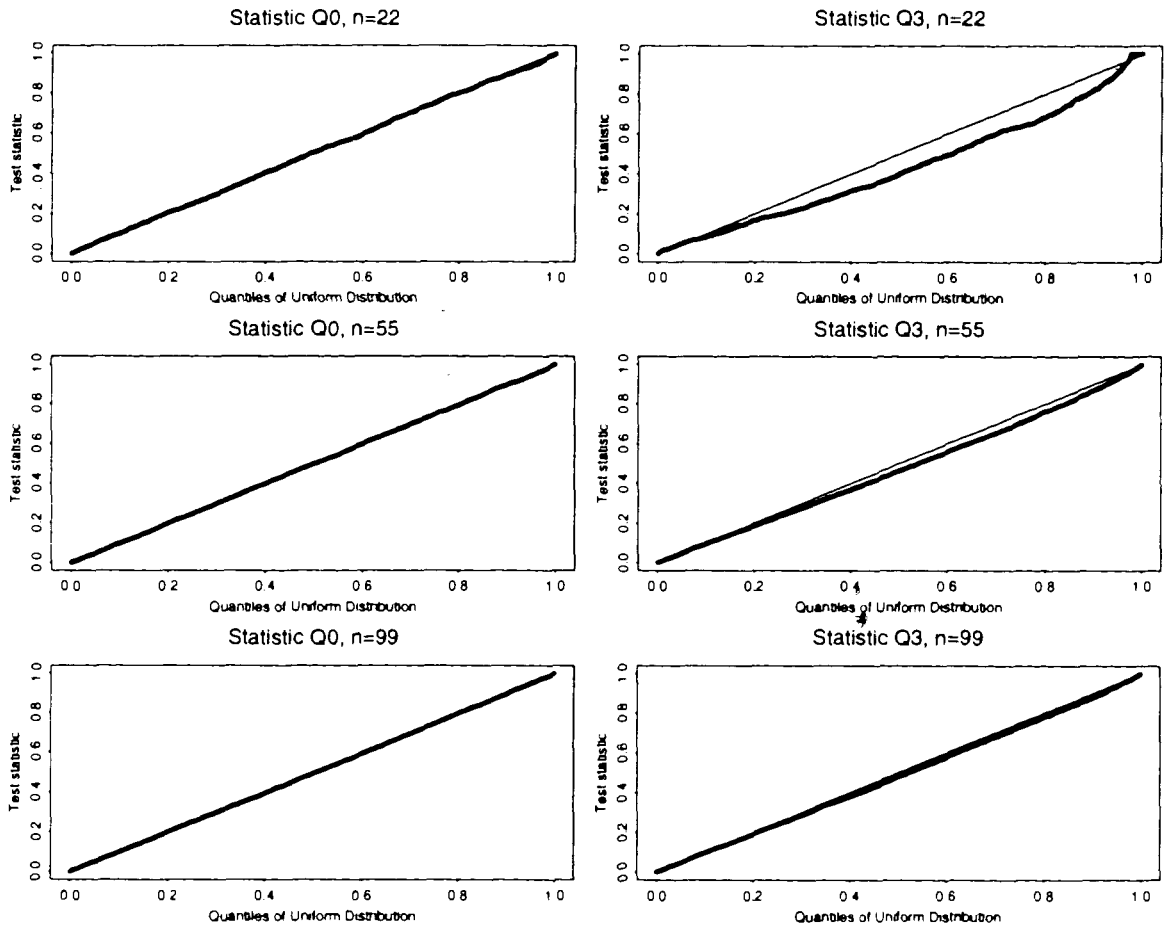

Figure 5.2: Uniform Q-Q plots for statistics Q_0 and Q_3, example 1

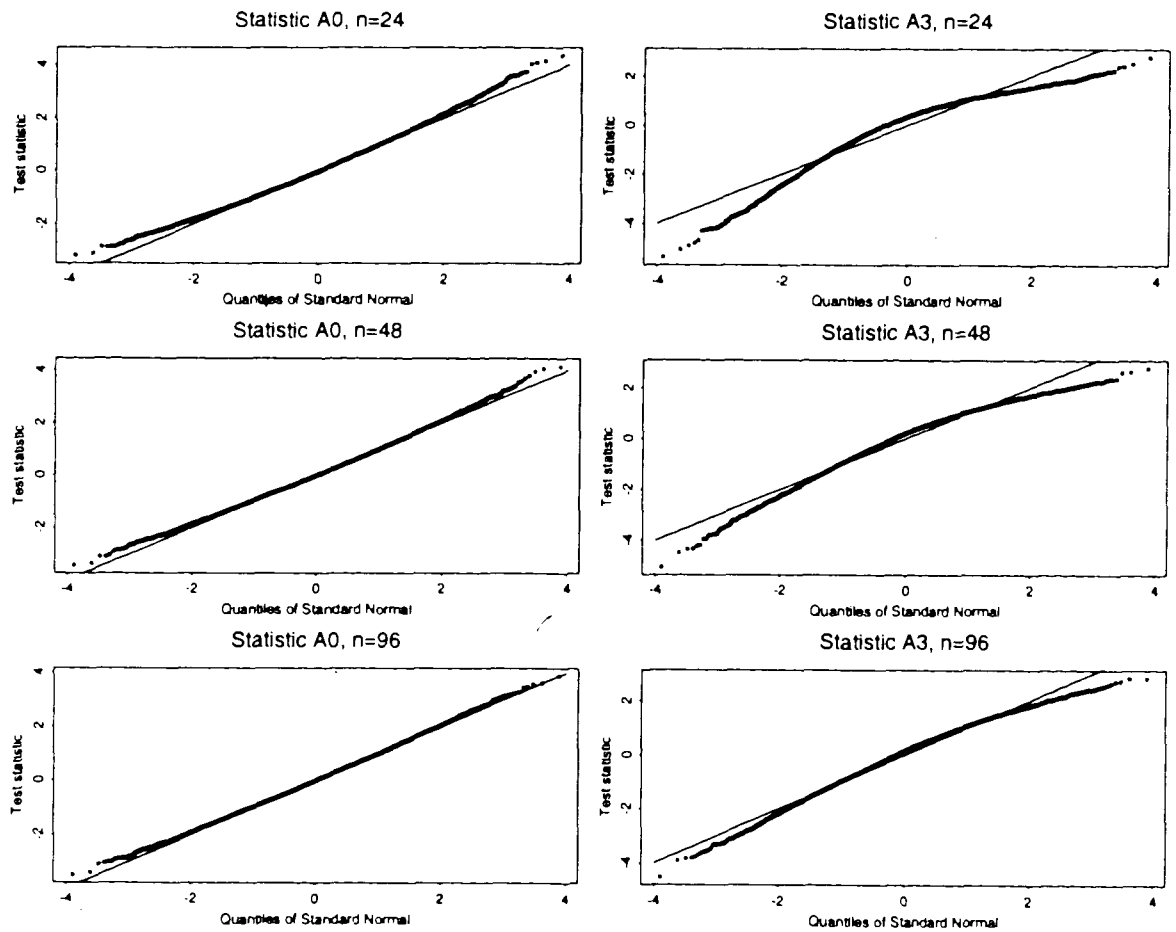

Figure 5.3: Normal Q-Q plots for statistics A_0 and A_3, example 2

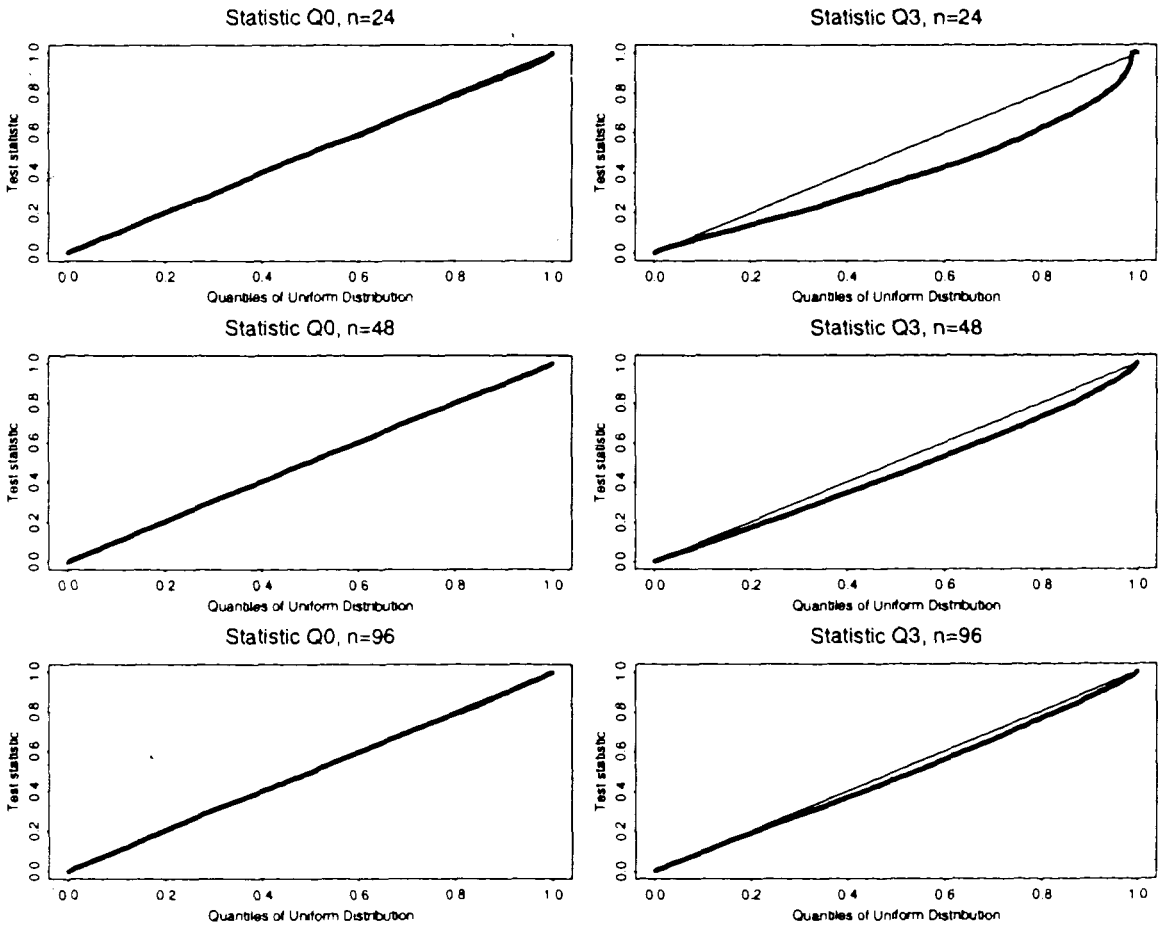

Figure 5.4: Uniform Q-Q plots for statistics Q_0 and Q_3, example 2

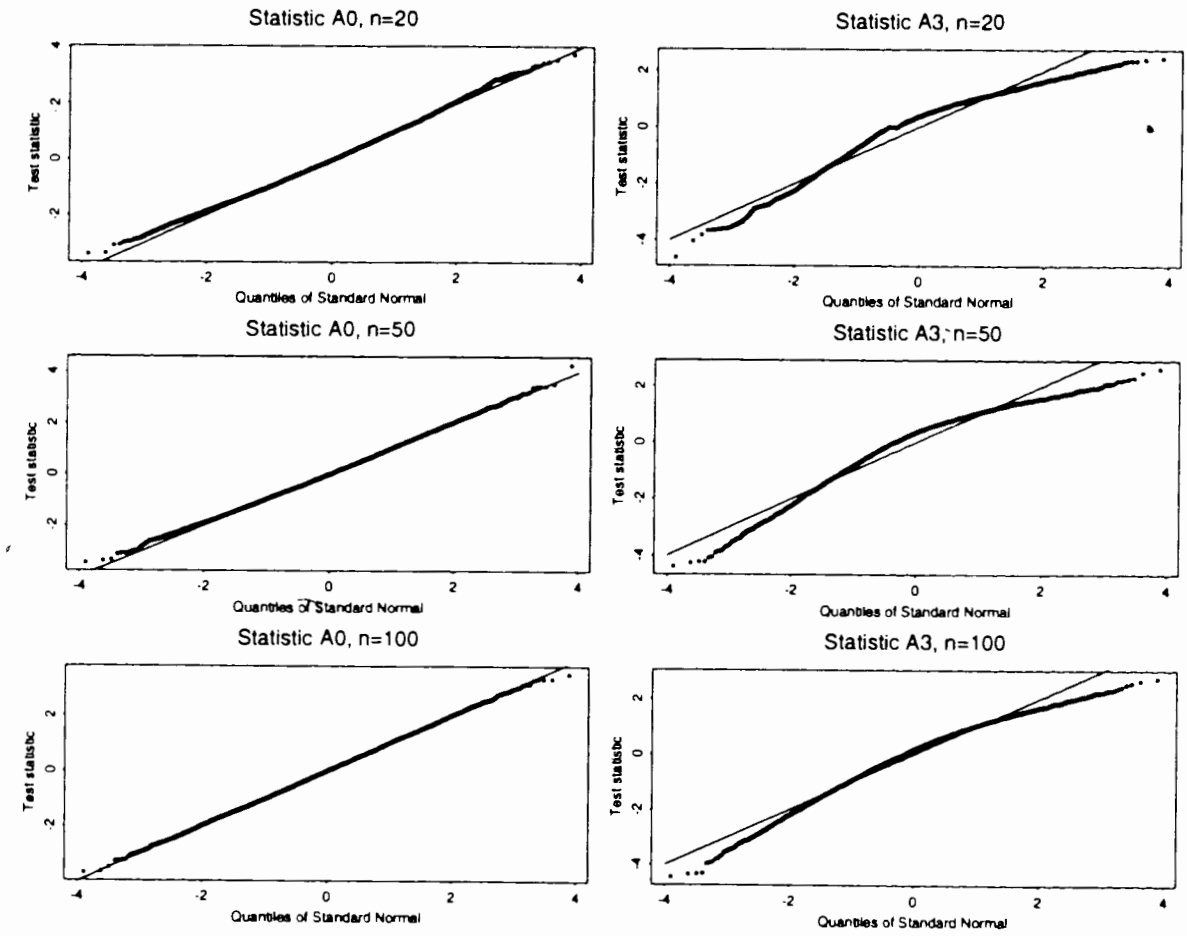

Figure 5.5: Normal Q-Q plots for statistics A_0 and A_3, example 3

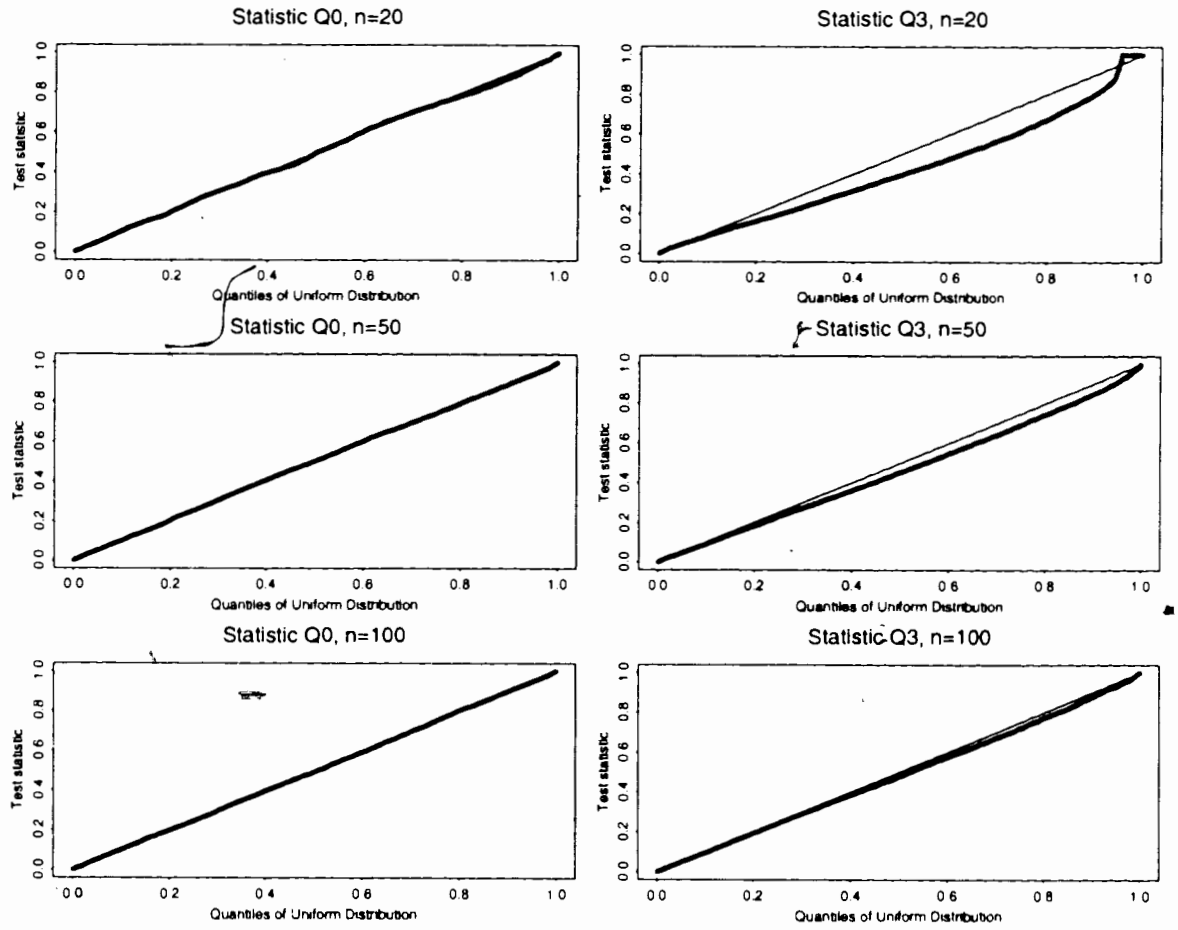

Figure 5.6: Uniform Q-Q plots for statistics Q_0 and Q_3, example 3

All three examples contain an intercept, and contain a true intercept of zero and true slopes of 1. (Provided that the slope coefficients are not exactly zero, in which case any model will fit equally well, the choice of coefficents is not crucial, since, for example, one can double a slope coefficient and halve the corresponding x-variable to leave the probabilities the same. Given our choice of intercept and slope, the values for the x-variables were chosen to give a mixture of high, low and intermediate success probabilities.) The first example contains one x-variable, taking values between -2 and 2 in steps of 0.4. This design contains 11 points, and is then replicated 2, 5 and 9 times to produce logistic regressions with $n = 22, 55, 99$. The second example has two x-variables: x_1 takes values between -2 and 1 in steps of 0.6, and x_2 takes the values $-1, 0, 1, 2$. Each possible combination of values of x_1 and x_2 is taken, producing a design with $n = 24$; this design is then replicated twice and four times to produce designs with $n = 48$ and $n = 96$. In these first two examples, the passage to infinity can be viewed as an increasing number of replications of a fixed design, so the simpler theory of Section 5.3 holds for the Q-statistics. The third example, on the other hand, has one x-variable whose values are random samples of sizes 20, 50 and 100 from a standard normal distribution (the *same* design is used for all the simulations for a particular value of n), and so the most general theory is needed to obtain an asymptotic distribution.

For each example, we give Q-Q plots for the statistics A_0 and Q_0, for which all parameters are known, and for the statistics called here A_3 and Q_3, which are A_2 and Q_2 with the parameters of their asymptotic distributions estimated from the data. These latter statistics are typically the ones that would be used in practice. The A-statistics are asymptotically normally distributed, and so normal Q-Q plots are shown. The Q-statistics have distributions that are weighted sums of chi-squared random variables; in the case of Q_3, the distribution is different on each simulation, because the parameters λ_i are estimated from the data. For this reason, we have adopted the attitude for Q_3 (and, for ease of comparison, for Q_0 also) that the P-value is the test statistic, and therefore a Q-Q plot against uniform order statistics is appropriate. These plots are shown in Figure 5.1–5.6.

Although the rate of convergence appears to differ in the three examples, with the two-x-variable Example 2 showing the lowest accuracy in the approximations, some patterns are evident. The statistics A_0 and Q_0 converge very rapidly to their asymptotic distributions; even for the smallest sample sizes shown here, the approximations are very good. The statistics A_3 and Q_3 converge rather more slowly, and in some cases the approximations are still poor even for n near 100; however, convergence does appear to be taking place, even if

slowly. We note also that the binary responses in a logistic regression do not, individually, convey much information (compare, for example, binary-response opinion polls in which a sample size of 1000 is typically required to achieve the desired accuracy), so that $n = 100$ is by no means an especially "large" sample in this kind of experiment. It is also worth pointing out that for Q_3, it is the lower tail that is of most interest, and this tail is approximated better than any other part of the distribution, at least in the examples considered here.

5.5 Power considerations

We have not, so far, carried out a power study to assess the ability of our proposed tests and their competitors to reject false null hypotheses. Until this is done, it is difficult to do more than speculate about the relative performances of the tests. Nonetheless, the quadratic statistics of Section 5.3 may be expected to perform well against a variety of alternatives in which the true p_i diverge from the hypothesized values in a smooth way, as will typically be the case when the link function has been misspecified.

The same may not be true of the area statistics of Section 5.2, since there may be positive and negative deviations of the process $X_n(p)$ from zero that cancel each other out when the statistics are calculated. For example, suppose that the true relationship between a single x-variable and logit p_i is hypothesized as linear but is actually quadratic. Then it may be that for large and small x, the true p_i are smaller than the hypothesized, while for the remaining x-values, the true p_i are larger. As a result, the process $X_n(p)$ will generally be negative for small x and positive for large x, and the area statistics will exhibit lesser power. On the other hand, a misspecified link function will usually result in the hypothesized p_i being too large in one tail and too small in the other; in this case, $X_n(p)$ will generally have the same sign for all p, and the area statistics can be expected to have reasonable power.

Bibliography

[1] Albert, A. and Anderson, J. A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10.

[2] Atkinson, A. C. and Donev, A. N. (1992) *Optimum Experimental Designs.* Oxford University Press.

[3] Billingsley, P. (1995) *Probability and measure.* Wiley, New York.

[4] Bradley, R. A. and Terry, M. E. (1952) The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345.

[5] Bühlmann, H. and Huber, P. J. (1963) Pairwise comparison and ranking in tournaments. *Ann. Math. Statist.* **34** 501–510.

[6] Chung, K. L. (1974) *A course in probability theory.* Academic Press, New York.

[7] Cook, R. D. and Nachtsheim, C. J. (1980) A comparison of algorithms for constructing exact *D*-optimum designs. *Technometrics* **22** 315–324.

[8] David, H. A. (1988) *The Method of Paired Comparisons.* Griffin, London.

[9] Davidson, R. R. (1970) On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Amer. Statist. Assoc.* **65** 317–328.

[10] Davidson, R. R. and Beaver, R. J. (1977) On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics* **33** 693–702.

[11] Davidson, R. R. and Farquhar, P. H. (1976) A bibliography on the method of paired comparisons. *Biometrics* **32** 241–252.

[12] Davidson, R. R. and Solomon, D. L. (1973) A Bayesian approach to paired comparison experimentation. *Biometrika* **60** 477–487.

[13] Dennis, J. E. and Schnabel, R. B. (1983) *Numerical methods for unconstrained optimization and nonlinear equations.* Prentice-Hall, Englewood Cliffs, New Jersey.

[14] Dinitz, J. H. and Stinson, D. R. (1981) A fast algorithm for finding strong starters. *SIAM J. Alg. Disc. Meth.* **2** 50–56.

[15] Dinitz, J. H. and Stinson, D. R. (1987) A hill-climbing algorithm for the construction of one-factorizations and Room squares. *SIAM J. Alg. Disc. Meth.* **8** 430–438.

[16] Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368.

[17] Feller, W. (1968) *An introduction to probability theory and its application.* Wiley, New York.

[18] Fienberg, S. E. (1979) Log linear representation for paired comparison models with ties and within-pair order effects. *Biometrics* **35** 479–481.

[19] Gillot, C. and Caussinus, H. (1966) Sur un modèle de comparaisons par paires avec une échelle de réponses à trois valeurs. *Revue de Statistique Appliquée* **14** 31–42.

[20] Grimmett, G. and Stirzaker, D. (1982) *Probability and random processes.* Oxford University Press, Oxford.

[21] Guttorp, P. and Lockhart, R. A. (1988) On the asymptotic distribution of quadratic forms in normal order statistics. *Ann. Statist.* **16** 433–449.

[22] Hosmer, D. W. and Lemeshow, S. (1980) Goodness-of-fit tests for the multiple logistic regression model. *Comm. Statist. Theor. Meth.* **9** 1043–1069.

[23] Kraitchik, M. (1953) *Mathematical Recreations.* Dover Publications, New York.

[24] Maurer, W. (1975) On most effective tournament plans with fewer games than competitors. *Ann. Statist.* **3** 717–727.

[25] Mendelsohn, E. and Rosa, A. (1985) One-factorizations of the complete graph — a survey. *J. of Graph Theory* **9** 43–65.

[26] Mitchell, T. J. (1974) An algorithm for the construction of 'D-optimum' experimental designs. *Technometrics* **16** 203–210.

[27] Mitchell, T. J. and Miller, F. L. (1970) Use of design repair to construct designs for special linear models. *Rep. ORNL-4661*, pp. 130–131, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

[28] Rosa, A. and Wallis, W. D. (1982) Premature sets of 1-factors, or How not to schedule round-robin tournaments. *Discrete Appl. Math.* **4** 291–297.

[29] Ross, R. T. (1939) Optimal orders in the method of paired comparisons. *J. Experimental Psychology* **25** 417–421.

[30] Russell, K. G. (1980) Balancing carry-over effects in round robin tournaments. *Biometrika* **67** 127–132.

[31] Santner, T. J. and Duffy, D. E. (1986) A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73** 755–758.

[32] Silvey, S. D. (1980) *Optimal Design.* Chapman & Hall, London.

[33] Spinelli, J. J. (1994) *Cramér-von Mises statistics for discrete distributions.* Unpublished Ph. D. thesis, Simon Fraser University.

[34] Starks, T. H. (1958) *Significance tests in experiments involving paired comparisons.* Unpublished Ph. D. dissertation, Virginia Poly. Inst.

[35] Stephens, M. A. (1986) *Tests based on EDF statistics.* In *Goodness of Fit Techniques*, eds. R. B. D'Agostino and M. A. Stephens, Marcel Dekker, New York, Ch. 4.

[36] Su, J. Q. and Wei, L. J. (1991) A lack-of-fit test for the mean function in a generalized linear model. *J. Amer. Statist. Assoc.* **86** 420–426.

[37] Thisted, R. (1988) *Elements of Statistical Computing.* Chapman & Hall, New York.

[38] Tsiatis, A. A. (1980) A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67** 250–251.

[39] Van Schalkwyk, D. J. (1971) *On the design of mixture experiments.* Unpublished Ph. D. thesis, University of London.

[40] Wallis, W. D. (1983) A tournament problem. *J. Austral. Math. Soc. Series B* **24** 289–291.

[41] de Werra, D. (1980) Geography, games and graphs. *Discrete Appl. Math.* **2** 327–337.

[42] de Werra, D. (1982) Minimizing irregularities in sports schedules using graph theory. *Discrete Appl. Math.* **4** 217–226.

[43] Wynn, H. P. (1970) The sequential generation of D-optimum experimental designs. *Ann. Math. Statist.* **41** 1655-1664.

[44] Wynn, H. P. (1972) Results in the theory and construction of D-optimum experimental designs. *J. Roy. Statist. Soc. B* **34** 133–147.