1994

Tests Of The Effect Of Treatment In Stratified Cluster Randomization Trials

Neil S. Klar

Tests of the Effect of Treatment in Stratified Cluster Randomization Trials

by

Neil S. Klar

Department of Epidemiology and Biostatistics

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Faculty of Graduate Studies

The University of Western Ontario

London, Ontario

October 1993

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

ISBN 0-315-90569-7

Abstract

It is becoming increasingly common for epidemiologists to consider randomizing intact social units (e.g. families, schools, communities) rather than individuals in experimental trials. Reasons are diverse, but include administrative convenience, a desire to reduce the effect of treatment contamination and the need to avoid ethical issues that might otherwise arise. Dependencies among cluster members typical of such designs must be considered when determining sample size and analysing the resulting data.

The primary focus of this thesis is on comparisons of tests of the effect of treatment in trials where clusters are randomly assigned to treatment groups after stratifying on cluster-level baseline risk factors (e.g. cluster size). Particular attention is paid to the analysis of binary outcome data.

Tests of the effect of treatment for such trials range in complexity from adaptations of standard statistical methods performed using the cluster as the unit of analysis to extensions of logistic regression adjusted for clustering. The validity of such extensions was shown to be assured if the average correlation among cluster members is fixed. This assumption can be relaxed by using robust variance estimators. Test statistics using these different approaches were shown to be asymptotically equivalent when there is no variability in cluster size.

Simulation studies were used to examine the small sample properties of test statistics assuming an average cluster size of 100 subjects and either two or four strata. These simulation studies indicated that exact permutation tests should be used to make inferences about the effect of treatment if there are 20 or fewer clusters per treatment group. Approximate test statistics using cluster-level analyses or extensions of the Mantel-Haenszel test statistic are appropriate if there are more than 20 clusters per treatment group. Valid rejection rates for methods using robust variance estimates can not be assured even if there are 40 clusters per treatment group. There is little need to employ such techniques, however, since the simulation studies also showed that typical violations of the common correlation assumption have no effect on the validity or power of test statistics.

Acknowledgements

The work presented in this thesis benefited from the intellectual and moral support of many people. I am particularly grateful to Dr. Allan Donner my academic advisor. His guidance and insight have been invaluable. I would also like to thank Dr. John Koval and Dr. Mike Eliasziw for taking the time to discuss a wide range of issues concerning the analysis of correlated binary outcome data and Dr. Theresa Gyorkos and Dr. David Murray for allowing me to use their data. Dr. Eliasziw was also kind enough to take the time to read and comment on a draft of this thesis. I am also indebted to my family and friends for their unflagging moral support.

Work on this thesis was supported in part by a National Health Research and Development Program grant from Health and Welfare Canada and by an Ontario Graduate Scholarship from the Ontario Ministry of Colleges and Universities.

TABLE OF CONTENTS

TABLES AND FIGURES

intervention trials. The most statistically oriented paper was by Cornfield (1978) in which he expanded the arguments put forward by Cornfield and Mitchell (1969) describing the statistical inefficiency of cluster randomization by drawing an analogy between cluster sampling and cluster randomization. The symposium stimulated much of the subsequent research developments in this field.

Many of the subsequent research papers (Donner, Birkett and Buck (1981), Donner (1982), Salonen et al (1986), Williams, Fortmann, Farquhar et al (1981)) describing the design and analysis of cluster randomization trials were published in epidemiologic journals. Publication of these papers did not, of course, immediately translate into any marked improvement in the methodological quality of cluster randomization trials. The continuing problem of poorly designed and analyzed cluster randomization trials is described in a recent methodological review of 16 studies of non-therapeutic interventions (Donner et al, 1990). This review found that eleven of the trials failed to justify the need for cluster randomization, only three trials correctly accounted for between-cluster variation in discussing sample size and power while eight accounted for between-cluster variation in the analysis. The quality of the more recently published papers tended to be higher than in earlier trials. These results are consistent with those of Altman and Bland (1991) who after reading 150 methodological reviews of articles from medical journals report that about half the published papers in the medical literature are statistically flawed. They found that problems are more likely to be due to inadequate descriptions of design, often due to the failure to explain how that treatment affects the behavior of control subjects could be avoided by randomly assigning treatment to the entire factory as was done in a trial evaluating educational strategies used to prevent heart disease (World Health Organization European collaborative Group, 1986).

Cluster randomization may also be used to increase participation rates or to avoid ethical issues which might otherwise arise. Physicians, for example, might feel ethically bound to offer all their patients the same treatment and be willing to participate in a trial only under these circumstances.

There are several analytic implications associated with randomly allocating treatment to clusters. These arise since subjects' responses within a cluster, whether for reasons related to genetics or environment, tend to be dependent. These dependencies increase the ratio of between to within cluster variability, reducing the effective sample size and increasing the variance of the estimated effect of treatment. Failure to adjust standard statistical methods for within-cluster dependence will result in studies with spuriously elevated type I errors. Since prognostic variables measured on cluster members also tend to be positively correlated, cluster randomization tends to reduce the probability of obtaining balance on these variables, thus increasing the importance of stratification in design and the need for multivariate methods in analysis.

None of these issues would arise if inferences were intended at the cluster level. The cluster would then be both the unit of randomization and the unit of analysis and the degree of dependence among cluster members would be irrelevant. However in many applications the inferences are aimed at individual subjects. For example, in a trial of hypertension screening and management (Bass et al, 1986) administrative convenience and the need to obtain physician cooperation dictated that intact medical practices be randomized, although the ultimate aim was to reduce cardiovascular morbidity and mortality among individual patients.

The source of dependencies among cluster members varies with the nature of the cluster and the outcome variable. Dependencies among family members, for example, are caused by a combination of genetics and shared environment, while similarities of students in a classroom are partially determined by differences among teachers and interactions among students. A unique feature of vaccine trials is described by Comstock (1978), in which dependencies among members of a cluster may be determined by the dynamics of the disease.

Pocock (1987, pp. 188-191) describes three classes of patient outcomes which occur in clinical trials: quantitative responses (e.g. blood pressure, weight loss), qualitative responses (e.g. alive/dead), and time to relapse. Standard statistical techniques such as generalized least squares (Donner, 1985b) can be used to analyze quantitative outcome data while methods for the analysis of correlated failure times are still in the early stages of development (Anderson, 1991).

The primary focus of this thesis is on the analysis of binary outcome data, reflecting an epidemiologic emphasis and the current interest in analysis of correlated categorical data (Zeger, 1988). The discussion also tends to focus on trials with two treatment groups in which the effect of treatment is evaluated at only one point in time, although all of the methods which will be discussed can be extended to more complicated designs.

The thesis is divided into five chapters. The theory and practice of cluster randomization trials are reviewed in the first chapter. This review includes a discussion of the history of cluster randomization to help place ideas described elsewhere into historical perspective. Methods for the design and analysis of cluster randomization tri^ls are discussed in the last two sections of the chapter.

A detailed algebraic examination of several methods is conducted in the second chapter of the thesis. The methods can all be used to test the effect of treatment in cluster randomization trials which stratify clusters prior to randomization. Such trials have received, perhaps, the least attention in the literature. The discussion focuses on designs in which there are few strata but where the number of clusters per stratum is large. Brief attention will also be given to community trials where there tend to be very few clusters.

The complexity of the methods described in Chapter 2 generally restricts algebraic comparisons to fairly simple and unrealistic situations in which there is no variability in cluster sizes. Simulation studies were therefore used to compare the validity and power of methods described in the previous chapter in the more realistic case where cluster size is variable. The simulation studies were also used to determine how robust these methods are to violations of their underlying assumptions. The results of this research are described in Chapter 3.

Data analyses from two cluster randomization trials are presented in Chapter 4. The presence of baseline risk factors, missing data, and unbalanced designs in such case studies tend to offer more complex analytic challenges than is practical to simulate using computer generated data. Case studies can therefore be helpful in seeing how well theoretical findings hold in particular applications. Large deviations from expected results may even suggest fruitful avenues for further research.

The principal findings of the thesis are summarized in Chapter 5, the last chapter of the thesis. Recommendations for the analysis of data from cluster randomization trials and suggestions for future research are also discussed in this chapter.

1.2 Historical Survey of Cluster Randomization Trials

Random allocation of treatment serves three functions in clinical trials. It ensures that subjects are, on average, alike at the start of the study, precludes any use of judgement in assigning treatment and provides a basis for the validity of statistical tests of treatment efficacy. The first two functions of randomization were known in the 19'th century (Stigler (1986, p. 253)). There are also several early examples in medical research of physicians and statisticians using or advocating randomization (Armitage (1985), Greenwood and Yule (1915)) to ensure internal validity. The role of random treatment allocation in tests of significance, however, appears not to have been appreciated before 1925 when R.A. Fisher published Statistical Methods for Research Workers (Cochran, 1989). Fisher expanded on these ideas (Fisher, 1926) in a paper which evolved into The Design of Experiments (Fisher, 1935).

The earliest known example of random allocation in clinical medicine was also in 1926 (Lilienfeld, 1982). Its use appears to have been independent of Fisher's influence. In this year (i.e. 1926) Amberson et al (1931) divided 24 patients into two matched groups of 12 subjects each and then assigned the groups to treatment on the basis of a single coin flip. Thus the first use of random allocation in medicine was in a trial using cluster randomization.

Although other early examples of random allocation are known (Armitage (1985), D'Arcy Hart (1972), Lilienfeld, (1982)) the modern era of randomized clinical trials is generally accepted to have started in 1948 with a trial of streptomycin as a treatment for pulmonary tuberculosis (Medical Research Council (1948), Pocock (1987, p. 17)). The streptomycin trial stands apart because of the care which was used in its design and analysis, the influence of Bradford Hill, the statistician who designed it (Armitage, 1991ab)), and, perhaps, due to the success of streptomycin in treating tuberculosis (D'Arcy Hart, 1972).

The success of the streptomycin trial in instilling the virtues of randomization in the minds of clinical researchers was at first quite modest. For example, none of the 29 clinical trials reported in the New England Journal of Medicine in 1953 used randomized controls (Chalmers and Schroeder, 1979). In spite of a fairly steady and dramatic increase only 50 percent of clinical trials published in the late 1970's could claim to have employed randomization. It should therefore come as little surprise that progress in the design and analysis of cluster randomization trials has been much slower.

An interesting early source of technical material on cluster randomization appears in Lindquist (1940), a book written as an interpretation of Fisher's (1935) Design of Experiments for educational researchers. Such researchers often evaluate new methods of instruction using the classroom as the unit of randomization and using test scores as the outcome variable. Lindquist suggested that the effect of such interventions can be tested using standard statistical methods applied to cluster means. His ideas were not initially well received

(Glass and Hopkins (1984), McNemar (1940)) and the confusion was still evident 40 years later (Barcikowski (1981), Hopkins (1982)).

In the medical research area, Mainland (1952, pp. 114-115) describes the importance of replication in trials where the unit of randomization is a cluster of subjects and not independent individuals. The discussion was timely since some trials (e.g. Ast and Schlesinger, 1956) assigned only one cluster to each treatment group confounding the effect of treatment with between cluster variability.

Although the impact of Mainland's book is difficult to assess, some communicable disease epidemiologists in the 1960's were aware of several analytic issues unique to cluster randomization trials. Pollock (1966), for example, recognized that trials randomizing clusters were less likely to be balanced for important prognostic variables than trials randomizing independent individuals but acknowledged that at times cluster randomization trials must, of necessity, be performed. The reasons for randomizing clusters, as cited by Pollock (1966), included administrative convenience, increased participation, reducing the risk of treatment contamination, and are similar to reasons given in current trials (Donner, Brown and Brasher, 1990).

The methodological sophistication of this early work is demonstrated by three double-blind, placebo controlled trials of isoniazid (Comstock (1962), Horwitz and Magnus (1974), Ferebee et al (1963)), a drug still used to prevent and treat tuberculosis (Berkow et al, 1987, pp. 113-114). In one such trial patients in 566 hospital wards received either placebo or isoniazid (Ferebee et al, 1963). The wards were selected as the unit of randomization to reduce the administrative complexity of the trial. The test of treatment effect was corrected for overdispersion by adapting a method described by Cochran (1953, pp. 124-127) for the analysis of cluster sample surveys.

Unfortunately, while familiarity with the analysis of data from cluster randomization trials grew among epidemiologists interested in communicable disease, chronic disease epidemiologists were slower to adopt the analytic techniques required by cluster randomization. Perceived differences in the methods used by communicable and chronic disease epidemiologists probably delayed the use of appropriate cluster randomization methodology among chronic disease epidemiologists (Comstock, 1978). Except for a few isolated papers (Cornfield and Mitchell (1969), Henderson and Meinert (1975), Spitzer, Feinstein and Sackett (1975)) no advice was targeted at these epidemiologists, even though cluster randomization was being used in trials of preventive measures for chronic diseases (World Health Organization European Collaborative Group, 1974).

A breakthrough occurred in 1978 with the publication of papers from a symposium on coronary heart disease prevention trials (Comstock (1978), Cornfield (1978), Farquhar (1978), Hulley (1978), Sherwin (1978), Syme (1978)). This was the first occasion on which extensive attention had focussed on methodologic challenges faced by medical researchers performing community intervention trials. The most statistically oriented paper was by Cornfield (1978) in which he expanded the arguments put forward by Cornfield and Mitchell (1969) describing the statistical inefficiency of cluster randomization by drawing an analogy between cluster sampling and cluster randomization. The symposium stimulated much of the subsequent research developments in this field.

Many of the subsequent research papers (Donner, Birkett and Buck (1981), Donner (1982), Salonen et al (1986), Williams, Fortmann, Farquhar et al (1981)) describing the design and analysis of cluster randomization trials were published in epidemiologic journals. Publication of these papers did not, of course, immediately translate into any marked improvement in the methodological quality of cluster randomization trials. The continuing problem of poorly designed and analyzed cluster randomization trials is described in a recent methodological review of 16 studies of non-therapeutic interventions (Donner et al, 1990). This review found that eleven of the trials failed to justify the need for cluster randomization, only three trials correctly accounted for between-cluster variation in discussing sample size and power while eight accounted for between-cluster variation in the analysis. The quality of the more recently published papers tended to be higher than in earlier trials. These results are consistent with those of Altman and Bland (1991) who after reading 150 methodological reviews of articles from medical journals report that about half the published papers in the medical literature are statistically flawed. They found that problems are more likely to be due to inadequate descriptions of design, often due to the failure to explain how sample size was determined, rather than a problem with the analysis.

1.3 Experimental Design of Cluster Randomization Trials

1.3.1 Issues Arising when Selecting the Unit of Randomization

The choice of the unit being randomized is largely determined by the scientific question. For example, in studies of the effect of water fluoridation on prevention of dental caries, the community is the natural unit since water supplies are usually controlled at the level of the community. Families, however, are a more natural unit in studies of the relationship between diet and health.

Researchers might, however, occasionally conclude that a variety of units are suitable. They would then have to consider the tradeoff between administrative convenience, expected participation rates, varying risks of treatment contamination and the variable costs likely to occur when treatment is randomly assigned to different types of clusters (e.g. families, schools, communities).

There is also an ethical concern which needs to be addressed when selecting the unit of randomization. Informed consent will usually only be possible when treatment is randomly assigned to smaller clusters such as families or schools. Although informed consent is still possible when physicians' practises are the unit of randomization it is not uncommon for patients to know only that their doctor is involved in research (Buck and Donner, 1982). Such tacit consent by patients is scientifically valuable in maintaining patient blinding. This is particularly important in trials of lifestyle interventions where knowledge of treatment can affect outcome (Buck and Donner, 1982). Informed consent becomes impossible in most community intervention trials requiring the use of community or proxy consent (Strasser, Jeanneret and Raymond, 1987)) from the local government.

Use of proxy consent by physicians on behalf of their patients or of local governments on behalf of their electorate is, however, ethically questionable. Strasser et al (1987) criticize the use of proxy consent arguing that the mandate of most governments does not include enrolling their community into randomized trials, no matter how seemingly innocuous the intervention might be.

The assumption that the intervention is innocuous also needs to be examined. Cluster randomization tends to be used far more often in trials of interventions aimed at preventing rather than treating disease (Buck and Donner, 1982). There seems to be a belief that non-therapeutic interventions are risk free. Skrabanek (1990) warns of situations where this is not the case and argues for the development of ethical guidelines for preventive medicine in general and for the importance of informed consent in particular.

Some ethical concerns about possible harm caused by treatment are addressed in phase I and phase II trials of therapeutic interventions (Pocock, 1987, pp. 2-3). Only treatments shown to have been helpful in these trials will be evaluated in full-scale phase III trials. While there are no accepted equivalent procedures for trials of non-therapeutic interventions, proposals for possible criteria are being discussed (ACS, (1992), Cullen (1990), Flay (1986), Piantadosi and Byar (1989)).

1.3.2 Choosing an Experimental Design

The survey discussed above (i.e. Donner et al, 1990) showed that cluster randomization designs adopted by medical researchers tend to fall into one of three categories: completely randomized, stratified, or pair-matched. In completely randomized designs clusters are randomly allocated to treatment without matching or stratification. This design is often satisfactory when many clusters are available to be randomized. The number of clusters which can be economically and feasibly allocated declines rapidly with cluster size (Koepsell et al, 1992), implying that completely randomized designs are practical only when relatively small clusters such as families are being allocated. Stratification by important baseline risk factors is recommended in order to reduce the probability of large imbalances on such variables when larger clusters (e.g. physicians' practices, communities) are the unit of randomization. Since precision is increased when there are equal numbers of subjects in treatment groups stratification on cluster size is recommended whenever the number of subjects per cluster is highly variable.

If there are many known risk factors requiring a finer stratification a pair-matched design can be performed by matching similar clusters and randomly assigning one member of each match to treatment. When there are few clusters in a trial it can become increasingly difficult to obtain close matches on all potential risk factors, increasing the likelihood that there will be only a small gain in precision obtained by using a pair-matched as opposed to a stratified design. Furthermore Martin et al (1992) have shown that ineffective matching can lead to a considerable loss of power relative to the completely randomized design, especially if the total number of clusters is small. Associated analytic limitations of this design are discussed in Section 1.4.

Two possible variants of these designs have been described in the literature and are worth noting. The first occurs when researchers have control over cluster size. Most clusters are naturally defined units like families or communities. Hauck et al (1991), however, describe a trial of nurse-managed instructional support for cardiac patients in which temporal clusters of patients were randomly assigned to one of two types of support groups. Control over cluster size allows a choice between relatively inexpensive designs involving a few large clusters and a more statistically powerful design in which many, small clusters are randomly assigned to treatment.

The second variant is a composite design developed for clinical trials in physician's practices (Simon, 1981). Some physicians may feel that ethical and administrative considerations require the use of cluster randomization while others may be willing to offer either treatment to their patients. The design is an attempt at increasing the power of cluster randomization trials balanced against accrual problems common to many clinical trials.

The difficulty of studying more than a few experimental units in community intervention trials has been previously noted (Blum and Feachem (1983), Koepsell et al (1992), Mickey et al (1991)). Two pragmatic solutions to the absence of replication have been put forward. Kramer (1988, pp. 83-84) suggested using standard statistical tests when no between cluster differences are found in a pretrial study period. This approach is potentially misleading, both because of the low power to detect significant between-cluster differences in a pretrial study period of reasonable length, and because, for large clusters, standard tests of the effect of treatment are strongly influenced by even very small within-cluster correlations (Donner, 1982).

The second solution to the replication problem has been to adopt a variant of the crossover design. Turpeinen et al (1979) assessed the effect of diet on heart disease by randomly assigning a cholesterol-lowering or a standard diet to all patients in one of two Finnish mental hospitals. After 6 years the hospitals switched meal plans. The power of crossover designs when individuals are randomized is only evident when patients can experience the outcome in both study periods, and in the absence of carryover effects (Jones and Kenward, 1989, Chapter 2). In the Finnish study mortality was the outcome and any patient who participated in the entire trial would be a highly selected survivor (Halperin, Cornfield and Mitchell, 1973).

1.3.3 The Determination of Sample Size

Since the scientific question will usually determine the choice of cluster (e.g. family, classroom, community) to be used in the trial, the average cluster size is frequently known in advance. Furthermore when cluster sizes are small or when routinely collected vital statistics are used to assess the effect of intervention all cluster members will be included in the study (Sherwin, 1978). In this case power analyses are needed only to determine the number of clusters needed to detect clinically relevant treatment effects.

When very large clusters are randomized sub-samples of individual cluster members may be necessary in order to assess the effect of treatment. For example, the end point in a community intervention trial of smoking cessation programs is the quit rate of 500 heavy smokers randomly selected from each cluster (Byar, 1988). In such a trial both the number of clusters and the number of subjects per cluster must be determined in advance (Donner (1992), Hsieh (1988)).

Sample size requirements for completely randomized and stratified cluster randomization trials can be calculated after adapting standard formulae developed for trials randomizing individuals. It is well known (e.g. Donner, Birkett and Buck, 1981) that when clusters consist of the same number of subjects, n, the appropriate trial size is obtained by multiplying the number of subjects required under individual randomization by a variance inflation factor, $1+(n-1)\rho$ where ρ is the expected intracluster correlation coefficient. This coefficient measures the degree of dependence between responses of cluster members. It equals zero, its minimum value, when subjects' responses are independent and increases towards one as the responses of cluster members become more alike.

An alternative approach used to determine sample size has been described by Hsieh (1988). Although both approaches assume that there is no variability in cluster size the method described by Hsieh (1988) can be used to calculate both the number of clusters per treatment group and the number of subjects per cluster.

As the expected variability in cluster size increases, stratification on cluster size may have to be considered in order to obtain a more balanced and hence more powerful design. Donner (1992) extended a formula due to Woolson et al (1986) to provide sample size requirements for stratified cluster randomization designs. Several methods have also been developed to determine sample size for pair-matched cluster randomization trials (Byar (1988), Freedman, Green and Byar (1990), Gail et al (1992), Hsieh (1988), Shipley, Smith and Dramaix (1989)).

Randomization assures that estimates of treatment effect are unbiased. When linear models are used the expected value of the estimated treatment effect is unchanged by the inclusion of variables which account for stratification or baseline risk factors (e.g. age, sex). What becomes relevant are the gains in power which will be obtained by stratification or by modelling baseline risk factors.

The discussion becomes more complicated when non-linear models are adopted for binary outcome data to examine the effect of treatment The most general models available to analyze correlated binary outcome data are extensions of logistic regression which summarizes the effect of treatment using odds ratios. Results obtained by Gart (1992) and Robinson and Jewell (1992) imply that the population odds ratio from completely randomized designs will tend to be closer to one than the population odds ratio from either stratified or pair-matched designs. This relationship will hold in the absence of effect modification and when there are about equal numbers of subjects in each treatment group within each stratum. As the variability in the numbers of subjects per treatment group in each stratum increases it becomes impossible to predict the differences in the size of the population odds ratios from different designs. Random assignment will therefore only assure that sample odds ratios are consistently estimating the population odds ratio for that particular design. It is not possible, in general, to compare odds ratio estimators between different designs.

There are three exceptions to this rule. Population odds ratios from the different designs will be equal when treatment has no effect, when risk does not vary across strata, or when risk is low (i.e. when the rare disease assumption holds). The rare disease assumption is most likely to hold in cluster randomization trials of preventive strategies aimed at reducing mortality in the general population. In general, however, the greater power of stratified and pair-matched designs arises, in part, because odds ratios tend to be larger for such designs. The difficulty of comparing odds ratios from different models will be discussed further in Section 1.4.2.

In all three designs power is more directly affected by varying the number of clusters than by varying cluster size. This is easily demonstrated by noting that for m clusters of n subjects the

$$\mathrm{var}(\ \bar{y}_{.}\ ) = \frac{\sigma^2}{mn}\ [\ 1 + (n-1)\rho\ ] \qquad (1.3.1)$$

where $\bar{y}_{.}$ is the average score for the m clusters and σ^2 is the variance for the response of any one of the mn subjects (Donner et al, 1981). As $m \rightarrow \infty$ the $\mathrm{var}(\ \bar{y}_{.}\ ) \rightarrow 0$ while if $n \rightarrow \infty$ $\mathrm{var}(\ \bar{y}_{.}\ ) \rightarrow \frac{\sigma^2}{m}\ \rho$. Thus the power to detect the effect of treatment can only reach 100 percent when the number of clusters becomes large.

Prior estimates of ρ typically required by sample size formulae for cluster randomization designs can be obtained from the literature, or, alternatively from a pilot study. Cornfield (1978) and Donner (1982) have urged investigators to publish empirical estimates of ρ that can help in sample size planning.

Estimates of ρ obtained from studies in which there are few clusters will likely be imprecise complicating the determination of sample size. In particular, a simulation study performed by Feng and Grizzle (1992) found that the number of clusters required to obtain stated levels of power would frequently be underestimated when the sample size was determined using imprecise estimates of ρ. It is therefore usually advisable to perform a sensitivity analysis to determine the effect of different values of ρ on sample size. Sensitivity analyses also need to be performed since random samples of clusters are not obtained in most cluster randomization trials so estimates of ρ may not be representative of the degree of intracluster correlation in the target population.

There can also be difficulties in using prior estimates of intracluster correlation from trials which used different experimental designs. An estimate of intracluster correlation obtained from a stratified cluster randomization trial, for example, is likely to be smaller than a coefficient from a trial using a completely randomized design since some of the between-cluster variability is likely explained by stratification.

An additional issue is that the degree of intracluster correlation may not be independent of disease risk. In toxicological trials, for example, intracluster correlation tends to increase as risk increases from 0 to 50 percent (Williams, 1988b). More recently a similar relationship was noted by Mickey and Goodwin (1993) between the degree of variance inflation due to clustering and mortality rates using county-level data from 44 states in the USA.

A theoretical explanation is offered by Kraemer (1979) and Thompson and Walter (1988) in the context of reliability studies which have two observations per cluster. Their model is dependent upon the idea that disease risk in a cluster is determined by the binomial distribution conditional on the true underlying risk which follows a Bernoulli distribution.

Estimates of ρ in cluster randomization trials are almost always positive and tend to be larger in smaller clusters. This empirical inverse relationship was first described in the context of survey sampling by Smith (1938) and is sometimes referred to as "Smith's Law" (e.g. Proctor, 1985). It has been further noted that the influence of cluster size on ρ tends to be less than linear (Hansen, Hurwitz and Madow (1953, pp. 306-309)). A similar correlation with cluster size was noted by Mickey and Goodwin (1993) using design effects calculated as a ratio of county-level sample variance estimates of mortality rates and estimates of variance obtained using simple random sampling.

An example of "Smith's Law" can be obtained using data from a cluster randomization trial which examined if intensive screening and management would lower the risk of stroke among patients of primary care physicians (Bass et al, 1986). Estimates of intracluster correlation were calculated for four dichotomized baseline risk factors (i.e. hypertension, smoking, drinking, and obesity) and the data were organized into three types of clusters: spouse pairs, physicians' practises, and counties (Donner, 1982). The resulting estimates of intracluster correlation have been plotted as a function of cluster size and are displayed in Figure 1.1. Nearly identical patterns were observed for all four risk factors.

A common feature of cluster randomization trials is that there tends to be a relatively small number of clusters included in most studies. An idea of the variability in size of such trials can be obtained by examining Table 1.1. These trials were previously discussed in the methodological review conducted by Donner et al (1990). The pair-matched and triplet-matched are the most common designs among these studies. In either design there is only one cluster for each combination of treatment and stratum. The principal finding in this table is that except for one notable exception (Sommer et al, 1986) the number of clusters in a trial tends to be inversely proportional to cluster size.

The possibility of loss to follow-up is potentially serious in all longitudinal studies. This can be a particularly severe problem in cluster randomization trials because of the relatively long follow-up time required to evaluate the non-therapeutic interventions common to such studies (Piantadosi and Byar (1989), Syme (1978)), because of the possibility that entire clusters may drop out (Byar, 1988), and because treatments are often applied at the cluster level, with little or no attention given to individual study participants. Note however that some studies, e.g. mass education or community intervention trials, have also to contend with immigration of new subjects after baseline, further complicating the problem (Gillum et al (1980), Jooste et al, (1990)). Sample size estimates need to be adjusted to account for loss to follow-up and immigration of new subjects after baseline Byar (1988), Gillum et al (1980)). A variety of issues related to loss to follow-up and data quality in longitudinal research are discussed in a book edited by Magnusson and Bergman (1990).

Table 1.1
How Big are Cluster Randomization Trials in Epidemiology?
after Donner, Brown and Brasher (1990)

Experimental Design	Reference	$\bar{m}$	Unit of Randomization
Completely Randomized	Black et al (1981)	2	Day-Care Centers
	* Wilson et al (1988)	24	Medical Practices
	McDonald et al (1984)	27?	Medical Practices
	Sommer et al (1986)	225	Villages
Stratified	*? Tabar et al (1985) 7 strata	24	Communities
	Evans et al (1986) 4 Strata	31	Medical Practises
	Farr et al (1988) * Trial #1 2 Strata	100	Families
	Trial #2 3 Strata	115	Families

* indicates trials with three treatment groups

$\bar{m}$ denotes the mean number of clusters per treatment group

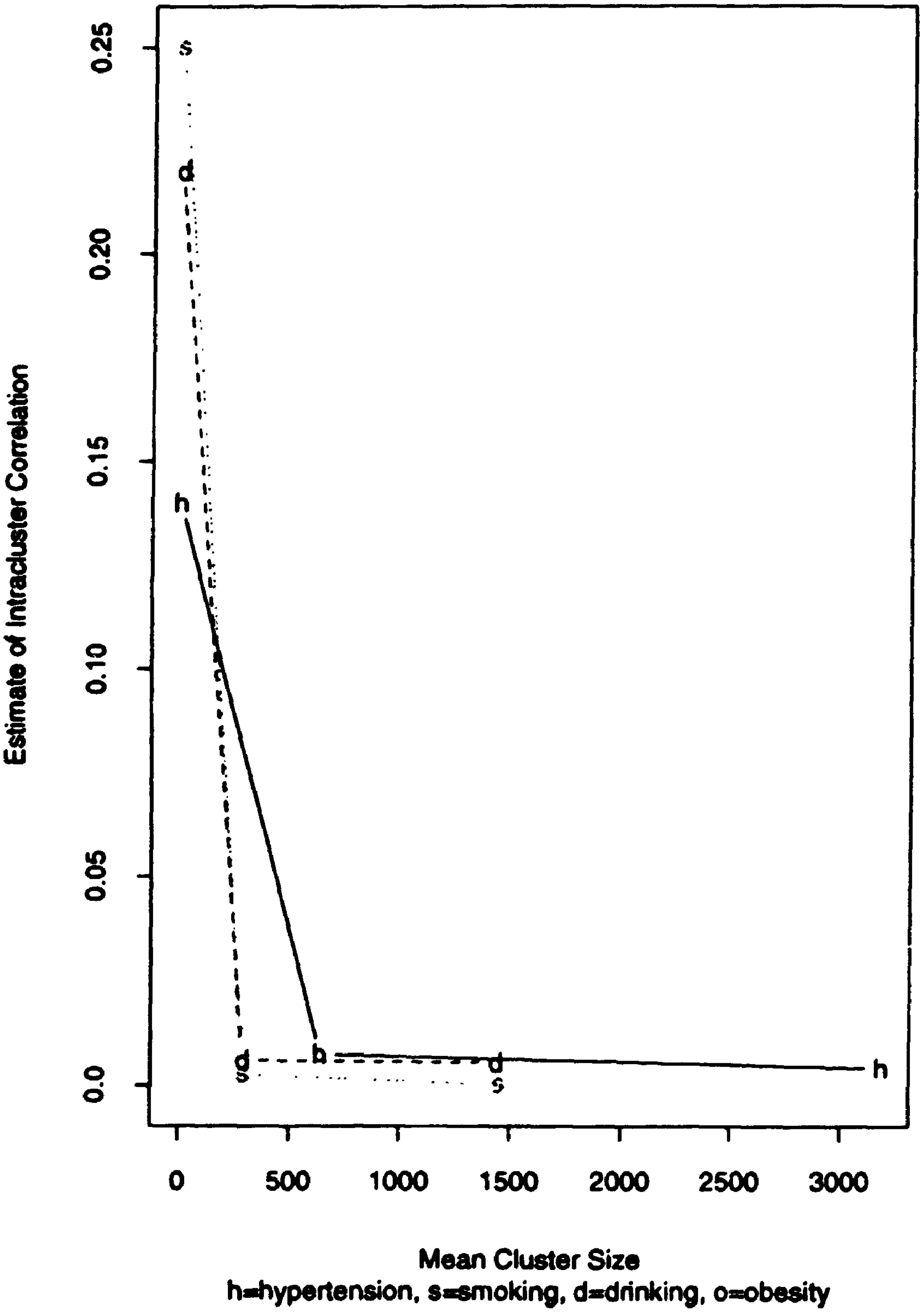
Table 1.1 continued ...
How Big arr Cluster Randomization Trials in Epidemiology?
after Donner, Brown and Brasher (1990)

Experimental Design	Reference	$\bar{m}$	Unit of Randomization
Pair-Matched and Triplet-Matched	* Bush et al (1989)	3	Schools
	* Dwyer et al (1983)	7	Classroom
	Bass et al (1986)	17	Medical Practices
	Stanton et al (1987)	25	Communities
	Grant et al (1989)	33	Practices,Clinics Hospitals
	WHO European Collaborative Group (1986)	40	Factories
	Lloyd et al (1983)	44	Schools

* indicates trials with three treatment groups.

$\bar{m}$ denotes the mean number of clusters per treatment group.

Fig 1.1 Empirical Demonstration of Smith's Law



1.4 Analysis of Data from Cluster Randomization Trials

1.4.1 Methods of Analysis for Continuous Outcome Data

Analysis of continuous outcome data from cluster randomization trials can often be accomplished using mixed-effects linear models, as fit by generalized least squares (Donner, 1985b). Data can also be fit using analysis of variance, maximum likelihood, restricted maximum likelihood, and MINQUE (Searle, 1988). For mixed-effects linear models maximum likelihood is equivalent to generalized least squares (Goldstein (1986), del Pino (1989)). None of these methods can be unreservedly recommended when cluster sizes vary and all methods are asymptotically equivalent when there is no variation in cluster size (Searle, 1988). For this reason the discussion is limited to ANOVA (i.e. analysis of variance), the simplest approach, and generalized least squares, perhaps the most general (del Pino, 1989). The discussion will also omit mention of hierarchical linear models, reviewed recently by Bryk and Raudenbush (1992).

Mixed-effects models can be used with stratified or completely randomized designs to estimate the effect of treatment, to test if the observed eff :t is due to chance, and to adjust for imbalance on baseline risk factors. Several relevant computer programs are described by Jennrich and Sampson (1988), Prosser (1991), Wolfinger et al (1991), and Wolter (1985, Appendix E).

It is well accepted that data should be analyzed in accordance with the underlying design (e.g. Fisher (1935, Section 33), Meier (1981)). This implies that data from pair-matched and stratified designs take into account these design features.

A mixed-effects model which includes covariates for treatment, strata and baseline risk factors is given by

$$y_{ijst}=\beta\, X_{ijst} + C_{(ij)s} + \varepsilon_{(ijs)t} \qquad (1.4.1)$$

where y_{ijst} is the score for the t'th subject, t=1,...,n_{ijs}, from cluster s, s=1,...,m_{ij}, treatment j, j=1,2 and strata i, i=1,...,k. The fixed effects of treatment, strata, and baseline risk factors are represented by the vector β while $C_{(ij)s}$ and $\varepsilon_{(ijs)t}$ denote the respective independent random effects of cluster nested in treatment and stratum, assumed to be i.i.d. $N(0,\sigma_C^2)$, and subject nested in cluster, assumed to be i.i.d. $N(0,\sigma^2)$.

An equivalent marginal model can be constructed after averaging over $C_{(ij)s}$, yielding the model

$$y_{ijst}=\beta\, X_{ijst} + w_{(ijs)t} \qquad (1.4.2)$$

where $w_{(ijs)t}=C_{(ij)s} + \varepsilon_{(ijs)t}$ is the t'th element of the vector w_{ijs}, each of which are i.i.d. multivariate normal random variables with a mean ot zero and a variance matrix denoted by $(\sigma_C^2+\sigma^2)\, V$. The matrix V is a correlation matrix in which all correlation coefficients are equal implying that responses of all cluster members are equally correlated. It is sometimes referred to as an exchangeable or common correlation matrix.

Note that the parameters, β, in the marginal model are identical to those in the mixed-effects model. This equivalence only holds because both model outcome as a linear function of the data. The distinction between mixed-effects and marginal models becomes important, however, when outcomes are binary requiring the use of nonlinear (e.g. logistic) models and will be discussed in the next section.

An intuitive understanding of mixed-effects linear models is complicated by two factors. The first complication occurs when cluster sizes are variable. In this case iterative solutions are required to obtain parameter estimates. Insight can be gained by considering the case of a stratified cluster randomization trial in which all clusters are the same size and in which there is no adjustment for baseline risk factors.

In this case test statistics are easily calculated using ANOVA methods. The test of the null hypothesis that there is no difference between treatment groups is equal to the square of a stratified t-test (Fleiss, (1986, Section 6.1), Schwartz, Flamant and Lellouch (1980, pp. 189-191)) which uses the cluster as the unit of analysis. In a completely randomized design this statistic simplifies to the square of an ordinary t-test, again using the cluster as the unit of analysis. These relationships are derived and discussed in some detail in Section 3 of the next chapter.

The identity between test statistics constructed using the individual as the unit of analysis, i.e. mixed-effects models, and tests performed on cluster means is not unexpected. Similar identities have been previously described for related designs by Hopkins (1982) and more recently by Koepsell et al (1991). The identitv is a validation of methods originally described by Lindquist (1940). It shows that an unweighted cluster-level analysis is fully informative when there is no variability in cluster size.

When cluster sizes are variable the test statistic can be adjusted using the weights n_{ijs} /[$1+(n_{ijs}-1)\hat{\rho}$] where the cluster size is denoted n_{ijs} and $\hat{\rho}$ is an estimate of ρ. The weighted, stratified t-test can be thought of as the first iteration towards the weights obtained using generalized least squares. In either case the exact distribution for the test statistic is unknown but can be approximated using a t-distribution with $\sum_{i=1}^{k} \sum_{j=1}^{2} (m_{ij} - 1)$ degrees of freedom. Alternatively the degrees of freedom could be obtained using Satterthwaite's (1946) approach as discussed by Giesbrecht and Burns (1985). The approximation might not hold when there are few clusters. Simulation studies are needed to examine the small sample properties of test statistics obtained from generalized least squares (Goldstein, 1987, p. 29) as well as from the other methods used to fit mixed-effects linear models (Searle, 1988).

An estimate of ρ can be obtained as $\hat{\rho}=\hat{\sigma}_C^2 / [\ \hat{\sigma}_C^2 + \hat{\sigma}^2\ ]$ where $\hat{\sigma}^2$ and $\hat{\sigma}_C^2$ are estimates of variance components within and between clusters, respectively. Negative estimates can occur. They are usually set to zero since values of ρ less than zero are generally considered implausible in the context of cluster randomization trials. The ANOVA approach in Donner (1985a) is easily adapted to calculate $\hat{\rho}$ in completely randomized cluster randomization designs. These results are extended to a stratified design in Section 2.3.

A distinction is sometimes made between tests performed using cluster-level data and tests performed using individual-level data (Haseman and Hogan (1975), Kalter (1974), Palmer (1974), Weil (1974)). The discussion in the last few paragraphs can be used to argue against this distinction for completely randomized and stratified cluster randomization trials since analyses using continuous outcome data at either level are equivalent if methods have been adjusted for the effect of intracluster conelation. The distinction might have arisen because individual-level analyses unadjusted for intracluster correlation were sometimes being compared to unweighted cluster-level analyses (Haseman and Hogan (1975), Kalter (1974)).

The second complication unique to mixed-effects models arises when adjusting for individual-level baseline risk factors (e.g. age, sex). Scott and Holt (1982) point out that parameter estimates of such covariates are weighted averages of slopes calculated with cluster-level data (i.e. an ecological analysis) and a common within cluster slope. The parameter being estimated is only interpretable when both slopes are at least approximately equal. Differences between parameter estimates can be examined by comparing the results obtained when individual-level (e.g. age, sex) and cluster-level versions (e.g. mean age, proportion male) of a covariate are included in the same model. Complications arising when analyzing multi-level data have recently been reviewed in an epidemiological context by von Korff, Koepsell, Curry and Diehr (1992).

The random allocation of clusters assures that the assumption of a common value of ρ between treatment groups is tenable, at least under the null hypothesis. However it is also commonly assumed that the within-cluster correlation is constant across strata. This assumption can be investigated using tests of significance described by Donner (1985a) and Munoz, Rosner and Carey (1986). Unfortunately, when there are few clusters, a not infrequent occurrence in community intervention trials, these tests have little power. An approach which relaxes this assumption has been described by Liang and Zeger (1986). Their approach is introduced in the next section of the thesis and then described in greater detail in Section 2.8. It should be noted, however, that the increased robustness of this approach also increases the asymptotic requirements for the validity of tests of significance. This can be an important issue in cluster randomization trials.

The methods which have been described cannot be used in the analysis of pair-matched designs. Since there are only two clusters per stratum, estimates of between-cluster variability within a stratum are confounded with the effect of treatment. Tests of the effect of treatment must therefore be calculated using between-stratum information. The resulting test statistic is the square of a paired t statistic obtained using the cluster as the unit of analysis and is only valid in the absence of any interaction; an untestable assumption. The same untestable assumption is required to estimate the degree of intracluster correlation. Alternatively an estimate of ρ can be calculated under the null hypothesis that treatment is unrelated to the outcome. Neither option is attractive.

Estimates of intracluster correlation calculated under the assumption that there is no treatment by stratum interaction will likely be quite imprecise relative to a completely randomized design with the same number of clusters. Such imprecision occurs because approximately half the available degrees of freedom are needed to maintain the pair-match. However, these estimates could be used to construct models which allow statistical inferences to be made concerning the effect of baseline individual-level risk factors on outcome. This ability must be weighed against the likely imprecision of the resulting variance estimates.

These difficulties do not arise if all baseline risk factos are measured at the cluster level. Adjusted inferences for the effect of treatment could then be constructed by adapting an approach described by Rosner and Hennekens (1978) for the analysis of matched case-control and cohort studies. Stratum specific differences in outcome are modelled as a function of similar differences among cluster-level covariates. Adjusted inferences concerning the effect of treatment are constructed using the estimate of the intercept. Inclusion of individual-level baseline risk factors is only possible if modelled at the cluster level. This approach must be used cautiously because cluster-level adjustment for imbalance does not necessarily imply adjustment at the individual-level (Greenland and Morgenstern (1989), von Korff et al (1992)).

1.4.2 Methods of Analysis for Binary Outcome Data

The development of methods for the analysis of correlated binary data has lagged behind and been more fragmented than research on correlated continuous data (Zeger, 1988). There are several reasons for this. The first and most obvious reason is that linear models were developed for continuous data earlier than the nonlinear models generally used for binary data. The fragmentary nature of the research probably arose because there is no unique multivariate extension of the binomial distribution which has the flexibility of the multivariate normal. The result has been a wide array of methods which use different approaches to adjust for the dependence among cluster members and which include extensions of linear, logit and Probit regression (Ashby et al, 1992). Attention will be restricted almost entirely to extensions of logistic regression because of the popularity of this approach among health science researchers (Hosmer and Lemeshow, 1989, pp. vii).

Unlike the situation for continuous outcome data the distinction between cluster-level and individual-level analyses becomes important in evaluating methods of analysis for binary outcome data. The simplest cluster-level analysis for binary outcome data would involve arbitrarily dividing clusters into different groups as a function of cluster-level responses and then to analyze the results on a cluster-level basis. Weil (1970) argues that the requirement of imposing an arbitrary categorization makes this simple approach quite unattractive. An exception is presented in Gyorkos (1985) where a cluster-level analysis was performed to determine if screening and treatment for parasitic infection of family members was successful. In this study it was quite natural to dichotomize families into those with and without any infected members at the end of the treatment period since the absence of infection could be considered a clinical success while families with even one infected member would require additional treatment.

More powerful cluster-level analyses are constructed using the proportion of subjects in a cluster who responded positively. The primary advantage of this approach is that standard methods can then be used to analyze data from any one of the three designs. For example, independent t-tests or the nonparametric Wilcoxon test can be used to analyze data from a completely randomized design. Such tests are easily extended to stratified designs using multiple regression or stratified rank tests (described in Section 2.4) while paired t-tests or the Wilcoxon signed rank test can be used to analyze data from pair-matched designs (Cochran (1954, p. 447), Donner (1987)).

There are, however, disadvantages to this approach. Tests of significance using the cluster as the unit of analysis will, in general, have less power than methods using the individual as the unit of analysis, although simulation studies have demonstrated that the loss in power is small (e.g. Shirley and Hickling (1981, Table 3)). The efficiency of these procedures can be increased using weighted least squares as first advocated by Cochran (1943) and more recently by Marubini et al (1988). Such analyses still lack appeal since they do not always yield interpretable estimates of treatment effect on an individual level and since they can not be extended to model individual-level baseline risk factors.

Methods used to test the effect of treatment in pair-matched designs invariably use the cluster as the unit of analysis, with estimates of variance derived using between-strata variability. As stated in the previous section, this is done because there are only two clusters per treatment group so that estimates of between-cluster variability within a stratum are confounded with the effect of treatment. Procedures capable of testing the effect of treatment and adjusting for cluster-level baseline risk factors are described by Liang, Beatty and Cohen (1986) and Donner (1987). The confounding of between-cluster variability with the effect of treatment, unique to pair-matched designs, does not allow the intracluster correlation to be estimated (except under the null hypothesis), and thus prevents modelling of individual-level covariates.

Unlike the case of a continuous outcome variable, the absence of treatment by stratum interaction in a pair-matched design with a binary outcome is not sufficient to allow a consistent estimate of ρ to be computed. For example, a logistic regression model which omits treatment by stratum interaction terms still implies interaction on an additive scale except under the null hypothesis or when the stratification variable is unrelated to the outcome. Standard estimates of intracluster correlation, however, assume that between-cluster variability is on an additive scale.

Four different methods can be distinguished for completely randomized and stratified cluster randomization trials when individuals are the unit of analysis. The first method adjusts standard tests for the effects of clustering (Donald and Donner (1987), Donner and Donald (1988), Rosner and Milton (1988)). Estimates of intracluster correlation used in these adjustments can be calculated using the previously described ANOVA approach adapted to binary data (Fleiss, 1981, pp. 225-227). These methods have the added advantage of not requiring complicated iterative solutions. Furthermore the resulting test statistics simplify to standard test statistics when $\hat{\rho}=0$ and to standard test statistics divided by the variance inflation factor, $1+(n-1)\hat{\rho}$, when all clusters are the same size. Rao and Scott (1992) described an alternative approach which can be used to adjust standard tests for the effect of clustering by adapting the theory of ratio estimation from the sample survey literature to the analysis of correlated binary outcome data.

The remaining three methods (i.e. population-averaged, cluster-specific, Rosner's model) are computationally intensive. Important examples of each method are, respectively, the generalized estimating equations approach of Liang and Zeger (1986), the logistic-normal (Anderson and Aitkin (1985), Stiratelli, Laird and Ware (1984)), and the logistic-binomial (Mauritsen, 1984) models and Rosner's polytomous logistic model (Rosner, 1984). Their primary advantage is that they can adjust for individual-level and cluster-level baseline risk factors while testing for the effect of treatment. Unfortunately the parameters estimated by these three approaches are only equal when risk of disease is rare, treatment is unrelated to outcome or in the absence of clustering; otherwise the interpretation of the effect of treatment is dependent upon which type of model is used (Neuhaus and Jewell (1990), Neuhaus et al (1991)). Thus tests of the effect of treatment are asymptotically equivalent for these different models (Neuhaus (1991,1992), Glynn and Rosner (1992)) when the null hypothesis is true.

The beta-binomial (Williams, 1975), Williams (1982) quasi-likelihood approach, and Liang and Zeger's generalized estimating equations approach (Liang and Zeger, 1986) are all examples of population-averaged models. They are nonlinear analogs of the marginal model defined in equation (1.4.2). A common feature of these models is that they tend to simplify to ordinary logistic regression when ρ=0 or cluster size equals unity.

The population-averaged model for a stratified cluster randomization trial can be expressed as

$$\text{logit}(p_{ijs})=\beta^{*}X_{ijs} \qquad (1.4.3)$$

when there are no baseline risk factors. In this model p_{ijs} denotes the risk for subjects from the s'th cluster, s=1,..., m_{ij}, in the i'th stratum, i=1,...,k, and j'th treatment group j=1,2 and β^{*} is the vector of k+1 parameters summarizing the effects of strata and treatment, parametrized using the matrix X_{ijs}.

One method of fitting this model is to assume that the between-cluster variability can be described using a beta distribution (Williams, 1975) so that the marginal distribution of risk within a cluster follows a beta-binomial distribution. This model can be fit using the computer program EGRET (1989). Two limitations of this model are that it is not robust to violations of the parametric assumptions (Williams, 1988a) and that it cannot be used to adjust for individual-level baseline risk factors.

The assumptions underlying a beta-binomial model can be avoided by using the quasi-likelihood approach developed by Williams (1982), which only requires specifying the first two moments of the distribution. Williams' (1982) approach can only be used to adjust for cluster-level risk factors and also assumes that the degree of intracluster correlation is fixed. When all clusters are the same size the model simplifies to logistic regression with variance estimators multiplied by the variance inflation factor, $1+(n-1)\hat{\rho}$.

Liang and Zeger (1986) describe an extension of generalized linear models used to fit multivariate data. For logistic regression models their approach is similar to Williams' (1982) but the validity of inferences about the effect of treatment no longer depend on the assumption that the degree of intracluster correlation is fixed. Furthermore this model is capable of fitting both individual-level and cluster-level risk factors. The model can be fit using a computer program privately distributed by Liang and Zeger or by using the computer package SPIDA (1992). A limitation of the method described by Liang and Zeger (1986) is its requirement of a large number of clusters (Liang, Zeger and Qaqish, 1992).

Cluster-specific models are constructed in an analogous manner to the mixed-effects linear models examined in the last section. Two approximately equivalent models are the logistic-normal (Anderson and Aitkin (1985), Stiratelli, Laird and Ware (1984)) and the logistic-binomial (Mauritsen, 1984). Using a logistic-normal model equation (1.4.3) can be expressed as

$$\text{logit}(p_{ijs})=\beta X_{ijs} + C_{(ij)s} \qquad (1.4.4)$$

where p_{ijs} again denotes the risk for subjects in the s'th cluster of the i'th stratum and j'th treatment group and $C_{(ij)s}$ is the independent random effect of cluster nested in treatment and stratum, assumed to be i.i.d. $N(0,\sigma_C^2)$. The logistic-binomial differs from the logistic normal in assuming that the random effect $C_{(ij)s}$ comes from a standardized binomial distribution with parameters K and r, where K is a positive integer and $0<r<1$. The computational burden of fitting a cluster-specific model can be reduced by using a small value for K. As K increases the standardized binomial becomes approximately normal. Both models simplify to logistic regression when there is no between-cluster variability in risk and allow inclusion of individual-level and cluster-level baseline risk factors. They can be fit using the computer package EGRET (1989).

Coefficients of cluster-level variables obtained from cluster-specific models are not as easy to interpret as are coefficients from population-averaged models. For example, if a completely randomized design is employed, the effect of treatment from population-averaged models is equal to the log of the odds of the average risk in the treated group relative to the odds of the average risk in the control group. Odds ratios from cluster-specific models however, are calculated conditional on a latent variable which measures cluster-specific risk. These latter odds ratios will tend to be further from unity than those from population-averaged models (Neuhaus et al, 1991). A very simple example which illustrates this point for a binary individual-level covariate is presented by Kenward and Jones (1992 Table 5).

Rosner (1984) developed the first model which allowed inclusion of both individual-level and cluster-level covariates. This polytomous logistic regression model has been criticized for requiring a possibly artificial conditioning on responses of all other cluster members, forcing dependence between cluster size and the proportion of subjects with the outcome of interest and being too computationally burdensome to fit clusters with more than 10 subjects when individual-level baseline risk factors are included in the model (Neuhaus and Jewell (1990), Rosner (1989)).

Random assignment provides assurance that the value of ρ will be the same in treatment and control groups when the null hypothesis holds. No such assurance exists under the alternative hypothesis. The value of ρ may also vary among strata in a stratified cluster randomization trial. These considerations affect both the construction of test statistics and of confidence intervals.

Tests of the common correlation assumption can be developed using the beta-binomial distribution (Williams (1975), Kupper et al (1986)) or following the approach described by Ganio and Schafer (1992). A more sophisticated approach has been developed by Liang, Zeger and Qaqish (1992) which allows simultaneous modelling of risk and intracluster correlation. Parameter estimates calculated using this newer model are asymptotically more precise than estimates obtained using their original approach but now depends on correctly specifying the degree of intracluster correlation. An alternative extension of Liang and Zeger's (1986) approach which also allows simultaneous modelling of risk and intracluster correlation has been described by Paik (1992). This approach will likely require a very large number of clusters to obtain valid tests of significance since robust variance estimators are employed to model both risk and intracluster correlation.

Many cluster randomization trials will lack the power to reject the common correlation assumption. However the assumption of a common correlation may be necessary in trials involving only a few clusters to ensure the validity of statistical procedures. Confidence intervals for the effect of treatment could then be constructed by inverting Wald type tests as described by Donner and Klar

(1993a). The generalized estimating equations methodology might yield valid confidence intervals if the common correlation assumption was used to estimate the variance of the log odds ratio. It is also possible to construct confidence intervals using cluster specific models or Rosner's model. These approaches can not be recommended, however, because of the difficulty of interpreting odds ratios for the effect of treatment from such models (Neuhaus and Jewell (1990), Neuhaus et al (1991)).

Most cluster randomization trials, as stated in the introduction, are used to evaluate non-therapeutic interventions and generally recruit healthy subjects. Risk for subjects will naturally tend to be low when the stated objective of the trial is to reduce disease incidence or mortality. Confidence intervals from any of the multivariate methods would then be asymptotically equivalent (Neuhaus et al, 1991) and the odds ratio would be approximately equal to the relative risk.

If the rare disease assumption does not hold the confidence intervals from the different multivariate models would likely be different and none of the odds ratio estimates would be similar to the relative risk. Greenland (1987) has argued that odds ratios are only valid measures of effect when they approximate relative risks. Certainly estimates of relative risk are preferable in prospective studies. Unfortunately relative risk models (Prentice and Farewell, 1986) have not yet been extended to correlated binary outcome data.

Adjustment for baseline risk factors in randomized trials is used to increase the power to detect treatment differences. Assessment of the gain in power obtained using logistic regression models is complicated by the difficulty obtained in comparing odds ratios from models including different risk factors (Hauck et al (1991), Gart (1992), Robinson and Jewell (1991)).

Additional complications arise when individual-level baseline risk factors are included in logistic regression models adjusted for clustering. As previously stated estimates of treatment effect from population-averaged models have a more direct interpretation than do estimates obtained from cluster-specific models (Neuhaus et al, 1991). Coefficients estimating the effect of individual-level baseline risk factors, however, are simpler to interpret if obtained from cluster-specific models. One solution suggested by Neuhaus (personal communication, 1992) is to use results from both types of models when estimated treatment effects are adjusted for individual-level baseline risk factors. An example of these difficulties in the context of an observational, longitudinal study is provided by the discussion between Galbraith (1991) and Zeger, Liang and Albert (1991).

The complications described in the last two paragraphs are a result of using logistic regression which summarizes treatment effects in terms of odds ratios. These problems are reduced when the rare disease assumption holds and might not occur if models for correlated binary outcome data were available which summarized the effect of treatment using risk differences or relative risks (Gail et al, 1984).

Any solution for the problems of logistic regression would also have to address the possibility that individual-level baseline risk factors might have separate effects at the level of the individual and the cluster. This possibility is easily handled when outcomes are continuous following the approach described by Scott and Holt (1982), reviewed in Section 1.4.1. For correlated binary outcome data the possible separate effects of baseline risk factors at the level of the individual and the cluster could, perhaps, be examined using the hierarchical logistic regression models described by Wong and Mason (1985) and Goldstein (1991). These models can be fit using the computer program ML3 (Prosser, 1991).

A final consequence of using multivariate models is that estimates of intracluster correlation are likely to become smaller as more baseline risk factors are used to adjust estimates of the treatment effect. Examples of this phenomenon were given by Prentice (1988) using baseline risk factors from a cluster randomization trials and by Bull and Pederson (1987) in the context of a complex survey. This result has also been noted by Gail, Tan and Piantadosi (1988), Leisenring and Ryan (1992) and Rotnitzky and Jewell (1990). The most likely explanation is that baseline covariates account for some of the between cluster variability which arises as a consequence of shared environmental exposures and genetic relationships among cluster members. These issues are explored algebraically in the context of correlated continuous outcome data by Stanish and Taylor (1983).

An alternative explanation is possible for binary outcome data. The expected risk for any two subjects in a cluster, denoted p_u and p_v where $u \neq v$, will tend to be different as the number of individual-level baseline risk factors in the model increases. The possible range of intracluster correlation can be expressed as (Prentice, 1988)

$$\max\left\{ -\left[\frac{p_u p_v}{q_u q_v}\right]^{1/2}, -\left[\frac{q_u q_v}{p_u p_v}\right]^{1/2} \right\} \leq \rho \leq \min\left\{ \psi^{-1/2}, \psi^{1/2} \right\} \qquad (1.4.5)$$

where $\psi=(p_v/q_v)/(p_u/q_u)$ denotes the odds ratio between the the u'th and v'th cluster members. As the odds ratio gets larger ρ is forced towards zero, especially if negative estimates of intracluster correlation are unlikely.

Tests of the effect of treatment for the methods which have been discussed are only asymptotically valid. None of the methods are applicable when there are very few clusters, as is the case of most community intervention trials (Koepsell et al, 1992). Exact tests could of course be constructed using randomization theory as suggested by Williams (1988b), but would be expected to have low power. Such tests are reviewed in Section 2.4. A two-stage method described by Gail, Tan and Piantadosi (1988), in which results from a preliminary analysis are used as response measures in a primary analysis, might also allow adjustment for baseline risk factors including pre-treatment estimates of risk (Duffy et al, 1992).

An additional concern is that cluster size is not necessarily unrelated to the risk of a positive outcome in a cluster. Haseman and Kupper (1979), for example, have noted that cluster size is sometimes correlated with risk in toxicological studies. Toxicologists randomly assign treatment to pregnant animals to determine the risk of birth defects or death among their offspring. Particularly potent chemicals might not only increase the mortality of animals after birth but could also induce resorption of the fetus in utero reducing the eventual cluster size (Catalano and Ryan, 1992).

Some work in this area has been done by Rai and van Ryzin (1985). Their model has been criticized by Williams (1987), however, for not sufficiently adjusting for overdispersion. Williams (1987) proposes including cluster size as an additional covariate as an alternative approach and acknowledges that more parametric methods which impose an underlying distribution on cluster size could also be developed.

1.4.3 Methods of Analysis for Data with Other Outcomes

The development of methods for the analysis of correlated continuous or dichotomous outcomes is fairly advanced in comparison to the development of methods for other types of response data. There has been little work on the analysis of correlated multinomial or ordinal data and there has been even less development of methods which could be used to analyze correlated failure times data. The near absence of methods for these outcomes reflects the greater complexity of the problem, and the relatively recent introduction of methods for uncorrelated multinomial, ordinal or failure time data. Recent publications (Ashby et al (1991), Binder (1992), Hougaard et al (1992), Liang, Self and Chang (1993), O'Hara Hines and Lawless (1993), Segal and Neuhaus (1993)) suggest that in spite of limitations (Andersen, 1991) progress is being made in the development of methods for some of these outcomes.

All of the methods reviewed in this chapter need to be extended to allow for multiple longitudinal measurements per subject. Koepsell et al (1991) review such methods for continuous outcome data but there has not yet been any examination of similar methods for other outcomes.

1.4.4 Diagnostic Methods for Correlated Outcome Data

There has been little development of methods for testing the assumptions underlying models used to analyze correlated outcome data (Liang, Zeger and Qaqish (1992), Moulton and Zeger (1989), Hocking, Green and Bremer (1989)). This is not surprising, given the scope of the problem. Not only do all of the diagnostic methods developed for uncorrelated observations (e.g. Cook and Weisberg, 1982) need to be extended to allow for intracluster correlation but additional methods are needed to examine the assumptions particular to correlated outcomes models.

An example of the difficulty is offered by considering the extension of methods needed to detect influential points in the data. For correlated outcomes data such points might be either particular subjects in a cluster or even entire clusters. Influential points could also affect both coefficients in regression models and estimates of intracluster correlation.

Several researchers have begun developing diagnostic methods for correlated outcome data. Hocking et al (1989), for example, have developed methods for correlated continuous data while Roberts et al (1987) discuss diagnostic methods for logistic regression analyses of complex sample surveys. Methods which allow joint modelling of risk and intracluster correlation (Liang et al (1992), Paik (1992)) or tests for the variation of intracluster correlation as a function of treatment, stratification variables or baseline risk factors (Ganio and Schafer, 1992) can also be useful diagnostic tools.

Methods are also available to test for the presence of overdispersion (e.g. Tarone (1979), Dean (1992), Ganio et al (1992)). Such tests should not be used in cluster randomization trials because they will tend to have low power when correlations are near zero (Donner and Klar, 1993b). Furthermore the results of Donner (1982), reviewed above, indicate that for large clusters even small correlations can have sizable effects on estimated variances. For example, consider a cluster randomization trial in which there were 100 subjects per cluster and where the intracluster correlation coefficient was 0.01. Then, using equation (1.3.1), the variance of the estimated treatment effect for this trial can be shown to be approximately twice that of the variance obtained under the assumption that responses of subjects from the same cluster are independent. Thus rejection rates for tests of the effect of treatment would be elevated if failure to reject the null hypothesis that $\rho=0$ was used as evidence for the absence of clustering.

2. Tests of the Effect of Treatment in Stratified Cluster Randomizatior Trials

2.1 Introduction

Detailed algebraic examination of several methods of analysis are presented in this chapter. Attention is restricted to methods which could be used in stratified cluster randomization trials. Particular attention is paid to those methods which will be used in the simulation study described in Chapter 3.

There are two asymptotic cases which could arise in stratified cluster randomization trials, one assuming a large number of strata, the other assuming a large number of clusters per stratum. The average cluster size is determined by the type of clusters (e.g. families, classrooms, communities) used in the study and so can be considered to be fixed by design. These asymptotic conditions are equivalent to the two cases presented by Hauck (1989) for uncorrelated binary outcome data.

A matched-pairs design arises as the limit of the first asymptotic case. Methods of inference for such designs are all performed at the cluster level using between-stratum information tc obtain estimates of variability, as described in Chapter 1. These methods are quite well known and have been reviewed by Donner (1987) and Donner and Klar (1993a).

There has been comparatively much less work on methods developed for designs in which there are few strata where estimates of variance are obtained using between cluster variability within each stratum. Therefore only this asymptotic case will be considered. Examples of trials in which such designs were used are given in Table 1.1 of the previous chapter.

The primary focus of this chapter is on comparisons of the different ways in which statistical tests can be constructed for correlated binary outcome data from stratified cluster randomization trials. Less attention is paid to confidence interval construction although methods which can be used for this purpose are identified. Some of the issues arising from variation in the degree of intracluster correlation across strata are also discussed.

The chapter begins with a discussion of the similarities and differences which exist between moment and ANOVA estimators of intracluster correlation. This discussion is required to understand the relationship between different tests of the effect of treatment in stratified cluster randomization trials.

The main body of the chapter is devoted to describing six different ways in which tests can be constructed for this design: adaptations of linear models, non-parametric tests, simple adjustments of methods originally developed for binomially distributed data, adaptations of methods developed for sample surveys, beta-binomial models and Liang and Zeger's (1986) generalized estimating equations approach. The first four approaches were selected because they are simple non-iterative methods. The beta-binomial model is investigated both because it is probably the most widely known method used to analyze correlated binary outcome data and because both it and the generalized estimating equations approach employ population-averaged models. Odds ratios of the effect of treatment from such models can be interpreted in the same way as simpler, but still valid estimates such as the Mantel-Haenszel odds ratio (see Section 1.4.2). Cluster-specific models and Rosner's model are omitted primarily because of difficulties found in interpreting estimates of the effect of treatment when using these approaches.

The size of stratified cluster randomization trials described in this chapter is determined by the number of clusters in the i'th stratum, i=1,...,k and j'th treatment group, j=1,2, denoted m_{ij}, and the number of subjects in each cluster, denoted n_{ijs}, s=1,..., m_{ij}. Algebraic comparisons between the six approaches are often only possible if simplifying assumptions are made about the number of clusters per treatment group and the variability in cluster size.

The least restrictive assumption is that clusters within a stratum are all the same size. This occurs fairly frequently since clusters are often stratified by size. Randor assignment within strata will usually assure that there will also be equal numbers of clusters in both treatment groups of the i'th stratum (i.e. $m_{ij} = m_i$, j=1,2, i=1,...,k). Studies are balanced when all the clusters in the i'th stratum are the same size and when there are equal numbers of clusters in both treatment groups of each stratum.

At times an additional assumption that there is no variability in cluster size between strata is required to draw out similarities between methods. The additional assumption that there are equal numbers of clusters in each treatment group and stratum is occasionally imposed for the sake of drawing insights. Designs which meet this last criterion (i.e. m_{ij}=m, n_{ijs}=n) are said to be completely balanced.

2.2 Introduction to Moment and ANOVA Estimates of Intracluster Correlation

There are at least three different methods which can be used to estimate coefficients of intracluster correlation for binary outcome data. These include fully parametric approaches employing distributions such as the beta-binomial, the use of methods of moments (Moore and Tsiatis, 1991), and adaptations of AN^OVA metnods to binary outcomes (Fleiss (1981 Section 13.2), Landis and Koch (1977)).

Fully parametric methods based on the beta-binomial or logistic-normal distributions, for example, require sophisticated software. Estimates derived using such models can be biased when the parametric assumptions are not met. Simpler and more robust methods are available as an alternative, with the consequent loss of precision when the parametric assumptions are true.

The moment and ANOVA estimators are consistent so long as the first moments are accurately estimated, a far less restrictive assump:ion. The moment method, ho.vever, requires an iterative solution when cluster sizes are variable. The similarity between these two approaches is most evident when estimating intracluster correlation in a single population.

Consider a single population from which a random sample of m clusters is drawn each of which contains n subjects. Let $\hat{p}_s$ denote the observed risk for the s'th cluster and $\hat{q}_s = 1 - \hat{p}_s$ so that the average risk for the m clusters is $\hat{p} = \sum_{s=1}^{m} \hat{p}_s / m$ and $\hat{q} = 1 - \hat{p}$. Following Moore and Tsiatis (1991) the moment estimate is derived by solving the equation,

$$\frac{n\sum_{s=1}^{m} (\hat{p}_s - \hat{p})^2}{\hat{p}\hat{q}\,[1 + (n-1)\rho]} = m-1, \qquad (2.2.1)$$

for ρ yielding

$$\hat{\rho}_M = \frac{1}{n-1}\left[\frac{n\,s^2}{\hat{p}\hat{q}} - 1\right] \qquad (2.2.2)$$

where $s^2 = \sum_{s=1}^{m} (\hat{p}_s - \hat{p})^2/(m-1)$. This statistic is identical to κ_C derived by Donner, Birkett and Buck (1981) except they used m rather than m-1 degrees of freedom in the denominator of s^2.

Note that as cluster size becomes large

$$\hat{\rho}_M \approx \frac{\hat{\text{var}}(\ \hat{p}\ )}{\hat{p}\hat{q}}\ , \qquad (2.2.3)$$

which is a consistent estimate of the intracluster correlation coefficient obtained for mixed-effects logistic regression models (Neuhaus, Kalbfleisch and Hauck, 1991). This formula for intracluster correlation was also derived by Kraemer (1979) in the context of population models for a coefficient of reliability.

The ANOVA estimator, denoted $\hat{\rho}_A$, is derived by adapting the estimator of ρ from a one-way random effects model (Fleiss, 1981 Section 13.2) to binary outcomes data. Following Landis and Koch (1977) it can be expressed as

$$\hat{\rho}_A = (\text{ BMS} - \text{WMS })/(\text{ BMS} + (n-1)\text{WMS })\qquad(2.2.4)$$

where the between cluster mean sum of squares, BMS = ns^2 and the within cluster mean sum of squares, WMS = $\sum_{s=1}^{m} n\hat{p}_s\hat{q}_s/m(n-1)$. This statistic differs from $\hat{k}$ derived in Section 13.2 of Fleiss (1981) by using m-1 rather than m degrees of freedom in the denominator of s^2. The effect of the choice of degrees of freedom on estimates of ρ was evaluated by Feng and Grizzle (1992) in their simulation study. They found that Fleiss's (1981) version of this statistic underestimates ρ even if there are as many as 30 clusters in the study (i.e. m = 30). The moment and ANOVA estimators of intracluster correlation are identical when $m/(m-1) \approx 1$.

It is important to note that estimates of p and ρ are not likely to be independent. The asymptotic covariance between $\hat{p}$ and the moment estimator derived by Donner, Birkett and Buck (1981) can be expressed as

$$\frac{n}{n-1}\left\{\frac{(1-2p)[1+(n-1)\rho]^2}{n^2} + \frac{E(\hat{p}_s - p)^3}{pq}\right\}\qquad(2.2.5)$$

Moore (1985, p. 50). A more explicit calculation requires additional assumptions about the distribution of the number of subjects from the s'th cluster, s=1,...,m, having a positive response.

Suppose that the number of subjects from the s'th cluster who responded positively can be described using the beta-binomial distribution introduced in Section 1.4.2 and described in detail in Section 7 of this chapter. In this case results from Moore (1985, p. 33) can be used to show that

$$\mathrm{cov}(\,\hat{p}\,,\hat{\rho}\,) = \frac{(1-2p)\rho(1-\rho)[1+(n-1)\rho]}{n(1+\rho)} \tag{2.2.6}$$

so that for $\rho > 0$

$$\mathrm{corr}(\,\hat{p}\,,\hat{\rho}\,) \begin{cases} >0 & \text{if } p<1/2 \\ =0 & \text{if } p=1/2 \\ <0 & \text{if } p>1/2. \end{cases} \tag{2.2.7}$$

Note that $\hat{p}$ and $\hat{\rho}$ are also uncorrelated when $\rho = 0$.

This dependence complicates inference. For example, several authors have noted that estimates of intracluster correlation tend to become larger as estimates of risk increase from 0 to 0.5 (e.g. Williams, 1988b). This empirical result has been used as an argument in favor of models which allow ρ to vary. An additional factor which might be associated with the phenomenon is the dependence between $\hat{p}$ and $\hat{\rho}$.

2.3 Adaptations of Linear Models to Correlated Binary Outcome Data

2.3.1 Derivation of Results when Outcomes are Continuous

Tests of the effect of treatment constructed using mixed-effects linear models were outlined in Section 1 of the thesis. The technical details needed to construct such tests are described in this section using analysis of variance. Difficulties arising with this approach when cluster sizes are variable are also discussed. The variance components from the analysis of variance are then used to derive an estimator of intracluster correlation.

A mixed-effects linear model which includes covariates for treatment and strata is given by

$$y_{ijst}=\mu+\alpha_i+\beta_j+\alpha\beta_{ij}+C_{(ij)s}+\varepsilon_{(ijs)t} \qquad (2.3.1)$$

where y_{ijst} is the score for the t'th subject, $t=1,\ldots,n_{ijs}$, from cluster s, $s=1,\ldots,m_{ij}$, treatment j, j=1,2 and stratum i, i=1,...,k. The fixed effects of strata, treatment and their interaction are represented by α_i, β_j, and $\alpha\beta_{ij}$ respectively, while $C_{(ij)s}$ and $\varepsilon_{(ijs)t}$ denote the respective independent random effects of cluster nested in treatment and stratum, assumed to be i.i.d. $N(0,\sigma_C^2)$, and subject nested in cluster, assumed to be i.i.d. $N(0,\sigma^2)$. The usual linear constraints

$$\sum_{i=1}^{k}\alpha_i=\sum_{j=1}^{2}\beta_j=\sum_{i=1}^{k}\alpha\beta_{ij}=\sum_{j=1}^{2}\alpha\beta_{ij}=0 \qquad (2.3.2)$$

are imposed to fit the model. This model can be fit to data from a stratified cluster randomization trial which includes $N = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} n_{ijs}$ subjects from $M = \sum_{i=1}^{k} \sum_{j=1}^{2} m_{ij}$ clusters.

The sums of squares for the fixed effects can be expressed as

$$\text{SS(Fixed Effects)} = \sum_{i=1}^{k} \sum_{j=1}^{2} n_{ij} (\bar{y}_{ij..} - \bar{y}_{....})^2 \quad \text{where} \tag{2.3.3}$$

$$\bar{y}_{ij..} = \sum_{s=1}^{m_{ij}} \sum_{t=1}^{n_{ijs}} y_{ijst} \Big/ \sum_{s=1}^{m_{ij}} n_{ijs}$$

$$\text{and } \bar{y}_{....} = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \sum_{t=1}^{n_{ijs}} y_{ijst} / N.$$

Imbalance in cluster size prevents decomposing SS(Fixed Effects) into separate sums of squares for strata, treatment, and their interaction complicating construction of tests of significance. The sums of squares for the fixed effects can, however, be written as a sum of sequential or Type I sum of squares (SAS (1989, page 115), Rawlings (1988, Section 4.5 and Section 16.5), Speed, Hocking and Hackney (1978)), i.e.

$$\begin{aligned} \text{SS(Fixed Effects)} = {} & \text{SS(Strata)} \\ & + \text{SS(Treatment } | \text{ Strata)} \\ & + \text{SS(Interaction } | \text{ Strata,Treatment).} \end{aligned} \tag{2.3.4}$$

These sums of squares, their degrees of freedom, and accompanying expected mean squares are displayed in Table 2.1. The weighted average of stratum specific treatment differences is denoted by

$$\bar{d}_w = \sum_{i=1}^{k} w_i \, \bar{d}_i/w. \text{ where} \tag{2.3.5}$$

$$\bar{d}_i = \bar{y}_{i2..} - \bar{y}_{i1..} \text{ and}$$

$$w_i = \left[\frac{1}{n_{i1.}} + \frac{1}{n_{i2.}}\right]^{-1}.$$

Statistical tests are constructed from ANOVA tables using ratios of mean sums of squares. Mean sums of squares are selected so that ratios of their expected values are equal under the null hypothesis and are larger than one under the alternative hypothesis. Furthermore when the null hypothesis is true the ratio of mean sums of squares follow an F distribution with degrees of freedom obtained from the numerator and denominator mean sums of squares respectively. These distributional properties are a mathematical consequence of the assumption that the random errors are normally distributed.

Such exact test statistics cannot be constructed for the effect of treatment in stratified cluster randomization trials when cluster size is variable. For example, under the null hypothesis that treatment is unrelated to outcome the expected mean squares for treatment given strata and for cluster are unequal (i.e. $A_3 \neq A_7$). A test of the effect of treatment could only be constructed in the extremely unlikely instance that there is no between-cluster variability (i.e. $\sigma_C^2 = 0$) using the ratio of mean sums of squares for treatment given strata and the mean sums of squares for error. In either case validity of the test statistic depends on the absence of any interaction since $E(\bar{d}_i) = \beta_2-\beta_1 + \alpha\beta_{i2}-\alpha\beta_{i1}$.

This difficulty in constructing tests of the effect of treatment is not unique to stratified cluster randomization trials. Dunn and Clark (1987, pp. 116-119) and Donner (1985a) noted similar difficulties in situations comparable to unbalanced completely randomized cluster randomization trials.

Even though exact tests of the effect of treatment cannot be constructed when cluster sizes are variable there are several approaches which could be used to construct approximate tests of significance. Dunn and Clark (1986, pp. 116-119), for example, suggest using Satterthwaite's (1946) procedure to construct an approximate F-test. Alternatively a cluster-level analysis could be constructed as outlined in Chapter 1 of the thesis or the generalized least squares method described by Donner (1985b) could be used.

The cluster-level analysis is the simplest approach. In this case the test statistic described by Schwartz, Flamant and Lellouch (1980, pp. 189-191) can be adapted to test the effect of treatment in stratified cluster randomization trials. It is denoted by

$$t_{M-2k} = \frac{\sum\limits_{i=1}^{k} \frac{m_{i1}m_{i2}}{m_{i.}} \bar{D}_i}{S_u \left[\sum\limits_{i=1}^{k} \frac{m_{i1}m_{i2}}{m_{i.}}\right]^{1/2}} \quad \text{where} \qquad (2.3.6)$$

$$\bar{D}_i = \bar{Y}_{i2} - \bar{Y}_{i1},$$

$$\bar{Y}_{ij} = \sum_{s=1}^{m_{ij}} \bar{y}_{ijs.} / m_{ij},$$

$$\bar{y}_{ijs.} = \sum_{t=1}^{n_{ijs}} y_{ijst} / n_{ijs} \quad \text{and}$$

$$S_u^2 = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \left[\bar{y}_{ijs.} - \bar{Y}_{ij}\right]^2 /(M-2k).$$

and is approximately distributed as a random variable from a Students t-distribution with M-2k degrees of freedom. Furthermore it reduces to the ordinary two sample t-test performed on the cluster means in completely randomized designs. An alternative weighting scheme could be developed which also adjusts for the degree of clustering.

The simplification which occurs when all clusters have the same number of subjects, $n_{ijs} = n$ and when there are the same number of clusters, $m_{ij} = m$, in each strata and treatment group (i.e. completely balanced designs) is displayed in Table 2.2. This table could alternatively have been constructed using the algorithms developed to obtain sums of squares and expected mean squares for analysis of variance (Montgomery, 1984 Chapter 8). Notice that under balance

$$\begin{aligned} \text{SS(Fixed Effects)} = {} & \text{SS(Strata)} \\ & + \text{SS(Treatment | Strata)} \\ & + \text{SS(Interaction | Strata,Treatment)} \end{aligned} \tag{2.3.7}$$

$$= SS(\text{ Strata }) + SS(\text{ Treatment }) + SS(\text{ Interaction }).$$

Tests of the null hypothesis that treatment has no effect can now be constructed using the ratio of mean sums of squares for treatment and cluster respectively. Under Ho this test statistic follows an F distribution with 1 and 2k(m-1) degrees of freedom and can be expressed as

$$F_{1,2k(m-1)} = \left\{ \frac{\bar{d}}{S_p\sqrt{2/km}} \right\}^2 \qquad (2.3.8)$$

where $\bar{d} = \sum_{i=1}^{k} \bar{d}_i/k$, a special case of equation (2.3.5), while

$$S_p^2 = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m} (\bar{y}_{ijs} - \bar{y}_{ij.})^2/2k(m-1)$$
$$= \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{1}{2k} s_{ij}^2$$

is an unbiased estimator of between cluster variability. Thus the test statistic is the square of a stratified t-test (Fleiss, 1986 Section 6.1) performed on the cluster means and simplifies to the square of a t-test in a completely randomized design.

An additional special case arises when there is no treatment by stratum interaction. The sums of squares for interaction can then be pooled with the sums of squares for estimating the between-cluster variability. The test statistic constructed as a ratio of mean sums of squares for treatment and the new pooled mean sums of squares is equal to the square of a paired t statistic with k-1 degrees of freedom in matched-pairs designs where there are only two clusters per stratum (i.e. m=1).

These results validate the cluster-level analyses discussed in Section 1.4.1 and originally described by Lindquist (1940), at least when there is no variability in cluster size. Similar identities have been described by Hopkins (1982) who, like Lindquist, was concerned with the analysis of data arising in educational research. Earlier proof that unweighted cluster-level analysis is fully informative is due to Greenhouse and Geisser (1959) who were concerned with the analysis of repeated measures on the same subjects. They also showed that such cluster-level analyses are valid for testing between-group differences even if the underlying correlation matrix was not exchangeable.

The sums of squares and their expected values displayed in Table 2.1 and Table 2.2 can also be used to estimate the degree of intracluster correlation. The intracluster correlation coefficient is defined to be $\rho_S = \sigma_C^2 / [\sigma_C^2 + \sigma^2]$ and, using model (2.3.1), can also be derived as the correlation between any two members of the same cluster since $\text{cov}(y_{ijst}, y_{ijsu}) = \sigma_C^2$, and $\text{var}(y_{ijst}) = \sigma_C^2 + \sigma^2$, for $t \neq u$, $t,u=1, \cdots ,n_{ijs}$. Thus model (2.3.1) implies that responses of cluster members are equally correlated.

The estimate of intracluster correlation denoted

$$\hat{\rho}_S = \hat{\sigma}_C^2/(\hat{\sigma}_C^2+\hat{\sigma}^2) \tag{2.3.9}$$

$$= \frac{\text{MS(Cluster)} - \text{MS(Error)}}{\text{MS(Cluster)} + (n_o - 1)\ \text{MS(Error)}}$$

since $\hat{\sigma}_C^2$ = (MS(Clusters) − MS(Error))/n_o , $\hat{\sigma}^2$ = MS(Error) , and

$$n_o = \frac{N - \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{n_{ijs}^2}{n_{ij.}}}{\sum_{i=1}^{k} \sum_{j=1}^{2} (m_{ij} - 1\ )}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{m_{ij} - 1}{\sum_{i=1}^{k} \sum_{j=1}^{2} (\ m_{ij} - 1\ )} \left\{ \bar{n}_{ij.} - \frac{\hat{v}_{ij}}{m_{ij}\ \bar{n}_{ij.}} \right\} \quad \text{where}$$

$\hat{v}_{ij}$ denotes the observed between-cluster variability in cluster size. Note that when there is only a single stratum (i.e. k=1) this estimator of intracluster correlation simplifies to the intraclass correlation coefficient derived by Donner (1985a) assuming only two populations.

A limitation of $\hat{\rho}_S$ is that it may be negative even though the parameter can only be positive. The lower limit of $\hat{\rho}_S$ is $-1/\max(\ n_{ijs} - 1\ )$ since, from model (2.3.1),

$$\hat{\text{var}}(\bar{y}_{ijs.}) = \frac{\hat{\sigma}_C^2 + \hat{\sigma}^2}{n_{ijs}} [1 + (\ n_{ijs} - 1\ )\ \hat{\rho}] \qquad (2.3.10)$$

which is greater than or equal to zero. Following Searle, Casella and McCulloch (1992, page 60) such negative estimates of ρ are usually set to zero in practice.

A common simplification used in estimating ρ involves replacing n_o with $\bar{n}_{....} = N/M$, where $n_o \le \bar{n}_{....}$. The second algebraic expression for n_o suggests this is reasonable when the number of clusters in each treatment group and stratum is large. Replacing n_o with $\bar{n}_{....}$ might otherwise result in underestimating the degree of intracluster correlation. For example, if the number of clusters per treatment group, $m_{ij} = m$ for all $i=1,...,k$ and $j=1,2$ then $n_o \le \bar{n}_{....}$ for small m. Note that n_o is also less than or equal to $\bar{n}_{....}$ when the number of clusters per treatment group and stratum is variable but $\bar{n}_{ij..} = \bar{n}_{....}$, $i=1,...,k$, $j=1,2$. No general relationship appears to exist between n_o and $\bar{n}_{....}$ when the number of clusters per treatment group and stratum is variable.

Thus imbalance complicates but does not otherwise affect estimation of ρ. The simplification which obtains for completely balanced designs allows $\hat{\rho}_S$ to be rewritten as a weighted average of stratum and treatment specific one-way random effect estimators of intracluster correlation (Fleiss, 1986, page 11) denoted $\hat{\rho}_{ij}$. The weights are functions of stratum and treatment specific estimates of between-cluster variability. That is,

$$\hat{\rho}_S = \frac{\text{MS(Cluster)} - \text{MS(Error)}}{\text{MS(Cluster)} + (n-1)\text{MS(Error)}} \tag{2.3.11}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{2} W_{ij}\, \hat{\rho}_{ij} \quad \text{where}$$

$$\hat{\rho}_{ij} = \frac{MSC_{ij} - MSE_{ij}}{MSC_{ij} + (n-1)MSE_{ij}},$$

$$W_{ij} = \frac{1}{2k} \frac{MSC_{ij} + (n-1)MSE_{ij}}{MS(\text{ Cluster }) + (n-1)MS(\text{ Error })},$$

$$MS(\text{ Cluster }) = \sum_{i=1}^{k} \sum_{j=1}^{2} MSC_{ij} / 2k,$$

$$MSC_{ij} = \sum_{s=1}^{m} n(\bar{y}_{ijs.} - \bar{y}_{ij..})^2 / (m-1),$$

$$MS(\text{ Error }) = \sum_{i=1}^{k} \sum_{j=1}^{2} MSE_{ij} / 2k \quad \text{and}$$

$$MSE_{ij} = \sum_{s=1}^{m} \sum_{t=1}^{n}(y_{ijst} - \bar{y}_{ijs.})^2 / m(n-1).$$

The mean square estimates are calculated using the sum of squares and degrees of freedom displayed in Table 2.2. A simpler summary estimate of intracluster correlation could be constructed as the unweighted average of stratum and treatment specific estimators $\hat{\rho}_{ij}$. This approximation is reasonable since the weights are asymptotically equal if the model holds.

2.3.2 Adaptations for Binary Outcome Data

The ANOVA methods described in Section 2.3.1 can also be used to construct a test of the effect of treatment and to estimate ρ for binary outcome data. When the design is completely balanced and y_{ijst} is binary and scored 1 for success and 0 for failure $\bar{y}_{ijs.} = \hat{p}_{ijs}$ which estimates the risk for subjects from the s'th cluster, i'th stratum and j'th treatment group. The average risk for subjects from the i'th stratum and j'th treatment group is $\hat{p}_{ij} = \sum_{s=1}^{m} \hat{p}_{ijs}/m$ and the average risk for subjects from the i'th stratum is $\hat{p}_i = \sum_{j=1}^{2} \sum_{s=1}^{m} \hat{p}_{ijs}/2m$.

Therefore the test statistic for the effect of treatment obtained using the results displayed in Table 2.2 is

$$F_{1,2k(m-1)} = \left\{ \frac{\bar{d}}{S_p\sqrt{2/km}} \right\}^2 \qquad (2.3.12)$$

$$= \frac{\left\{ \sum_{i=1}^{k} \frac{mn}{2} (\hat{p}_{i2} - \hat{p}_{i1}) \right\}^2}{\sum_{i=1}^{k} \frac{mn}{2} \cdot \hat{p}_i \hat{q}_i [1+(n-1)\bar{\rho}]}$$

$$= \frac{\chi_C^2}{1+(n-1)\bar{\rho}}$$

where χ_C^2 is Cochran's (1954) version of the Mantel-Haenszel statistic (Mantel and Haenszel, 1959) and $\bar{\rho}$ is an estimate of intracluster correlation. Statistical significance is determined by comparing the test statistic to an F distribution with 1 and 2k(m-1) degrees of freedom which can be approximated by a χ_1^2 distribution as m gets large.

This relationship arises because $\bar{d} = \sum_{i=1}^{k} (\hat{p}_{i2} - \hat{p}_{i1}) / k$ and

$$kn \frac{S_P^2}{\sum_{i=1}^{k} \hat{p}_i \hat{q}_i} = n \frac{\sum_{i=1}^{k} \sum_{j=1}^{2} \frac{1}{2} s_{ij}^2}{\sum_{i=1}^{k} \hat{p}_i \hat{q}_i} \qquad (2.3.13)$$

$$= \sum_{i=1}^{k} \frac{\hat{p}_i \hat{q}_i}{\sum_{i=1}^{k} \hat{p}_i \hat{q}_i} \left\{ \sum_{j=1}^{2} \frac{1}{2} \frac{n\, s_{ij}^2}{\hat{p}_i \hat{q}_i} \right\}$$

$$= \sum_{i=1}^{k} \frac{\hat{p}_i \hat{q}_i}{\sum_{i=1}^{k} \hat{p}_i \hat{q}_i} \sum_{j=1}^{2} \frac{1}{2} \{1 + (n-1)\hat{\rho}_{ij}]$$

$$= 1 + (n-1)\bar{\rho}$$

where $\hat{\rho}$ is a moment estimator of intracluster correlation calculated using m-1 degrees of freedom but estimating p_{ij} in the denominator under the null hypothesis (see Section 2.2). A similarly weighted estimator can be obtained by adapting the estimate of intracluster correlation derived by Donner (1985a) to binary outcome data.

If cluster sizes are variable this simple relationship does not hold. A stratified t-tes' on the cluster means may still be valid. The properties of such a test have been examined by simulation and are described in Chapter 3 of the thesis.

An alternative estimator of intracluster correlation for stratified cluster randomization trials is provided by adapting $\hat{\rho}_S$ to binary outcome data. This estimator can be used even if the design is unbalanced. It is theoretically plausible to estimate intracluster correlation coefficients for binary outcome data using variance components originally proposed for continuous outcome data because the expected mean squares can be calculated without the need for normality assumptions (Scheffe, 1959 page 229).

When m is large and the design is completely balanced

$$\hat{\rho}_S = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{\hat{p}_{ij}\, \hat{q}_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{2} \hat{p}_{ij}\, \hat{q}_{ij}}\, \hat{\rho}_{ij} \qquad (2.3.14)$$

where $\hat{\rho}_{ij}$ is the ANOVA estimator of intracluster correlation discussed in Section 2.2 calculated using m degrees of freedom to estimate between-cluster variability.

Donald and Donner (1987) advocated using the arithmetic average of the 2k estimators $\hat{\rho}_{ij}$. This simpler statistic will tend to be less precise than $\hat{\rho}_S$. An indication of the loss of precision is obtained by observing that when balance holds ρ_S can be written as a weighted average of ρ_{ij} and noting that these weights will likely vary between strata when outcomes are binary.

The tests of significance and estimates of intracluster correlation which have been described were calculated allowing for interaction between treatment and strata. If such interaction seems unlikely the precision of these procedures could be increased by pooling the sums of squares from the interaction with the between-cluster sums of squares. Even small increases in precision can be important when there are few clusters in the trial.

The absence of interaction in stratified samples of binary outcome data is usually described in terms of equality of odds ratios across strata, i.e.

$$\begin{aligned}\psi_i &= (p_{i2}q_{i1}/p_{i1}q_{i2}) \\ &= \psi\end{aligned} \tag{2.3.15}$$

The parametric models introduced in the last section, however, define the absence of interaction on a linear rather than a multiplicative scale, i.e. $p_{i2} - p_{i1} = \beta_2 - \beta_1$ (see Table 2.1). Equality between odds ratios across strata is likely to result in variable risk differences unless the stratification variables are unrelated to the outcome or when the common odds ratio equals one. Therefore pooling the interaction and between-cluster sums of squares increases the precision of tests of the effect of treatment when the null hypothesis holds but reduces the power of the test under the alternative hypothesis when it is defined in terms of odds ratios. In this instance estimates of intracluster correlation calculated by pooling the interaction and between-cluster sums of squares will be positively biased.

For testing the effect of treatment in randomized studies the precision of estimates of intracluster correlation can be increased further by calculating them under the null hypothesis. This is accomplished by pooling the sums of squares from the effect of treatment with the sums of squares for the treatment and strata interaction and the sums of squares for between-cluster variability. An equivalent statistic can be obtained by adapting the estimate of intracluster correlation described by Donner (1985a). Such estimates of ρ will be positively biased under the alternative hypothesis whether it is defined in terms of risk differences or odds ratios.

Table 2.1
ANOVA Table for Unbalanced Stratified Cluster Randomization Trial
Fixed Effects Estimated using Sequential (Type I) Sums of Squares

Source	df	SS	E(MS)
Strata	k-1	$\sum_{i=1}^{k} n_{i..} (\bar{y}_{i...} - \bar{y}_{....})^2$	$\sigma^2 + A_1 \sigma_C^2$ $+A_2$
Treatment (Rx) Given Strata	1	$\left[\sum_{i=1}^{k} w_i \bar{d}_i\right]^2 / w.$	$\sigma^2 + A_3 \sigma_C^2$ $+ A_4$
Strata x Rx Given Strata, Rx	k-1	$\sum_{i=1}^{k} w_i(\bar{d}_i - \bar{d}_w)^2$	$\sigma^2 + A_5 \sigma_C^2$ $+ A_6$
Cluster	M − 2k	$\sum_{ijs} n_{ijs}(\bar{y}_{ijs.} - \bar{y}_{ij..})^2$	$\sigma^2 + A_7 \sigma_C^2$
Error	N − M	$\sum_{ijst} (y_{ijst} - \bar{y}_{ijs.})^2$	σ^2

Table 2.1 continued ...

Coefficients for E(MS) of the Variance Components

$$A_1 = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{n_{ijs}^2}{k-1} \left[\frac{1}{n_{i..}} - \frac{1}{N} \right]$$

$$A_2 = \sum_{i=1}^{k} \frac{n_{i..}}{k-1} \left[E\left(\bar{y}_{i...} - \bar{y}_{....} \right) \right]^2$$

$$A_3 = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{w_i^2}{w.} \left(\frac{n_{ijs}}{n_{ij.}} \right)^2, \quad \text{where } w_i^{-1} = \frac{1}{n_{i1.}} + \frac{1}{n_{i2.}}$$

$$A_4 \quad \left(\sum_{i=1}^{k} w_i E\left[\bar{d}_i \right] \right)^2 / w.$$

$$A_5 = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{w_i}{k-1} \left(1 - \frac{w_i}{w.} \right) \left(\frac{n_{ijs}}{n_{ij.}} \right)^2$$

$$A_6 = \sum_{i=1}^{k} \frac{w_i}{k-1} \left[E\left[\bar{d}_i - \bar{d}_w \right] \right]^2$$

$$A_7 = \frac{N - \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{n_{ijs}^2}{n_{ij.}}}{\sum_{i=1}^{k} \sum_{j=1}^{2} (m_{ij} - 1)}$$

Table 2.2
ANOVA Table for Balanced Stratified Cluster Randomization Trial

Source	df	SS	E(MS)
Strata	k-1	$2mn\sum_{i=1}^{k}(\bar{y}_{i...} - \bar{y}_{....})^2$	$\sigma^2 + n\sigma_C^2$ $+ 2mn\sum_{i=1}^{k}\frac{\alpha_i^2}{k-1}$
Treatment (Rx)	1	$\frac{kmn}{2}(\bar{d}.)^2$	$\sigma^2 + n\sigma_C^2$ $+ kmn\sum_{j=1}^{2}(\beta_j)^2$
Strata x Rx	k-1	$\frac{mn}{2}\sum_{i=1}^{k}(\bar{d}_i - \bar{d}.)^2$	$\sigma^2 + n\sigma_C^2$ $+ nm\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{(\alpha\beta)_{ij}^2}{k-1}$
Cluster	2k(m-1)	$n\sum_{ijs}(\bar{y}_{ijs.} - \bar{y}_{ij..})^2$	$\sigma^2 + n\sigma_C^2$
Error	2km(n-1)	$\sum_{ijst}(y_{ijst} - \bar{y}_{ijs.})^2$	σ^2

2.4 Nonparametric Tests of the Effect of Treatment

2.4.1 Introduction

Nonparametric tests are the simplest and most robust methods available to evaluate the effect of treatment in cluster randomization trials. Their primary advantage is that they can be constructed without requiring any distributional assumptions. Weil (1970) was the first to suggest using such tests in cluster randomization trials. His advice was foreshadowed, however, by Wilcoxon (1945) who used, what is likely, a cluster randomization trial as an example to motivate derivation of the Wilcoxon rank sum test. More recently nonparametric tests have been recommended by Crump et al (1991), Edgington (1987, Section 2.3) and Williams (1988b) for use in cluster randomization trials.

The statistical significance of nonparametric tests can be determined using randomization theory. Test statistics are calculated for each permutation of the data and exact p-values are defined as the proportion of test statistics calculated using permuted data which are at least as large as the test statistic calculated using the observed data (Edgington, 1987 p. 1).

The nonparametric methods described in this chapter use either $\hat{p}_{ijs}$, the observed cluster risks in the i'th stratum, i=1,...,k, j'th treatment group, j=1,2, and s'th cluster, s=1,..., m_{ij}, or rank transformations of them. Standard textbooks of nonparametric methods (e.g. Conover (1980), Lehmann (1975)) tend to focus on methods using rank transformed data downplaying the importance of tests using the observed data. There are several reasons for this preference for rank transformed data. Tables of the exact statistical significance of tests can only be constructed for tests performed on ranked data. Tests performed using the permutation distribution of the observed cluster risks, however, require sophisticated computer programs to calculate the exact statistical significance. Rank transformed data are also more widely applicable to problems in which outcomes can only be measured on ordinal rather than ratio scales and are less affected by extreme values.

There are three objectives in this section. First, the usual advantages just noted for rank transformed data will be explored to determine if they are likely to occur in stratified cluster randomization trials. Second, several different approximate methods are investigated as less computationally intensive alternatives to the exact tests. These approximate methods use the central limit theorem to derive the asymptotic distribution of the test statistics. Third, algebraic comparisons are drawn between the methods described in this section and methods described elsewhere in the thesis in the hope of gaining additional insight.

2.4.2 Randomization Tests

Several statisticians have strong reservations about using statistical tests derived from randomization theory (e.g. Basu (1980), Royall (1991, Section 5)). Royall (1991), for example, has arg ied that any test statistic derived from a randomization distribution is ancillary. The conditionality principal from likelihood theory would then suggest using probability distributions conditioning on the ancillary statistic. The resulting distribution would be degenerate eliminating the possibility of constructing tests of significance or confidence intervals. Welch (1990), however, has argued that these philosophical concerns can be avoided by viewing the design as fixed and the observations as random.

These ideas can be made more concrete by displaying results from the i'th stratum as in Table 2.3, adapted from Mantel (1963) and Mehta et al (1992). The columns of the table are headed by U unique estimates of risk denoted $\hat{p}_{iu}$, u=1,...,U, from the $m_{i.}$ clusters in the stratum. There are A_{i1u} control and A_{i2u} treated clusters sharing the same estimate of risk in the i'th stratum and m_{i1} control and m_{i2} treated clusters in all. The multivariate hypergeometric distribution (Lehmann, 1975 pp. 381-385) arises as the permutation distribution for the vector $(A_{i1}, \ldots, A_{iU})$ after conditioning on the marginals. Although controversies remain about the value of using conditional versus unconditional distributions in randomization tests the former approach is the more common (Agresti, 1992 Section 1.3), has been used in other exact stratified tests (Mehta et al, 1992) and does lend itself to comparisons among a wide variety of methods. A corollary of conditioning is that the observed risks are treated as fixed and only the numbers within a table are permuted.

In either case the maximum number of permutations is only obtained when all clusters in the i'th stratum have different estimates of risk. The maximum number of possible permutations equals

$$\prod_{i=1}^{k} \binom{m_{i.}}{m_{i1}} . \tag{2.4.1}$$

(Lehmann, 1975 p. 134). The maximum number of permutations is reduced when there are equal number of clusters in each treatment and stratum and two-tailed tests are employed because there will then be mirror image summary statistics (Edginton, 1987, p. 42). The maximum number of permutations is then

$$\frac{1}{2} \binom{2m}{m}^{k} . \tag{2.4.2}$$

A relatively small trial with two strata and 10 clusters per stratum could still require 31,752 separate permutations.

The number of possible permutations decreases as the number of clusters with tied estimates of cluster-specific risk increases. There are several factors unique to cluster randomization trials which will tend to affect the expected number of ties occurring in a stratum. For example as the degree of intracluster correlation approaches one the observed cluster risk will tend to be either near zero or one increasing the probability of ties. The probability of ties will also increase as cluster size and the variability in cluster size decreases. In a cluster randomization trial in ophthalmology, for example, there are only three possible responses depending on the number of eyes per subject having a particular outcome.

The computational burden can also be reduced somewhat by recognizing that many different statistics may be monotonically equivalent (Edgington 1987, Section 3.5). That is, they will yield equal rank ordering of the permutations and so produce identical p-values. For example most tests statistics used to compare two treatments can be written as an estimate of the effect of treatment divided by an estimate of the square root of its variance. Variances of test statistics constructed using randomization theory are derived under the assumption that the null hypothesis holds (e.g. Mantel, 1963) and so are identical for all permutations. Thus valid test statistics for randomization tests usually only involve estimates of the effect of treatment.

An early attempt at resolving the remaining computational burden suggested by Dwass (1957) and also discussed by Edgington (1969) involved using a random sample of permutations to estimate the exact p-value. The precision of the estimated p-value is a function of the sample size, denoted n_p. For a given rejection rate, α, 99% of the estimated significance levels would be within the limits $\alpha \pm 2.58 \sqrt{\alpha\,(1-\alpha)/n_p}$ (Manly, 1991 Section 3.3). Using as many as 1000 permutations would still result in a 99% confidence interval of (0.032, 0.068) when the true rejection rate was 0.05. Thus a deficiency of this approach is that two analyses of the same data set using the same test statistic could produce different inferences.

Mehta, Agresti, Patel and their colleagues have developed computer algorithms which reduce the time needed to calculate exact p-values. These methods have recently been reviewed by Agresti (1992), have been applied to stratified data by Mehta et al (1992) and are available in the computer package StatXact (1991). Estimates provided by Agresti (1992) and Agresti et al (1990) suggest that these algorithms allow calculation of exact p-values within a few minutes on personal computers even with data sets requiring tens of thousands of permutations. Although such tests are now possible less computationally intensive approximations which use central limit theorem results to determine statistical significance are still preferable as the number of clusters gets large.

In addition to their computational complexities exact tests have been criticized for being overly conservative when sample sizes are small (Agresti, 1992, p. 147). This conservatism arises because the data are discrete so small changes in results can cause large jumps in p-values when data are sparse. A solution proposed by Lancaster (1961) calculates one-tailed p-values as one half the observed probability plus the probability of more extreme values. Two-tailed tests can then be calculated by doubling the one-tailed value. Cluster randomization trials have been conducted, however, in which it was impossible to obtain statistically significant results even using Lancaster's mid-p approach.

Black et al (1981), for example, reported on results of a study in which day-care centers were randomly assigned to hand-washing or control programs to determine the effect on diarrhea among the children. There were only two clus-

ters in each treatment group which is too few to guarantee the validity of approximate test statistics. The smallest p-value which can occur using a two-tailed randomization test, however, is 0.33. In fact at least four clusters per treatment group are required in a completely randomized design to allow for the possibility of statistically significant randomization tests. This problem has been previously alluded to by Koepsell et al (1992).

2.4.3 Stratified Rank Tests

Following Conover and Iman (1981) there are several ways in which data from stratified cluster randomization trials can be ranked. A natural approach is to rank clusters separately in each stratum. A stratified extension of Wilcoxon's rank sum test can then be used to determine the effect of treatment.

The literature on such stratified rank tests is somewhat confused. Similar statistics have been independently rediscovered by several authors (Mantel (1963), Shirley (1987), Fry and Lee (1988)) since first discussed by van Elteren (1960) ironically mirroring earlier multiple discoveries of the Wilcoxon rank sum test (Kruskal, 1957). These approaches differ in the choice of weights selected when the number of clusters varies over treatment and stratum. These differences do not appear to be widely appreciated (Fry and Lee (1988), Kuritz, Landis and Koch (1988)).

The test statistics for the different stratified rank tests can be expressed as

$\sum_{i=1}^{k} w_i R_{i1.}$ where w_i is the i'th stratum weight and $R_{i1.}$ is the sum of the ranks for clusters from the control group. van Elteren (1960) proposed two statistics using the weights $1/(m_{i.} + 1)$ and $1/(m_{i1}m_{i2})$. Tests using the first weight have locally optimal properties when there are few strata and many clusters per stratum and treatment group (van Elteren, 1960). Statistics using this weight are also discussed in detail by Lehmann (1975 pp. 135-138). Tests using the latter weight are more appropriate as the number of strata gets large. The test statistic suggested by Shirley (1987) differs from van Elteren's (1960) in selecting identical weights for all strata while Mantel (1963) and Fry and Lee (1988) argue for the use of weights proportional to the inverse of the number of clusters per stratum. All these methods are equivalent when there are the same number of clusters in each stratum and treatment group.

Exact tests for these stratified rank statistics are likely to be nearly as computationally intensive as more powerful tests using the untransformed cluster risks. This difficulty occurs because the high probability of ties would require calculating statistical significance separately for each data set rather than being able to construct tables of p-values as is commonly done for rank transformed data. Even if there were no ties tables of the distribution might not be practical to construct since the distribution of the test statistic is a function of the number of strata in addition to the number of clusters per stratum and treatment group (Lehmann, 1975 p. 135). Separate tables would then have to be constructed for a very large number of possible designs to be of any general use.

If there are at least 30 clusters in the trial claims have been made (Koch, Imrey and Singer et al (1985, p. 239), Mehta et al (1992)) that an asymptotically χ_1^2 test can be used to approximate the exact permutation distribution. The resulting stratified Wilcoxon test statistic denoted

$$\chi^2_{SW} = \frac{\left\{\sum_{i=1}^{k} w_i\left[R_{i1.} - E(\, R_{i1.}\,)\right]\right\}^2}{\sum_{i=1}^{k} w_i^2 \operatorname{var}(\, R_{i1.}\,)}, \qquad (2.4.3)$$

$$= \frac{\left\{\sum_{i=1}^{k} w_i \frac{m_{i1}\, m_{i2}}{m_{i.}} \left[\frac{R_{i2.}}{m_{i2}} - \frac{R_{i1.}}{m_{i1}}\right]\right\}^2}{\sum_{i=1}^{k} w_i^2 \operatorname{var}(\, R_{i1.}\,)} \quad \text{where}$$

$$\operatorname{var}(R_{i1.}) = \sum_{i=1}^{k} \sum_{j=1}^{2} \left[R_{ijs} - \frac{m_{i.}\,(\, m_{i.} + 1\,)}{2}\right]^2 /\, (\, m_{i.} - 1\,)$$

is calculated under the null hypothesis and R_{ijs} is the rank for the ijs'th cluster. The algebraic simplification in the numerator of χ^2_{SW} arises because

$E(R_{ij.}) = \dfrac{m_{ij}\,(\, m_{i.} + 1\,)}{2}$, j=1,2, (Lehmann, 1975 p. 14) and

$$E(R_{i1.}) + E(R_{i2.}) = R_{i1.} + R_{i2.} \qquad (2.4.4)$$

$$= \frac{m_{i.}\,(\, m_{i.} + 1\,)}{2} \quad \text{so}$$

$$E(R_{i1.}) = m_{i1} (R_{i1.} + R_{i2.})/m_{i.}.$$

The presence of ties complicates both the accuracy of the χ_1^2 approximation and calculation of the variance estimator. The approximation is still valid so long as the proportion of tied observations denoted $A_{i.u}/m_{i.}$ in Table 2.3, i=1,...,k, u=1,...,U, are not too near one (Lehmann, 1975 pp. 20-21). The resulting "lumpiness" of the distribution might still increase the asymptotic requirements of the approximation. In the absence of ties the

$$\mathrm{var}(R_{ij.}) = m_{i1}m_{i2}(m_{i.} + 1)/12 \qquad (2.4.5)$$

(Lehmann, 1975 pp. 137).

Two limitations of stratified rank tests are that there does not appear to be any obvious extension which allows inclusion of baseline risk factors nor are any useful summary estimates of the effect of treatment available. A solution to both problems might depend on proportional odds models (McCullagh, 1980). A score test of the null hypothesis that there is no difference between two populations obtained using this ordinal model is equivalent to Wilcoxon's rank sum test (McCullagh (1980, pp. 116-117), McCullagh and Nelder (1989 p. 188)). The odds ratios from proportional odds models thus seems to be a natural nonparametric summary statistic. Additional research is needed to see if score tests from ordinal models adjusted for stratification variables are equivalent to stratified rank tests before attempting additional extensions to include baseline risk factors.

Lehmann (1975, pp. 138-140) argues that methods which employ separate rankings for each stratum can lack power when there are few clusters per stratum and suggests ranking data from all strata after adjusting for between stratum differences. The Wilcoxon rank sum test can then be performed on the aligned ranks. Alternatively standard parametric methods such as analysis of variance can be used after ranking all clusters in the study even when original scores are not aligned (Conover and Iman, 1981). The central limit theorem ensures asymptotic normality will hold under the null hypothesis for such test statistics.

2.4.4 Stratified Randomization Tests using the Observed Cluster Risks

Some gain in power can be obtained using randomization tests with the observed cluster risks rather than their ranks. The asymptotic gain in power might, however, be small. Hodges and Lehmann (1962) have shown that the asymptotic efficiency of the van Elteren's (1963) test statistic constructed using the optimal weights is equal to

$$\frac{3}{\pi} \sum_{i=1}^{k} \frac{m_{i1}\; m_{i2}}{m_{i.} + 1} \quad \sum_{i=1}^{k} - \frac{m_{i1}\; m_{i2}}{m_{i.}}, \tag{2.4.6}$$

relative to a stratified t-test when the data are normally distributed. Thus, in a trial with m_{ij}=m clusters per stratum and treatment group the asymptotic relative efficiency is never less than 63 percent and can be as large as 96 percent. Asymptotic comparisons with other distributions can be constructed using the formulae in Puri (1965). Results may, however, be decidedly different in finite samples (Lehmann, 1975 p. 81).

Mantel (1979) offers additional arguments against using rank transformed data for correlated binary outcome data. He points out that observed cluster risk is a more reasonable measure and that rank transformations can be inappropriate when cluster sizes are highly variable because small clusters will tend to have more extreme ranks than larger clusters.

There are two types of randomization tests which use the observed cluster risks. The first type of test is performed at the cluster level ignoring any variation in cluster size and can not be extended to include individual-level covariates. The stratified rank tests can be derived as the special case of these tests which arise when the data are rank transformed. The second type of test employs summary statistics which take varying cluster sizes into account. These methods are more computationally intensive and can be extended to include both cluster-level and individual-level baseline risk factors.

2.4.4.1 Type 1: Cluster-Level Randomization Tests

The numerator of the extended Mantel-Haenszel test can be used as a test statistic to demonstrate the first type of randomization test. Birch (1965) derived the following equivalent version of this statistic

$$\left\{ \sum_{i=1}^{k} \frac{m_{i1}\, m_{i2}}{m_{i.}} (\bar{p}_{i2} - \bar{p}_{i1}) \right\}^2 \qquad (2.4.7)$$

where $\bar{p}_{ij}$ is the average risk for clusters in the i'th stratum and j'th treatment group. The exact statistical significance is again determined as the proportion of tests statistics calculated using the permuted data which are at least as large as the test statistic evaluated using the observed data.

A one degree of freedom χ^2 approximation to the exact permutation distribution is given by

$$\chi^2_{EMH} = \frac{\left\{ \sum_{i=1}^{k} \frac{m_{i1}\, m_{i2}}{m_{i.}} (\bar{p}_{i2} - \bar{p}_{i1}) \right\}^2}{\sum_{i=1}^{k} \frac{m_{i1}\, m_{i2}}{m_{i.}} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{(\hat{p}_{ijs} - \bar{p}_i)^2}{m_{i.} - 1}} \qquad (2.4.8)$$

where $\bar{p}_i = \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \hat{p}_{ijs}/m_{i.}$. The variance estimate is obtained using the multivariate hypergeometric distribution.

As mentioned in Section 2.4.2 exact p-values for stratified permutation tests can be obtained using the computer package StatXact (1991). This computer package uses the sum of the observed cluster risks for one of the treatment groups summed over all strata as the test statistic. The extended Mantel-Haenszel statistic is the square of a standardized version of this test statistic. It is constructed by subtracting off the expected value of the stratum-specific sum and then dividing by the square root of its variance. The expected value and variance are the same for all permutations since they are calculated after conditioning in each stratum on the number of clusters per treatment group and the observed cluster risks. This algebraic relationship implies that StatXact can be used to obtain exact p-values for the extended Mantel-Haenszel test statistic.

The proposed randomization test simplifies to the stratified rank test discussed by Shirley (1987) when the observed cluster risks in a stratum are rank transformed. Alternative tests which employ other weighting schemes could be developed. Such weights might be helpful if there is considerable variation in the number of clusters per stratum.

Note that χ^2_{EMH} differs from the stratified t-test, described in Section 2.3, only in how the stratum-specific variance estimates are calculated. Variance estimates in the approximate randomization test are calculated assuming that the null hypothesis holds while stratum-specific pooled estimates of variance are used in the stratified t-test.

An important specia' case of χ^2_{EMH} occurs when all clusters are the same size, i.e. $n_{ijs} = n$. Then the extended Mantel-Haenszel χ^2 test statistic equals Cochrans' (1954) version of the classical Mantel-Haenszel χ^2 test statistic (Mantel and Haenszel, 1959) divided by the variance inflation factor

$$1 + (n-1)\bar{\rho} \quad \text{where} \qquad (2.4.9)$$

$$\bar{\rho} = \sum_{i=1}^{k} \frac{\dfrac{m_{i1}\, m_{i2}}{m_{i.}}\, \hat{p}_i(1-\hat{p}_i)}{\displaystyle\sum_{i=1}^{k} \frac{m_{i1}\, m_{i2}}{m_{i.}}\, \hat{p}_i(1-\hat{p}_i)}\, \hat{\rho}_i,$$

$\hat{p}_i = \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \hat{p}_{ijs} \,/\, nm_{i.}$ and $\hat{\rho}_i$ is the moment estimator of intracluster correlation for the i'th stratum.

The extended Mantel-Haenszel test can be adapted to adjust for variations in cluster size. This adaptation requires the stratified design to be further stratified by cluster size in each of the original strata. This approach has been suggested for completely randomized designs by Paul and Mantel (1989) and independently by Zucker and Wittes (1992). That is, if there were F_i different cluster sizes in the i'th original stratum, i=1,...k, the data would be divided into $F. = \sum_{i=1}^{k} F_i$ strata containing clusters of identical size. The extended Mantel-Haenszel statistic is then calculated on these F. strata. In trials with few clusters there will now likely be strata which would include only control or only treated clusters. Such strata would not contribute to the test statistic reducing its precision possibly outweighing any advantage obtained by only comparing estimates of risk for clusters of equivalent size.

2.4.4.2 Type 2: Individual-Level Randomization Tests

The permutation tests which have been discussed so far are all cluster-level analyses. An alternative approach, favoured by Crump et al (1991), considers the permutation disuibution of summary statistics which are functions of both cluster-specific risk and cluster sizes. Examples of such statistics for stratified cluster randomization trials include the numerator of the classical Mantel-Haenszel's χ^2 statistic (Mantel and Haenszel, 1959), or estimates of summary statistics like the odds ratio. The exact statistical significance for tests using such statistics is calculated the same way as for the other randomization tests.

It is helpful, however, to use log transformations of parameters like the odds ratio or relative risk so that the distribution of the test statistic is symmetric under the null hypothesis simplifying calculation of statistical significance. One-tailed tests of the null hypothesis that the log odds ratio is zero could then be constructed by calculating the summary statistic for each permutation of the data and determining the proportion of times that statistics calculated using the permuted data are as large as the statistic calculated using the original data. Two-tailed p-values are defined to be twice this probability. Alternatively two-tailed tests could be directly calculated by squaring estimates of the log odds ratio (Edgington, 1987, p. 35).

Greater precision can be obtained using summary statistics which also attempt to adjust for intracluster correlation. Permutation tests, for example, could be constructed using estimates of odds-ratios from iterative procedures such as Liang and Zeger's (1986) generalized estimating equations. The permutation test is then asymptotically equivalent to the bootstrap method proposed for the analysis of correlated binary outcome data and can be extended to incorporate baseline risk factors (Moulton and Zeger (1989), Welch (1990)). Moulton and Zeger (1989) have also described an asymptotic relationship which exists between the bootstrap approach of estimating variance and the robust variance estimator proposed by Liang and Zeger (1986). The application of bootstrap methods to cluster randomization trials in particular have been discussed by Rao and Colin (1991).

An extension of a jackknife test proposed by Gladen (1979) can be used as an approximation of the second type of permutation test (Crump et al. 1991). Let the pseudovalue (Hinkley, 1983) for the i'th stratum, j'th treatment group and s'th cluster be denoted

$$\hat{p}^{J}_{ijs} = m_{ij}\,\hat{p}_{ij} - (m_{ij} - 1)\hat{p}_{-ijs} \qquad (2.4.10)$$

where $\hat{p}_{ij} = \sum_{s=1}^{m_{ij}} n_{ijs}\hat{p}_{ijs} \,/\, n_{ij.}$ and $\hat{p}_{-ijs}$ is the crude risk in the i'th stratum and j'th treatment group calculated after eliminating the s'th cluster and can be expressed as

$$\sum_{v\neq s} n_{ijv}\hat{p}_{ijv} \,/\, (\, n_{ij.} - n_{ijs} \,). \qquad (2.4.11)$$

The jackknife estimator for the i'th stratum and j'th treatment group is then

$p_{ij}^{J} = \sum_{s=1}^{m_{ij}} \hat{p}_{ijs}^{J} \,/\, m_{ij}$ and its variance denoted

$$\hat{\mathrm{var}}(\, \hat{p}_{ij}^{J} \,) = \sum_{s=1}^{m_{ij}} (\, \hat{p}_{ijs}^{J} - \hat{p}_{ij}^{J} \,)^2/m_{ij}(\, m_{ij}-1 \,). \qquad (2.4.12)$$

A jackknife estimator of the weighted sum of stratum-specific treatment differences is then given by $\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i..}}(\hat{p}_{i1}^{J} - \hat{p}_{i2}^{J})$ and its variance can be estimated as

$$\sum_{i=1}^{k} \left\{ \frac{n_{i1.}\, n_{i2.}}{n_{i..}} \right\}^2 \sum_{j=1}^{2} \hat{\mathrm{var}}(\hat{p}_{ij}^{J}\,). \qquad (2.4.13)$$

Under the null hypothesis the resulting test denoted

$$\chi_J^2 = \frac{\left[\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i..}}(\hat{p}_{i1}^{J} - \hat{p}_{i2}^{J}) \right]^2}{\sum_{i=1}^{k} \left\{ \frac{n_{i1.}\, n_{i2.}}{n_{i..}} \right\}^2 \sum_{j=1}^{2} \hat{\mathrm{var}}(\hat{p}_{ij}^{J}\,)} \qquad (2.4.14)$$

is asymptotically χ_1^2.

An alternative and simpler test statistic, denoted χ_{JC}^2 is obtained by using the square of a weighted sum of the crude stratum-specific treatment differences, $\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i..}}(\hat{p}_{i1} - \hat{p}_{i2})$, in the numerator. This alternative test statistic is valid because both summary estimates of treatment differences have the same expected value and since $\sum_{i=1}^{k} \left\{ \frac{n_{i1.}\, n_{i2.}}{n_{i..}} \right\}^2 \sum_{j=1}^{2} \hat{\mathrm{var}}(\hat{p}_{ij}^{J}\,)$ is a valid estimate of variance for either statistic (Gladen, 1979). The two test statistics are identical when n_{ijs} is constant for all clusters in the same stratum and treatment group and can be seen to differ from stratified t-tests on the cluster means only in using cluster and treatment specific estimates of variance rather than the more precise pooled estimate of variance. A related simplification occurring with completely randomized designs was previously noted by Rao and Colin (1991).

The alternative test statistic, χ^2_{JC} is preferable not only because it is easier to calculate but also because of the possibility of aberrant estimates of treatment differences possible with χ^2_J. This possibility arises because p^J_{ij} is not constrained to be between 0 and 1 when there are few clusters in each treatment group and stratum, and when cluster sizes are highly variable (Gladen, 1979).

More complicated jackknife estimators could also have been developed. For example estimators could have been calculated as an average of all $m_{.}$ pseudovalues by consecutively eliminating single clusters from all strata and treatment groups. Alternatively jackknifed odds ratio or log odds ratio estimators could have been proposed. The primary advantage of the relatively simple method which has been proposed is that it simplifies to approaches previously described for completely randomized cluster randomization trials (Crump et al (1991), Gladen (1979)).

Table 2.3
Frequency of Unique Cluster-Specific Estimates of Risk in the i'th Stratum

Treatment Group	$\hat{p}_{i1}$	...	$\hat{p}_{iU}$	Total
Control	A_{i11}	...	A_{i1U}	m_{i1}
Treated	A_{i21}	...	A_{i2U}	m_{i2}
Total	$A_{i.1}$	...	$A_{i.U}$	$m_{i.}$

2.5 Simple Adjustments of Mantel-Haenszel and Stratified Woolf Test Statistics

2.5.1 Introduction

Statistical methods for testing the effect of treatment in cluster randomization trials can be derived as relatively simple extensions of methods originally developed for binomially distributed data. Examples of such methods as applied to stratified cluster randomization trials are described by Donald and Donner (1987). Such test statistics share several properties such as being noniterative and simplifying to standard methods when estimates of intracluster correlation equal zero.

Two test statistics which could be used in stratified cluster randomization trials are distinguished by using either weighted sums of stratum-specific risk differences or log odds ratios. The test statistic using risk differences was originally proposed by Donald and Donner (1987) and is an extension of the Mantel-Haenszel-Cochran statistic while the latter method is an extension of the stratified Woolf (1955) estimator previously adapted to matched-pairs designs (Donner and Klar, 1993a). Test statistics can also be distinguished by whether or not stratum specific estimates of the effect of treatment are constructed using asymptotically optimal weights.

2.5.2 Methods of Inference using Non-Optimal Weights

Let y_{ijst} scored one or zero denote the response of the t'th subject, t=1,..., n_{ijs}, from the s'th cluster, s=1,..., m_{ij} , j'th treatment group, j=1,2 and i'th stratum, i=1,...,k. Also, assume that

$$\begin{aligned} E(\ y_{ijst}\ ) &= p_{ij}, \qquad\qquad\qquad\qquad (2.5.1)\\ \mathrm{var}(\ y_{ijst}\ ) &= p_{ij}q_{ij} \quad \text{where } q_{ij} = 1 - p_{ij} \text{ and}\\ \mathrm{corr}(\ y_{ijsu}, y_{ijsv}\ ) &= \rho_{ijs(uv)}, \quad u \neq v. \end{aligned}$$

Then the observed risk for the ijs'th cluster,

$$\hat{p}_{ijs} = \sum_{t=1}^{n_{ijs}} y_{ijst} \ / \ n_{ijs} \qquad (2.5.2)$$

and an estimate of the effect of treatment in the i'th stratum can be expressed as

$$\hat{p}_{i2} - \hat{p}_{i1} = \sum_{s=1}^{m_{i2}} \left[n_{i2s}\ \hat{p}_{i2s}/n_{i2.} \right] - \sum_{s=1}^{m_{i1}} \left[n_{i1s}\ \hat{p}_{i1s}/n_{i1.} \right]. \qquad (2.5.3)$$

Note that (2.5.3) is an unbiased estimate of the effect of treatment since

$$\begin{aligned} E(\ \hat{p}_{ij}\ ) &= \sum_{s=1}^{m_{ij}} n_{ijs}\ E(\ \hat{p}_{ijs}\ )/n_{ij.} \qquad (2.5.4)\\ &= \sum_{s=1}^{m_{ij}} n_{ijs}\ p_{ij}/n_{ij.}\\ &= p_{ij}. \end{aligned}$$

A limitation of (2.5.3) is that it lacks precision since $\hat{p}_{i2} - \hat{p}_{i1}$ does not use asymptotically optimal weights except when subjects' responses within a cluster are uncorrelated or when there are the same number of subjects in each cluster in the i'th stratum, i=1,...,k. More precise estimators are described in the next section.

The variance of $\hat{p}_{i2} - \hat{p}_{i1}$, denoted by var($\hat{p}_{i2} - \hat{p}_{i1}$), can be expressed as

$$\sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \left[\frac{n_{ijs}}{n_{ij.}} \right]^2 \mathrm{var}(\ \hat{p}_{ijs}\ ) \qquad (2.5.5)$$

$$= \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \left[\frac{n_{ijs}}{n_{ij.}} \right]^2 \left[\sum_{t=1}^{n_{ijs}} \frac{\mathrm{var}(\ y_{ijst}\ )}{n_{ijs}^2} + \sum_{u=1}^{n_{ijs}} \sum_{v \neq u} \frac{\mathrm{cov}(\ y_{ijsu}, y_{ijsv}\ )}{n_{ijs}^2} \right]$$

$$= \sum_{j=1}^{2} \left[\frac{p_{ij}q_{ij}}{n_{ij.}} \sum_{s=1}^{m_{ij}} \frac{n_{ijs}}{n_{ij.}} [\ 1 + (n_{ijs}-1)\bar{\rho}_{ijs}\ ] \right]$$

$$= \left[\frac{p_{i1}q_{i1}}{n_{i1.}} + \frac{p_{i2}q_{i2}}{n_{i2.}} \right] C_i$$

$$\text{where } C_i = \sum_{j=1}^{2} \frac{p_{ij}q_{ij}/n_{ij.}}{\sum_{j=1}^{2}(\ p_{ij}q_{ij}/n_{ij.}\ )} \sum_{s=1}^{m_{ij}} \frac{n_{ijs}}{n_{ij.}} [\ 1 + (n_{ijs}-1)\bar{\rho}_{ijs}\ ],$$

and $\bar{\rho}_{ijs}$ is the average intracluster correlation among the n_{ijs} subjects in the ijs'th cluster. Thus the variance of $\hat{p}_{i2} - \hat{p}_{i1}$ equals the variance derived under the assumption that responses of cluster members are uncorrelated, multiplied by a correction factor C_i to adjust for the effect of clustering. The correction factor is a weighted average of the variance inflation factors from each cluster allowing for the possibility that the average degree of intracluster correlation may vary from cluster to cluster.

There are two important special cases which need to be considered. First, suppose that all clusters within a stratum have $n_{ijs}=n_i$ subjects and that there are $m_{ij}=m_i$ clusters in the j'th treatment group of the i'th stratum. Then

$$\mathrm{var}(\,\hat{p}_{i2}-\hat{p}_{i1}\,)=\left\{\frac{p_{i1}q_{i1}}{n_i m_i}+\frac{p_{i2}q_{i2}}{n_i m_i}\right\}[\,1+(n_i-1)\bar{\rho}_{i..}\,] \qquad (2.5.6)$$

$$\text{where } \bar{\rho}_{i..}=\sum_{j=1}^{2}\sum_{s=1}^{m_i}\left[\frac{p_{ij}q_{ij}}{\sum\limits_{j=1}^{2}p_{ij}q_{ij}}\;\frac{1}{m_i}\;\bar{\rho}_{ijs}\right].$$

Therefore balance assures that valid inferences about treatment effects in the i'th stratum can be constructed using simple adjustment techniques which do not make any particular assumptions about intracluster correlation. For example the average intracluster correlation in the i'th stratum could then be estimated using the ANOVA approach discussed by Donner (1985a) and described in Section 2.3 of this thesis.

In general some imbalance in cluster size is expected. Then the $\mathrm{var}(\,\hat{p}_{i2}-\hat{p}_{i1}\,)$ can only be estimated using the robust estimates described in Sections 2.6 and 2.8 or by making additional assumptions about intracluster correlation. The simplest assumption is that $\bar{\rho}_{ijs}=\bar{\rho}_{ij.}$ in the i'th stratum and j'th treatment group. Separate ANOVA estimators for each treatment group would then have to be used to estimate the variance in the i'th stratum. Such estimators would likely be quite imprecise when there were few clusters in the j'th treatment group, j=1,2, and the i'th stratum, i=1,...,k. Some simplification occurs under the null hypothesis that $p_{i1} = p_{i2} = p_i$ when treatment has been randomly assigned. In this case $\bar{\rho}_{i1.} = \bar{\rho}_{i2.}$ and the ANOVA approach could again be used to estimate the common intracluster correlation in the i'th stratum.

The second special case arises upon extending this assumption and supposing that all responses of individuals within a cluster are equally correlated. This assumption usually arises indirectly as a consequence of assuming that the binary responses of subjects from the ijs'th cluster, y_{ijst}, follows a Bernoulli distribution after conditioning on their expected value, p_{ijs}. Furthermore it is supposed that p_{ijs} varies at random within each treatment group and stratum as a common multiple of the variance of a Bernoulli distributed random variable (e.g. Donald and Donner (1987), Moore and Tsiatis (1991), Williams (1982)). The beta-binomial distribution also arises as a consequence of these assumptions when it is further assumed that the between-cluster variability of p_{ijs} can be described using a beta distribution.

Let $E(\, p_{ijs}\,) = p_{ij}$ and let

$$\text{var}(\, p_{ijs}\,) = \sigma^2\, p_{ij}q_{ij} \quad \text{where } 0 \leq \sigma^2 \leq 1. \qquad (2.5.7)$$

Then for $u \neq v$

$$\text{cov}(\, y_{ijsu}\,,\, y_{ijsv}) = E(\, y_{ijsu}\, y_{ijsv}\,) - E(\, y_{ijsu}\,)\, E(\, y_{ijsv}\,)$$

$$= E(E(\ y_{ijsu},y_{ijsv} \mid p_{ijs}\ )) - E[E(\ y_{ijsu} \mid p_{ijs}\ )]E[E(\ y_{ijsv} \mid p_{ijs}\ )]$$

$$= E(\ p_{ijs}^2\ ) - [E(\ p_{ijs}\ )]^2$$

$$= \sigma^2\ p_{ij}q_{ij}$$

since y_{ijsu} and y_{ijsv} are conditionally independent. Furthermore

$$\mathrm{var}(\ y_{ijst}\ ) = \mathrm{var}[E(\ y_{ijst} \mid p_{ijs}\ )] + E[\mathrm{var}(\ y_{ijst}\ )] \qquad (2.5.8)$$

$$= \mathrm{var}[\ p_{ijs}\ ] + E[\ p_{ijs}(1-p_{ijs})\ ]$$

$$= p_{ij}q_{ij}$$

using theorem 7 of Mood, Graybill, and Boes (1974, p. 159). Therefore the $\mathrm{corr}(\ y_{ijsu},y_{ijsv}\ ) = \sigma^2$, $u \neq v$.

Results from (2.5.5) suggest that this approach makes more assumptions than are strictly needed for most inference problems. It is sufficient to assume only that the average correlation within all clusters is common across strata and treatment groups. Furthermore (2.5.6) implies that in completely balanced designs no assumptions about differences in the $\bar{\rho}_{ij.}$ are required. For these reasons the methods used to test the effect of treatment in this section will all assume only that $\bar{\rho}_{ij}=\rho$ for all $i=1,\ldots,k$ and $j=1,2$.

There are several ways in which this assumption can be relaxed to allow ρ to vary by strata or across treatment groups. The simplest approach is to use the methods described in this section but using separate estimates of ρ as needed. A second approach, proposed by Rao and Scott (1992), adapts the theory of ratio estimation from survey sampling to cluster randomization trials. Their approach is quite flexible and can used to make inferences in a wide variety of design settings. It is de... ...ed in Section 2.6. A third possibility is to use the generalized estimating equation approach described by Liang and Zeger (1986). This last approach is perhaps the most general since it incorporates all the generalized linear models (McCullugh and Nelder, 1989). This approach was introduced in Chapter 1 and is described in detail in Section 2.8 of this chapter. All three approaches require large numbers of clusters in each stratum and treatment group, i.e. are asymptotically valid as the numbers of clusters in each cell becomes large.

The assumption that $\bar{\rho}_{ij} = \rho$ allows (2.5.5) to be expressed as

$$\text{var}(\ \hat{p}_{i2} - \hat{p}_{i1}\ ) = \sum_{j=1}^{2} \frac{p_{ij}q_{ij}}{n_{ij.}} B_{ij} \quad \text{where} \tag{2.5.9}$$

$$B_{ij} = \sum_{s=1}^{m_{ij}} \frac{n_{ijs}}{n_{ij.}} [\ 1 + (n_{ijs}-1)\rho\ ].$$

The variance can be estimated using $\hat{p}_{ij}$ and any consistent estimate of ρ , e.g. the ANOVA estimator derived in Section 2.3. Then

$$\hat{\text{var}}(\ \hat{p}_{i2} - \hat{p}_{i1}\ ) = \sum_{j=1}^{2} \frac{\hat{p}_{ij}\hat{q}_{ij}}{n_{ij.}} \hat{B}_{ij} \quad \text{where} \tag{2.5.10}$$

$$\hat{B}_{ij} = \sum_{s=1}^{m_{ij}} \frac{n_{ijs}}{n_{ij.}} [\ 1 + (n_{ijs}-1)\hat{\rho}\ ].$$

If stratum-specific weights inversely proportional to $\hat{var}(\hat{p}_{i2} - \hat{p}_{i1})$ under the null hypothesis are used to obtain a summary statistic for the effect of treatment then an extension of Cochran's (1954) statistic, denoted χ^2_{CA}, can be used to test the effect of treatment in stratified cluster randomization trials. The statistic is expressed as

$$\chi^2_{CA} = \frac{\left[\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i1.}\hat{B}_{i2} + n_{i2.}\hat{B}_{i1}} (\hat{p}_{i2} - \hat{p}_{i1})\right]^2}{\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i1.}\hat{B}_{i2} + n_{i2.}\hat{B}_{i1}} \hat{p}_i\hat{q}_i} \qquad (2.5.11)$$

where $\hat{p}_i = \sum_{s=1}^{m_{ij}} \sum_{j=1}^{2} n_{ijs}\hat{p}_{ijs} / n_{i..}$. Note that χ^2_{CA}, like all the other test statistics described in this section, is asymptotically χ^2_1 under the null hypothesis. An asymptotically equivalent statistic derived by Donald and Donner (1987) is an extension of the Mantel-Haenszel χ^2 test statistic (Mantel and Haenszel, 1959). This statistic denoted χ^2_{MHA} is expressed as

$$\chi^2_{MHA} = \frac{\left[\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i1.}\hat{B}_{i2} + n_{i2.}\hat{B}_{i1}} (\hat{p}_{i2} - \hat{p}_{i1})\right]^2}{\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i1.}\hat{B}_{i2} + n_{i2.}\hat{B}_{i1} - 1} \hat{p}_i\hat{q}_i} \qquad (2.5.12)$$

Some insight into why these statistics are reasonable can be gained by considering three special cases: $\hat{\rho} = 0$, $\hat{\rho} = 1$, and the simplification which occurs when there are $n_{ijs} = n$ subjects in each cluster. If $\hat{\rho} = 0$ χ^2_{MHA} reduces to the standard Mantel-Haenszel χ^2 statistic

$$\chi^2_{MH} = \frac{\left[\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i..}} \left(\hat{p}_{i2} - \hat{p}_{i1} \right)\right]^2}{\sum_{i=1}^{k} \frac{n_{i1.}\, n_{i2.}}{n_{i..}-1} \hat{p}_i \hat{q}_i} . \qquad (2.5.13)$$

At the other extreme χ^2_{MHA} becomes equal to

$$\frac{\left[\sum_{i=1}^{k} \frac{m_{i1}\, m_{i2}}{m_{i.}} \left(\hat{p}_{i2} - \hat{p}_{i1} \right)\right]^2}{\sum_{i=1}^{k} \frac{m_{i1}\, m_{i2}}{m_{i.}-1} \hat{p}_i \hat{q}_i} \qquad (2.5.14)$$

at least when there is no variability in cluster size among clusters in the i'th stratum (i.e. $n_{ijs} = n_i$). Equation (2.5.14) is the standard Mantel-Haenszel χ^2 test statistic, but using the number of clusters rather than the number of subjects to calculate the weights. This is appropriate since when $\hat{\rho} = 1$ each cluster effectively contributes only a single independent response. Finally if all clusters have exactly n subjects in each cluster $\chi^2_{MHA} \approx \chi^2_{MH} / [1 + (n-1)\hat{\rho}\ ]$.

Day and Byar (1979) have shown that the Mantel-Haenszel test statistic can be derived as a score test from logistic regression in the absence of clustering. The statistic proposed by Donald and Donner (1987) is therefore an extension of this score test adjusted for clustering.

An alternative approach based on extending a Wald statistic proposed by Woolf (1955) is more easily adapted to either hypothesis testing or confidence interval construction. The classical Woolf's (1955) estimator of the log of the odds ratio of the effect of treatment is given by

$$\hat{\gamma}_W = \sum_{i=1}^{k} \hat{W}_i\hat{\gamma}_i \,/\, \sum_{i=1}^{k}\hat{W}_i \text{ where} \qquad (2.5.15)$$

$$\hat{\gamma}_i = \log_e(\, \hat{p}_{i2}\hat{q}_{i1} \,/\, \hat{p}_{i1}\hat{q}_{i2} \,) \text{ and}$$

$$\hat{W}_i = (n_{i1.}\hat{p}_{i1}\hat{q}_{i1})^{-1} + (n_{i2.}\hat{p}_{i2}\hat{q}_{i2})^{-1}.$$

This estimator is consistent for correlated binary outcome data but lacks precision, at least asymptotically.

The imprecision arises because the weights used by Woolf (1955) are only optimal when $\rho = 0$ or when there are the same numbers of subjects in each cluster. Optimal weights are proportional to the inverse of a statistic's variance (Rao, 1975, p. 308, Example 2.2). Thus optimal weights for the statistic $\hat{\gamma}_i$ are equal to the inverse of

$$\text{var}(\,\hat{\gamma}_i\,) \approx \frac{B_{i1}}{n_{i1.}p_{i1}q_{i1}} + \frac{B_{i2}}{n_{i2.}p_{i2}q_{i2}} \qquad (2.5.16)$$

which is obtained using the delta method (Agresti 1990, Chapter 12).

A similar approach was previously used by Donner and Hauck (1988) to obtain odds ratio estimates from case-control studies of familial aggregation. Such designs are the observational equivalent of pair-matched cluster randomization trials.

A modified version of the Woolf estimator is then denoted by

$$\hat{\gamma}_{WW} = \sum_{i=1}^{k} \hat{W}_{iW}\hat{\gamma}_i \,/\, \sum_{i=1}^{k}\hat{W}_{iW} \text{ where} \tag{2.5.17}$$

$$\hat{W}_{iW}^{-1} = \left[\frac{\hat{B}_{i1}}{n_{i1.}\hat{p}_{i1}\hat{q}_{i1}} + \frac{\hat{B}_{i2}}{n_{i2.}\hat{p}_{i2}\hat{q}_{i2}}\right] \text{ and so the}$$

$\hat{var}(\, \hat{\gamma}_{WW} \,) = 1 \,/\, \hat{W}_{.}$

Therefore a weighted version of the classical Woolf test is given by

$$\chi^2_{WW} = \left[\sum_{i=1}^{k}\hat{W}_{iW}\,\hat{\gamma}_i\right]^2 \hat{W}_{.} \tag{2.5.18}$$

Simplifications previously noted for the adjusted Mantel-Haenszel χ^2 test (Donald and Donner, 1987) when $\hat{\rho} = 0$, when $\hat{\rho} = 1$, and when there are $n_{ijs} = n$ subjects in each cluster also occur for the weighted Woolf test. That is the test statistic reduces to the classical Woolf test when $\hat{\rho} = 0$, and to the classical Woolf test divided by the variance inflation factor $1 + (n-1)\hat{\rho}$ when $n_{ijs} = n$. Finally when $\hat{\rho} = 1$ and $n_{ijs} = n_i$ the weights reduce to

$$\left[\frac{1}{m_{i1}\hat{p}_{i1}\hat{q}_{i1}} + \frac{1}{m_{i2}\hat{p}_{i2}\hat{q}_{i2}}\right]^{-1}. \tag{2.5.19}$$

2.5.3 Methods of Inference using Asymptotically Optimal Weights

The test statistics described in the previous section were constructed using simple estimates of p_{ij} calculated as $\hat{p}_{ij} = \sum_{s=1}^{m_{ij}} n_{ijs}\hat{p}_{ijs} \,/\, n_{ij.}$. Such estimates are only optimal when there is no variability in cluster size within each stratum and treatment group, i.e. $n_{ijs} = n_{ij}$, or when $\rho = 0$. Optimally weighted estimates are given by

$$\hat{P}_{ij} = \sum_{s=1}^{m_{ij}} \frac{\hat{f}_{ijs}}{\hat{f}_{ij.}} \hat{p}_{ijs} \text{ where} \tag{2.5.20}$$

$$\hat{f}_{ijs} = n_{ijs} \,/\, [\ 1 + (n_{ijs}-1)\hat{\rho}\ ].$$

The weights $\hat{f}_{ijs}$ are consistent estimates of $n_{ijs} / [\ 1 + (n_{ijs}-1)\rho]$ which are inversely proportional to $\mathrm{var}(\hat{p}_{ijs})$.

The greater precision of $\hat{P}_{ij}$ relative to $\hat{p}_{ij}$ can be demonstrated by noting that

$$\mathrm{var}(\ \hat{P}_{ij}\ ) = \frac{p_{ij}q_{ij}}{n_{ij.}} \; \frac{n_{ij.}}{\sum_{s=1}^{m_{ij}} n_{ijs} \,/\, [1+(n_{ijs}-1)\rho]}$$

which is the standard binomial variance multiplied by the weighted harmonic mean of the variance inflation factors $1+(n_{ijs}-1)\rho$. The variance of $\hat{p}_{ij}$, however, can be expressed as the standard binomial variance multiplied by the weighted arithmetic mean of the variance inflation factors. A corollary of Hardy et al's (1991, p. 17) Theorem 9 is that weighted harmonic means are less than or equal to weighted arithmetic means with equality holding only if terms being averaged are equal. Therefore the $\mathrm{var}(\hat{P}_{ij}) \leq \mathrm{var}(\hat{p}_{ij})$ with equality holding if there are the same number of subjects in each cluster or when $\rho = 0$. This relationship will also hold for estimates of each variance.

Under the null hypothesis the

$$\mathrm{var}(\ \hat{P}_{i2} - \hat{P}_{i1}\ ) = p_i q_i \left[\frac{1}{r_{i1.}} + \frac{1}{r_{i2.}} \right] \qquad (2.5.21)$$

and so following the same arguments used to derive the adjusted Mantel-Haenszel statistic (Donald and Donner, 1987) an asymptotically more precise test statistic is given by

$$\chi^2_{MHO} = \frac{\left[\sum_{i=1}^{k} \frac{\hat{r}_{i1.}\ \hat{r}_{i2.}}{\hat{r}_{i..}} (\ \hat{P}_{i2} - \hat{P}_{i1}\ ) \right]^2}{\sum_{i=1}^{k} \frac{\hat{r}_{i1.}\ \hat{r}_{i2.}}{\hat{r}_{i..}} \hat{P}_i \hat{Q}_i} \text{ where} \qquad (2.5.22)$$

$$\hat{P}_i = \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{\hat{r}_{ijs}}{\hat{r}_{i..}} \hat{p}_{ijs} \quad \text{and} \quad \hat{Q}_i = 1 - \hat{P}_i.$$

The optimally weighted Woolf estimator denoted

$$\hat{\gamma}_{OW} = \sum_{i=1}^{k} \hat{W}_{iO}\hat{\gamma}_{iO} \,/\, \sum_{i=1}^{k} \hat{W}_{iO} \text{ where} \qquad (2.5.23)$$

$$\hat{\gamma}_{iO} = \log_e(\ \hat{P}_{i2}\ \hat{Q}_{i1}/\hat{Q}_{i2}\hat{P}_{i1}\ ) \text{ and}$$

$$\hat{W}_{iO}^{-1} = \left[\frac{1}{\hat{r}_{i1.}\ \hat{P}_{i1}\ \hat{Q}_{i1}} + \frac{1}{\hat{r}_{i2.}\ \hat{P}_{i2}\ \hat{Q}_{i2}} \right] \text{ and so the}$$

$$\hat{\mathrm{var}}(\hat{\gamma}_{OW}) \approx 1 / \hat{W}_{.O}.$$

Therefore an asymptotically optimal weighted version of the classical Woolf test is given by

$$\chi^2_{WO} = \left[\sum_{i=1}^{k} \hat{W}_{iO}\, \hat{\gamma}_{iO}\right]^2 \hat{W}_{.O}. \qquad (2.5.24)$$

Both of these optimally weighted test statistics are equal to χ^2_{CA} and χ^2_{WW} respectively when $\rho = 0$ and when there are the same number of subjects in each cluster.

2.6 Adaptations of Methods Developed for Complex Sample Surveys

2.6.1 Introduction

In Chapter 1 many early ideas concerning the design and analysis of cluster randomization trials were traced back to earlier work on cluster sampling. In spite of these common beginnings there are many differences between surveys and experiments so methods originally developed for complex surveys may not always be applicable in randomized trials.

The most important difference is the use of random allocation in experimental studies. Random allocation assures equality of intracluster correlation parameters across treatment groups under the null hypothesis and eliminates the possibility of confounding. Survey samples, on the other hand, are characterized by the selection of a sample from a specified population using a random mechanism. The focus here is on estimation of unknown parameters rather than tests of significance.

The analysis of data from national surveys is also often complicated by having to account for the use of stratified, multi-stage cluster sampling. Such complicated designs are required for three reasons. First, no national sampling frame may exist (Korn and Graubard, 1991) eliminating the possibility of a simple random sample. Second, travel costs and time to complete the study are reduced when subjects are selected by household within neighbourhoods. Third, stratified samples allow surveys to be designed to obtain precise estimates of health behaviours at both the national and local level. This requirement for obtaining precise estimation at both national and local levels increases the overall sample size. Therefore most complex surveys sample large numbers of clusters (e.g. usually households or subjects within a home).

The nature of the clusters selected in cluster randomization trials (e.g. families, schools, communities) is largely driven by the scientific question. Furthermore this choice will affect the number of clusters which are needed to have sufficient power to detect clinically relevant effects while satisfying practical considerations. For example, community intervention trials are usually limited in practice to a relatively small number of clusters. Cluster randomization trials also tend to use one of three fairly simple designs (i.e. completely randomized, stratified, pair-matched) reviewed in Chapter 1.

Finally, analytic methods for complex surveys often include finite population corrections (Cochran, 1977, p. 24) which can become important when a high proportion of members of the sampled population are included in the study. No such corrections are needed for methods used to analyze data from cluster randomization trials since the clusters are generally assumed to be a sample from an infinite population of clusters.

Curiously these differences have received little attention in the statistical literature. Methods developed for sample surveys have on occasion been applied to data obtained from randomized trials without any consideration of the benefits of random assignment or the limitations arising from a small number of available clusters (e.g. Rao and Scott, 1992). This is exceedingly ironic since the main thrust of statistical advice offered to medical researchers using data collected from cluster randomization trials (e.g. Donner, Brown and Brasher, 1991) or from complex surveys (Korn and Graubard, 1991) has been to consider the design in the analysis.

2.6.2 Ratio Estimators in Stratified Cluster Randomization Trials

The only sample survey analytic technique which will be discussed is a simple method described by Rao and Scott (1992). They adapt the theory of ratio estimation from the sample survey literature (e.g. Cochran 1977, Chapter 6) to adjust methods used to analyze uncorrelated binary outcome data for the effect of clustering. Their method is of interest because its simplicity allows it to be easily adapted to a variety of problems and because it can be shown to be closely related to methods discussed elsewhere in this thesis.

Let the observed cluster risk in the i'th stratum, i=1,...,k, j'th treatment group, j=1,2, and s'th cluster, s=1,..., m_{ij} be denoted $\hat{p}_{ijs} = y_{ijs} / n_{ijs}$ where y_{ijs} equals the number of subjects in that cluster who responded positively and n_{ijs} denotes the cluster size. Then the observed risk in the i'th stratum and j'th treatment group denoted

$$\hat{p}_{ij} = \sum_{s=1}^{m_{ij}} \frac{n_{ijs}\hat{p}_{ijs}}{n_{ij}} \qquad (2.6.1)$$

can be reexpressed as a ratio of two sample means

$$\frac{\sum_{s=1}^{m_{ij}}(y_{ijs}/m_{ij})}{\sum_{s=1}^{m_{ij}}(n_{ijs}/m_{ij})} \qquad (2.6.2)$$

as suggested by Rao and Scott (1992). An estimator of the variance of this ratio obtained using the delta method (Agresti (1990), Mood, Graybill and Boes (1963, p. 181) is given by

$$\hat{\mathrm{var}}_R(\ \hat{p}_{ij}\ ) = \frac{m_{ij}}{(m_{ij} - 1)n_{ij.}^2} \sum_{s=1}^{m_{ij}} n_{ijs}^2\ (\ \hat{p}_{ijs} - \hat{p}_{ij}\ )^2 \qquad (2.6.3)$$

using results from the sample survey literature (Rao and Scott (1992), Cochran (1977, pp. 31-32 and 155)), and omitting the finite population correction (Cochran, 1977, p. 155). The ratio of this variance estimate to $\hat{p}_{ij}\hat{q}_{ij}/n_{ij.}$, the variance obtained in the absence of clustering, represents the variance inflation due to clustering and is denoted by d_{ij}. This approach does not explicitly involve the notion of an intracluster correlation coefficient ρ .

Standard methods for uncorrelated binary outcomes data can be used if $y_{ij.}$ is replaced by $\tilde{y}_{ij.}=y_{ij.}/d_{ij}$ and $n_{ij.}$ is replaced by $\tilde{n}_{ij.}=n_{ij.}/d_{ij}$ since $\tilde{y}_{ij.}$ is approximately binomial with parameters $\tilde{n}_{ij.}$ and p_{ij} (Rao and Scott, 1992). Note that $\tilde{p}_{ij} = \hat{p}_{ij}$ in this case. The principal disadvantage of using this approach in stratified cluster randomization trials is that the asymptotic distributional properties of statistical tests or confidence intervals depends on there being large numbers of clusters in each stratum and treatment group.

Similarly restrictive asymptotic requirements are made when using the robust variance estimates described by Liang and Zeger (1986). Liang and Zeger's (1986) generalized estimating equations approach is discussed in Section 2.8.

In general, inference procedures adjusted for the effect of clustering using the approach described by Rao and Scott (1992) simplify in the absence of clustering to standard methods when $d_{ij}=1$ in all strata and across treatment groups. Fung et al (1993) recommend setting $d_{ij}=1$ if d_{ij} is computed as less than 1, which is essentially equivalent to truncating negative estimates of ρ.

In randomized comparisons the variance inflation factors will be homogenous across treatment groups under the null hypothesis. In this case Rao and Scott (1992) suggest methods of estimating a common inflation factor although they offer little advice as to how this can best be accomplished in stratified designs. One possibility is to use

$$d = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{N-n_{ij.}}{(2k-1)N} \, \frac{\hat{p}_{ij}\hat{q}_{ij}}{\hat{p}\hat{q}} \, d_{ij} \qquad (2.6.4)$$

where $\hat{p}$ denotes the proportion of positive responses for subjects in the study and $\hat{q} = 1-\hat{p}$. Equation (2.6.4) simplifies to equation (5) from Rao and Scott (1992) when there is only a single stratum and two treatment groups. Test statistics which use a pooled estimate of the variance inflation due to clustering will equal standard tests divided by the common variance inflation factor. A similar result arises as a special case of a robust test of significance derived using the GEE methodology and will be discussed in Section 2.8. The focus of the remaining discussion in this section is on the more general approach using separate adjustments for clustering in each stratum and treatment group.

Consider the special case where there are $n_{ijs} = n_{ij}$ subjects in each cluster in the i'th stratum and j'th treatment group. Then the estimated variance of $\hat{p}_{ij}$ expressed by equation (2.6.3) reduces to $\sum_{s=1}^{m_{ij}} (\hat{p}_{ijs} - \hat{p}_{ij})^2 / m_{ij} (m_{ij}-1)$. The variance inflation factor d_{ij} is then recognizable as the formula used to obtain a moment estimator of intracluster correlation previously described in Section 2.2 so $d_{ij} = 1 + (n_{ij}-1)\hat{\rho}_{ij}$ where $\hat{\rho}_{ij}$ is a moment estimator of intracluster correlation. This identity will be used below to point out some algebraic simplifications. Notc that tests of significance constructed using d_{ij} will be somewhat imprecise in this case because the equality of intracluster correlation parameters across treatment groups is ignored.

As the imbalance in cluster size increases the ratio estimator of variance,

$\hat{var}_R(\hat{p}_{ij})$, is likely to be biased downward unless there are at least 30 clusters per treatment group and stratum and the coefficient of variation for cluster size is less than ten percent (Cochran 1977, pp. 162-164). When these conditions are not met the variance inflation will be underestimated resulting in test statistics which reject the null hypothesis too often.

2.6.3 Adjusting Mantel-Haenszel Methods for the Effect of Clustering

The Mantel-Haenszel test statistic (Mantel and Haenszel, 1959) can be adjusted for clustering after substituting $\tilde{y}_{ij.}$ for $y_{ij.}$ and $\tilde{n}_{ij.}$ for $n_{ij.}$. The statistic can then be expressed as

$$\chi^2_{MH} = \frac{\left[\sum_{i=1}^{k} \frac{\tilde{n}_{i1.}\tilde{n}_{i2.}}{\tilde{n}_{i..}} (\hat{p}_{i2} - \hat{p}_{i1}) \right]^2}{\sum_{i=1}^{k} \frac{\tilde{n}_{i1.}\tilde{n}_{i2.}}{\tilde{n}_{i..}-1} \tilde{p}_i \tilde{q}_i} \quad \text{where} \qquad (2.6.5)$$

$$\tilde{p}_i = \sum_{j=1}^{2} \tilde{y}_{ij.} / \sum_{j=1}^{2} \tilde{n}_{ij.}$$

The method described by Rao and Scott (1992) can also be used to adjust Wald type test statistics for the effect of clustering. These test statistics have the advantage of being easily adapted to confidence interval construction. Rao and Scott (1992) consider using a cluster adjusted version of the natural log of the Mantel-Haenszel odds ratio expressed as

$$\bar{\gamma}_{MH} = \log_e(\bar{\psi}_{MH}) \qquad (2.6.6)$$

$$= \log_e \left[\frac{\sum_{i=1}^{k} \frac{\bar{n}_{i1.}\bar{n}_{i2.}}{\bar{n}_{i..}} \hat{p}_{i2}\, \hat{q}_{i1}}{\sum_{i=1}^{k} \frac{\bar{n}_{i1.}\bar{n}_{i2.}}{\bar{n}_{i..}} \hat{p}_{i1}\, \hat{q}_{i2}} \right]$$

whose variance estimate is

$$\bar{\mathrm{var}}(\bar{\gamma}_{MH}) = \sum_{i=1}^{k} \left[\frac{\frac{\bar{n}_{i1.}\bar{n}_{i2.}}{\bar{n}_{i..}} \hat{p}_{i1}\hat{q}_{i2}}{\sum_{i=1}^{k} \frac{\bar{n}_{i1.}\bar{n}_{i2.}}{\bar{n}_{i..}} \hat{p}_{i1}\hat{q}_{i2}} \right]^2 \left[\frac{1}{\bar{n}_{i1.}\hat{p}_{i1}\hat{q}_{i1}} + \frac{1}{\bar{n}_{i2.}\hat{p}_{i2}\hat{q}_{i2}} \right] \qquad (2.6.7)$$

after adapting the variance formula given by Hauck (1979). An approximate test of the null hypothesis that $\log_e(\psi) = 0$ is then

$$\bar{\chi}^2_{IMH} = \frac{\bar{\gamma}^2_{MH}}{\bar{\mathrm{var}}(\bar{\gamma}_{MH})} \qquad (2.6.8)$$

and the $(1-\alpha)100\%$ confidence interval is

$$\exp\left[\bar{\gamma}_{MH} \pm z_{1-\alpha/2}(\bar{\mathrm{var}}(\bar{\gamma}_{MH}))^{1/2} \right] \qquad (2.6.9)$$

The simplification noted for most other methods when all clusters are the same size does not occur for statistics constructed using the approach described by Rao and Scott (1992). This lack of simplification results because even when there is complete balance intracluster correlation is allowed to vary not only between strata but also between treatment groups. Furthermore the variation in intracluster correlation is used when estimating the effect of treatment and its variance. None of the other methods discussed in this chapter allows ρ to vary when estimating treatment effects. Since under the null hypothesis randomization ensures that ρ is constant at least within each stratum χ^2_{MH} and χ^2_{IMH} are likely imprecise.

2.7 The Beta-Binomial Distribution: A Parametric Approach to Modelling Correlated Binary Outcome Data

2.7.1 Introduction

The beta-binomial distribution was the first parametric model proposed for use in the analysis of over-dispersed binomial data. It is derived as the marginal distribution of a random variable which is binomially distributed conditional on a random risk parameter which follows a beta distribution. The beta distribution is used to model the distribution of risk since it is the conjugate prior of the binomial and so is the mathematically simplest choice (Lee, 1989 Chapter 3). The beta distribution is also very flexible and can resemble a surprisingly wide array of distributions (Johnson and Kotz 1970 pp. 42-44).

The beta-binomial distribution was first derived by statisticians interested in Bayesian methods of inference (Dale (1991), Pearson (1925)). This distribution was introduced by Skellam (1948) as a useful method for the analysis of over-dispersed binomial data. Following its introduction the distribution was proposed for a variety of different problems in which correlated binary outcome data were likely to occur (Chatfield and Goodhart (1970), Griffiths (1973)) and was first compared to other less parametric methods by Kleinman (1973).

Williams (1975) can be credited with demonstrating the usefulness of this distribution for testing hypotheses in teratological studies. His work arose as part of the debate among toxicologists (Luning et al (1966), Weil (1970), Healy (1972), Weil (1974), Palmer (1974), Haseman and Hogan (1975), Haseman and Soares (1976), Kupper and Haseman (1978)) as to the appropriate analysis of data from trials in which pregnant animals are randomly assigned to treatment or control groups and their offspring compared for risks of birth defects or death. The beta-binomial distribution was subsequently discussed by Crowder (1978) as an extension of logistic regression for over-dispersed binomial data independently of Williams (1975) (Crowder, 1979).

The beta-binomial distribution is derived in the next section of the thesis. Its relationship with logistic regression is then explored in Section 2.7.3 in the context of likelihood based inferences.

2.7.2 Derivation of the Beta-Binomial Distribution

Let y_{ijs} denote the number of subjects having positive responses in a cluster of size n_{ijs} in the i'th stratum, i=1,...,k, j'th treatment group, j=1,2, and s'th cluster s=1,..., m_{ij}. Now suppose that y_{ijs} is binomially distributed conditional on the unknown, true risk denoted p_{ijs} so

$$f(\, y_{ijs} \mid p_{ijs}\,) = \begin{bmatrix} n_{ijs} \\ y_{ijs} \end{bmatrix} p_{ijs}^{y_{ijs}}\, q_{ijs}^{n_{ijs}-y_{ijs}}. \qquad (2.7.1)$$

If p_{ijs} follows a beta distribution with parameters $a_{ij}>0$ and $b_{ij}>0$ then

$$g(\, p_{ijs}\,) = p_{ijs}^{a_{ij}-1}\, q_{ijs}^{b_{ij}-1} / B(\, a_{ij}\,,\, b_{ij}\,) \qquad (2.7.2)$$

where $B(a_{ij} , b_{ij})$ is the beta function (Mood, Graybill and Boes 1974, pp. 534-535). Therefore

$$h(y_{ijs}) = \int_{p_{ijs}=0}^{1} f(y_{ijs} \mid p_{ijs}) g(p_{ijs})dp_{ijs} \tag{2.7.3}$$

$$= \binom{n_{ijs}}{y_{ijs}} \frac{B(a_{ij}+y_{ijs} , n_{ijs}+b_{ij}-y_{ijs})}{B(a_{ij} , b_{ij})}$$

$$= \binom{n_{ijs}}{y_{ijs}} \frac{\dfrac{\Gamma(a_{ij}+y_{ijs})}{\Gamma(a_{ij})} \dfrac{\Gamma(b_{ij}+n_{ijs}-y_{ijs})}{\Gamma(b_{ij})}}{\dfrac{\Gamma(a_{ij}+b_{ij}+n_{ijs})}{\Gamma(a_{ij}+b_{ij})}}$$

$$= \binom{n_{ijs}}{y_{ijs}} \frac{\prod_{v=0}^{y_{ijs}-1} (a_{ij}+v) \prod_{v=0}^{n_{ijs}-y_{ijs}-1} (b_{ij}+v)}{\prod_{v=0}^{n_{ijs}-1} (a_{ij}+b_{ij}+v)}$$

since $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(a) = (a-1)\ \Gamma(a-1)$ (Mood, Graybill and Boes, 1974, pp. 534-535).

Expected values and variances for y_{ijs} can be obtained using the expected values and variances of variables from the binomial and beta distributions (Mood, Graybill and Boes 1974 Appendix B) and identities relating conditional and marginal expected values and variances respectively (Mood, Graybill and Boes (1974, pp. 158-159), Moran (1968, pp. 76-77)). Thus the

$$E(\ y_{ijs}\ ) = E(\ E[\ y_{ijs} \mid p_{ijs}\ ]) \tag{2.7.4}$$

$$= n_{ijs}a_{ij}\ /\ (\ a_{ij} + b_{ij}\ )$$

$$= n_{ijs}\ p_{ij}\ \text{and}$$

$$\text{var}(\ y_{ijs}\ ) = E(\text{var}[\ y_{ijs} \mid p_{ijs}\ ]) + \text{var}(E[\ y_{ijs} \mid p_{ijs}\ ])$$

$$= n_{ijs}\ p_{ij}\ q_{ij}\left[1 + \frac{n_{ijs}-1}{1 + a_{ij} + b_{ij}}\right]$$

$$= n_{ijs}\ p_{ij}\ q_{ij}\ [\ 1 + (n_{ijs}-1)\rho_{ij}\ ]$$

where ρ_{ij} is the intracluster correlation coefficient.

Proof of the last step in this equation is obtained by calculating the correlation between any two subjects within the same cluster. Let y_{ijsu} and y_{ijsv} denote the responses for the u'th and v'th subjects within a cluster, $u,v=1,\ldots,n_{ijs},u\neq v$. Then the

$$\text{cov}(\ y_{ijsu},y_{ijsv}\ ) = E(\ y_{ijsu}y_{ijsv}\ )-E(\ y_{ijsu}\ )E(\ y_{ijsv}\ ) \tag{2.7.5}$$

$$= E(\ p_{ijs}^{2}\ ) - \left[E(\ p_{ijs}\ )\right]^{2}$$

$$= \text{var}(\ p_{ijs}\ )$$

$$= p_{ij}\ q_{ij}\ /\ (\ 1 + a_{ij} + b_{ij}\ )$$

using results from Appendix B in Mood, Graybill and Boes (1974). Thus the corr(y_{ijsu},y_{ijsv}) = 1/(1 + a_{ij} + b_{ij}) since var(y_{ijsu}) = p_{ij} q_{ij}; obtained as a special case of var(y_{ijs}) where n_{ijs}=1. Notice that negative intracluster correlation is not possible with this beta-binomial model since a_{ij} and b_{ij} must be greater than zero.

These results suggest a more convenient expression for the beta-binomial distribution. Let $p_{ij} = a_{ij}/(a_{ij}+b_{ij})$ and let $\theta_{ij} = 1/(a_{ij}+b_{ij})$ then $a_{ij} = p_{ij}/\theta_{ij}$, $b_{ij} = q_{ij}/\theta_{ij}$ and $\theta_{ij} = \rho_{ij} / (1 - \rho_{ij})$. Therefore h(y_{ijs}) can be expressed as

$$\begin{bmatrix} n_{ijs} \\ y_{ijs} \end{bmatrix} \frac{\prod_{v=0}^{y_{ijs}-1} (a_{ij}+v) \prod_{v=0}^{n_{ijs}-y_{ijs}-1} (b_{ij}+v)}{\prod_{v=0}^{n_{ijs}-1} (a_{ij}+b_{ij}+v)} \tag{2.7.6}$$

$$= \begin{bmatrix} n_{ijs} \\ y_{ijs} \end{bmatrix} \frac{\prod_{v=0}^{y_{ijs}-1} \left[\frac{p_{ij} + v\theta_{ij}}{\theta_{ij}} \right] \prod_{v=0}^{n_{ijs}-y_{ijs}-1} \left[\frac{q_{ij} + v\vartheta_{ij}}{\theta_{ij}} \right]}{\prod_{v=0}^{n_{ijs}-1} \left[\frac{1 + v\theta_{ij}}{\theta_{ij}} \right]}$$

$$= \begin{bmatrix} n_{ijs} \\ y_{ijs} \end{bmatrix} \frac{\prod_{v=0}^{y_{ijs}-1} (p_{ij} + v\theta_{ij}) \prod_{v=0}^{n_{ijs}-y_{ijs}-1} (q_{ij} + v\theta_{ij})}{\prod_{v=0}^{n_{ijs}-1} (1 + v\theta_{ij})}$$

where $\prod_{v=0}^{-1}(c+dv) = 1$ by definition for any constants c and d.

There are three important special cases of this distribution. First, if $n_{ijs}=1$ for all clusters then $h(\ y_{ijs}\ )$ reduces to the Bernoulli distribution. More generally the beta-binomial distribution reduces to the binomial distribution when $\rho_{ij} = 0$.

The second important reduction occurs when there are exactly two subjects in each cluster. In this case y_{ijs} can take on three values and

$$h(\ y_{ijs} = 0\ ) = q_{ij}^2 + \rho_{ij}p_{ij}q_{ij} \tag{2.7.7}$$

$$h(\ y_{ijs} = 1\ ) = 2p_{ij}q_{ij}(\ 1 - \rho_{ij}\ )$$

$$h(\ y_{ijs} = 2\ ) = p_{ij}^2 + \rho_{ij}p_{ij}q_{ij}$$

It is worth noting that this is the only parametric marginal distribution which can be constructed when there are exactly two subjects in each cluster and only cluster-level covariates (Prentice, 1988). Other approaches (e.g. Kupper and Haseman's (1978) correlated binomial model) differ only in the restrictions imposed on ρ_{ij} in this case. The bewildering array of methods developed to analyze correlated binary outcomes data (Ashby et al, 1991) have arisen, in part, because of the absence of any similarly unique distribution for larger clusters.

There are, however, very few cluster randomization trials in which there would be two observations per cluster. A notable exception occurs in ophthalmological trials (Rosner (1982), Dallal (1988), and Donner (1989)). Applications are far more common in reliability studies (e.g. Kraemer (1979), Mak (1988), and Donner and Eliasziw (1992)) where inferences are concerned with the degree of intracluster correlation and the risk is a nuisance variable in contrast to cluster randomization trials.

The third special case is likely to occur in community intervention trials evaluating methods to prevent disease or mortality. In su:h trials clusters tend to be large while risk is low. The negative binomial distribution is the limiting form of the beta-binomial distribution in this case (Skellam, 1948). The negative binomial distribution can also be derived by ass:uming that the number of positive outcomes in a cluster follow a Poisson distribution conditional on the expected risk which is sampled from a gamma distribution (Breslow, 1990). The negative binomial distribution reduces to the Poisson distribution when there is no between-cluster variability in risk (Moore and Tsiatis, 1991).

2.7.3 Likelihood Based Inferences

Inferences about the effect of treatment in stratified cluster randomization trials can be conducted using likelihood methods when a parametric distribution, such as the beta-binomial, is assumed. In order to construct tests of the effect of treatment p_{ij}, the expected risk in the i'th stratum and j'th treatment group, is rewritten using the logistic model

$$\text{logit}(\ p_{ij}\ ) = \alpha_i + \gamma x_j \quad \text{where} \tag{2.7.8}$$

$$x_j = \begin{cases} 0 & j=1 \\ 1 & j=2 \end{cases}$$

$\alpha_i = \text{logit}(p_{i1})$ and $\gamma = \log_e(\psi)$, is the natural log of the odds ratio. Therefore the null hypothesis that γ, the log of the odds ratio, equals zero is used to evaluate the effect of treatment.

The likelihood can be written as a multiple of the density functions $h(y_{ijs})$ so that

$$L = \prod_{i=1}^{k}\left[\prod_{s=1}^{m_{i1}} h(y_{i1s}) \prod_{s=1}^{m_{i2}} h(y_{i2s})\right] \tag{2.7.9}$$

$$= \prod_{i=1}^{k}\left\{\prod_{s=1}^{m_{i1}} \frac{\prod_{v=0}^{y_{i1s}-1}\left[p_{i1} + v\theta\right] \prod_{v=0}^{n_{i1s}-y_{i1s}-1}\left[q_{i1} + v\theta\right]}{\prod_{v=0}^{n_{i1s}-1}\left[1 + v\theta\right]}\right.$$

$$\left.\prod_{s=1}^{m_{i2}} \frac{\prod_{v=0}^{y_{i2s}-1}\left[p_{i2} + v\theta\right] \prod_{v=0}^{n_{i2s}-y_{i2s}-1}\left[q_{i2} + v\theta\right]}{\prod_{v=0}^{n_{i2s}-1}\left[1 + v\theta\right]}\right\}.$$

Therefore the log-likelihood (2.7.10), denoted $\log_e L$ can be expressed as

$$\sum_{i=1}^{k}\sum_{s=1}^{m_{i1}}\left\{\sum_{v=0}^{y_{i1s}-1}\log_e\left[p_{i1} + v\theta\right] + \sum_{v=0}^{n_{i1s}-y_{i1s}-1}\log_e\left[q_{i1} + v\theta\right] - \sum_{v=0}^{n_{i1s}-1}\log_e\left[1 + v\theta\right]\right\}$$

$$+\sum_{i=1}^{k}\sum_{s=1}^{m_{i2}}\left\{\sum_{v=0}^{y_{i2s}-1}\log_e\left[p_{i2} + v\theta\right] + \sum_{v=0}^{n_{i2s}-y_{i2s}-1}\log_e\left[q_{i2} + v\theta\right] - \sum_{v=0}^{n_{i2s}-1}\log_e\left[1 + v\theta\right]\right\}.$$

Notice that the simplifying assumption that $\theta_{ij} = \theta$ has been made. This assumption is equivalent to assuming that the degree of intracluster correlation is constant across strata and treatment groups. These likelihood equations reduce to standard equations for logistic regression when $\theta = 0$.

There are three approaches to developing tests of the effect of treatment which can be constructed using likelihood theory: likelihood ratio, score and Wald tests. These three test statistics are asymptotically equivalent when the null hypothesis holds and are approximately distributed according to a χ^2_{T-1} distribution when there are T treatment groups (Cox and Hinkley 1974 Chapter 9).

The likelihood ratio statistic denoted χ^2_{LRB} is given by

$$-2\log_e\left[\frac{L(\hat{\alpha}_{1N}, \ldots, \hat{\alpha}_{kN}, \gamma=0, \hat{\beta}_N)}{L(\hat{\alpha}_{1A}, \ldots, \hat{\alpha}_{kA}, \hat{\gamma}_A, \hat{\beta}_A)}\right] \tag{2.7.11}$$

where $L(\hat{\alpha}_{1N}, \ldots, \hat{\alpha}_{kN}, \gamma=0, \hat{\beta}_N)$ is the likelihood maximized under the null hypothesis and $L(\hat{\alpha}_{1A}, \ldots, \hat{\alpha}_{kA}, \hat{\gamma}_A, \hat{\beta}_A)$ is the likelihood maximized under the alternative hypothesis. This test statistic is the most commonly proposed. A problem with this test statistic is that it tends to reject the null hypothesis too often when it is true, at least in completely randomized designs in which there are few clusters (Haseman and Kupper (1979), Donner, Eliasziw and Klar (1993)).

The poor small sample properties of likelihood ratio statistics might arise, in part, because it uses maximum likelihood estimates obtained under both the null and alternative hypotheses. The score test, however, only uses maximum likelihood estimates calculated under the null hypothesis and so might have better small sample properties. It is denoted χ^2_{SB} and expressed as

$$(\hat{U}_1, \ldots, \hat{U}_k, \hat{U}_g, \hat{U}_r)' \hat{\Sigma}^{-1} (\hat{U}_1, \ldots, \hat{U}_k, \hat{U}_g, \hat{U}_r) \qquad (2.7.12)$$

where $(\hat{U}_1, \ldots, \hat{U}_k, \hat{U}_g, \hat{U}_r)$ is a vector of score equations for the parameters $\alpha_1, \ldots, \alpha_k, \gamma$ and ρ respectively and $\hat{\Sigma}$ is an estimate of the (k+2)x(k+2) variance-covariance matrix for the score equations.

Calculation of the score test can be simplified upon recognizing that the score equations $\hat{U}_1, \ldots, \hat{U}_k$ and $\hat{U}_r$ are equal to zero when the γ is set to zero and the maximum likelihood estimates for the other parameters are estimated under the null. The score test then simplifies to

$$\hat{U}_g^2 / \hat{\sigma}^{gg} \qquad (2.7.13)$$

where $\hat{\sigma}^{gg}$ is the element from $\hat{\Sigma}$ corresponding to the variance of $\hat{U}_g$.

The third statistic which can be used to make inferences about treatment is the Wald test. This test statistic denoted χ^2_{WB} is expressed as

$$\hat{\gamma}^2 / \hat{\mathrm{var}}(\hat{\gamma}) \qquad (2.7.14)$$

where $\hat{\mathrm{var}}(\hat{\gamma})$ is an estimate of the variance of $\hat{\gamma}$ calculated by maximizing the likelihood assuming that the alternative hypothesis holds.

The principal advantage of the Wald test is the ease with which they can be inverted to construct confidence intervals. These tests tend, however, to require larger sample sizes to be valid than do score or likelihood ratio tests. Wald tests for parameters from logistic regression models which assume independence have also been shown to exhibit aberrant behavior by several authors (Hauck and Donner (1977), Vaeth (1985), Mantel (1987)). In particular the test statistic tends to move towards zero as the odds ratio gets very large instead of getting consistently larger! The same behavior is likely to affect Wald tests of γ from beta-binomial regression.

The variance of score equations and maximum likelihood estimates are equal to functions of the expected information matrix. Estimates of variance can be calculated by substituting in the maximum likelihood estimates. A simpler estimate can be calculated using the observed information matrix. These two approaches are asymptotically equivalent and will yield identical estimates when responses of subjects from the same cluster are uncorrelated (Collett, 1991, p. 343). Variance estimates calculated using the observed information is, however, generally believed to have better small sample properties (Efron and Hinkley (1978), Buse (1982)).

An interesting consequence of using a parametric model is that the variance of $\hat{\gamma}$ depends on the variance of $\hat{\rho}$. This stands in contrast to the adjusted χ^2 methods discussed in the last section or the test statistics discussed in Section 2.8. These other methods calculate variance estimates assuming that ρ is known. This may account, in part, for the inability to obtain simple algebraic expressions for test statistics from the beta-binomial model even when there is no variability in cluster size.

All three tests require iterative procedures to obtain the appropriate maximum likelihood estimates. These estimates can be obtained using the Newton-Raphson procedure which requires calculation of the first and second derivatives of the log-likelihood.

The first derivatives of the log-likelihood (2.7.10) can be expressed as

$$\frac{\delta l}{\delta \alpha_i} = \sum_{j=1}^{2} p_{ij} q_{ij} A_{ij}, \quad i=1,\ldots,k, \tag{2.7.15}$$

$$\frac{\delta l}{\delta \gamma} = \sum_{i=1}^{k} p_{i2} q_{i2} A_{i2},$$

$$\frac{\delta l}{\delta \theta} = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \left\{ \sum_{v=0}^{y_{ijs}-1} \frac{v}{p_{ij} + v\theta} + \sum_{v=0}^{n_{ijs}-y_{ijs}-1} \frac{v}{q_{ij} + v\theta} - \sum_{v=0}^{n_{ijs}-1} \frac{v}{1 + v\theta} \right\},$$

$$\text{where } A_{ij} = \sum_{s=1}^{m_{ij}} \left\{ \sum_{v=0}^{y_{ijs}-1} \frac{1}{p_{ij} + v\theta} - \sum_{v=0}^{n_{ijs}-y_{ijs}-1} \frac{1}{q_{ij} + v\theta} \right\}.$$

The second derivatives can then be expressed as

$$\frac{\delta^2 l}{\delta \alpha_i^2} = -\sum_{j=1}^{2} p_{ij} q_{ij} [(p_{ij} - q_{ij}) A_{ij} + p_{ij} q_{ij} B_{ij}], \tag{2.7.16}$$

$$\frac{\delta^2 l}{\delta \gamma^2} = -\sum_{i=1}^{k} p_{i2} q_{i2} [(p_{i2} - q_{i2}) A_{i2} + p_{i2} q_{i2} B_{i2}],$$

$$\frac{\delta^2 l}{\delta\theta^2} = -\sum_{i=1}^{k}\sum_{j=1}^{2}\sum_{s=1}^{m_{ij}}\left\{\sum_{v=0}^{y_{ijs}-1}\frac{v^2}{(\ p_{ij} + v\theta\ )^2}\right.$$

$$\left. + \sum_{v=0}^{n_{ijs}-y_{ijs}-1}\frac{v^2}{(\ q_{ij} + v\theta\ )^2} - \sum_{v=0}^{n_{ijs}-1}\frac{v^2}{(\ 1 + v\theta\ )^2}\right\},$$

$$\frac{\delta^2 l}{\delta\alpha_i\delta\alpha_{i'}} = 0,$$

$$\frac{\delta^2 l}{\delta\alpha_i\delta\gamma} = -p_{i2}q_{i2}[(p_{i2}-q_{i2})A_{i2} + p_{i2}q_{i2}B_{i2}\ ],$$

$$\frac{\delta^2 l}{\delta\alpha_i\delta\theta} = -\sum_{j=1}^{2} p_{ij}q_{ij}D_{ij},$$

$$\frac{\delta^2 l}{\delta\gamma\ \delta\theta} = -\sum_{i=1}^{k} p_{i2}q_{i2}D_{i2},$$

$$\text{where } B_{ij} = \sum_{s=1}^{m_{ij}}\left\{\sum_{v=0}^{y_{ijs}-1}\frac{1}{(\ p_{ij} + v\theta\ )^2} + \sum_{v=0}^{n_{ijs}-y_{ijs}-1}\frac{1}{(\ q_{ij} + v\theta\ )^2}\right\},$$

$$\text{and } D_{ij} = \sum_{s=1}^{m_{ij}}\left\{\sum_{v=0}^{y_{ijs}-1}\frac{v}{(\ p_{ij} + v\theta\ )^2} - \sum_{v=0}^{n_{ijs}-y_{ijs}-1}\frac{v}{(\ q_{ij} + v\theta\ )^2}\right\},$$

for $i,i'=1,\ldots,k,\ i \neq i'$.

The matrix of second derivatives can be expressed as a (k+2)x(k+2) partitioned matrix

$$\begin{bmatrix} \mathbf{D} & \mathbf{U} \\ \mathbf{U}' & \mathbf{G} \end{bmatrix} \tag{2.7.17}$$

where the kxk submatrix

$$\mathbf{D} = \mathrm{diag}\left[\frac{\delta^2 l}{\delta\alpha_i^2} \right] \quad i=1,\dots,k,$$

$$\mathbf{U}' = \begin{bmatrix} \mathbf{u}'_1 \\ \mathbf{u}'_2 \end{bmatrix},$$

$$\mathbf{u}'_1 = \left[\frac{\delta^2 l}{\delta\alpha_1 \delta\gamma}, \dots, \frac{\delta^2 l}{\delta\alpha_k \delta\gamma} \right],$$

$$\mathbf{u}'_2 = \left[\frac{\delta^2 l}{\delta\alpha_1 \delta\theta}, \dots, \frac{\delta^2 l}{\delta\alpha_k \delta\theta} \right] \quad \text{and}$$

$$\mathbf{G} = \begin{bmatrix} \frac{\delta^2 l}{\delta\gamma^2} & \frac{\delta^2 l}{\delta\gamma\, \delta\theta} \\ \frac{\delta^2 l}{\delta\gamma\, \delta\theta} & \frac{\delta^2 l}{\delta\theta^2} \end{bmatrix}.$$

The negative of the inverse of the matrix of second derivatives is needed to calculate the observed information and is also required to obtain maximum likelihood estimates when using the Newton-Raphson procedure. The inverse of this matrix can be expressed as (2.7.18)

$$\begin{bmatrix} \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{U}(\,\mathbf{G} - \mathbf{U}'\mathbf{D}^{-1}\mathbf{U}\,)^{-1}\mathbf{U}'\mathbf{D}^{-1} & -\mathbf{D}^{-1}\mathbf{U}(\,\mathbf{G} - \mathbf{U}'\mathbf{D}^{-1}\mathbf{U}\,)^{-1} \\ -(\,\mathbf{G} - \mathbf{U}'\mathbf{D}^{-1}\mathbf{U}\,)^{-1}\mathbf{U}'\mathbf{D}^{-1} & (\,\mathbf{G} - \mathbf{U}'\mathbf{D}^{-1}\mathbf{U}\,)^{-1} \end{bmatrix}$$

using the formula for the inverse of a partitioned matrix (Searle 1982, p. 260). Note that the advantage of this formula is that it reduces the problem to having to invert nothing more complicated than a 2x2 symmetric matrix. The observed information is then calculated as the negative of (2.7.18) after replacing the parameters by their maximum likelihood estimates.

The Wald test of the effect of treatment can therefore be expressed as

$$\chi^2_{Wbb} = \hat{\gamma}^2 \;/\; \hat{\mathrm{var}}(\,\hat{\gamma}\,) \qquad (2.7.19)$$

where $\hat{\mathrm{var}}(\,\hat{\gamma}\,)$ is

$$\frac{\dfrac{\delta^2 l}{\delta\hat{\theta}^2} - \sum_{i=1}^{k} \dfrac{\left(\dfrac{\delta^2 l}{\delta\hat{\alpha}_i \delta\hat{\theta}}\right)^2}{\dfrac{\delta^2 l}{\delta\hat{\alpha}_i^2}}}{\left[\dfrac{\delta^2 l}{\delta\hat{\gamma}^2} - \sum_{i=1}^{k} \dfrac{\left(\dfrac{\delta^2 l}{\delta\hat{\alpha}_i \delta\hat{\gamma}}\right)^2}{\dfrac{\delta^2 l}{\delta\hat{\alpha}_i^2}}\right]\left[\dfrac{\delta^2 l}{\delta\hat{\theta}^2} - \sum_{i=1}^{k} \dfrac{\left(\dfrac{\delta^2 l}{\delta\hat{\alpha}_i \delta\hat{\theta}}\right)^2}{\dfrac{\delta^2 l}{\delta\hat{\alpha}_i^2}}\right] - \left[\dfrac{\delta^2 l}{\delta\hat{\gamma}\,\delta\hat{\theta}} - \sum_{i=1}^{k} \dfrac{\dfrac{\delta^2 l}{\delta\hat{\alpha}_i \delta\hat{\gamma}}\;\dfrac{\delta^2 l}{\delta\hat{\alpha}_i \delta\hat{\theta}}}{\dfrac{\delta^2 l}{\delta\hat{\alpha}_i^2}}\right]^2}$$

2.8 The Generalized Estimating Equations Approach

2.8.1 Introduction

Liang and Zeger (1986) described a family of models which extended the generalized linear models (McCullagh and Nelder, 1989) to include multivariate outcomes. Their method requires specification of an assumed or working correlation matrix to describe the association between outcomes, in addition to the link and error functions, needed for fitting generalized linear models. This correlation matrix is assumed to be common to all clusters in the study.

The consistency o^f parameter estimates obtained using this approach depends only on correctly specifying the first moment, i.e. all confounders have been included in the model. Misspecification of the assumed correlation matrix increases the variance of the parameter estimates although they remain consistent. Model dependent estimates of variance calculated using the working correlation matrix can be employed if there is little likelihood that the correlation is misspecified. In general such certitude is unlikely requiring the use of less precise robust estimates of variance calculated using a combination of model dependent and empirically calculated between-cluster information. Earlier applications of robust variance estimators is described by Royall (1986) where these ideas have been traced back primarily to Huber (1967).

Parameter estimates are obtained by solving a set of equations which Liang and Zeger (1986) have called generalized estimating equations to distinguish them from the estimating equations obtained using ordinary generalized linear models. The asymptotic optimality of parameter estimates obtained by solving these equations is assured by the theory of estimating functions (Godambe and Kale, 1991), assuming that the working correlation is correctly specified. The theory of estimating functions combines the properties of least squares and maximum likelihood theory (Godambe and Kale, 1991) and also bears some resemblance to the method of moments.

Generalized estimating equations have properties similar to score equations obtained from maximum likelihood theory. For example, Godambe (1960) showed that the estimating function for a parametric model with a single unknown parameter is the score equation. More generally the expected value of estimating functions are asymptotically equal to zero. Parameter estimates are obtained by setting the equations equal to their asymptotic expected value and solving for the unknown parameters as is done using the method of moments. The optimality of parameter estimates obtained by this approach is due to their minimizing the mean square error, at least asymptotically (Godambe and Kale, 1991).

An alternative justification for the asymptotic optimality of parameter estimates obtained using generalized estimating equations has been put forward by Zhao, Prentice and Self (1992). These authors show that when the working correlation is correctly specified the generalized estimating equations are equivalent to maximum likelihood score equations obtained using a partly exponential model and so are assured of yielding asymptotically optimal estimates.

One of the models provided by Liang and Zeger's (1986) approach is an extension of logistic regression adjusted for clustering. This model and some special cases of it will be described in this section of the thesis.

Different methods of fitting such models have been described by Williams (1982), Moore (1986), Liang and Zeger (1986) and Moore and Tsiatis (1991). The similarities and differences of these approaches are also explored. The primary focus will be on models used to make inferences about the effect of treatment in stratified cluster randomization trials. Attention is restricted to models where all variables are measured at the cluster-level.

Four different tests of the effect of treatment are discussed. They are distinguished by simplifying asymptotically to Wald and score tests in the absence of clustering and may use either model dependent (i.e. naive) or robust estimates of variance. Relationships between these different test statistics are derived.

2.8.2 Application to Stratified Cluster Randomization Trials

Consider a stratified cluster randomization trial in which there are $M=\sum_{i=1}^{k}\sum_{j=1}^{2} m_{ij}$ clusters overall and m_{ij} clusters in the i'th stratum, i=1,...,k, and j'th treatment group, j=1,2. The effect of treatment in such a trial can be estimated using the logistic regression model,

$$\text{logit}(\ p_{ij}\ ) = \theta_1' \mathbf{X}_{ij} \qquad (2.8.1)$$

where p_{ij} is the expected risk in the i'th stratum and j'th treatment group. The vector of parameters

$$\theta_1' = (\ \alpha_1, \cdots, \alpha_k, \gamma\ ) \qquad (2.8.2)$$

where $\alpha_i = \text{logit}(p_{i1})$ and $\gamma = \log_e(\ \psi\ )$, is the natural log of the odds ratio. Furthermore $\mathbf{X}_{ij}$ is a vector of length k+1 where

$$x_{iju} = \begin{cases} 1 & \text{if } u{=}i \\ 0 & \text{otherwise} \end{cases}$$

and for u=k+1

$$x_{ij(k+1)} = \begin{cases} 1 & \text{if } j{=}2 \\ 0 & \text{if } j{=}1. \end{cases}$$

This model can also be expressed as

$$\begin{aligned} \text{logit}(\ p_{ij}\ ) &= \alpha_1 x_{ij1} + \cdots + \alpha_k x_{ijk} + \gamma x_{ij(k+1)} \\ &= \alpha_i + \gamma x_{ij(k+1)} \end{aligned}$$

Both formulations will be used as required. This is the same model used in Section 2.7 which discussed beta-binomial regression, a parametric approach.

Following Moore and Tsiatis (1991) the common or exchangeable correlation matrix, which assumes that responses of all cluster members are equally correlated, will be used as the working correlation. The estimating equations for model (2.8.1) then arise as a special case of the generalized estimating equations used by Moore and Tsiatis (1991). Moore (1986) has shown that parameter estimates obtained by solving these equations are equivalent to those found using Williams' (1982) model II.

The estimating equations for the k+2 parameters $\alpha_1, \cdots, \alpha_k, \gamma$ and the intracluster correlation coefficient, denoted ρ can be expressed as

$$f_1 = \sum_{j=1}^{2} \sum_{s=1}^{m_{1j}} r_{1js} \, (\hat{p}_{1js} - p_{1j}) = 0 \qquad (2.8.3)$$

$$f_2 = \sum_{j=1}^{2} \sum_{s=1}^{m_{2j}} r_{2js} \, (\hat{p}_{2js} - p_{2j}) = 0$$

$$\vdots$$

$$f_k = \sum_{j=1}^{2} \sum_{s=1}^{m_{kj}} r_{kjs} \, (\hat{p}_{kjs} - p_{kj}) = 0$$

$$f_{k+1} = \sum_{i=1}^{k} \sum_{s=1}^{m_{i2}} r_{i2s} \, (\hat{p}_{i2s} - p_{i2}) = 0$$

$$f_{k+2} = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{r_{ijs}(\hat{p}_{ijs} - p_{ij})^2}{p_{ij}q_{ij}} - (M - k - 1) = 0,$$

where $\hat{p}_{ijs}$ denotes the observed cluster risk for subjects in the i'th stratum, j'th treatment group and s'th cluster, $q_{ij} = 1 - p_{ij}$ and $r_{ijs} = n_{ijs} / [1 + (n_{ijs}-1)\rho]$.

The terms denoted r_{ijs} are the effective cluster sizes which represent the number of effectively independent observations from each cluster. Observed and effective cluster sizes are equal if $\rho = 0$. The effective cluster size approaches one as the degree of intracluster correlation increases. The first k+1 generalized estimating equations are equivalent to score equations for ordinary logistic regression but replacing the observed by the effective number of independent observations. Furthermore these k+1 equations simplify to score equations for ordinary logistic regression when there is no variation in cluster size (Moore (1986), Williams, (1982)).

The last estimating equation also reduces to a familiar form when there is no variation in cluster size. In this case

$$f_{k+2} = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{n(\hat{p}_{ijs} - p_{ij})^2}{p_{ij}q_{ij}[1 + (n-1)\rho]} - (M - k - 1) = 0, \qquad (2.8.4)$$

This equation can be solved to obtain

$$\rho \approx \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{1}{2k} \left\{ \frac{1}{n-1} \left[\frac{\sum_{s=1}^{m_{ij}} n(\hat{p}_{ijs} - p_{ij})^2}{\overline{M}p_{ij}q_{ij}} - 1 \right] \right\} \qquad (2.8.5)$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{1}{2k} \rho_{ij}$$

where ρ_{ij} is a moment estimator of intracluster correlation. This simplification depends on replacing M-k-1 by $2k\overline{M}$ where $\overline{M}$=M/2k, is the average number of clusters over the 2k treatment groups and strata, which is reasonable when there are many clusters in each stratum and treatment group.

The estimating equations must be solved iteratively whenever there are two or more strata, i.e. whenever k>1. An algorithm for this is described below.

A slightly different approach was taken by Liang and Zeger (1986). They proposed using a non-iterative estimator of intracluster correlation rather than using an additional estimating equation. A parameter ϕ was also included to account for any residual over-dispersion. One advantage of their approach is that when k=1 (i.e. for completely randomized designs) the two estimating equations can be solved analytically. The estimating equations approach, however, is likely to provide more precise estimates of intracluster correlation, at least asymptotically, assuming that the correlation matrix has been correctly specified. An alternative, non-iterative estimator of intracluster correlation could be constructed using the ANOVA approach described in Section 2.3.

2.8.3 Asymptotic Distribution of Estimating Equations and Parameter Estimates

The k+2 estimating equations are each sums of independent functions of the data and the unknown parameters. The central limit theorem guarantees that such sums will be asymptotically normally distributed, under some regularity conditions (Moore, 1986). Following the arguments given by Moore (1986)

$$M^{-1/2}\ \mathbf{f}\ \overset{D}{\rightarrow} MVN_{k+2}(\ \mathbf{0}\ ,\ \Lambda\ ) \tag{2.8.6}$$

where $\mathbf{f}' = (\mathbf{f}'_1\ ,\ f_{k+2})$

$$= (f_1\ ,\ f_2\ ,\ \cdots\ ,f_k\ ,\ f_{k+1}\ ,\ f_{k+2})$$

and Λ is the (k+2)x(k+2) covariance matrix. Note that the $\mathrm{cov}(\ f_u\ ,\ f_v\ ) = E(\ f_u f_v\ )$ since $E(\ f_u\ ) = E(\ f_v\ ) = 0$, $u \neq v, u, v = 1, \ldots, k+2$.

Since the estimating equations are asymptotically normally distributed one-to-one transformations of them are also asymptotically normally distributed. Variance estimates for these transformed variables are obtained using a multivariate extension of the delta method (Moore and Tsiatis, 1991). In particular for the k+2 vector of parameter estimates denoted $\hat{\boldsymbol{\theta}} = (\ \hat{\boldsymbol{\theta}}_1\ ,\ \hat{\boldsymbol{\beta}}\ )'$

$$M^{1/2}\ (\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\ )\ \overset{D}{\rightarrow} MVN_{k+2}(\ 0\ ,\ \Gamma^{-1}\Lambda\Gamma^{-1}\ ) \tag{2.8.7}$$

where Γ is the matrix of derivatives of the estimating equations.

Inferences in cluster randomization trials are generally restricted to $\boldsymbol{\theta}_1$ since ρ is generally considered to be a nuisance parameter. The distribution of the k+1 subvector of parameter estimates $\hat{\boldsymbol{\theta}}_1$ can be expressed as

$$M^{1/2} \; (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \overset{D}{\rightarrow} MVN_{k+1}(0 , \Gamma_{11}^{-1}\Lambda_{11}\Gamma_{11}^{-1}) \tag{2.8.8}$$

$$= MVN_{k+1}(0 , \Lambda_{11}^{-1})$$

where Γ_{11} and Λ_{11} are the upper left-hand submatrices of Γ and Λ respectively and $\Gamma_{11} = -\Lambda_{11}$ (Moore, 1986). The asymptotic distribution of $\mathbf{f}_1$ can be similarly expressed as

$$M^{-1/2} \, \mathbf{f}_1 \overset{D}{\rightarrow} MVN_{k+1}(\mathbf{0} , \Lambda_{11}) \tag{2.8.9}$$

This result from Moore (1986) is very important. It proves that the asymptotic covariance of $\hat{\boldsymbol{\theta}}_1$ is not affected by estimation of ρ and so proves the asymptotic equivalence between Liang and Zeger's (1986) approach in which a simple, non-iterative estimator of ρ is recommended, and the approach used by Moore (1986) or Moore and Tsiatis (1991) where estimating equations were preferred. In either case the variance of $\hat{\boldsymbol{\theta}}_1$ can then be calculated as if ρ were known.

Moore's (1986) result also adds insight into algorithms used to obtain parameter estimates. In particular rather than having to solve the k+2 estimating equations simultaneously a two stage procedure could be used. For example, following Williams (1982) and Moore (1985, pp. 25-26)

1. Solve $\mathbf{f}_1$ using some initial estimator of ρ.

2. Obtain a new estimator of ρ using the new fitted values of p_{ij}.

3. Iterate until convergence.

The uv'th term in A_{11} is expressed as

$$\frac{1}{M}\sum_{i=1}^{k}\sum_{j=1}^{2}\sum_{s=1}^{m_{ij}}\left\{ r_{ijs}p_{ij}q_{ij}\ x_{iju}\ x_{ijv} \right\} \qquad (2.8.10)$$

where u,v=1,...,k+1 (Moore, 1986). When there is no variability in cluster size (i.e. $n_{ijs} = n$) it is possible to factor out $1 + (n-1)\rho$. In this case A_{11} is equal to the negative of the information matrix from ordinary logistic regression divided by the variance inflation factor, $1 + (n-1)\rho$. Parameter estimates in this case are also identical to those obtained by logistic regression. This result was proved by Williams (1982) and Moore (1986) and was earlier used in the context of Probit regression by Finney (1947).

Variance estimates are obtained by replacing parameters by estimates obtained by solving the generalized estimating equations. The validity of this approach for estimating variance is dependent upon the common correlation assumption. If, for example, the degree of intracluster correlation varies across strata then alternative estimators of variance are required.

Moore and Tsiatis (1991), following Liang and Zeger (1986), suggested replacing Λ_{11} by a robust estimator, i.e. an estimator which will be consistent deperding only on correctly specifying the terms in the logistic regression model. In a stratified cluster randomization trial the estimator described by Liang and Zeger (1986) and Moore and Tiatis (1991), denoted $\hat{\Lambda}_{11C}$, is a (k+1)x(k+1) covariance matrix whose uv'th element, u,v=1,...,k+1, is expressed as

$$\frac{1}{M}\sum_{i=1}^{k}\sum_{j=1}^{2}\sum_{s=1}^{m_{ij}}\left\{\hat{r}_{ijs}^{2}\,(\,\hat{p}_{ijs}-\hat{p}_{ij}\,)^{2}\,x_{iju}\,x_{ijv}\right\} \qquad (2.8.11)$$

where $\hat{p}_{ijs}$ is the observed risk in the s'th cluster from the i'th stratum and j'th treatment group while $\hat{p}_{ij}$ and $\hat{r}_{ijs} = n_{ijs} / [1 + (n_{ijs}-1)\hat{\rho}]$ are obtained by solving the estimating equations. The distribution of the k+1 subvector of parameter estimates $\hat{\theta}_1$ can then be expressed as

$$M^{1/2}\;(\,\hat{\theta}_1-\theta_1\,)\xrightarrow{D} MVN_{k+1}(\,0\,,\,\Gamma_{11}^{-1}\hat{\Lambda}_{11C}\Gamma_{11}^{-1}\,) \qquad (2.8.12)$$

$$\rightarrow MVN_{k+1}(\,0\,,\,\Lambda_{11}^{-1}\,)$$

asymptotically when the common correlation assumption is true.

2.8.4 Wald Type Inference Procedures

The first approach used to make inferences about the effect of treatment is an extension of a procedure first described by Wald (1943) for problems where the likelihood function is known. It employs $\hat{\gamma}$, the log odds ratio estimate and an estimate of its variance. The simpler of the two variance estimates is derived assuming that the correlation among cluster members is correctly specified.

This model dependent variance estimator can be found using model (2.8.1) and equation (2.8.10) to show that Λ_{11} can be expressed as

$$\begin{bmatrix} \mathbf{D} & \mathbf{c} \\ \mathbf{c}' & \mathbf{e} \end{bmatrix} \text{ where} \tag{2.8.13}$$

$$\mathbf{D} = \text{diag}(\ u_{1..}, \cdots, u_{k..}\ ),$$

$$\mathbf{c}' = (\ u_{12.}, \cdots, u_{k2.}\ ),$$

$$\mathbf{e} = u_{.2.} \quad \text{and}$$

$$u_{ij.} = p_{ij}\, q_{ij} r_{ij.}\ , \qquad r_{ij.} = \sum_{s=1}^{m_{ij}} r_{ijs}.$$

The kxk upper left-hand diagonal sub-matrix of Λ_{11} is the covariance matrix for the first k estimating equations, $f_1, \ldots, f_k$ while **c** contains elements of the covariance between f_u and f_{k+1}, u=1,...,k. Of primary interest is e = var(f_{k+1}) since f_{k+1} is used to estimate the effect of treatment.

The matrix Λ_{11} needs to be inverted to construct Wald or score type test statistics. The inverse of Λ_{11} can be found using the identity described by Searle (1982, p. 260) so that Λ_{11}^{-1} can be expressed as

$$\begin{bmatrix} \mathbf{D}^{i} & -\mathbf{c}^{i} \\ (\ -\mathbf{c}^{i}\ )' & \mathbf{e}^{i} \end{bmatrix} \tag{2.8.14}$$

$$\text{where } \mathbf{D}^1 = \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{c}(\, e - \mathbf{c}'\mathbf{D}^{-1}\mathbf{c}\,)^{-1}\mathbf{c}'\mathbf{D}^{-1},$$

$$\mathbf{c}^1 = \mathbf{D}^{-1}\mathbf{c}(\, e - \mathbf{c}'\mathbf{D}^{-1}\mathbf{c}\,)^{-1} \text{ and,}$$

$$e^1 = (\, e - \mathbf{c}'\mathbf{D}^{-1}\mathbf{c}\,)^{-1}$$

$$= \left\{ \sum_{i=1}^{k} \frac{u_{i1.}u_{i2.}}{u_{i..}} \right\}^{-1} .$$

A model dependent Wald type test of the null hypothesis that $\gamma = 0$ can thus be calculated as

$$\chi^2_{NW} = \hat{\gamma}^2 / \hat{\text{var}}_N(\, \hat{\gamma}\,) \tag{2.8.15}$$

$$= \hat{\gamma}^2 \sum_{i=1}^{k} \frac{\hat{u}_{i1.}\hat{u}_{i2.}}{\hat{u}_{i..}}$$

$$= \hat{\gamma}^2 \sum_{i=1}^{k} \left[\frac{1}{\hat{p}_{i1}\hat{q}_{i1}\,\hat{r}_{i1.}} + \frac{1}{\hat{p}_{i2}\hat{q}_{i2}\,\hat{r}_{i2.}} \right]^{-1}$$

where χ^2_{NW} is asymptotically distributed as a χ^2_1 random variable. If all clusters are the same size the model dependent test statistic χ^2_{NW} reduces to the Wald test obtained from ordinary logistic regression divided by the variance inflation factor $1 + (n-1)\hat{\rho}$.

Note that χ^2_{NW} is written in the same form as the non-iterative test statistic χ^2_{WO} , which was derived in Section 2.5 as an extension of Woolf's (1955) method which adjusted for clustering. Furthermore both these test statistics are equivalent to the maximum likelihood tesı statistic derived by Gart (1962) for uncorrelated binary outcome data obtained after replacing the observed cluster sizes, n_{ijs}, by the effective cluster sizes, $\hat{r}_{ijs}$.

It is also worth noting that the simpler non-iterative statistic (i.e. χ^2_{WO}) was derived by explicitly assuming ρ was known and then substituting in a consistent estimate of intracluster correlation. The asymptotic equivalence is thus a particular example of Moore's (1986) proof that the asymptotic covariance of $\hat{\theta}_1$ is unaffected by estimation of ρ. This asymptotic equivalence is analogous to the earlier asymptotic equivalences found by Gart (1962) for stratified analyses of uncorrelated binary outcome data.

There are two parts to the common correlation assumption which are usually believed to be needed if the model dependent test statistic, χ^2_{NW}, is to give valid type I errors. The first part of the assumption is that responses of all subjects within a cluster must be equally correlated. It arises as a direct consequence of using an exchangeable correlation matrix to describe the correlation between responses among cluster members. The intracluster correlation coefficient can, however, be derived as the average correlation among cluster members when all variables are measured at the cluster-level since then the expected value for responses of any two cluster members are equal (see Section 2.5). This type of misspecification of the correlation therefore only affects the asymptotic precision with which the parameters are estimated.

The second and more important part of the common correlation assumption concerns the requirement that ρ not vary between strata or across treatment groups. Randomization only assures equality of ρ across treatment groups under the null hypothesis. A robust Wald test statistic, denoted χ^2_{RW} could then be constructed using (2.8.11).

The robust test statistic is expressed as

$$\chi^2_{RW} = \hat{\gamma}^2 / \hat{\text{var}}_R(\ \hat{\gamma}\ ) \text{ where} \tag{2.8.16}$$

$\hat{\text{var}}_R(\ \hat{\gamma}\ )$ is the k+1,k+1 element of $\hat{\Gamma}_{11}^{-1}\hat{\Lambda}_{11C}\hat{\Gamma}_{11}^{-1}$ which is estimated by substituting parameter estimates in the formulae for $\Gamma_{11}^{-1}\hat{\Lambda}_{11C}\Gamma_{11}^{-1}$. Equation (2.8.11) and the formula for $\Gamma_{11}{}^{-1}=-\Lambda_{11}{}^{-1}$ can be used to prove that

$$\text{var}_R(\ \hat{\gamma}\ ) = ((c^1)',e^1)\ \ \hat{\Lambda}_{11C}\ \begin{bmatrix} c^1 \\ e^1 \end{bmatrix} \tag{2.8.17}$$

$$= K\ \text{var}_N(\ \hat{\gamma}\ ) \text{ where}$$

$$K = \sum_{i=1}^{k} W_i \frac{v_i}{\left[\hat{u}_{i1.}\hat{u}_{i2.} / \hat{u}_{i..}\right]^{-1}},$$

$$W_i = \frac{\hat{u}_{i1.}\hat{u}_{i2.} / \hat{u}_{i..}}{\sum_{i=1}^{k}\hat{u}_{i1.}\hat{u}_{i2.} / \hat{u}_{i..}} \quad \text{and}$$

$$v_i = \sum_{j=1}^{2} \frac{\sum_{s=1}^{m_{ij}} \left[\frac{\hat{f}_{ijs}}{\hat{f}_{ij.}} \left(\hat{f}_{ijs} - \hat{p}_{ij} \right) \right]^2}{\left(\hat{p}_{ij}\hat{q}_{ij} \right)^2}.$$

Thus the robust variance estimator of $\hat{\gamma}$ is equal to the model dependent variance estimator multiplied by the correction factor K which adjusts for misspecification of the correlation matrix.

The manner by which such corrections are made can be clarified by noting that the correction factor K is a weighted average of the ratio of v_i and $\left(\hat{a}_{i1.}\hat{a}_{i2.} / \hat{a}_{i..} \right)^{-1}$, i=1,....,k. The denominator of the i'th ratio, i.e. $\left(\hat{a}_{i1.}\hat{a}_{i2.} / \hat{a}_{i..} \right)^{-1}$, i=1,....,k, is equal to $\hat{\text{var}}\left[\log_e \left\{ \frac{\hat{p}_{i2}/\hat{q}_{i2}}{\hat{p}_{i1}/\hat{q}_{i1}} \right\} \right]$ under the common correlation assumption. The numerator is a robust estimate of the same variance. This can be demonstrated by noting that

$$\hat{\text{var}}\left[\log_e \left\{ \frac{\hat{p}_{i2}/\hat{q}_{i2}}{\hat{p}_{i1}/\hat{q}_{i1}} \right\} \right] \approx \sum_{j=1}^{2} \frac{\hat{\text{var}}(\hat{p}_{ij})}{\left(\hat{p}_{ij}\hat{q}_{ij} \right)^2} \tag{2.8.18}$$

$$\approx v_i$$

using the delta method and an adaptation of the ratio estimator of var($\hat{p}_{ij}$)

applied by Rao and Scott (1992) to the analysis of correlated binary outcome data. The robust estimate of var($\hat{p}_{ij}$) used in the correction factor K differs from variance estimates of ratio estimators by using modelled estimates of p_{ij}, by using the effective rather than the actual cluster sizes, and by assuming that $m_{ij}/(m_{ij}-1)=1$. The last difference is likely to yield estimates of variance which are too small in cluster randomization trials with few clusters compounding the bias noted by Cochran (1977, p. 162) for variance estimates of ratio estimators.

A second difficulty with robust variance estimates is that they will likely be less precise than model dependent variance estimates when the common correlation assumption holds. In this case the correction factor, K, would be equal to 1, on average, at least asymptotically. The robust variance estimator would then be less precise than the model dependent variance estimator owing to the need to estimate K.

Additional insight can be obtained by considering two special cases. The first special case arises when using a working correlation matrix which assumes independence among responses of cluster members (Liang and Zeger, 1986). Tests of hypotheses and confidence intervals could then be constructed using robust variance estimators. The advantage of this approach is that standard statistical packages can be used to fit correlated outcome data. For example, any computer program capable of fitting ordinary logistic regression models could be used to model correlated binary outcome data. Specialized programs would still be needed to obtain variance estimates. The only disadvantage of this approach is that parameter estimates will be asymptotically less precise than estimators obtained by correctly specifying the correlation among cluster members.

Parameter estimates obtained by assuming that responses among cluster members are independent arises as the special case of the exchangeable correlation matrix where ρ is set to zero. In this case the model dependent variance estimator, denoted $\mathrm{var}_N(\hat{\gamma})$, is identical to $\mathrm{var}_{OL}(\hat{\gamma}_{OL})$, the variance estimator obtained using ordinary logistic regression and the correction factor, K, now adjusts for the effect of clustering. The robust test statistic is therefore equal to the Wald test statistic calculated using ordinary logistic regression divided by a correction factor for the effect of clustering.

Suppose there is only a single stratum, i.e. a completely randomized design. Then the robust test statistic constructed using an independent working correlation matrix, is written in a form very similar to an approximation to Rosner's (1984) model described by Rosner and Milton (1988). The main difference is that Rosner and Milton (1988) assume a common correlation among responses of cluster members.

The second special case which will be considered is the simplification which occurs when all clusters in the trial are the same size. In this case

$$\mathrm{var}_R(\hat{\gamma}) = \mathrm{var}_N(\hat{\gamma})\, K \qquad (2.8.19)$$

$$= \mathrm{var}_{OL}(\ \hat{\gamma}_{OL}\ )[\ 1 + (n-1)\bar{\rho}_W]$$ since balance ensures that

$$\mathrm{var}_N(\ \hat{\gamma}\ ) = \mathrm{var}_{OL}(\ \hat{\gamma}_{OL}\ )[1+(n-1)\rho].$$

$$K = \sum_{i=1}^{k} W_i \frac{v_i}{\left[\hat{u}_{i1.}\hat{u}_{i2.}\ /\ \hat{u}_{i..}\right]^{-1}}$$

$$= \sum_{i=1}^{k} W_i^* \frac{\displaystyle\sum_{j=1}^{2} \frac{n \displaystyle\sum_{s=1}^{m_{ij}} (\ \hat{p}_{ijs} - p_{ij}\ )^2\ /\ m_{ij}p_{ij}q_{ij}}{nm_{ij}p_{ij}q_{ij}}}{\left[u_{i1}^* u_{i2}^*\ /\ u_{i.}^*\right]^{-1}\ [1 + (n-1)\rho]}$$

$$= \sum_{i=1}^{k} W_i^* \frac{1 + (n-1)\bar{\rho}_{i.}}{1 + (n-1)\rho}$$

$$= \frac{1 + (n-1)\bar{\rho}_W}{1 + (n-1)\rho}$$ where

$$W_i^* = \frac{u_{i1}^* u_{i2}^*\ /\ u_{i.}^*}{\displaystyle\sum_{i=1}^{k} u_{i1}^* u_{i2}^*\ /\ u_i^*}$$ and

$$u_{ij}^{*} = m_{ij}np_{ij}q_{ij}.$$

That is, under balance, the robust variance estimator equals the variance estimator for ordinary logistic regression multiplied by a variance inflation factor calculated using a weighted average of intracluster correlation coefficients over the 2k strata and treatment groups. In this case the only difference between model dependent and robust variance estimators is in how ρ is estimated. Since both estimators are weighted averages of stratum and treatment specific estimators of intracluster correlation there is no asymptotic difference. Therefore when all clusters are the same size the model dependent variance estimator is consistent even if the working correlation matrix is misspecified. This result also follows from the discussion in Section 2.5. A similar result has been noted by Dean (1993) for over-dispersed count data.

The estimates of ρ_{ij} obtained using the robust variance approach are likely to be too small since the moment estimators of ρ_{ij} are calculated using m_{ij} degrees of freedom for the between-cluster variability. Thus the size of tests of significance are likely to be too large and confidence intervals too narrow when there are small numbers of clusters in each treatment group and stratum. A similar problem has been noted by Thornquist and Anderson (1992) for robust variance estimators of coefficients from linear models.

Underestimation of the effect of clustering is likely to be most severe when the common correlation assumption holds and ρ is near zero. In this case negative estimates of ρ are likely to occur. Methods which assume a common correlation usually set such negative estimates to zero since ρ is assumed to be positive in cluster randomization trials. It is not as clear how such corrections could be effected using robust variance estimators. One possibility would be to set the correction factor, K, equal to one whenever smaller estimates are obtained. This approach could only be used when the working correlation matrix assumes independence among responses of cluster members.

2.8.5 Score Tests

An alternative approach which can be used to construct tests of the effect of treatment exploits the asymptotic normality of the estimating equations. The resulting statis al tests are extensions of the score tests first described by Rao (1947). Although score and Wald tests are asymptotically equivalent under the null hypothesis their small sample properties are quite different (Vaeth, 1985). These differences arise, in part, because variances are estimated under the alternative for Wald tests and under the null for score tests. The model dependent and robust score tests derived in this section are based on the discussions presented by Breslow (1990), Sharples (1989), and Sharples and Breslow (1992).

Kent (1982) discusses an alternative procedure which could be used to obtain robust score tests. Both procedures are compared by Boos (1992). This latter approach will not be discussed further since the test statistics for the effect of treatment in stratified cluster randomization trials obtained by either approach can be shown to be identical.

Under the null hypothesis that $\gamma = 0$ the first k estimating equations in (2.8.3) can be solved to yield

$$\hat{\rho}_i = \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \frac{\hat{f}_{ijs}}{\hat{f}_{i..}} \hat{\rho}_{ijs}. \tag{2.8.20}$$

Using these results the k+1'th estimating equation

$$f_{k+1} = f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\rho}) \tag{2.8.21}$$

$$= \sum_{i=1}^{k} \frac{\hat{f}_{i1.}\hat{f}_{i2.}}{\hat{f}_{i..}} (\hat{\rho}_{i2W} - \hat{\rho}_{i1W}) \text{ where}$$

$$\hat{\rho}_{ij}W = \sum_{s=1}^{m_{ij}} \frac{\hat{f}_{ijs}}{\hat{f}_{ij.}} \hat{\rho}_{ijs} \text{ , } j=1,2$$

and where the parameter estimates of $\alpha_1, \ldots, \alpha_k$ are calculated under the null hypothesis and some estimate of ρ is used to adjust for the effect of clustering.

The model dependent variance estimator of $f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\rho})$ is equal to the inverse of the variance estimator for $\hat{\gamma}$ (Breslow, 1990). This is just a special case of the variance estimator of score equations from fully parametric models (Cox and Snell, 1989, p. 182). That is

$$\text{var}_N(M^{-1/2} f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\rho})) = 1 / \text{var}_N(\hat{\gamma}) \tag{2.8.22}$$

$$= \sum_{i=1}^{k} \frac{u_{i1.}u_{i2.}}{u_{i..}}.$$

This variance can be estimated by setting $p_{ij} = \hat{p}_i$. In this case

$$\hat{u}_{ij} = \hat{p}_i\hat{q}_i\hat{r}_{ij.} \qquad \text{and so} \qquad (2.8.23)$$

$$\hat{var}_N(M^{-1/2}f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0|\hat{\rho})) = \sum_{i=1}^{k} \frac{\hat{r}_{i1.}\hat{r}_{i2.}}{\hat{r}_{i..}} \hat{p}_i\hat{q}_i.$$

The model dependent score test, denoted χ^2_{NS}, can therefore be expressed as

$$\frac{\left\{\sum_{i=1}^{k} \frac{\hat{r}_{i1.}\hat{r}_{i2.}}{\hat{r}_{i..}} (\hat{p}_{i2w}-\hat{p}_{i1w})\right\}^2}{\sum_{i=1}^{k} \frac{\hat{r}_{i1.}\hat{r}_{i2.}}{\hat{r}_{i..}} \hat{p}_i\hat{q}_i}. \qquad (2.8.24)$$

This test statistic is an iterative version of the adjusted Mantel-Haenszel statistic, denoted χ^2_{MHO} which was derived in Section 2.5. Both statistics were derived using identical assumptions and differ only in how the parameters are estimated.

Furthermore both test statistics reduce to Cochrans's (1954) version of the Mantel-Haenszel statistic (Mantel and Haenszel, 1959) divided by the variance inflation factor, $1 + (n-1)\hat{\rho}$, when there is no variability in cluster size. This simplification occurs because the estimating equations can then be solved analytically under the null hypothesis yielding $\hat{p}_i = \sum_{j=1}^{2} \sum_{s=1}^{m_{ij}} \hat{p}_{ijs} / m_{i.}$. The relationship between Mantel-Haenszel methods and score tests from logistic regression were previously noted by Day and Byar (1979) in the context of uncorrelated binary outcome data.

The robust variance estimator for

$f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\beta})$, denoted $\mathrm{var}_R(M^{-1/2} f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\beta}))$,

can be directly obtained by substituting $\hat{\Lambda}_{11C}$ and terms from $\Lambda_{11}{}^{-1}$ into equation (9) of Breslow (1990). That is $\mathrm{var}_R(M^{-1/2} f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\beta}))$ can be expressed as

$$(-c'D^{-1} \quad 1)\hat{\Lambda}_{11C}\begin{bmatrix} -D^{-1}c \\ 1 \end{bmatrix} \tag{2.8.25}$$

$$= K \, \mathrm{var}_N(M^{-1/2} f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\beta}))$$

$$= K \sum_{i=1}^{k} \frac{r_{i1.} r_{i2.}}{r_{i..}} p_i q_i$$

where K is the correction factor which adjusts for the effect of misspecifying the working correlation matrix.

Estimates of $\mathrm{var}_R(M^{-1/2} f_{k+1}(\hat{\alpha}_1, \cdots ,\hat{\alpha}_k,\gamma=0 | \hat{\beta}))$ are calculated in the same manner as for the model dependent variance estimator. The correction factor, K, is now estimated by setting $p_{ij} = \hat{p}_i$ yielding

$$Ko = \sum_{i=1}^{k} \hat{W}_i \frac{\hat{v}_i}{\hat{p}_i\hat{q}_i \left[\dfrac{1}{\hat{r}_{i1.}} + \dfrac{1}{\hat{r}_{i2.}} \right]} \quad \text{where} \tag{2.8.26}$$

$$\hat{W}_i = \frac{\dfrac{\hat{r}_{i1.}\hat{r}_{i2.}}{\hat{r}_{i..}} \; \hat{p}_i\hat{q}_i}{\displaystyle\sum_{i=1}^{k} \frac{\hat{r}_{i1.}\hat{r}_{i2.}}{\hat{r}_{i..}} \; \hat{p}_i\hat{q}_i} \quad \text{and}$$

$$\hat{v}_i = \sum_{s=1}^{m_{ij}} \left[\frac{\hat{r}_{ijs}}{\hat{r}_{ij.}} (\hat{p}_{ijs} - \hat{p}_i) \right]^2 .$$

Therefore the robust score test, denoted χ^2_{RS}, can be expressed as

$$\chi^2_{NS} \,/\, Ko \tag{2.8.27}$$

In balanced trials estimates of the correction factor Ko can be expressed as (2.8.28)

$$Ko = \sum_{i=1}^{k} \frac{\dfrac{m_{i1}m_{i2}}{m_{i.}} \; \hat{p}_i\hat{q}_i}{\displaystyle\sum_{i=1}^{k} \frac{m_{i1}m_{i2}}{m_{i.}} \; \hat{p}_i\hat{q}_i} \; \frac{\displaystyle\sum_{j=1}^{2} \frac{\dfrac{1}{m_{ij}}}{\dfrac{1}{m_{i1}} + \dfrac{1}{m_{i2}}} \sum_{s=1}^{m_{ij}} \frac{n\,(\hat{p}_{ijs}-\hat{p}_i)^2}{m_{ij}\hat{p}_i\hat{q}_i}}{1 + (n-1)\hat{\rho}}$$

$$= \sum_{i=1}^{k} \frac{\dfrac{m_{i1}m_{i2}}{m_{i.}}\hat{p}_i\hat{q}_i}{\displaystyle\sum_{i=1}^{k} \frac{m_{i1}m_{i2}}{m_{i.}}\hat{p}_i\hat{q}_i} \quad \frac{1 + (n-1)\hat{\rho}_i}{1 + (n-1)\hat{\rho}}$$

$$= \frac{1 + (n-1)\hat{\rho}_{W_o}}{1 + (n-1)\hat{\rho}}.$$

The robust test statistic is now, again, equal to Cochran's (1954) version of the Mantel-Haenszel test statistic divided by a variance inflation factor. The variance inflation factor uses a weighted average of intracluster correlation coefficients calculated under the null hypothesis ensuring greater precision than was the case for the robust Wald test stat'^tic under balance, at least when the null hypothesis is true. Therefore the asymptotic requirements of the robust score test are likely less stringent than the requirements for the robust Wald test.

Robust score tests could also be developed using a working correlation matrix which assumes that responses of cluster members are independent, as was described for robust Wald tests in the previous section. Such test statistics are a special case of χ^2_{RS} calculated by setting $\rho = 0$. If there is only a single stratum χ^2_{RS} would then reduce to the test statistic derived by Boos (1992, Example 7).

2.9 Summary of Methods

Four principal results were obtained in this chapter. First, an approximate equivalence was demonstrated for test statistics when there are the same number of subjects in each cluster and the same number of clusters in each stratum. Second, the common correlation assumption was shown to be overly restrictive. Third, arguments were put forward suggesting that the methods which use robust estimates of variance would require larger numbers of clusters per treatment group to ensure their validity than other approaches. Fourth, tests of significance constructed using the generalized estimating equations approach were shown to be relatively simple algebraic extensions of standard procedures.

The approximate equivalence between test statistics in completely balanced designs includes methods as seemingly diverse as stratified t-tests constructed using cluster-level estimates of risk, Mantel's (1963) extended Mantel-Haenszel test statistic, simple adjustments of standard procedures such as Donald and Donner's (1987) adjusted Mantel-Haenszel test statistic, and more complicated iterative techniques based on extensions of logistic regression which account for the effects of clustering (e.g. Moore and Tsiatis (1991)). Equivalence occurs because in completely balanced designs these procedures are asymptotically equal to test statistics constructed under the assumption that the responses of cluster members are independent divided by the variance inflation factor $1+(n-1)\bar{\rho}$ where $\bar{\rho}$ denotes the average degree of correlation between cluster members.

A consequence of the equivalence between test statistics is that the distinction between cluster-level and individual-level analyses seems somewhat arbitrary, especially in completely balanced designs. This result mirrors the earlier discussion in Section 1.4.1 concerning the algebraic equivalence between cluster-level and individual-level analyses of correlated continuous outcome data.

The equivalence between cluster-level and individual-level analyses also suggests that the use of the t or F distribution to determine the statistical significance of s .atified t-tests and of χ_1^2 to determine the statistical significance of all other test statistics is somewhat arbitrary. The choice will of course have no effect in cluster randomization trials in which there are large numbers of clusters per treatment group but can be important in smaller trials.

This is not a completely novel idea. For example, Williams (1991) and McCullagh and Nelder (1983) have considered using the F distribution to determine the statistical significance of test statistics obtained from logistic regression models adjusted for the effect of clustering to account for the degrees of freedom used t estimate ρ .

An additional consequence of the asymptotic equivalence between test statistics is that there is no reason to use robust test statistics in completely balanced stratified cluster randomization trials. Robust methods were developed to avoid having to make possibly incorrect assumptions about the degree of correlation between responses of cluster members. This is not a concern in completely balanced designs since then variance estimates of the effect of treatment are a function of the average intracluster correlation coefficient from each stratum and treatment group. In this case robust variance estimates described by Liang and Zeger (1986) and Rao and Scott (1992) will likely be less precise since they allow for variation in the degree of dependence among responses of cluster members across strata and treatment groups.

Assumptions about the degree of correlation between responses of cluster members are only required when there is variability in cluster size. Several approaches begin by imposing the common correlation assumption. This assumption is satisfied if the correlation between any two cluster members is fixed and if this correlation does not vary across clusters, strata or treatment groups. It was shown in Sections 2.5 and 2.8 that this assumption is overly restrictive. The validity of methods such as Donald and Donner's (1987) adjusted Mantel-Haenszel test statistic is assured if at least the average correlation in each cluster is fixed. This same assumption is sufficient to assure the validity of model dependent test statistics constructed using Liang and Zeger's (1986) generalized estimating equations approach, at least in the absence of baseline risk factors measured at the level of the individual (e.g. age, sex).

The imprecision of the robust methods is not limited to completely balanced designs. Randomization assures ρ will be identical in both treatment groups under the null hypothesis, although it may vary across strata. Therefore test statistics constructed using robust variance estimates which ignore this feature of stratified cluster randomization trials will likely require a greater number of clusters per treatment group to ensure that rejection rates will be near nominal levels. Simulation studies will be used to determine how great this effect will be. Conversely methods which do not allow ρ to vary across strata may be invalid, even in very large trials.

The identification of the assumptions required to assure the validity of the model dependent test statistics and the insight gained into the potential imprecision of the robust test statistics were enhanced by the algebraic work presented in Section 2.8. In this section of the thesis the model dependent test statistics were shown to be asymptotically equivalent to relatively simple extensions of either Cochran's (1954) version of the Mantel-Haenszel test statistic (Mantel and Haenszel, 1959) or Woolf's (1955) test statistic. Robust test statistics were shown to equal the product of model dependent test statistics and a correction factor which accounts for misspecification of the correlation between responses of cluster members. Correction factors are a function of variance estimates constructed using the theory of ratio estimation. Such variance estimates are known to be imprecise when there are few clusters per treatment group (Cochran, 1977, p. 162).

3. A Simulation Study of Tests of Significance from Stratified Cluster Randomization Trials

3.1 Introduction

The discussion of tests of significance in the previous chapter was limited to algebraic comparisons of the different approaches which could be used to analyze data collected from stratified cluster randomization trials. Algebraic comparisons are, however, not helpful alone in identifying the factors affecting the type I error rates or in determining the power of tests of significance. These characteristics of test statistics are needed to provide epidemiological researchers with advice on how to analyze their data and can best be determined using finite-sample simulation studies.

Simulation studies are experiments performed on computers in which data are generated to approximate the properties of random variables likely to arise in practice. The Type I error rates and power of tests of significance can then be calculated as a function of the parameters of these pseudo-random variables. Estimates of rejection rates are calculated by repeating each experiment several hundred times and calculating the relative frequency with which the null hypothesis is rejected for a particular test statistic. Simulation studies can also be helpful in suggesting additional avenues of research.

The results of a simulation study designed to compare the small sample properties of test statistics which could be used in stratified cluster randomization trials are discussed in this chapter. Simulations were performed using both random and fixed cluster sizes. The average cluster size in each iteration of the simulation was set to 100 which is approximately equal to the size of clusters in trials which have randomized schools (e.g. Murray et al, 1992).

Attention is restricted to eleven tests of significance which were selected as representative of the methods discussed in the previous chapter. These tests are applicable in trials corresponding to Hauck's (1989) second asymptotic case in which there are few strata and many clusters per stratum. This case was selected because it corresponds most closely to how stratified cluster randomization trials are designed in practice.

The methods include the stratified t-test (Schwartz et al, 1980), the extended Mantel-Haenszel χ^2 test (Mantel, 1963), Donner and Donald's (1987) adjusted Mantel-Haenszel test, Rao and Scott's (1992) version of the Mantel-Haenszel test, a likelihood ratio test constructed using the beta-binomial distribution, and four tests derived using Liang and Zeger's (1986) generalized estimating equations approach with an exchangeable working correlation matrix. These four test statistics are distinguished by their use of either model dependent or robust variance estimates and by being either Wald or score tests. The classical Mantel-Haenszel test statistic was also included to demonstrate the effect of falsely assuming that cluster members' responses are independent.

The chapter is divided into eight sections, including the introduction. The next two sections contain the objectives and rationale for the simulation study. A detailed discussion of the parameters used to define the study is presented in Section 3.4. The methods used to generate the pseudo-random variables are discussed in Section 3.5 while the different test statistics being evaluated are reviewed in Section 3.6. Finally the results of the simulation are presented and discussed in the last two sections of the chapter.

3.2 Objectives

This simulation study was designed to compare methods of testing the effect of treatment in stratified cluster randomization trials. There were 3 specific objectives.

1. To determine the number of clusters required per treatment group to obtain valid type I error rates.
2. To compare the power of the test statistics.
3. To examine how robust the methods are to violations of the assumption that the average correlation between responses of cluster members is fixed.

These three objectives were selected to allow recommendations to be made regarding the choice of test statistics to be used when analyzing data obtained from stratified cluster randomization trials. Attention was restricted to trials where there were approximately 100 subjects per cluster on average while varying the number of strata, the range of risk across strata, the degree of imbalance in cluster size and the number of clusters per treatment group and stratum. Such cluster sizes are typical of school based trials (e.g. Murray et al, 1992) although subsampling from larger intact social units (e.g. physicians' practises) could result in clusters of similar size.

All of the test statistics being compared are approximate methods so that rejection rates for finite samples are best determined by simulation. The first objective will allow determination of the minimum number of clusters per treatment group required to obtain valid Type I error rates.

The sample size needed for obtaining valid type I error rates may not be the same for all methods. It was postulated in the previous chapter that the validity of type I error rates of methods using robust variance estimators (i.e. Rao and Scott's (1992) ratio estimator approach and Liang and Zeger's (1986) robust test statistics) require a greater number of clusters per treatment group than do other approaches. The possibility that some test statistics have more stringent sample size requirements than do other procedures was examined by varying both the number of clusters in each treatment group and stratum and the number of strata.

For the first two objectives the data were generated so that the degree of correlation between cluster members was the same for each cluster satisfying the common correlation assumption. In the previous chapter it was shown that the validity of methods such as the adjusted Mantel-Haenszel test statistic (Donald and Donner, 1987) depends only upon the less restrictive assumption that the average correlation between responses of cluster members is fixed. The same result was shown to hold for the model dependent test statistics derived using Liang and Zeger's (1986) generalized estimating equations, at least when all variables are measured at the cluster level.

The effect of allowing the average degree of correlation within a cluster to vary was examined by generating the data according to "Smith's Law" (Proctor, 1985) which states that the degree of intracluster correlation is inversely related to cluster size. These simulations will help to develop a sense of how robust these statistical tests are to violations of their underlying assumptions.

3.3 Rationale

Numerous simulation studies comparing the properties of statistical methods used in the analysis of correlated binary outcome data have been published. Many of these are summarized in the annotated bibliography put together by Ashby et al (1992).

The present simulation differs from these earlier studies in four ways. Firstly, very few of these earlier studies focused on the small sample properties of inferences concerning the effect of treatment in stratified cluster randomization trials (e.g. Donald and Donner, 1990). Rather their focus has been on completely randomized (e.g. Donner et al (1993), Kupper et al (1986)) or matched pairs designs (e.g. Donner and Donald (1987), Donner and Hauck (1989), Liang (1985)) or have only been interested in individual-level covariates (Smith (1993), Weerasekera and Bennett (1992), Wickens (1993)).

A second difference is that most of the earlier simulation studies have been concerned with study designs in which there are few subjects per cluster and where the degree of intracluster correlation is likely to range between 0.2 and 0.8 (e.g. Donner et al (1993), Kupper et al (1986)). The present simulation, however, is concerned with trials where fairly large clusters are randomized and where the degree of intracluster correlation is small (i.e. less than or equal to 0.1). The asymptotic requirements of test statistics might well be different in this latter situation.

The third difference between this simulation and earlier studies is that the effect of variable cluster sizes was examined. Cluster sizes were generated using the truncated negative binomial distribution described by Donner and Koval (1987). This approach allows cluster sizes to be randomly generated with a specified degree of imbalance. Most earlier simulation studies restricted attention to situations where cluster sizes were fixed (e.g. Donald and Donner (1990)) or generated cluster sizes using empirically determined distributions (e.g. Fung et al (1993)).

The final distinct feature of the simulation studies described in this chapter is the examination of the effect of violating the assumption that the average correlation between the responses of cluster members is fixed. Most earlier simulation studies have only generated data under the common correlation assumption. Williams (1988a), one of the few researchers who was interested in violations of this assumption, limited his examination to the effect on a beta-binomial model which incorrectly assumed that the degree of correlation was fixed across treatment groups.

3.4 Parameters used to Generate the Random Variables

3.4.1 Type I Error Rates Under the Common Correlation Assumption

The number of parameters used in the design of the simulation study varied for each of the three objectives. There were 96 different parameter combinations used to examine the first objective which concerns the minimum number of clusters required to obtain valid rejection rates for the different test statistics. These 96 different parameter combinations arose by varying five different parameters: the number of strata, the number of clusters per treatment group, the degree of imbalance in cluster size, the range of risk across strata, and the degree of intra-cluster correlation.

Data were generated using either two or four strata and 20, 40 or 80 clusters per treatment group. The same number of clusters were used in each treatment group and stratum. For example in simulations using two strata and 20 clusters per treatment group there were m = 10 clusters in each of the four treatment group by stratum combinations, using the notation encountered in the previous chapter. This combination of the number of strata and the number of clusters was selected to determine if methods which use robust estimates of variance have more stringent asymptotic requirements than other approaches.

The smallest number of clusters per treatment group is approximately equivalent to the number of clusters used in many of the trials examined by

Donner et al (1990) in their review of cluster randomization trials. Larger numbers of clusters were required to assure the validity of some test statistics. Few trials would, of course, have as many as 80 clusters per treatment group. A notable exception are the trials of vitamin A supplementation described in a meta-analysis by Fawzi et al (1993). There were at least 100 clusters per treatment group in half of the eight community intervention trials which they reviewed.

There tends to be considerable variability in cluster size in most cluster randomization trials. Donner and Koval (1987), following the arguments put forward by Ahrens and Pincus (1981), suggested measuring such variability using a statistic denoted by

$$\hat{\kappa} = \frac{1}{1 + CV^2} \qquad (3.1)$$

where CV is the coefficient of variation for cluster size. This statistic equals one when there is no variability in cluster size and decreases as imbalance in cluster size increases.

Values of $\kappa = 0.8$ were used to generate the variable cluster sizes employed to study the empirical rejection rates. This degree of imbalance corresponds to the variability in cluster size found using data from two recent cluster randomization trials. Gyorkos et al (1989) conducted a cluster randomization trial in which families were the unit of randomization while Murray et al (1992) reported on a trial in which schools were randomized to treatment and control groups. The data from these two trials are analyzed in Chapter 4. An idea of the degree of variability in cluster size corresponding to $k = 0.8$ can be obtained by examining the data displayed in Tables 4.1 and 4.2.

Simulations were also performed in which all clusters had exactly 100 subjects In Chapter 3 it was shown that all of the test statistics used in the simulation study are asymptotically equivalent when there are the same number of subjects in each cluster. It is therefore of some interest to determine how many clusters are required to assure equal performance among the test statistics. i.e. to determine the minimum number of clusters per treatment group required to assure the application of asymptotic results.

The proportion of affected subjects in a stratum was determined using either a narrow or wide risk range. The narrow risk range varied from 40 to 60 percent while the wide band allowed risk to vary from 30 to 70 percent. Values of risk were equally spaced across strata. Similar levels of risk were noted by Best et al (1984 Figure 3) for various adolescent smoking behaviors (e.g. percent of students who had never smoked) and by Gyorkos et al (1989) for risk of parasite infection.

Four different values of the intracluster correlation coefficient (i.e. $\rho = 0, 0.025, 0.050, 0.100$) were used to generate the proportion of affected individuals per cluster. Binomially distributed data (i.e. $\rho = 0$) were included to examine how the test statistics behaved when the responses of cluster members were independent as well as to help identify any patterns which might occur as the degree of correlation approaches zero. The other values were selected to be representative of correlation coefficients obtained from trials in which there are about 100 subjects per cluster as well as to examine the effect of increasing degrees of correlation on the empirical type I error rates.

There are very few published estimates of intracluster correlation representative of studies in which there are 100 subjects per cluster. Estimates of intracluster correlation between 0.01 and 0.02 were calculated for the prevalence of smoking using data collected by Murray et al (1992) and LaPrelle et al (1992). There were, on average, approximately 140 subjects per cluster in these studies.

Estimates of comparable size were also found by interpolation using "Smith's Law" (Hansen, Hurwitz and Madow (1953, p. 309), Proctor (1985)). This empirical law can be expressed as

$$\rho = A\ n^{-B} \qquad (3.2)$$

where A and B are positive constants and n denotes cluster size. Data needed to fit this model were provided by Donner (1982) who calculated intracluster correlation coefficients for four different outcomes (i.e. the prevalence of hypertension, smoking, drinking and obesity) and three differe t types of clusters (i.e. spouse pairs, physicians practises, counties). Parameter estimates for this model were obtained by regressing the log of the intracluster correlation coefficient against the log of the average number of people per cluster.

It is less common for correlations much larger than 0.05 to be found in social units with as many as 100 members. Humphreys and Carr-Hill (1991 Table 2), however, did find correlations near 0.10 for a variety of health outcomes using random samples of subjects from 206 electoral wards in the United Kingdom.

3.4.2 Power Comparisons Under the Common Correlation Assumption

The same set of parameters used to determine the type I error rates of the test statistics were also used to compare their power. Comparisons were however limited to trials in which there were 40 clusters per treatment group. It was not possible to perform such comparisons when there were 20 clusters per treatment group since then all of the statistical methods being compared tended to be overly liberal (i.e. they rejected the null hypothesis too often when the null hypothesis is true). The empirical rejection rates for both the classical Mantel-Haenszel χ^2 test statistic and Rao and Scott's (1992) version of this statistic were not tabulated for the same reason. The classical Mantel-Haenszel χ^2 test statistic was always overly liberal when $\rho > 0$ while the rejection rates for the test statistic constructed using Rao and Scott's (1992) approach was often greater than the nominal five percent level even when there were 40 clusters per treatment group (see Tables 3.1-3.4).

Power was determined for detecting a common odds ratio of 1.5 as statistically significant at $\alpha = 0.05$ (two-sided). Odds ratio of this size correspond to small effects in each stratum using the definition of effect size put forward by Cohen (1988, pp. 180-185) for uncorrelated binary outcome data. This effect size was selected because small effects are typical in preventive trials, particularly those which attempt to affect behavioral changes among participants. As noted earlier, many cluster randomization trials are in fact preventive trials (Donner et al, 1990). An odds ratio of 1.5 was used previously by Donner and Donald (1987) to compare the power of statistical tests used in pair-matched cluster randomization trials. These authors were also interested in trials in which there were about 100 subjects per cluster.

3.4.3 Intracluster Correlation as a Function of Cluster Size

The type I error rates and power of the different test statistics were first determined by generating the number of positive responses in a cluster under the common correlation assumption. The effect of relaxing this assumption was investigated by allowing the degree of correlation between cluster members to vary as a function of cluster size, as suggested by "Smith's Law". This was accomplished by generating the random number of positive responses per cluster from a different beta-binomial distribution separately for each cluster size and as a function of the expected risk in each stratum and treatment group. Parameters were selected to reflect the relationship between cluster size and intracluster correlation described by Donner (1982) and to ensure that $\rho = 0.025$ when there were 100 subjects in a cluster.

Four combinations of the parameters A and B in equation (3.2) were used. The four pairs are (0.025, 0.0), (0.1, 0.3010), (0.5, 0.6505), and (0.9, 0.7782). The first pair of parameters fixes the degree of correlation at 0.025 for all clusters and was included to serve as a baseline comparison. The variation in the degree of intracluster correlation as a function of cluster size was greatest when A=0.9 and B=0.7782. Examples of the variation in the degree of intracluster correlation for clusters of three different sizes are presented in Table 3.7 for each of the A,B parameter pairs.

It is quite common to stratify by cluster size in cluster randomization trials. To correspond to this design feature the average cluster size was varied across strata. This also had the effect of varying the average degree of intracluster correlation across strata.

There were 55 subjects per cluster, on average, in the stratum where the risk was lowest while in the stratum with the highest risk there were, on average, 145 subjects per cluster. Average cluster size varied evenly between these two extremes in the remaining strata. Therefore there were still 100 subjects per cluster, after averaging across strata.

Some reduction in the variability of cluster size would likely result in practice when stratifying by average cluster size, i.e. stratification would not be "perfect". Therefore cluster sizes were generated by setting the imbalance parameter, κ equal to 0.9.

$\sum_{i=1}^{k} w_i R_{i1.}$ where w_i is the i'th stratum weight and $R_{i1.}$ is the sum of the ranks for clusters from the control group. van Elteren (1960) proposed two statistics using the weights $1/(m_{i.} + 1)$ and $1/(m_{i1}m_{i2})$. Tests using the first weight have locally optimal properties when there are few strata and many clusters per stratum and treatment group (van Elteren, 1960). Statistics using this weight are also discussed in detail by Lehmann (1975 pp. 135-138). Tests using the latter weight are more appropriate as the number of strata gets large. The test statistic suggested by Shirley (1987) differs from van Elteren's (1960) in selecting identical weights for all strata while Mantel (1963) and Fry and Lee (1988) argue for the use of weights proportional to the inverse of the number of clusters per stratum. All these methods are equivalent when there are the same number of clusters in each stratum and treatment group.

Exact tests for these stratified rank statistics are likely to be nearly as computationally intensive as more powerful tests using the untransformed cluster risks. This difficulty occurs because the high probability of ties would require calculating statistical significance separately for each data set rather than being able to construct tables of p-values as is commonly done for rank transformed data. Even if there were no ties tables of the distribution might not be practical to construct since the distribution of the test statistic is a function of the number of strata in addition to the number of clusters per stratum and treatment group (Lehmann, 1975 p. 135). Separate tables would then have to be constructed for

3.5 Generation of the Pseudo-Random Data

There are three approaches which have been used to generate variable cluster sizes for simulation studies of correlated binary outcome data. The simplest approach is to deterministically select variable cluster sizes. This approach was used by Donner and Donald (1987) in their simulation study which was designed to compare tests of the effect of treatment in pair-matched cluster randomization trials. Three levels of imbalance were used. If there were 120 subjects per cluster the study was said to be balanced. A mildly imbalanced study had clusters of 60, 120 or 180 subjects while severely imbalanced studies had 20, 120 or 220 subjects per cluster. The principal attraction of this approach is that the degree of variability and mean cluster size can be varied to suit particular problems.

An alternative approach is based on taking random samples from distributions of cluster sizes obtained from earlier studies. This approach has been primarily used by researchers investigating methods applicable to dose response modelling in teratology (e.g. Kupper et al, 1986). The results of simulation studies which use this approach are more generally applicable than when the variable cluster sizes are restricted to only a few predetermined values. Of course it is no longer possible to vary either the mean cluster size or the degree of variability in cluster size when this technique is used.

The third approach combines the advantages of both these methods. It requires specification of a parametric distribution from which cluster sizes can be randomly sampled. Both the average cluster size and the degree of imbalance can be varied to suit particular problems while not restricting clusters to, a few, predetermined sizes.

Donner and Koval (1987) suggested using a negative binomial distribution truncated below one to model cluster sizes. This distribution can be expressed as

$$P(n) = \frac{(s+n-1)!\ (1+R)^{-s}\ (R/[1+R])^{n}}{(s-1)!\ n!\ (1-[1+R]^{-s})},\quad n=1,2... \tag{3.3}$$

where the mean and imbalance parameters can be expressed as

$$E(n) = sR\ /\ (\ 1 + [1+R]^{-s}\ )\ \text{ and} \tag{3.4}$$

$$\kappa = E(n)\ /\ (\ 1 + R + sR\ )\ \text{ respectively.} \tag{3.5}$$

Values of s (s>0) and R (R>0) for given mean and imbalance parameters of cluster size can be obtained by solving the two nonlinear equations (3.4) and (3.5).

The truncated negative binomial distribution was earlier shown to fit the distribution of family sizes in a variety of countries by Brass (1958) and has also been used to generate litter sizes in a recent simulation study which compared methods used for testing homogeneity of proportions in teratologic studies (Donner et al, 1993). This model was also used to generate random cluster sizes for the simulation study discussed in this chapter of the thesis.

The number of affected individuals in a cluster were generated from a beta-binomial distribution, conditional on the cluster size. This distribution has been used in a number of simulation studies which compared methods of analysis for correlated binary outcome data (e.g. Kupper et al (1986), Paul et al (1989)). The distribution is described in greater detail in Section 2.7.

The cluster sizes and the number of positive responses per cluster were generated using the cumulative distribution function algorithm described by Kennedy and Gentle (1980, p. 177). This technique is implemented by first generating a random number between 0 and 1 from a uniform distribution. A pseudo-random number from any specified distribution can then be obtained as the p'th percentile from the relevant cumulative distribution function. Pseudo-random numbers from a uniform (i.e. u(0,1)) distribution were generated using an adaptation of Wichman and Hill's (1982) procedure.

Random variables from a truncated negative binomial distribution include all integers greater than or equal to one. The cumulative distribution function algorithm can however only be implemented by specifying an upper limit. No clusters were generated which were greater than or equal to 500. This likely had little effect since there is only a 0.0001497 percent chance of obtaining a cluster larger than 500 from a truncated negative binomial distribution with mean cluster size of 100 and imbalance parameter of $\kappa = 0.80$.

A continuing problem with the design of many simulation studies is that little justification is given for the number of iterations used (Robey and Barcikowski, 1992). In this simulation study 500 iterations were used to estimate rejection rates for each combination of parameters. This number of iterations was selected so that the approximate 95% confidence limits for a five percent rejection rate were (0.031,0.069). Therefore statistical tests which have empirical type I error rates less than 0.03 can be considered to be overly conservative while tests which have empirical type I error rates greater than 0.07 are overly liberal (Bradley, 1978).

All of the computer programs used for this simulation study were written in FORTRAN 77 and run on a UNIX based network of SUN Workstations. Much of the code was adapted from earlier programs written by colleagues in the Department of Epidemiology and Biostatistics at the University of Western Ontario.

3.6 Test Statistics

The 11 test statistics which are being evaluated in this chapter were selected as being representative of the procedures discussed in the previous chapter. Ten of these test statistics adjust for the effect of clustering. These ten procedures include the stratified t-test (Schwartz et al, 1980), the extended Mantel-Haenszel χ^2 test (Mantel, 1963), Donald and Donner's (1987) adjusted Mantel-Haenszel test, Rao and Scott's (1992) version of the Mantel-Haenszel test, a likelihood ratio test constructed using the beta-binomial distribution, and four tests derived using Liang and Zeger's (1986) generalized estimating equations approach.

The classical Mantel-Haenszel test statistic was also included to demonstrate the effect of ignoring the effect of clustering on the empirical rejection rates. This test statistic is denoted CMH in the tables displaying the results of the simulation study.

Two of the ten test statistics which adjust for the effect of clustering use the cluster as the unit of analysis. These two methods are the stratified t-test discussed by Schwartz et al (1980, pp. 189-191) and the extended Mantel-Haenszel test statistic described by Mantel (1963) and discussed in detail in Section 2.4. They are denoted TST and EMH respectively. The stratified t-test is a special case of equation (2.3.6), found in Section 2.3 of this thesis, which arises when outcomes are binary and there are m clusters in each stratum and treatment group.

It was pointed out in Section 2.4 that these two test statistics differ primarily in that the sample variance is calculated under the null hypothesis for the extended Mantel-Haenszel test and under the alternative hypothesis for the stratified t-test. An additional difference arises in how the statistical significance for these two test statistics is calculated. The stratified t-test was declared to be statistically significant at the 5% level if the absolute value of the test statistic was greater than the 97.5'th percentile from a $t_{2k(m-1)}$ distribution. Extended Mantel-Haenszel tests, and each of the remaining test statistics were declared to be statistically significant at the 5% level if the test statistic was greater than 3.841, the 95'th percentile of a χ_1^2 distribution. Thus all these procedures employ two-tailed tests of significance.

Several authors have proposed using the F rather than the χ^2 distribution to determine critical values for statistical tests constructed using correlated binary outcome data (Collett (1991 pp. 196-204), Lipsitz et al (1991), Williams (1991)). The rationale for this, somewhat ad hoc approach, is to account for the degrees of freedom used to estimate the coefficient of intracluster correlation.

This approach was not followed for the test statistics used in this simulation study and probably would have had little effect if it had. There were always at least 32 degrees of freedom (= 4 strata x 2 treatment groups x (5-1) clusters in each stratum and treatment group) available to estimate ρ. The critical value used to construct an F test with 1 and 32 degrees of freedom would be 4.149 if the α error were 5 percent. This is not much larger than 3.841, the equivalent χ_1^2 critical value.

The next two test statistics being compared by simulation are two versions of the adjusted Mantel-Haenszel test statistic described by Donald and Donner (1987). These two versions of the adjusted Mantel-Haenszel test statistics are distinguished by how the intracluster correlation coefficient is estimated. Following Donald and Donner (1987) the first version of this test statistic, denoted MH1, is calculated using the average of the 2k ANOVA estimators from each stratum and treatment group. The second version of the adjusted Mantel-Haenszel test statistic, denoted MH2, is calculated using an ANOVA estimator obtained using the variance components from a mixed-effects linear model with terms for the effect of the strata, the treatment and their interaction. This is the statistic denoted $\hat{\rho}_S$ described in Section 2.3.

These two ANOVA estimators of intracluster correlation were shown to be asymptotically equivalent in Section 2.3, at least in the special case where there are exactly n subjects in each cluster. It was also hypothesized in this section of the thesis that $\hat{\rho}_S$ would be the more precise estimator since it is uses more efficient weights. Both versions of the adjusted Mantel-Haenszel test statistic were included to determine if the hypothesized greater precision of $\hat{\rho}_S$ affects either the type I error rates or the power to reject the null hypothesis.

Rao and Scott (1992) suggested an alternative approach which could be used to adjust the Mantel-Haenszel test statistic for the effect of clustering. Their adjusted Mantel-Haenszel test statistic, denoted RSM, is presented in equation (2.6.5) of Section 2.6 of this thesis. The adjustment is based on estimating the effective number of independent observations in each stratum and treatment group using the theory of ratio estimation. The classical Mantel-Haenszel test statistic can then be calculated using these adjusted data. Such test statistics are robust in the sense that they make no assumptions about the degree of correlation between responses of cluster members.

The seventh test statistic used in the simulation study is a likelihood ratio test constructed using the beta-binomial distribution. This test statistic is included primarily as a benchmark against which to compare the other test statistics since this is the only procedure which is based upon the same distribution used to generate the data. Maximum likelihood estimates required to construct the test were obtained using the Newton-Raphson procedure with observed information as described by Collett (1991 Appendix B) and outlined in Section 2.7 of the thesis. Problems with convergence which occurred when $\rho = 0$ led to the use of the likelihood ratio test assuming independence in this case.

The last four test statistics (i.e. numbers 8 through 11) being compared by simulation are based on Liang and Zeger's (1986) generalized estimating equations approach using an exchangeable working correlation matrix. The four generalized estimating equations procedures are the model dependent Wald test, the robus, Wald test, the model dependent score test and the robust score test. They are denoted WZN, WZR, SZN, and SZR respectively in the tables displaying the results of the simulation study.

Parameter estimates needed to calculate the model dependent and robust Wald tests were obtained using the iterative procedure outlined in Section 2.8 of this thesis. Model dependent and robust score tests were calculated analytically using estimates of intracluster correlation obtained when constructing the Wald tests. Such estimates of intracluster correlation are preferable because they are consistent under both the null and alternative hypothesis (Breslow, 1990). An algebraic justification for this approach is presented in Section 2.8.

Parameters for the simulations were selected primarily to address the objectives listed in Section 2 of this chapter. Some attention was also given to reducing the possibility that there would be any convergence problems for the iterative procedures. Any iterations where the iterative procedures failed to converge were replaced.

It was noted in Section 1.4.1 that negative intracluster correlation coefficients are generally considered implausible in the context of cluster randomization trials. For this reason negative estimates of ρ are usually set to zero. This practice was also followed in the present study. For a similar reason the design effects required when calculating the Mantel-Haenszel test statistic using

Rao and Scott's (1992) ratio estimator approach were set to one if they were less than unity. This practice was also followed by Fung et al (1993) and Donner et al (1993).

3.7 Results

There were very few problems encountered when running the computer programs used in this simulation study. There was never more than one iteration out of 500 which failed to converge. As mentioned in the previous section any such iterations were replaced. Lack of convergence only occurred when attempting to fit the beta-binomial logistic regression model, and then only when generating data in which the intracluster correlation coefficient equals 0.025, the smallest non-zero correlation used in the simulation study. The failure to converge was most likely a consequence of attempting to maximize the likelihood equation when the best estimate of intracluster correlation was less than zero, which is outside of the allowable range for the parameters of the model.

Failure to converge when fitting beta-binomial logistic regression models can also occur when all of the subjects in a stratum and treatment group have the same response. If this situation occurs finite maximum likelihood estimates might not exist causing the lack of convergence. The combination of large cluster sizes and small correlation coefficients eliminated the possibility that similar situations could occur in this simulation study.

The results from the simulation study are presented separately for the three objectives of the study. The type I error rates calculated using data generated under the common correlation assumption are discussed first. Results from these simulations are displayed in Table 3.1 through Table 3.4. These tables are distinguished by the number of strata and the range in risk across the strata.

The power of the different test statistics are compared next. Results from simulations in which there were two strata are displayed in Table 3.5 and results from simulations in which there were four strata are displayed in Table 3.6. The data from these simulations were also generated under the common correlation assumption.

The third set of simulations were concerned with the effect of allowing the degree of correlation to vary as a function of cluster size. The type I error rates and power of the test statistics in such circumstances are presented in Tables 3.8 though 3.11.

3.7.1 Type I Error Rates Under the Common Correlation Assumption

The empirical type I error rates for the classical Mantel-Haenszel test (CMH) were only approximately equal to 0.05 when the responses of cluster members were uncorrelated. The rejection rates for this test statistic increased dramatically for even very small correlation coefficients.

The stratified t-test (TST) and the extended Mantel-Haenszel test statistic (EMH) were the two techniques whose type I error rates were the closest to the nominal 5% level. Even these two methods, however, tended to reject the null hypothesis too often when there were 20 clusters in each treatment group.

There is very little difference in the Type I error rates of the two adjusted Mantel-Haenszel test statistics even when there are only 20 clusters per treatment group and the clusters sizes are allowed to vary. The rejection rates tended to be overly liberal for these two test statistics and all the remaining test statistics unless there were at least 40 clusters per treatment group.

Rao and Scott's (1992) adjusted Mantel-Haenszel test statistic tended to have the most liberal rejection rates when responses of cluster members were correlated. The rejection rates remained too high unless there were at least 20 clusters per each treatment group and stratum. Note that rejection rates were still too high on occasion even with this number of clusters. This method had the most severe asymptotic requirements of all the test statistics being compared.

The beta-binomial likelihood ratio test also tended to reject the null hypothesis too often when the responses of cluster members were correlated unless there were at least 40 clusters per treatment group. For a given number of clusters per treatment group the problem tended to be more severe when there were four as opposed to two strata.

There was little difference in the size of the model dependent Wald and score tests. Rejection rates from the robust Wald tests were consistently larger than the robust score tests which were, in general, larger than the rejection rates of the model dependent tests. The robust Wald and score tests were also the only two out of the eleven procedures examined in the simulation which tended to reject the null hypothesis too often when the responses of cluster members were uncorrelated.

3.7.2 Power Comparisons Under the Common Correlation Assumption

The most striking feature of the power comparisons of the different test statistics is the inverse relationship between power and the degree of intracluster correlation. Power equalled 100 percent when responses of subjects were uncorrelated and decreased to about 70 percent as the degree of intracluster correlation rose to 0.100.

Power differences between methods tended to be less extreme. The greatest differences occurred when $\rho = 0.100$. In this case the range in power was as great as 10 percent. The highest powers were associated with the beta-binomial likelihood ratio test while the lowest powers were found using the stratified t-test, the extended Mantel-Haenszel test statistic or either adjusted Mantel-Haenszel test statistics. Only marginal differences in power were found between the two adjusted Mantel-Haenszel test statistics.

A curious pattern emerged when comparing the difference in power arising when there was no variability in cluster size (i.e. $\kappa = 1.0$) and and when $\kappa = 0.8$. The power was consistently larger for balanced data only for the adjusted Mantel-Haenszel test statistics, approaching a difference of 6 percent when $\rho = 0.100$. The effect of imbalance was inconsistent for the cluster-level test statistics while the power was greater for the remaining test statistics when cluster sizes were variable. The differences were never more than 3 or 4 percent, however.

3.7.3 Intracluster Correlation as a Function of Cluster Size

Type I error rates were not affected when the degree of intracluster correlation was allowed to vary by cluster size. Rao and Scott's (1992) version of the Mantel-Haenszel test statistic was the only test statistic which was overly liberal. However the problems with this test statistic only occurred when there were four strata and 10 clusters in each stratum and treatment group (see Table 3.8). There was also almost no decrease in power accompanying a greater dependence of ρ on cluster size.

Table 3.1
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates ($\alpha = 5\%$)
Wide Risk Range
4 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.000			0.025		
		m			m		
		5	10	20	5	10	20
1. CMH	.8	.050	.052	.064	.282	.348	.352
	1	.068	.060	.054	.284	.302	.274
2. TST	.8	.062	.056	.062	.056	.060	.068
	1	.070	.062	.052	.072	.062	.052
3. EMH	.8	.052	.058	.060	.054	.060	.068
	1	.068	.062	.050	.070	.060	.048
4. MH1	.8	.038	.048	.058	.050	.058	.072
	1	.064	.052	.046	.086	.064	.056
5. MH2	.8	.038	.048	.058	.050	.058	.072
	1	.064	.052	.046	.084	.064	.056
6. RSM	.8	.044	.046	.046	.130	.108	.086
	1	.052	.044	.034	.126	.088	.056
7. XLR	.8	.050	.054	.064	.066	.064	.074
	1	.068	.060	.054	.092	.068	.056
8. WZN	.8	.042	.052	.058	.056	.062	.070
	1	.064	.052	.046	.084	.062	.060
9. WZR	.8	.072	.068	.072	.078	.066	.074
	1	.094	.072	.056	.098	.072	.062
10. SZN	.8	.042	.052	.058	.058	.062	.070
	1	.064	.052	.046	.084	.062	.060
11. SZR	.8	.060	.064	.072	.064	.064	.074
	1	.082	.068	.054	.092	.062	.060

Table 3.1 Continued ...
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates (α = 5%)
Wide Risk Range
4 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.050			0.100		
		m			m		
		5	10	20	5	10	20
1. CMH	.8	.406	.486	.480	.536	.634	.608
	1	.406	.440	.406	.556	.540	.562
2. TST	.8	.056	.064	.068	.050	.066	.068
	1	.074	.062	.054	.072	.066	.052
3. EMH	.8	.058	.064	.066	.050	.066	.068
	1	.070	.058	.052	.066	.062	.050
4. MH1	.8	.048	.056	.066	.042	.064	.062
	1	.084	.064	.060	.084	.066	.060
5. MH2	.8	.044	.056	.066	.042	.058	.060
	1	.082	.064	.060	.082	.068	.058
6. RSM	.8	.158	.108	.088	.158	.112	.086
	1	.136	.086	.062	.136	.088	.058
7. XLR	.8	.072	.066	.072	.070	.068	.074
	1	.092	.068	.058	.092	.066	.050
8. WZN	.8	.062	.066	.070	.064	.064	.066
	1	.078	.062	.054	.078	.066	.058
9. WZR	.8	.076	.068	.080	.072	.074	.076
	1	.098	.072	.062	.092	.076	.060
10. SZN	.8	.062	.066	.072	.064	.064	.066
	1	.078	.062	.054	.078	.066	.060
11. SZR	.8	.068	.068	.074	.066	.066	.068
	1	.088	.068	.058	.080	.068	.060

Table 3.2
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates ($\alpha = 5\%$)
Narrow Risk Range
4 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.000			0.025		
		m			m		
		5	10	20	5	10	20
1. CMH	.8	.052	.066	.062	.284	.344	.366
	1	.068	.058	.046	.288	.302	.268
2. TST	.8	.062	.056	.058	.056	.060	.062
	1	.076	.062	.050	.076	.060	.050
3. EMH	.8	.056	.062	.058	.052	.058	.062
	1	.076	.058	.046	.072	.058	.050
4. MH1	.8	.044	.056	.060	.058	.058	.060
	1	.066	.054	.040	.086	.060	.052
5. MH2	.8	.044	.058	.060	.050	.058	.062
	1	.066	.054	.040	.082	.060	.052
6. RSM	.8	.050	.050	.050	.136	.108	.082
	1	.048	.044	.034	.124	.090	.062
7. XLR	.8	.052	.066	.062	064	.066	.068
	1	.068	.058	.046	.096	.066	.052
8. WZN	.8	.042	.056	.060	.056	.062	.068
	1	.066	.054	.040	.084	.060	.050
9. WZR	.8	.078	.072	.072	.072	.072	.076
	1	.096	.070	.060	.104	.074	.054
10. SZN	.8	.042	.056	.060	.056	.062	.068
	1	.066	.054	.040	.084	.060	.050
11. SZR	.8	.062	.070	.068	.062	.064	.070
	1	.090	.066	.054	.090	.064	.052

Table 3.2 Continued ...
Data Generated Under the Common Correlation Assumption
Empirica. Type I Error Rates (α = 5%)
Narrow Risk Range
4 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.050			0.100		
		m			m		
		5	10	20	5	10	20
1. CMH	.8	.408	.488	.482	.548	.612	.612
	1	.408	.438	.406	.554	.548	.554
2. TST	.8	.054	.064	.076	.056	.066	.068
	1	.074	.060	.048	.074	.062	.050
3. EMH	.8	.058	.060	.068	.056	.066	.068
	1	.072	.060	.046	.074	.064	.048
4. MH1	.8	.050	.058	.066	.048	.058	.064
	1	.086	.062	.050	.084	.068	.056
5. MH2	.8	.048	.060	.066	.044	.058	.064
	1	.088	.062	.048	.082	.068	.056
6. RSM	.8	.154	.112	.084	.158	.112	.080
	1	.134	.090	.058	.140	.090	.058
7. XLR	.8	.072	.068	.070	.070	.068	.074
	1	.092	.072	.054	.092	.072	.054
8. WZN	.8	.064	.066	.070	.062	.068	.072
	1	.082	.062	.052	.078	.066	.052
9. WZR	.8	.076	.070	.076	.072	.070	.074
	1	.100	.072	.058	.104	.076	.056
10. SZN	.8	.066	.066	.070	.064	.068	.072
	1	.082	.062	.052	.078	.066	.052
11. SZR	.8	.066	.066	.070	.066	.068	.074
	1	.088	.064	.054	.086	.068	.052

Table 3.3
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates (α = 5%)
Wide Risk Range
2 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.000			0.025		
		m			m		
		10	20	40	10	20	40
1. CMH	.8	.054	.048	.056	.396	.318	.314
	1	.058	.048	.064	.306	.280	.318
2. TST	.8	.070	.050	.060	.080	.050	.052
	1	.060	.054	.062	.062	.046	.066
3. EMH	.8	.066	.050	.060	.078	.050	.052
	1	.060	.052	.062	.058	.044	.066
4. MH1	.8	.042	.042	.050	.090	.046	.052
	1	.048	.044	.058	.070	.058	.068
5. MH2	.8	.042	.042	.050	.082	.046	.052
	1	.048	.044	.058	.068	.056	.068
6. RSM	.8	.042	.034	.046	.122	.052	.064
	1	.044	.040	.052	.082	.060	.070
7. XLR	.8	.054	.048	.056	.082	.050	.058
	1	.058	.048	.064	.068	.062	.062
8. WZN	.8	.038	.042	.050	.082	.048	.056
	1	.048	.042	.060	.074	.058	.068
9. WZR	.8	.100	.060	.062	.094	.054	.060
	1	.082	.062	.068	.076	.060	.070
10. SZN	.8	.038	.042	.050	.082	.048	.056
	1	.048	.042	.060	.074	.058	.068
11. SZR	.8	.074	.044	.060	.078	.048	.058
	1	.062	.058	.062	.066	.052	.068

Table 3.3 Continued ...
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates ($\alpha = 5\%$)
Wide Risk Range
2 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.050			0.100		
		m			m		
		10	20	40	10	20	40
1. CMH	.8	.518	.436	.464	.636	.598	.614
	1	.446	.420	.432	.558	.560	.550
2. TST	.8	.068	.048	.054	.080	.050	.056
	1	.064	.046	.066	.068	.050	.064
3. EMH	.8	.068	.048	.054	.070	.048	.056
	1	.062	.046	.066	.064	.048	.064
4. MH1	.8	.094	.048	.052	.094	.050	.052
	1	.074	.054	.070	.074	.056	.068
5. MH2	.8	.082	.048	.052	.084	.050	.052
	1	.068	.052	.068	.074	.054	.064
6. RSM	.8	.132	.060	.060	.136	.060	.060
	1	.082	.052	.074	.088	.058	.066
7. XLR	.8	.080	.056	.056	.086	.050	.056
	1	.066	.058	.064	.066	.056	.062
8. WZN	.8	.082	.050	.058	.080	.050	.056
	1	.072	.052	.068	.072	.054	.068
9. WZR	.8	.092	.054	.060	.092	.054	.058
	1	.080	.062	.072	.084	.058	.070
10. SZN	.8	.082	.050	.058	.082	.050	.058
	1	.072	.052	.068	.074	.054	.068
11. SZR	.8	.080	.048	.056	.082	.050	.056
	1	.066	.050	.068	.072	.052	.064

Table 3.4
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates ($\alpha = 5\%$)
Narrow Risk Range
2 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.000			0.025		
		m			m		
		10	20	40	10	20	40
1. CMH	.8	.056	.044	.050	.400	.310	.316
	1	.058	.050	.066	.314	.292	.326
2. TST	.8	.068	.052	.058	.076	.050	.052
	1	.062	.054	.066	.062	.054	.062
3. EMH	.8	.060	.050	.058	.074	.050	.052
	1	.058	.050	.066	.060	.054	.062
4. MH1	.8	.042	.038	.046	.090	.046	.054
	1	.052	.044	.060	.066	.060	.068
5. MH2	.8	.044	.038	.048	.084	.048	.054
	1	.052	.044	.060	.066	.060	.068
6. RSM	.8	.046	.038	.046	.124	.054	.060
	1	.046	.040	.056	.080	.064	.068
7. XLR	.8	.056	.044	.050	.080	.048	.058
	1	.058	.050	.066	.066	.058	.064
8. WZN	.8	.044	.038	.048	.082	.046	.060
	1	.052	.044	.060	.066	.058	.066
9. WZR	.8	.098	.052	.064	.094	.050	.062
	1	.084	.062	.068	.074	.058	.070
10. SZN	.8	.044	.038	.048	.082	.046	.060
	1	.052	.044	.060	.066	.058	.066
11. SZR	.8	.076	.046	.058	.078	.046	.062
	1	.064	.056	.066	.066	.056	.064

Table 3.4 Continued ...
Data Generated Under the Common Correlation Assumption
Empirical Type I Error Rates (α = 5%)
Narrow Risk Range
2 Strata

Test	κ	Intracluster Correlation Coefficient, ρ					
		0.050			0.100		
		m			m		
		10	20	40	10	20	40
1. CMH	.8	.526	.438	.464	.632	.582	.622
	1	.440	.422	.424	.560	.556	.542
2. TST	.8	.072	.054	.054	.072	.048	.050
	1	.062	.048	.064	.064	.048	.060
3. EMH	.8	.066	.052	.054	.068	.046	.052
	1	.056	.046	.062	.060	.048	.060
4. MH1	.8	.086	.046	.052	.088	.046	.050
	1	.070	.056	.064	.072	.052	.060
5. MH2	.8	.084	.046	.052	.082	.046	.052
	1	.070	.054	.066	.070	.052	.060
6. RSM	.8	.122	.054	.060	.132	.054	.058
	1	.080	.064	.066	.082	.060	.064
7. XLR	.8	.086	.054	.056	.088	.060	.058
	1	.064	.056	.066	.064	.056	.062
8. WZN	.8	.080	.054	.064	.078	.048	.058
	1	.068	.054	.064	.070	.052	.060
9. WZR	.8	.096	.056	.064	.094	.052	.060
	1	.080	.058	.070	.080	.056	.064
10. SZN	.8	.080	.054	.064	.082	.048	.058
	1	.068	.054	.064	.070	.052	.060
11. SZR	.8	.076	.052	.062	.078	.046	.058
	1	.066	.052	.064	.068	.052	.060

Table 3.5
Data Generated Under the Common Correlation Assumption
Power to Detect An Odds Ratio of 1.5
m=10
4 Strata

Test	κ	Risk Range							
		Narrow				Wide			
		ρ				ρ			
		0.000	0.025	0.050	0.100	0.000	0.025	0.050	0.100
2. TST	.8	1.000	.994	.944	.750	1.000	.992	.920	.722
	1	1.000	1.000	.934	.746	1.000	.996	.916	.720
3. EMH	.8	1.000	.996	.944	.746	1.000	.992	.922	.718
	1	1.000	1.000	.938	.742	1.000	.996	.914	.720
4. MH1	.8	1.000	.998	.924	.694	1.000	.988	.902	.666
	1	1.000	1.000	.940	.758	1.000	.996	.922	.728
5. MH2	.8	1.000	.998	.920	.692	1.000	.990	.900	.666
	1	1.000	1.000	.940	.758	1.000	.998	.922	.728
7. XLR	.8	1.000	.998	.948	.792	1.000	.996	.942	.766
	1	1.000	1.000	.944	.770	1.000	.996	.930	.744
8. WZN	.8	1.000	.998	.944	.758	1.000	.996	.934	.732
	1	1.000	1.000	.940	.752	1.000	.996	.916	.726
9. WZR	.8	1.000	.998	.950	.786	1.000	.998	.940	.750
	1	1.000	1.000	.944	.762	1.000	.998	.926	.738
10. SZN	.8	1.000	.998	.944	.758	1.000	.996	.934	.732
	1	1.000	1.000	.940	.752	1.000	.996	.916	.726
11. SZR	.8	1.000	.998	.944	.768	1.000	.998	.932	.740
	1	1.000	1.000	.942	.754	1.000	.998	.918	.730

Table 3.6
Data Generated Under the Common Correlation Assumption
Power to Detect An Odds Ratio of 1.5
m=20
2 Strata

Test	κ	Risk Range							
		Narrow				Wide			
		ρ				ρ			
		0.000	0.025	0.050	0.100	0.000	0.025	0.050	0.100
2. TST	.8	1.000	.990	.940	.758	1.000	.986	.910	.706
	1	1.000	.992	.936	.732	1.000	.992	.906	.664
3. EMH	.8	1.000	.990	.938	.754	1.000	.986	.910	.700
	1	1.000	.992	.934	.726	1.000	.992	.904	.658
4. MH1	.8	1.000	.984	.914	.700	1.000	.978	.870	.640
	1	1.000	.992	.940	.744	1.000	.992	.914	.684
5. MH2	.8	1.000	.986	.914	.694	1.000	.978	.872	.642
	1	1.000	.992	.940	.742	1.000	.992	.912	.682
7. XLR	.8	1.000	.992	.938	.788	1.000	.988	.922	.736
	1	1.000	.992	.940	.764	1.000	.992	.922	.706
8. WZN	.8	1.000	.992	.938	.768	1.000	.988	.916	.710
	1	1.000	.992	.938	.740	1.000	.992	.910	.674
9. WZR	.8	1.000	.992	.940	.780	1.000	.988	.922	.724
	1	1.000	.992	.942	.758	1.000	.992	.918	.694
10. SZN	.8	1.000	.992	.938	.768	1.000	.988	.916	.710
	1	1.000	.992	.940	.740	1.000	.992	.910	.678
11. SZR	.8	1.000	.992	.938	.768	1.000	.988	.914	.712
	1	1.000	.992	.938	.738	1.000	.992	.912	.676

Table 3.7
Intracluster Correlation As a Function of Cluster Size using Smith's Law
($\rho = A\, n^{-B}$)

n	Values for the Parameters A,B			
	0.025,0.0	0.1, 0.3010	0.5, 0.6505	0.9, 0.7782
2	.025	.081	.319	.525
100	.025	.025	.025	.025
500	.025	.015	.009	.007

Table 3.8
Intracluster Correlation Allowed to Vary with Cluster Size

$$\left[\rho = A\ n^{-B}\right]$$

Empirical Type I Error Rates (α = 5%)

m=10

4 Strata

Test	Risk Range							
	Narrow				Wide			
	A,B				A,B			
	.025,0	.1,.301	.5,.651	.9,.778	.025,0	.1,.301	.5,.651	.9,.778
1. CMH	.338	.316	.304	.290	.336	.320	.300	.304
2. TST	.066	.068	.070	.064	.066	.066	.066	.066
3. EMH	.068	.066	.070	.062	.062	.064	.062	.066
4. MH1	.060	.060	.064	.060	.056	.064	.062	.062
5. MH2	.060	.062	.064	.064	.054	.060	.058	.062
6. RSM	.096	.094	.100	.096	.090	.096	.096	.094
7. XLR	.066	.066	.072	.070	.058	.068	.066	.070
8. WZN	.064	.066	.068	.066	.058	.062	.062	.064
9. WZR	.072	.068	.074	.070	.064	.068	.068	.076
10. SZN	.066	.066	.068	.066	.058	.062	.062	.064
11. SZR	.064	.066	.068	.070	.060	.066	.068	.068

Table 3.9
Intracluster Correlation Allowed to Vary with Cluster Size

$$\left[\rho = A\ n^{-B}\right]$$

Empirical Type I Error Rates (α = 5%)
m=20
2 Strata

Test	Risk Range							
	Narrow				Wide			
	A,B				A,B			
	.025,0	.1,.301	.5,.651	.9,.778	.025,0	.1,.301	.5,.651	.9,.778
1. CMH	.322	.298	.284	.288	.316	.302	.288	.280
2. TST	.048	.050	.046	.046	.048	.046	.042	.048
3. EMH	.048	.050	.042	.046	.050	.044	.044	.048
4. MH1	.054	.046	.042	.036	.054	.040	.038	.040
5. MH2	.050	.048	.050	.046	.050	.046	.044	.050
6. RSM	.048	.052	.048	.044	.058	.050	.048	.048
7. XLR	.054	.054	.050	.046	.052	.052	.044	.044
8. WZN	.052	.050	.044	.044	.054	.048	.046	.046
9. WZR	.058	.054	.050	.048	.054	.056	.050	.046
10. SZN	.052	.050	.044	.044	.054	.048	.046	.046
11. SZR	.048	.052	.048	.046	.050	.048	.046	.046

Table 3.10
Intracluster Correlation Allowed to Vary with Cluster Size
$$\left[\rho = A\, n^{-B}\right]$$
Power to Detect An Odds Ratio of 1.5
m=10
4 Strata

Test	Risk Range							
	Narrow				Wide			
	A,B				A,B			
	.025,0	.1,.301	.5,.651	.9,.778	.025,0	.1,.301	.5,.651	.9,.778
2. TST	.994	.994	.994	.994	.994	.994	.994	.994
3. EMH	.994	.994	.994	.994	.994	.994	.994	.994
4. MH1	.998	.998	.996	.994	.998	.996	.996	.996
5. MH2	.998	.998	.998	.996	.998	.998	.998	.998
7. XLR	.996	.994	.996	.994	.994	.994	.994	.996
8. WZN	.996	.994	.996	.994	.998	.994	.994	.994
9. WZR	.996	.996	.998	.996	.998	.996	.996	.996
10. SZN	.996	.994	.996	.994	.998	.994	.994	.994
11. SZR	.996	.996	.998	.994	.998	.994	.996	.996

Table 3.11
Intracluster Correlation Allowed to Vary with Cluster Size

$$\left[\rho = A\, n^{-B}\right]$$

Power to Detect An Odds Ratio of 1.5
m=20
2 Strata

Test	Risk Range							
	Narrow				Wide			
	A,B				A,B			
	.025,0	.1,.301	.5,.651	.9,.778	.025,0	.1,.301	.5,.651	.9,.778
2. TST	.992	.988	.982	.980	.986	.982	.978	.976
3. EMH	.992	.988	.982	.980	.984	.982	.978	.974
4. MH1	.988	.990	.988	.990	.980	.980	.980	.978
5. MH2	.988	.990	.990	.992	.982	.984	.984	.986
7. XLR	.992	.992	.992	.992	.990	.990	.990	.986
8. WZN	.992	.992	.992	.990	.990	.986	.982	.984
9. WZR	.992	.992	.992	.992	.990	.990	.988	.984
10. SZN	.992	.992	.992	.990	.990	.986	.982	.984
11. SZR	.992	.992	.992	.990	.988	.986	.982	.984

3.8 Discussion

3.8.1 Type I Error Rates Under the Common Correlation Assumption

Cluster randomization trials in which there are about 100 subjects per cluster will rarely have as many as 20 clusters per treatment group. The results of these simulation studies suggests that even this number of clusters is insufficient to assure the validity of most approximate test statistics. Researchers will either be forced to include a greater number of clusters in their trials or limit attention to exact tests.

Exclusive reliance upon exact tests is however not practical. Such tests require sophisticated computer programs (e.g. StatXact (1991)) which are still not widely available. Furthermore there may still be situations were there are too few clusters to assure the validity of asymptotic methods and yet too many clusters to allow exact p-values to be calculated. The exact p-values would then have to be estimated by taking random samples of the possible permutations. An additional limitation of most exact tests is that they ignore any variation in cluster size resulting in a loss in power (see Section 2.4).

There was considerable variability in the asymptotic requirements of the different test statistics. Surprisingly the rejection rates nearest the nominal five percent nominal level were found using the stratified t-test and the extended Mantel-Haenszel χ^2 test statistic. It is difficult to explain why these methods performed so well. One possibility is that estimates of variance for these methods were calculated using the between-cluster variability rather than reiying upon estimates of intraclustei correlation to correct for the effect of clustering. Such estimates can be very imprecise when there are few clusters (e.g. Feng and Grizzle, 1992).

Difficulties in assuring the validity of approximate statistical tests has been noted in earlier simulation studies. Overly liberal rejection rates for likelihood ratio tests based on the beta-binomial distribution have been noted previously by Shirley and Hickling (1981) and by Donner et al (1981) when there were few clusters per treatment group.

Curiously Donald (1984) found that adjusted Mantel-Haenszel test statistics were overly conservative rather than rejecting the null hypothesis too often. The differences could be due to the very different cluster sizes in the two simulation studies. Donald's (1984) simulation study examined stratified cluster randomization trials in which there were five strata and 240 subjects. These subjects came from clusters of fixed size with either 2, 4, or 6 subjects per cluster and 60, 30, or 20 clusters per treatment group respectively.

The Mantel-Haenszel test statistic calculated using Rao and Scott's (1992) ratio estimator approach (i.e. RSM) was the most consistently liberal statistic examined. It rejected the null hypothesis as much as three times as often as it should when there were fewer than 20 clusters in each treatment group and stratum (i.e. m<20). Clearly this method can not be recommended for cluster randomization trials.

Alternative versions of this test statistic which used pooled estimates of design effects could be calculated. The rejection rates for such test statistics would then likely be much closer to the nominal level (Donner et al (1993), Fung et al (1992)). However, it is difficult to see what advantage they offer as compared to other simple adjusted test statistics like the adjusted Mantel-Haenszel χ^2 statistics (i.e. MH1 and MH2). Furthermore Rao and Scott (1992) provide no guidelines to determine how best to estimate such statistics in stratified designs.

The overly liberal rejection rates encountered with both robust Wald tests and robust score tests was not unexpected. Algebraic results in the previous chapter predicted that such tests would be overly liberal when there were few clusters in each treatment group and stratum and also predicted that these problems would be more severe for Wald than for score tests.

These problems arose because of the way in which robust estimates of variance correct for misspecification of the working correlation matrix. In Section 2.8 robust variance estimates were shown to be functions of variance estimates determined using the theory of ratio estimation. Variance estimates of ratio estimators tend to be negatively biased when there are few clusters and when there is considerable imbalance in cluster size (Cochran, 1977, pp. 162-163).

Furthermore the robust estimates of variance were shown to be negatively biased in Section 2.8 even when there are the same number of subjects in each cluster.

An additional problem with these correction factors occurs when the true correlation is very near zero. Negative estimates of correlation are quite likely to occur in such situations. These estimates are usually truncated at zero. A similar adjustment is applied to the design effects estimates required when using Rao and Scott's (1992) ratio estimator approach (Fung et al, 1993). Unfortunately no similar adjustment is available for robust score and Wald tests. This likely contributed to the overly liberal rejection rates which occurred for the robust test statistics when $\rho = 0$.

The problem with the robust test statistics arising when there were few clusters per treatment group and stratum was likely compounded by the large number of subjects per cluster. For example Lipsitz et al (1991) and Prentice (1988) point out that robust variance estimates are inaccurate when there are more subjects per cluster than there are clusters in the study.

Previous simulations have found similar problems with robust test statistics. Donner, Eliasziw and Klar (1993) reported that robust Wald tests tended to be overly liberal in completely randomized cluster randomization trials in which there were 10 subjects per cluster, on average. A similar problem occurred in the simulation studies reported by Moore and Tsiatis (1991). Curiously no such problems were found in a simulation study described by Sharples and Breslow

(1992). It is likely that the rejection rates were near the nominal levels because there were only two subjects per cluster and relatively many clusters.

The possibility of obtaining overly liberal Type I error rates when using robust estimates of variance is not restricted to correlated binary outcome data but will also occur with other types of outcomes and models when there are few clusters. Overly liberal rejection rates or negatively biased variance estimates were reported by Thornquist and Anderson (1993) when fitting linear models to continuous outcome data, by Breslow (1990) when fitting log-linear models to count data, and by both Lin and Wei (1989) and Segal and Neuhaus (1993) who were concerned with modelling correlated survival times data. Several of these researchers have also found that robust Wald tests tended to be more liberal than robust score tests (Breslow (1990), Lin and Wei (1989)).

3.8.2 Power Comparisons Under the Common Correlation Assumption

Power was lowest for the two cluster-level test statistics and the two adjusted Mantel-Haenszel test statistics, especially when cluster s:zes were variable. The low power of the cluster-level test statistics probably occurred because these test statistics ignore any variations in cluster size. An additional reason for their lower power is that the data were generated under the assumption of a common odds ratio. The cluster-level statistics, however, are more sensitive to a common risk difference. This latter reason might not be very important since risk differences tend to be constant when risk across strata is between 0.3 and

0.7, as was generally the case in these simulation studies.

It is more difficult to understand why the adjusted Mantel-Haenszel test statistics also had lower power than most other procedures, at least when cluster sizes were variable. One possible explanation is that the risk estimates used to construct these statistics are relatively imprecise. Asymptotically optimal estimates are weighted averages of cluster-specific risks where the weights are functions of both cluster size and the degree of intracluster correlation. Risk estimates used in both the numerator and denominator of the adjusted Mantel-Haenszel test statistics do not incorporate estimates of intracluster correlation. A comparison of the different ways in which risk estimates can be calculated is presented in Section 2.5.

Several researchers have derived formulae which can be used to determine sample size or power for cluster randomization trials. These are reviewed in Section 1.3.2 of this thesis. The approach suggested by Donner (1992) is the only one applicable to stratified cluster randomization trials.

The power obtained by simulation using the adjusted Mantel-Haenszel test statistics were compared to the power obtained using the methods described by Donner (1992). This comparison was used to see how well the two different approaches agreed. The adjusted Mantel-Haenszel test statistics were selected as being the two test statistics which correspond most closely to Donner's (1992) sample size formulae. Estimates of power obtained using Donner's (1992) method never differed from the empirically determined rejection rates by more than three percentage points.

It was pointed out in the results section of this chapter that the power tended to be greatest for the beta-binomial likelihood ratio test statistic. This is not surprising since this statistic makes use of the parametric distribution from which the data were generated. Similar results were found in the simulation study conducted by Donner et al (1993) who also generated cluster-level responses from a beta-binomial distribution.

3.8.3 Intracluster Correlation as a Function of Cluster Size

The rejection rates were not affected when the degree of correlation among the responses of cluster members varied as a function of cluster size. The lack of any effect might have occurred because of the opposing influences on ρ and the variance inflation factor caused by the combination of using "Smith's Law" and relatively large clusters. In situations where there are many subjects per cluster, as in this simulation study, increasing variability in ρ is accompanied by decreasing variability in the variance inflation factor.

The degree of correlation was determined using the equation $\rho_{ijs} = An_{ijs}^{-B}$. Since cluster size is negatively correlated with the degree of intracluster correlation the parameter B was always set to be greater than or equal to zero. Variation in the size of ρ is greatest for large values of B. As B approaches 1 the variance inflation factor can be expressed as

$$1 + A\frac{n_{ijs} - 1}{n_{ijs}} \approx 1+A \qquad (3.6)$$

if there are large numbers of subjects per cluster. Greater variability in ρ caused by further increases of the parameter B might still have some influence on the rejection rates of these test statistics but would, perhaps, be greater than could reasonably occur.

This result stands in stark contrast to Williams (1988a) observation that parameter estimates of beta-binomial models can be biased when the degree of correlation is incorrectly assumed to be the same in all clusters. Williams (1988a) however was interested in teratological studies where ρ has been noted to vary across treatment groups (Catalano and Ryan, 1992). Simulations described in this thesis were restricted to the case where the degree of correlation varied across strata but, on average, was the same in each treatment group.

Pregnant animals are randomly assigned to treatment groups in teratological studies. The mortality rates of newborn litter mates from the different treatment groups are then compared. Variations in ρ across treatment groups can occur in such studies as a function of litter size since some chemicals affect both the number of animals per litter and the mortality rates of animals after birth (Catalano and Ryan, 1992). This is quite different from the epidemiological studies for which the present simulation was designed. Cluster sizes are typically fixed at baseline in epidemiological studies. It is therefore not at all likely that cluster sizes could vary across treatment groups so that the same degree of intracluster correlation would be expected in both treatment groups (provided random assignment was used).

Even in teratological studies, however, variations in ρ across treatment groups could only occur under the alternative hypothesis. Thus one only needs to be concerned about the effect on the power of test statistics. The power of methods which assume that the average degree of correlation in a cluster was fixed (i.e. the adjusted Mantel-Haenszel test statistics and both the model dependent test statistics) would presumably be reduced when this assumption is violated.

4. Examples

4.1 Introduction

Data analyses from two cluster randomization trials are presented in this chapter of the thesis. These case studies are included to illustrate the methods discussed in the previous two chapters. They will also help in examining how well theoretical findings hold in particular applications.

The first data set comes from a trial conducted in 1982-1983 which enrolled Southeast Asian refugees upon their arrival in Montreal. The trial was designed to examine if a parasite screening and treatment program would reduce the risk of infection below the expected spontaneous loss of infection attained following immigration to Canada (Gyorkos et al, 1989). The rationale for the trial stemmed, in part, from the Canadian government's 1981 decision to eliminate the requirement for parasite screening of potential immigrants from tropical and subtropical countries.

Families were matched for size and age of family members separately during each month of the study. One family of each pair was then randomly assigned to the treatment group. If there were an odd number of families in any month the remaining unmatched family was randomly assigned either to the treatment or control group. Matching was used to increase similarity of subjects at baseline and was not retained for the analysis. Families randomly assigned to the treatment group were screened and treated at the beginning of the trial and again after six months while families randomly assigned to the control group were only screened and treated after being enrolled in the trial for six months. Subjects were identified as infected if any one of *Ascaris lumbricoides, Entamoeba histolytica, Giardia lamblia, Strongyloides stercoralis* or hookworm were found in their stool samples. Analyses focus on the subset of 119 subjects from 31 control families and 130 subjects from 35 screened families. These subjects were Kampuchean immigrants whose six month infection status was known.

Data for the second case study result from a trial conducted in Minnesota which randomly assigned 12 schools to one of three different adolescent tobacco prevention programs or to a control group (Murray et al, 1992). The three interventions were selected as representative of the most widely adopted tobacco use prevention programs in Minnesota following a 1985 state initiative. Students attending the control schools received the existing tobacco education programs. Students attending 6'th grade in 1987 were enrolled in the study and they were surveyed again in each of the next three years. For illustrative purposes the analyses focus on the comparison of weekly prevalence of smokeless tobacco use between the subset of 1338 grade 9 students in the Smoke Free Generation intervention group and the 1483 grade 9 students in the control group.

Comparisons between different methods of analysis focus on three aspects of models: statistical significance of associations, direction and strength of association as measured by size of odds ratios, and width of confidence intervals on odds ratios. These comparisons allow examination of how inference is affected by the choice of a method of analysis.

The comparisons of methods are performed in two ways. Comparisons are first drawn between results obtained using different methods within each of the two studies. More general comparisons are also drawn between the two studies since they represent the two extremes of cluster randomization trials. The first trial has a relatively large number of small clusters while Murray et al's (1992) tobacco use prevention trial has a few quite large clusters. The differences between methods will likely be more extreme in the latter case.

4.2 Materials and Methods

Analyses for the parasite screening and treatment trial were performed after stratifying by family size, a variable which has been shown to be a risk factor for se⁾eral infectious diseases (Anderson and May (1991, p. 315), Green and Zaaide (1989)) and is also believed to be a determinant of intestinal parasite infection Gyorkos (1985, p. 149). Families with 3 or fewer members were placed in one stratum and all other families were placed in a second stratum. This stratification maintained roughly equal numbers of clusters in each stratum and treatment group and was nearest to the median cluster size. The data from this trial are displayed in Table 4.1 stratified by family size to distinguish between cluster size and the number of family members who participated in the study.

A similar approach was taken with data from the second cluster randomization trial. Schools with 100 or fewer participants were placed in one stratum and schools with more participants were placed in a second stratum (see Table 4.2). The number of participants was used as a proxy variable for the actual cluster size which was not available. School size might be correlated with community size or the location of the school (i.e. urban or rural) both of which are predictive of adolescent smokeless tobacco use (Surgeon Generals Report, 1986, p. 20).

Several simple summary statistics were calculated for the data from the two randomized trials and displayed in Tables 4.3 and 4.4. Estimates of intracluster correlation were calculated in each stratum and treatment group to help in identification of any patterns in the dependencies among cluster members. These estimates were obtained by adapting one-way random effects models to binary outcome data as described in Sections 2.2 and 2.3 and then used to estimate the degree of variance inflation induced by clustering.

The degree of imbalance in cluster size in each treatment group was calculated using the statistic $\tilde{k}$ which was introduced in the previous chapter. This statistic equals one, its maximum value, when there are the same number of subjects in each cluster and then declines as the imbalance in cluster sizes increases.

Selected additional statistics were calculated for the parasite screening trial using baseline data. These data were available only for subjects randomly assigned to the intervention program (Gyorkos et al, 1989) and were used to examine temporal trends in intracluster correlation coefficients, and to determine the degree to which cured patients tend to cluster within families.

Results from 11 stratified, two-tailed tests of the effect of treatment are compared in this chapter (see Tables 4.5 and 4.6). These methods include the standard Mantel-Haenszel χ_1^2 test, and ten methods which are appropriate for the analysis of correlated binary outcome data. All ten test statistics which adjust for the effect of clustering follow an approximate χ_1^2 distribution under the null hypothesis when there are many clusters in each treatment group and stratum. The standard Mantel-Haenszel test statistic requires the additional assumption that the intracluster correlation coefficient, $\rho = 0$, for it to be approximately χ_1^2. When $\rho > 0$ the standard Mantel-Haenszel test will tend to reject the null hypothesis too often.

The first three methods which adjust for the effect of clustering (i.e. methods 2, 3 and 4) all employ cluster-level data. They include a cluster-level F test and both approximate and exact tests calculated using Mantel's (1963) extended Mantel-Haenszel test. The cluster-level F test is equal to the square of the stratified t test discussed by Schwartz et al (1980, pp. 189-191) and the exact test was performed using the statistical package StatXact (1991).

The exact test for the parasite screening and treatment trial had to examine

$$\binom{27}{13}\binom{39}{18} = 1.2508 \times 10^{18}$$

permutations since 13 out of 27 clusters in the first stratum and 18 out of 39 clusters in the second stratum were randomly assigned to the screening and treatment program (see Table 4.1). An exact test could not be calculated using StatXact because of the very large number of possible permutations. The exact p-value was approximated by drawing a random sample of 1,000,000 permutations from the permutation distribution. This sample size was selected so that the 99% confidence limits were (0.049,0.051), assuming a 5% rejection rate.

It was possible to calculate the exact p-value for the data collected in the adolescent smoking prevention trial since there were only

$$\binom{11}{7}\binom{13}{5} = 424,710$$

permutations which had to be examined. This number of permutations occurred since seven out of the eleven clusters in the first stratum and five out of thirteen clusters in the second stratum were randomly assigned to the Smoke Free Generation program.

Methods 5 and 6 employ simple adjustments of the Mantel-Haenszel χ^2 test for the effect of clustering. Rao and Scott's (1992) approach (see Section 2.6) uses the theory of ratio estimation to obtain a correction factor used to adjust the number of affected subjects and the cluster sizes for each cluster in the study. Standard tests of significance can then be constructed using these adjusted data. Following Fung et al (1993) correction factors were truncated at one. This is analogous to truncating negative estimates of intracluster correlation. Simulation studies described in the last chapter found that this test statistic was overly liberal when there were few clusters in each treatment group and stratum.

An alternative approach which employs a simple adjustment of a standard test is Donald and Donner's (1987) adjusted Mantel-Haenszel χ^2 test. This test statistic was described in Section 2.5. It is derived under the assumption that the average degree of correlation between cluster members is the same for all clusters in the trial.

The beta-binomial likelihood ratio test provided by the computer package EGRET was also used to test for the effect of treatment. This test is asymptotically most powerful when the data follow a beta-binomial distribution. Otherwise its validity may be in question (e.g. Williams, 1988a).

Liang and Zeger's (1986) generalized estimating equations approach, described in Section 2.8, is used to construct the last four tests of significance, i.e. methods 7 through 11. Model dependent and robust Wald tests were constructed following Moore and Tsiatis (1991) while score tests were obtained as described by Breslow (1990). The validity of the two model dependent test statistics is assured if the average correlation between cluster members is fixed, at least in the absence of baseline risk factors measured at the level of the individual (e.g. age, sex). The robust Wald and robust score tests will, at least asymptotically, have valid type I error rates even if this assumption is violated. Their robustness is offset, however, by the imprecision of the test statistics and confidence intervals when there are few clusters.

The estimate of intracluster correlation needed for the score test was obtained by solving the generalized estimating equations under the alternative hypothesis, as suggested by Breslow (1990). This approach was taken to increase power. Two additional benefits of this approach are that the same estimate of ρ is then used in both Wald and score tests and the parameter estimates for the score test can be obtained analytically.

Summary estimates of intracluster correlation were not provided for a number of procedures. Such estimates are not used when calculating the standard Mantel-Haenszel test statistics listed in the first row of Table 4.5 and Table 4.6 or when constructing confidence limits for stratified odds ratios as described by Woolf (1955) (see Table 4.7 and Table 4.8) because these methods assume that the true correlation is equal to zero. The other test statistics which omit estimates of intracluster correlation do so because they are either calculated at the cluster level and do not rely on such estimates (i.e. cluster-level F test, extended Mantel-Haenszel procedure, exact permutation test) or because the theory of ratio estimation is used to construct tests of significance.

The magnitude of the effect of the interventions in the two trials was summarized using stratified odds ratio estimates and their accompanying confidence intervals. These are displayed in Tables 4.7 and 4.8. The classical Woolf (1955) odds ratio estimate and confidence limits were included for the sake of comparison and to examine the effect of failing to adjust for the effect of clustering. The three other approaches are all extensions of this basic method. In all three cases confidence intervals were constructed as

$$\left[\exp\left\{\hat{\gamma} - 1.96\sqrt{\hat{v}}\right\}, \exp\left\{\hat{\gamma} + 1.96\sqrt{\hat{v}}\right\}\right]$$

where $\hat{\gamma}$ is the log of the odds ratio estimate and $\hat{v}$ is an estimate of the variance of $\hat{\gamma}$.

Weighted Woolf odds ratio estimates and confidence intervals were constructed as described in equation (2.5.19) of Section 2.5. Intracluster correlation was calculated as the average of the four treatment and stratum specific estimates of ρ as described by Donald and Donner (1987). This estimate of ρ was also used to construct the adjusted Mantel-Haenszel χ^2 test. Odds ratio estimates were also obtained using the beta-binomial and generalized estimating equations extensions of logistic regression.

4.3 Results

4.3.1 Parasite Screening and Treatment Program

Summary statistics for the parasite screening and treatment trial are displayed in Table 4.3. These data are strongly supportive of the benefits of the program. In the first stratum the odds of infection among control subjects was 4.67 times the odds of infection for screened and treated subjects. The odds ratio for the effect of screening and treatment fell to 2.16 among subjects from families with more than 3 members. Family size, however, had little effect on the risk of infection and may not be an important predictor. Summary estimates of risk in Table 4.3 are equal to the totals in Table 4 of Gyorkos et al (1989).

Estimates of intracluster correlation coefficient were highly variable. Responses from control subjects from smaller families were negatively correlated while the degree of correlation from subjects in the three other treatment and stratum combinations ranged from 0.04 to 0.38. Correlation coefficients among the treated subjects were approximately the same size at the start of the study. This high variability is probably due, in part, to the size and number of clusters in each category. Variance inflation factors were far less variable suggesting that variations in ρ 's may have been related to cluster size.

There were approximately four subjects per family in the 31 families in the control group and in the 35 families randomly assigned to the screening and treatment program. There was also a fair degree of imbalance in family size as measured by the statistic $\hat{k}$. In both the control and treatment groups the statistic was approximately equal to 0.8.

Finally note that there was little difference in the average family size and in the number of subjects participating per family. This occurred because participation rates were high among families in which at least one member participated.

Summary estimates of intracluster correlation, listed in Table 4.5, ranged from 0.058 to 0.084. Since cluster sizes were also relatively small there was little difference in the size of the test statistics or accompanying p-values. Although it is of no practical consequence in the interpretation of these data it is still worth noting that the p-value from the standard Mantel-Haenszel χ^2 test is at least half that of all other procedures. It clearly overestimates the statistical significance of the treatment effect.

Stratified odds ratio estimates displayed in Table 4.7 were all between 2.51 and 2.63. As expected the narrowest confidence intervals arose using the classical Woolf confidence limits (Woolf, 1955) which do not adjust for the effect of clustering. The width of adjusted confidence limits was as much as 17% wider although all confidence limits excluded one.

The beta-binomial logistic regression model was also used to examine the relationship between cluster size (the stratification variable) and the risk of parasite infection. After adjusting for the effect of treatment the χ_1^2 likelihood ratio test statistic was 1.10 (p=0.30) suggesting that there was little evidence for any such relationship. This is not at all surprising since approximately 54 percent of control subjects were infected in each stratum (see Table 4.3). Neither the size of the effect of treatment nor its statistical significance were much affected by fitting a simpler beta-binomial model which omitted the stratification variable.

4.3.2 Adolescent Tobacco Use Prevention Program

Results from the tobacco use prevention trial are far less striking than those from the first study. The odds for the prevalence of smokeless tobacco use among students in the control group (i.e. Existing Curriculum) relative to students receiving the intervention program (i.e. Smoke Free Generation) was 1.98 in schools with 100 or fewer participants and 1.24 in schools with more than 100 participants (see Table 4.4).

Estimates of intracluster correlation were smaller and less variable in this study. They ranged from 0.0003 to 0.0204. The larger cluster sizes led to variance inflation factors ranging from 1.02 to 4.30 indicating that failure to adjust for the effect of clustering will have far more serious consequences than in the first trial.

There were, on average, approximately 123 subjects participating per school in the control group but 112 subjects per school among subjects in the schools receiving the intervention program. The imbalance statistic was approximately 0.80 in both groups. Note that this was the same degree of imbalance in cluster size as in the parasite screening and treatment program.

Use of the standard Mantel-Haenszel χ^2 statistic displayed in Table 4.6 is indicative of a weakly statistically significant result (i.e. p=0.073). This p-value overestimates the statistical significance of the effect of the intervention. The other p-values displayed in Table 4.6 ranged from 0.121 for Rao and Scott's ratio estimator approach to 0.300 obtained using the beta-binomial likelihood ratio test. The second smallest p-value from methods which adjusted for the effect of clustering was obtained using the robust Wald test.

Stratified odds ratio estimates for the prevalence of smokeless tobacco use were between 1.3 and 1.5. All of the confidence intervals included 1 but the confidence interval for the classical Woolf odds ratio (Woolf, 1955) was again spuriously narrow. This confidence limit was only 62% of the widest confidence limit. Failure to adjust for the effect of clustering will severely overestimate the precision of the effect of the intervention.

Weekly prevalence of smokeless tobacco use tended to be slightly higher in schools with more than 100 participants than in schools with 100 or fewer participants. Prevalence increased from 5.57% to 6.29% among students from schools in the control group and increased from 2.90% to 5.13% among students in schools randomly assigned to receive the Smoke Free Generation program. These differences were not statistically significant. After adjusting for the effect of treatment the beta-binomial likelihood ratio test χ_1^2 statistic was 1.46 (p=0.23). Omitting the stratification variable did not affect inferences concerning the effect of treatment. Tests of the effect of treatment still failed to reach nominal levels of statistical significance (e.g. beta-binomial likelihood ratio $\chi_1^2 = 1.81$, p=0.16).

Table 4.1
Parasite Screening and Treatment Trial
Subjects Classified by Stratum, Intervention, Family, and Outcome

Stratum	Family Size	Intervention Group	
		Control	Screened
≤3 Per Family	1	0/1, 0/1, 1/1, 1/1	0/1, 0/1, 0/1
	2	0/2, 1/2, 1/2, 1/2	0/2, 0/2, 1/2, 2/2
	3	1/3, 2/3, 2/3, 2/3, 2/2	0/2, 0/3, 0/3, 0/3, 0/3, 1/2, 2/3
>3 Per Family	4	1/4, 1/4, 3/4, 3/4, 4/4	0/1, 0/4, 1/4, 1/4, 2/4, 3/4
	5	2/5, 2/5, 3/5, 4/5	0/2, 0/5, 1/5, 1/5, 1/4, 2/5, 4/5, 4/4
	6	1/6, 2/6, 3/6, 4/6, 3/4, 5/6	0/2, 1/6, 2/6, 3/6, 3/6
	8-10	1/5, 2/4, 6/10	1/8, 5/10

Table 4.2
Weekly Prevalence of Smokeless Tobacco Use Among Adolescents
Subjects Using Smokeless Tobacco
Classified by Stratum, Intervention, School, and Outcome

Stratum	Intervention Group	
	Existing Curriculum	Smoke Free Generation
≤ 100 Participants per School	1/55, 3/74, 6/83, 6/75	0/42, 1/96, 1/84 1/55, 2/63, 4/58, 5/85
>100 Participants per School	2/152, 3/174, 5/103, 12/207, 7/104, 7/102, 23/225, 16/125	4/160, 10/219, 10/194, 9/149, 11/136

Table 4.3
Parasite Screening and Treatment Trial
Summary Statistics Classified by Stratum, Intervention, and Outcome

Stratum	Variable	Intervention Group		Odds Ratio
		Control	Screened	
≤3 Per Family	Number of Families	13	14	4.67
	% Infected Subjects	53.85	20.00	
	Intracluster Correlation Coefficient	-0.26	0.38	
	Variance Inflation Factor	0.66	1.54	
>3 Per Family	Number of Families	18	21	2.16
	% Infected Subjects	53.76	35.00	
	Intracluster Correlation Coefficient $\hat{\rho}$	0.04	0.12	
	Variance Inflation Factor	1.12	1.53	
Total	Number of Families	31	35	2.53
	% Infected Subjects	53.78	31.54	
	Mean Number of Participants/Family	3.84	3.71	
	$\hat{k}$	0.79	0.77	

Table 4.4
Weekly Prevalence of Smokeless Tobacco Use Among Adolescents
Summary Statistics Classified by Stratum, Intervention, and Outcome

Stratum (Number of Participants)	Variable	Intervention Group		
		Existing Curriculum	Smoke Free Generation	Odds Ratio
≤ 100 per School	Number of Schools	4	7	1.98
	Prevalence (%)	5.57	2.90	
	Intracluster Correlatioa Coefficient, β	0.0087	0.0003	
	Variance Inflation Factor	1.63	1.02	
>100 per School	Number of Schools	8	5	1.24
	Prevalence (%)	6.29	5.13	
	Intracluster Correlation Coefficient, β	0.0016	0.0204	
	Variance Inflation Factor	1.29	4.31	
Total	Number of Schools	12	12	1.45
	Prevalence (%)	6.15	4.33	
	Mean Number of Participants/School	123.25	111.75	
	$\hat{k}$	0.78	0.83	

Table 4.5
Parasite Screening and Treatment Trial
Stratified Two-Tailed Tests of the Null Hypothesis
Ho: Treatment and Screening Does Not
Affect the Risk of Parasitic Infection

Method	ρ	Test Statistic (p-value)
1. Mantel-Haenszel	-	12.45 (0.0004)
2. Cluster-Level F Test	-	12.85 (0.0007)
Extended Mantel-Haenszel	-	10.88
3. Approximate Test		(0.0010)
4. Exact Test		(0.0008)
5. Ratio-Estimator Approach	-	9.58 (0.0020)
6. Adjusted Mantel-Haenszel	0.070	11.20 (0.0008)
7. Beta-binomial Likelihood Ratio	0.058	10.88 (0.0010)
8. Model Dependent Wald Test	0.084	10.24 (0.0014)
9. Robust Wald Test		10.81 (0.0010)
10. Model Dependent Score Test		10.46 (0.0012)
11. Robust Score Test		10.25 (0.0014)

Table 4.6
Weekly Prevalence of Smokeless Tobacco Use Among Adolescents
Stratified Two-Tailed Tests of the Null Hypothesis
Ho: There is no Difference between
the Two Smoking Prevention Programs

Method	$\hat{\rho}$	Test Statistic (p-value)
1. Mantel-Haenszel	-	3.22 (0.073)
2. Cluster-Level F Test	-	1.63 (0.216)
Extended Mantel-Haenszel	-	1.63
3. Approximate Test		(0.201)
4. Exact Test		(0.210)
5. Ratio-Estimator Approach	-	2.40 (0.121)
6. Adjusted Mantel-Haenszel	0.0077	1.82 (0.177)
7. Beta-binomial Likelihood Ratio	0.0096	1.07 (0.300)
8. Model Dependent Wald Test	0.0095	1.56 (0.212)
9. Robust Wald Test		2.10 (0.147)
10. Model Dependent Score Test		1.57 (0.211)
11. Robust Score Test		1.77 (0.184)

Table 4.7
Parasite Screening and Treatment Trial
Stratified Odds Ratio Estimates of the Effect of Treatment and Accompanying 95% Confidence Intervals

Method	ρ	Odds Ratio	95% Confidence Interval
1. Woolf	-	2.51	(1.49 , 4.22)
2. Weighted Woolf	0.070	2.57	(1.43 , 4.61)
3. Beta-binomial	0.058	2.63	(1.48 , 4.66)
4. Model Dependent Wald	0.084	2.63	(1.45 , 4.75)
5. Robust Wald			(1.48 4.68)

Table 4.8
Weekly Prevalence of Smokeless Tobacco Use Among Adolescents
Stratified Odds Ratio Estimates of the Effect of Treatment
and Accompanying 95% Confidence Intervals

Method	ρ	Odds Ratio Estimate	95% Confidence Interval
1. Woolf	-	1.37	(0.98 , 1.93)
2. Weighted Woolf	0.0077	1.42	(0.87 , 2.32)
3. Beta-binomial	0.0096	1.32	(0.79 , 2.20)
4. Model Dependent Wald	0.0095	1.39	(0.82 , 2.35)
5. Robust Wald			(0.89 , 2.19)

4.4 Discussion

Donner et al (1990) point out that too many cluster randomization trials are designed without ensuring that there are sufficient numbers of clusters to allow detection of clinically relevant effects. This problem was not apparent in either the parasite screening and treatment trial or the adolescent tobacco use prevention trial. The adolescent tobacco prevention trial was designed to detect a 50% reduction in weekly smoking incidence between the most effective intervention group and the control group assuming an intracluster correlation coefficient of 0.02 (Murray et al, 1992, p. 460). Post-hoc power analyses confirmed that the inability to detect a statistically significant difference arose as a consequence of the small difference between treatment groups and not because of a failure to plan for the effect of clustering (Murray et al, 1992, p. 469).

The analyses presented in the previous section demonstrate the importance of adjusting for the effect of clustering when making inferences about the effect of treatment using data derived from cluster randomization trials. Specifically, the differences between the size of test statistics or between the width of confidence limits is always greater when comparing methods which do and do not adjust for the effect of clustering than when contrasting approaches which make such adjustments.

Among those methods which adjusted for the effect of clustering there was still considerable variability in the size of the test statistics and the width of confidence intervals constructed using data from the adolescent smoking prevention trial. The largest test statistics and widest confidence intervals occurred when inferences were made using Rao and Scott's approach or using robust Wald statistics.

These results were likely due, at least in part, to there never being more than 8 clusters in any stratum or treatment group. Both test statistics were found to be overly liberal when there were few clusters per treatment group in the simulation studies described in the previous chapter. An algebraic explanation for this is provided in Chapter 2.

The p-value calculated using the permutation test is exact. All of the other procedures are based on asymptotic theory. These other p-values and confidence intervals must be cautiously interpreted when there are few clusters such as is the case for the adolescent smoking prevention trial. One limitation of the permutation test provided by the computer package StatXact is that it does not take the variable cluster sizes into account. This will reduce the power of such tests when cluster sizes are highly variable. A further limitation is that it is not possible to make inferences about odds ratios. Exact tests can be constructed which do not have these limitations but are not presently available in the computer package StatXact (see Section 2.4).

The estimates of intracluster correlation obtained using data from Murray et al's (1992) adolescent smoking prevention trial were of the same order of magnitude as coefficients from similar studies. For example, estimates of intracluster correlation could be calculated using data provided by LaPrelle et al (1992 Table 1) who collected data on adolescent smoking behavior from students in 10 different communities. The average of 6 estimates of intracluster correlation calculated using data from different intervention groups and timepoints in the trial was 0.0077. In another study Feng and Grizzle (1992), found an intracluster correlation of 0.0052 for smoking prevalence among 2458 subjects from 24 worksites in Florida.

Estimates of intracluster correlation were also calculated using variance components provided by Koepsell et al (1992 Table 1). These data were obtained from three studies of the prevalence of smoking. The resulting estimates of intracluster correlation ranged from 0.002 to 0.02.

It is generally accepted that infectious diseases spread more rapidly within families than between families (Becker, 1989, p. 11). This heterogeneity in the spread of disease is what creates the possibility that parasitic infections will tend to cluster within families. For example a correlation of 0.24 can be obtained using data from 40 households in which subjects were screened for serological evidence of the parasite *Trypanasoma cruzi* (Smith and Pike, 1976). It is also well accepted that infection from *Ascaris* , one of the parasites studied by Gyorkos et al (1989), clusters within families (Williams et al, 1974). Furthermore the usual inverse relationship between cluster size and the degree of intracluster correlation often results in coefficients which are between 0.1 and 0.5 when families are the unit of randomization (e.g. Donner, 1982). It was therefore somewhat surprising to find such small correlations using the data described by Gyorkos et al (1989).

Why then are the correlations obtained using data from Gyorkos et al (1989) so small? There are several possible explanations.

First, consider that both hookworms and *Ascaris* are not directly transmitted from person to person but rather that the parasites must spend some time in soil before they can infect a new host (Benenson, 1990). Both the climate in Montreal and the availability of more hygienic surroundings than was available in the refugee camps would combine to reduce the spread of disease between family members. This was investigated using baseline data from the screened subjects collected soon after they arrived in Montreal. The correlation averaged across strata was 0.19 which was slightly smaller than the average correlation for screened subjects obtained using data collected at six months. It therefore seems unlikely that the low overall correlation was a consequence of a diminishing dependence between cluster members over time.

An alternative explanation arises upon recognizing that the small summary estimates of correlation displayed in Table 4.5 were primarily a consequence of small correlations among subjects in the control group. A low correlation could result if the natural loss of infectivity among these subjects tended not to cluster in families to the same degree that occurred among subjects in the intervention group. This hypothesis could not be examined since baseline data was not available from subjects in the control group. An estimate of intracluster correlation was calculated using baseline data from the treated subjects. It was only 0.0895. Although it was very small it does not rule out the possibility that the difference in dependence of the risk of infection among control and treated subjects at the six month examination was not a consequence of differential patterns of loss of infection.

Subjects in this study were probably among the healthier members of the refugees camps in Southeast Asia. Less healthy people might not have survived or might not have been permitted into Canada. The low correlation between responses of cluster members might therefore have resulted because the between-family variability for subjects enrolled in the trial was much less than would have occurred if random samples of families from the refugee camps were selected.

In these analyses there was little evidence that cluster size was predictive of the outcome. The stratification was maintained only to illustrate the methods discussed in this thesis. In general stratification should be maintained in the analysis whenever it is used during randomization, following the policy first put forward by Fisher (1935, p. 83).

Similarly the recommended approach to take when a.ıalyzing data derived from cluster randomization trials is to assume that responses of subjects from the same cluster are positively correlated. The assumption of a positive correlation is, after all, just another version of Fisher's policy to incorporate features particular to the design of an experiment in its analysis. This policy is particularly important in cluster randomization trials because even very small correlations can dramatically inflate the variance of the estimated treatment effect when there are many subjects per cluster.

Several authors have developed tests of significance which could be used to determine if the true correlation is greater than zero. Tests of significance for continuous outcome data are described by Donner (1985a) while tests for binary outcome data have recently been summarized by Dean (1992). A limitation of all such approaches is that they will have little power to detect very small correlations (e.g. Donner and Klar, 1993b) especially when there are few clusters in the trial (Paul, Liang and Self, 1989). Given the low power in this instance failure to reject the null hypothesis should not be interpreted as equivalent to accepting the hypothesis that responses of cluster members are uncorrelated.

5. Summary

The primary focus of this thesis was on comparisons of statistical tests of the effect of treatment in stratified cluster randomization trials. This decision to focus on stratified cluster randomization trials was taken because of the relative lack of attention paid to such experimental designs. The purpose of the comparisons was to derive recommendations for researchers faced with the analysis of data collected from stratified cluster randomization trials. The present chapter was included to summarize the findings of this thesis, to present a brief list of recommendations to help select an appropriate method of analysis for stratified cluster randomization trials and to suggest areas for future research.

The methods of analysis which were compared used one of six approaches to construct inferences about the effect of treatment. They include adaptations of linear models, nonparametric tests, simple adjustments of methods originally developed for binomially distributed data, adaptations of methods developed for sample surveys, beta-binomial models and Liang and Zeger's (1986) generalized estimating equations approach.

These methods were compared in three ways: using algebraic comparisons, by simulation and via case studies. The algebraic comparisons were presented in Chapter 2. Four principal results were obtained from the algebraic comparisons. First, all methods were shown to be approximately equivalent when there are the same number of subjects in each cluster and the same number of clusters in each stratum. Second, the validity of methods such as Donald and Donner's (1987) adjusted Mantel-Haenszel test and the model dependent test statistics derived from the theory of generalized estimating equations were shown to depend only on the average correlation between cluster members' responses being fixed rather than depending on the more restrictive common correlation assumption. Third, arguments were put forward suggesting that the methods which use robust estimates of variance would require larger numbers of clusters per treatment group to ensure their validity than other approaches. Fourth, tests of significance constructed using the generalized estimating equations approach were shown to be relatively simple algebraic extensions of standard procedures.

The complexity of most of the methods restricts algebraic comparisons to fairly simple and unrealistic situations in which there is no variability in cluster size. Their small sample properties were therefore compared by simulation. The simulation studies were limited to comparisons of eleven test statistics and trials in which there were either 2 or 4 strata and 100 subjects per cluster, on average. Results of the simulation study were presented in Chapter 3. Similarities and differences among these 11 test statistics were illustrated in Chapter 4 using data from two cluster randomization trials.

The simulation study indicated that the type 1 error rates of all of these test statistics were overly liberal, albeit to varying degrees, if there were only 20 clusters per treatment group. Rejection rates near nominal levels were obtained when there were 40 clusters per treatment group. Methods which employed cluster-level analyses and relatively simple non-iterative methods such as the adjusted Mantel-Haenszel χ^2 test statistic (Donald and Donner, 1987) tended to have rejection rates nearest the five percent nominal level. One cost associated with using such methods, however, is that there can be up to a 10 percent difference in power in comparison to parametric methods such as the beta-binomial likelihood ratio test, at least when cluster sizes are variable and $\rho = 0.1$.

Researchers have advocated for the use of test statistics which use robust variance estimators to avoid having to make, possibly unrealistic, assumptions about the degree of correlation between cluster members. For example, procedures such as the adjusted Mantel-Haenszel χ^2 test assume that the average correlation between responses of cluster members does not vary across clusters, treatment groups or strata. The results of the simulation study demonstrate that violations of this assumption does not necessarily invalidate these test statistics, at least when the degree of correlation between cluster members' responses varies as a function of cluster size in accordance with "Smith's Law" (Proctor, 1985).

These results lead to the following recommendations to researchers testing the effect of treatment in stratified cluster randomization trials.

1. Exact tests constructed using randomization theory should be used to make inferences concerning the effect of treatment when there are 20 or fewer clusters per treatment group.

2. Approximate test statistics using cluster-level analyses or simple extensions of Mantel-Haenszel tests are appropriate if there are more than 20 clusters per treatment group.
3. If there are 40 or more clusters per treatment group then more sophisticated and powerful methods such as the model dependent score tests constructed using Liang and Zeger's (1986) generalized estimating equations approach can be used.

The simulation studies upon which these recommendations were based only examined trials in which there were 100 subjects per cluster, on average, were limited to stratified trials with two or four strata and either 20, 40 or 80 clusters per treatment group. Furthermore no comparisons were drawn between the rejection rates or power differences likely to arise as a consequence of using pair-matched or completely randomized designs instead of a stratified cluster randomization trial. Finally there was no consideration given to the possible gains in power obtained when adjusting for individual-level baseline risk factors. These limitations suggest that additional research is needed.

In particular comparisons need to be drawn between the three different cluster randomization designs (completely randomized, pair-matched, stratified) with the focus being fixed on community intervention trials. Economic and administrative considerations usually constrain such trials to randomly assigning treatment to very few clusters (Koepsell, Wagner, Cheadle et al, 1992) complicating any analyses of data collected from them.

There are few appropriate methods of analysis for such data. Methods which might prove appropriate in community intervention trials are t-tests calculated on the cluster means and permutation tests, as suggested by Williams (1988). There is some evidence that such t-tests have valid type I error rates in matched-pairs and completely randomized designs even when there are very few clusters per teatment group (Donner and Donald (1987), Donner and Klar (1993b)).

There are two ways of increasing the power of tests of the effect of treatment. One approach attempts to adjust at the design stage by using stratified or matched-pairs designs rather than completely randomized designs. Conflicting results about the gain in power obtained by using a matched-pairs rather than a completely randomized design were found by Freedman, Green and Byar (1990) and by Martin, Diehr, Perrin and Koepsell (1992). The conflict might be due to whether or not adjustment is made for the difference in degrees of freedom between the two designs. Comparisons need to be extended to include stratified designs.

An alternative approach to increasing power is to adjust during analysis. Differences between these approaches are likely to be irrelevant when there are large numbers of clusters (Grizzle, 1982). Such similarities may not hold in community intervention trials.

Inferences about the effect of treatment can be enhanced by calculating confidence limits. Such limits focus attention on the likely size of the effect of treatment which is often of greater scientific interest than whether or not the null hypothesis should be rejected. The work done in this thesis on comparisons of test statistics needs to be extended to comparisons of the bias and precision of estimates of treatment effect and should also compare the coverage for associated methods of calculating confidence limits.

The simplest approach used to calculate confidence limits is based on inverting Wald test statistics. This approach was discussed in Chapter 2 and Chapter 4 of this thesis. Donner and Klar (1993a) summarize a number of different methods used to calculate estimates of confidence limits in cluster randomization trials concentrating on completely randomized and pair-matched cluster randomization trials. All of these methods follow the approach of inverting Wald test statistics.

Several researchers have pointed out that Wald tests can exhibit aberrant behavior when the alternative hypothesis is true (Hauck and Donner (1977), Mantel (1987), Vaeth (1985)). For example Wald tests of coefficients from logistic regression models are not strictly increasing as a function of the odds ratio. It is possible for such test statistics to get smaller as the odds ratio increases. This problem will also affect the calculation of confidence intervals.

An additional difficulty with Wald statistics is that they can require a greater sample size to assure their validity as compared to score tests. This was seen in the simulation studies described in Chapter 3 for comparisons of robust Wald and score tests.

Both of these concerns can be addressed if confidence limits are calculated by inverting score tests as suggested by Moore and Tsiatis (1991). The resulting confidence limits might have more accurate coverage for smaller sample sizes (Minkin, 1993). Such a result was noted by Vollset (1993) for the calculation of confidence intervals for a binomial proportion.

References

ACS Workshop. Methodological Issues in the Evaluation of Community Intervention Programs. Newcastle, Australia September 1992.

Agresti A. Categorical Data Analysis. Wiley NY 1990.

Agresti A. A survey of exact inference for contingency tables. Statistical Science 1992;7(1):131-177.

Agresti A, Mehta CR, Patel NR. Exact inference for contingency tables with ordered categories. Journal of the American Statistical Association 1990;85:453-458.

Ahrens H, Pincus R. On two measures of unbalancedness in a one-way model and their relation to efficiency. Biometrical Journal 1981;23(3):227-35.

Aitchison J, Shen SM. Logistic-normal distributions: some properties and uses. Biometrika 1980;67(2):261-272.

Altman DG, Bland JM. Improving doctor's understanding of statistics. Journal of the Royal Statistical Society (Series A) 1991;154(2):223-67.

Amberson Jr. JB., McMahon BT, Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. American Review of Tuberculosis 1931;24:401-35.

Anderson DA, Aitkin M. Variance component models with binary response: interviewer variability. Journal of the Royal Statistical Society (Series B) 1985;47(2):203-210.

Anderson RM, May RM. Infectious Diseases of Human. Dynamics and Control. Oxford University Press Oxford 1991.

Anderson PK. Survival analysis 1982-1991: The second decade of the proportional hazards regression model. Statistics in Medicine 1991;10:1931-41.

Armitage P. Biometry and medical statistics. Biometrics 1985;41(4):823-33.

Armitage P. Fifty years of statistics: some reminiscences and reflections. Liason 1991a;35-40.

Armitage P. Obituary: Sir Austin Bradford Hill, 1897-1991. Journal of the Royal Statistical Society (Series A) 1991b;154(3):482-4.

Ashby M, Neuhaus JM, Hauck WW, Bacchetti P, Heilbron DC, Jewall NP, Segal MR, Fusaro RE. An annotated bibliography of methods for analyzing correlated categorical data. Statistics in Medicine 1992;11:67-99.

Ast DB, Schlesinger ER. The conclusion of a ten-year study of water fluoridation. American Journal of Public Health 1956;45(3):265-271.

Barcikowski RS. Statistical power with group mean as the unit of analysis. Journal of Educational Statistics 1981;6(3):267-85.

Bass MJ, McWhinney, Donner A. Do family physicians need medical assistants to detect and manage hypertension? Canadian Medical Association Journal 1986;134:1247-55.

Basu D. Randomization analysis of expe.imental data: the fisher randomization test. Journal of the American Statistical Association 1980;75(371):575-595.

Becker N. Analysis of Infectious Disease Data. Chapman and Hall London 1989.

Benenson AS (Ed.). Control of Communicable Diseases in Man 15'th Ed American Public Health Association Washington DC 1990.

Berkow R, Fletcher AJ et al. The Merck Manual 15'th Ed. Merck and Company, Rahway, NJ 1987.

Best JA, Flay BR, Towson SMJ, Ryan KB, Ferry CL, Brown KS, Kersell MW, D'Avernas JR. Smoking prevention and the concept of risk. Journal of Applied Social Psychology 1984;14(3):257-273.

Bradley JF. Robustness? British Journal of Mathematical and Statistical Psychology 1978;31:144-152.

Binder DA. Fitting Cox's proportional hazards models from survey data. Biometrika 1992;79(1):139-47.

Birch MW. The detection of partial association, II: the general case. Journal of the Royal Statistical Society (Series B) 1965;27:111-124.

Black RE, Dykes AC, Anderson KE, Wells JG, Sinclair SP, Gary W Jr, Hatch MH, Gangarosa EJ. Handwashing to prevent diarrhea in day-care centers. American Journal of Epidemiology 1981;113(4):445-451.

Blum D, Feachem RG. Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology. International Journal of Epidemiology 1983;12(3):357-65.

Boos DD. On generalized score tests. American Statistician 1992;46(4):327-333.

Brass W. Models of birth distributions in human populations. Bulletin of the International Statistical Institute 1958;36:165-178.

Breslow N. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. Journal of the American Statistical Association 1990;85:565-71.

Bryk AS, Raudenbush SW. Hierarchical Linear Models. Applications and Data Analysis Methods. SAGE Publications Newbury Park 1992.

Buck C, Donner A. The design of controlled experiments in the evaluation of non-therapeutic interventions. Journal of Chronic Diseases 1982;35: 531-38.

Bull SB, Pederson LL. Variances for polytomous logistic regression using complex survey data. Proceedings of the Survey Methods Section, ASA Annual Meetings 1987.

Buse A. The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note. American Statistician 1982;36(3):153-157.

Bush PJ, Zuckerman AE, Theiss PK, Taggart VS, Horowitz C, Sheridan MJ, Walter HJ. Cardiovascular risk factor prevention in black schoolchildren: Two-year results of the "Know Your Body" program. American Journal of Epidemiology 1989;129(3):466-482.

Byar DP. The design of cancer prevention trials. Recent Results in Cancer Research 1988;111:34-48.

Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. Journal of the American Statistical Association 1992;87:651-8.

Chalmers TC, Schroeder B. Controls in journal articles. New England Journal of Medicine 1979;301(23): 1293.

Chatfield C, Goofheart GJ. The beta-binomial model for consumer purchasing behaviour. Applied Statistics 1970;19:240-250.

Cochran WG. Analysis of variance for percentages based on unequal numbers. Journal of the American Statistical Association 1943;38:287-301.

Cochran WG. Fisher and the Analysis of Variance. in R.A. Fisher: An Appreciation. Fienberg SE, Hinkley (Eds.). Springer-Verlag Berlin 1989.

Cochran WG. Sampling Techniques. John Wiley and Sons. New York 1953.

Cochran WG. Sampling Techniques 3'rd Ed. John Wiley and Sons. New York 1977.

Cochran WG. Some methods of strengthening the common χ^2 test. Biometrics 1954;10:417-451.

Cohen J. Statistical Power Analysis for the Behavioral Sciences 2'nd Ed. Lawrence Erlbaum Assoc., Hillsdale 1988.

Collett, D. Modelling Binary Data. Chapman and Hall, London 1991.

Comstock GW. Isoniazid prophylaxis in an undeveloped area. American Review of Respiratory Diseases 1962;86:810-22.

Comstock GW. Uncontrolled ruminations on modern controlled trials. American Journal of Epidemiology 1978;108(2):81-84.

Conover WJ. Practical Nonparametric Statistics 2'nd Ed. John Wiley & Sons, New York 1980.

Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. American Statistician 1981;35(3):124-33.

Cook RD, Weisberg S. Residuals and influence in regression. Chapman and Hall, London 1982.

Cornfield J, Mitchell S. Selected risk factors in coronary disease. Possible intervention effects. Archives of Environmental Health 1969;19:382-94.

Cornfield J. Randomization by group: a formal analysis. American Journal of Epidemiology 1978;108(2): 100-102.

Cox DR, Hinkley DV. Theoretical Statistics. Chapman and Hall, London 1974.

Cox DR, Snell EJ. Analysis of Binary Data. 2'nd Ed. Chapman and Hall London 1989.

Crowder MJ. Beta-binomial ANOVA for proportions. Applied Statistics 1978;27(1): 34-37.

Crowder MJ. Inference about the intraclass correlation coefficient in the Beta-binomial ANOVA for proportions. Journal of the Royal Statistical Society (Series B) 1979;41(2):230-4.

Crump KS, Howe RB, Kodell RL. Permutation Tests for detecting Teratogenic Effects, in Chapter 17 of Krewski D. and Franklin C. (Eds.) Statistics in Toxicology. Gordon and Breach Science Publishers, New York 1991.

Cullen JW. Phases in cancer control: intervention research. pages 1-11 in Evaluating Effectiveness of Primary Prevention of Cancer Hakama M et al (Eds.) Lyon, IARC 1990.

Dale AI. A History of Inverse Probability, From Thomas Bayes to Karl Pearson. Springer-Verlag NY 1991.

Dallal GE. Paired Bernoulli trials. Biometrics 1988;44:253-7.

D'Arcy Hart, P. History of randomised control trials. Lancet 1972;1:965.

Day NE, Byar DP. Testing hypotheses in case-control studies - equivalence of Mantel-Haenszel statistics and logit score tests. Biometrics 1979;35:623-30.

Dean CB. A robust property of pseudo-likelihood estimation for count data. Journal of Statistical Planning and Inference 1993;35:309-317.

Dean CB. Testing for overdispersion in Poisson and binomial models. Journal of the American Statistical Association 1992;87(418):451-7.

Diwan VK, Eriksson B, Sterky G, Tomson G. Randomization by group in studying the effect of drug information in primary care. International Journal of Epidemiology 1992;21(1):124-30.

Donald A. The Analysis of Clustered Data in Sets of 2x2 Contingency Tables. Ph.D. Thesis, University of Western Ontario, 1984.

Donald A, Donner A. Adjustments to the Mantel-Haenszel Chi-square statistic and odds ratio variance estimator when the data are clustered. Statistics in Medicine 1987;6:491-499.

Donald A, Donner A. Simulation study of the analysis of sets of 2x2 contingency tables under cluster sampling: estimation of a common odds ratio. Journal of the American Statistical Association 1990; 85(410):537-543.

Donner A. An empirical study of cluster randomization. International Journal of Epidemiology 1982;11(3):283-286.

Donner A. The analysis of intraclass correlation in multiple samples. Annals of Human Genetics 1985a;49;75-82.

Donner A. A regression approach to the analysis of data arising from cluster randomization. International Journal of Epidemiology 1985b;14(2):322-6.

Donner A. Odds ratio inference with dependent data: a relationship between two procedures. Biometrika 1987;74(1):220.

Donner A. Sample size requirements for stratified cluster randomization designs. Statistics in Medicine 1992;11:743-750.

Donner A. Statistical methodology for paired cluster designs. American Journal of Epidemiology 1987;126(5):972-79.

Donner A. Statistical methods in ophthalmology: an adjusted chi-square approach. Biometrics 1989;45:605-611.

Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. American Journal of Epidemiology 1981;114(8):906-914.

Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989.

International Journal of Epidemiology 1990;19(4):795-800.

Donner A, Donald A. Analysis of data arising from a stratified design with the cluster as unit of randomization. Statistics in Medicine 1987;6:43-52.

Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Statistics in Medicine 1992;11:1511-1519.

Donner A, Eliasziw M, Klar N. A comparison of methods for testing homogeneity of proportions in teratologic studies. Statistics in Medicine 1993 In Press.

Donner A, Hauck W. Estimation of a common odds ratio in paired-cluster randomization designs. Statistics in Medicine 1989;8:599-607.

Donner A, Klar N. Confidence interval construction for effect measures arising from cluster randomization trials. Journal of Clinical Epidemiology 1993a;46(2):123-131.

Donner A, Klar N. Statistical considerations in the design of community intervention trials. Unpub. 1993b.

Donner A, Koval JJ. A procedure for generating group sizes from a one-way classification with a specified degree of imbalance. Biometrical Journal 1987;2:181-187.

Donner A, Koval JJ. The effect of imbalance on significance-testing in one-way model. II Analysis of Variance. Communications in Statistics - Theory and Methods 1989;18(4):1239-50.

Duffy SW, South MC, Day NE. Cluster randomization in large public health trials: The importance of antecedent data. Statistics in Medicine 1992;11:307-16.

Dunn OJ, Clark VA. Applied Statistics. Analysis of Variance and Regression 2'nd Ed. Wiley 1987.

Dwass M. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 1957;28:181-187.

Dwyer T, Coonan WE, Leitch DR, Hetzel BS, Baghurst RA. An investigation of the effects of daily physical activity on the health of primary school students in South Australia. International Journal of Epidemiology 1983;12(3):308-313.

Edgington ES. Approximate randomization tests. Journal of Psychology 1969;72:143-149.

Edgington E.S. Randomization Tests 2'nd Ed. Marcel Dekker, New York 1987.

Efron B, Hinkley DV. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika 1978;65(3):457-487.

EGRET: Epidemiological, Graphics, Estimation and Testing Package. Statistics and Epidemiology Research Corporation, Seattle WA 98105, 1989.

van Elteren PH. On the combination of independent two sample tests of Wilcoxon. Bulletin of the International Statistical Institute 1960;37:351-361.

Evans CE, Haynes RB, Birkett NJ, Gilbert JR, Taylor DW, Sackett DL, Johnston ME, Hewston SA. Does a mailed continuing education program improve physician performance? Journal of the American Medical Association 1986;255(4):501-504.

Farquhar JW. The community-based model of life style intervention trials. American Journal of Epidemiology 1978;108(2):103-111.

Farr BM, Hendley JO, Kaiser DL, Gwaltney JM. Two randomized controlled trials of virucidal nasal tissues in the prevention of natural upper respiratory infections. American Journal of Epidemiology 1988;128(5):1162-1172.

Fawzi WW, Chalmers TC, Herrera MG, Mosteller F. Vitamin A supplementation and child mortality. Journal of the American Medical Association 1993;269(7):898-903.

Feng Z, Grizzle JE. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. Statistics in Medicine 1992;11:1607-1614.

Ferebee SH, Mount FW, Murray FJ, Livesay VT. A controlled triai of isoniazid prophylaxis in mental institutions. American Review of Respiratory

Diseases 1963; 88(2):161-175.

Finney DJ. Probit Analysis. A Statistical Treatment of the Sigmoid Response Curve. At The University Press Cambridge 1947.

Fisher RA. The arrangement of field experiments. Journal of Ministry of Agriculture 1926;33:503-13

Fisher RA. The Design of Experiments. Edinburgh: Oliver and Boyd 1935.

Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Preventive Medicine 1986;15:451-474.

Fleiss JL. Statistical Methods for Rates and Proportions. 2'nd Ed. Wiley 1981.

Fleiss JL. The Design and Analysis of Clinical Experiments. New York: Wiley 1986.

Freedman LS, Green SB, Byar DP. Assessing the gain in efficiency due to matching in a community intervention study. Statistics in Medicine 1990;9:943-952.

Fry JS, Lee PN. Stratified rank tests. Applied Statistics 1988;37(2):264-6.

Fung KY, Kreski D, Rao JNK, Scott AJ. Tests for trend in developmental toxicity experiments with correlated binary data. Journal of Risk Analysis 1993 In Press.

Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the community intervention trial for smoking cessation. Controlled Clinical Trials 1992;13:6-21.

Gail MH, Tan WY, Piantadosi S. Tests for no treatment effect in randomized clinical trials. Biometrika 1988;75(1):57-64.

Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984;71(3):431-444.

Galbraith JI. The interpretation of a regression coefficient. Biometrics 1991;47:1593-1595.

Ganio LM, Schafer DW. Diagnostics for overdispersion. Journal of the American Statistical Association 1992;87:795-804.

Gart JJ. On the combination of relative risks. Biometrics 1962;18:601-610.

Gart JJ. Pooling 2x2 tables: asymptotic moments of estimators. Journal of the Royal Statistical Society (Series B) 1992;54(2):531-539.

Giesbrecht FG, Burns JC. Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. Biometrics 1985;41:477-486.

Gillum RF, Williams PT, Sondik E. Some considerations for the planning of total-community prevention trials - When is sample size adequate? Journal of Community Health 1980;5(4):270-8.

Gladen B. The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects. Journal of the American Statistical Association 1979;74(366):278-283.

Glass GV, Hopkins KD. Statistical Methods in Education and Psychology. 2'nd Ed. Prentice-Hall. 1984.

Glynn RJ, Rosner B. Comparison of alternative logistic regression models for paired ophthalmologic data. Abstract from the Annual Meeting of the American Statistical Association Boston, 1992.

Godambe VP. An optimum property of regular maximum likelihood estimation. Annals of Mathematical Statistics 1960;31:1208-1211.

Godambe VP, Kale BK. Estimating functions: an overview. Chapter 1, pp. 3-20, in Godambe VP (Ed.). Estimating Functions. Clarendon Press, Oxford 1991.

Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 1986;73(1):43-56.

Goldstein H. Multilevel Models in Educational and Social Research. Charles Griffin & Company Ltd. London 1987.

Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika 1991;78(1):45-51.

Grant A, Valentin L, Elbourne D, Alexander S. Routine formal fetal movement counting and risk of antepartum late death in normally formed singletons. Lancet 1989;345-349.

Green MS, Zaaide Y. Sibship size as a risk factor for Hepatitis A infection. American Journal of Epidemiology 1989;129(4):800-805.

Greenhouse SW, Geisser S. On methods in the analysis of profile data. Psychometrika 1959;24(2):95-112.

Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. American Journal of Epidemiology 1987;125(5):761-768.

Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. International Journal of Epidemiology 1989; 18(1):269-74.

Greenwood M, Yule GU. The statistics of anti-typhoid and anti-cholera inoculations, and the interpretations of such statistics in general. Proceedings of the Royal Society of Medicine. Section of Epidemiology and State Medicine 1915;8(2):113-194.

Griffiths DA. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics 1973;29:637-648.

Gyorkos TW. A Comparative Study to Determine the Effects of Screening for Intestinal Parasites in Newly-Arrived Southeast Asian Refugees. PhD Thesis. McGill University 1985.

Gyorkos TW, Frappier-Davignon L, MacLean JD, Viens P. Effect of screening and treatment on imported intestinal parasite infections: results from a randomized controlled trial. American Journal of Epidemiology 1989;129(4):753-761.

Halperin M, Cornfield J, Mitchell SC. Effect of diet on coronary-heart-disease mortality. Lancet 1973;7826:438-9.

Hansen MH, Hurwitz WN Madow WG. Sample Survey Methods and Theory. Vol I. Methods and Applications. New York: Wiley 1953.

Hardy G, Littlewood, Polya G. Inequalities 2'nd Ed. Cambridge University Press, Cambridge 1991.

Haseman JK, Hogan MD. Selection of the experimental unit in teratology studies. Teratology 1975;12(2):165-71.

Haseman JK, Kupper LL. Analysis of dichotomous response data from certain toxicological experiments. Biometrics 1979;35:281-93.

Haseman JK, Soares ER. The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects. Mutation Research 1976;41:277-88.

Hauck WW. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics 1979;35:817-819.

Hauck WW. Odds ratio inference from stratified samples. Communications in Statistics - Theory and Methods 1989;18(2):767-800.

Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association 1977;72:851-853.

Hauck WW, Gilliss CL, Donner A, Gortner S. Randomization by cluster. Nursing Research 1991;40(6):356-8.

Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. Journal of Clinical Epidemiology 1991;44(1):77-81.

Healy MJR. Animal litters as experimental units. Applied Statistics 1972;21(2):155-9.

Henderson MM, Meinert CL. A plea for a discipline of health and medical evaluation. International Journal of Epidemiology 1975;4(1):11-23.

Hinkley DV. Jackknife Methods in Kotz S (Ed.), Encyclopedia of Statistical Sciences New York Wiley 1984 Volume 4.

Hocking RR, Green JW, Bremer RH. Variance-component estimation with model-based diagnostics. Technometrics 1989;31(2):227-239.

Hodges JL, Lehmann EL. Rank methods for combination of independent experiments in analysis of variance. Annals of Mathematical Statistics 1962;33:482-97.

Hopkins KD, The unit of analysis: group means versus individual observations. American Educational Research Journal 1982;19(1):5-18.

Horwitz O, Magnus K. Epidemiologic evaluation of chemoprophylaxis against tuberculosis. American Journal of Epidemiology 1974;99(5):333-342.

Hosmer Jr. DW, Lemeshow S. Applied logistic regression. Wiley New York 1989.

Hougaard P, Harvald B, Holm NV. Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930. Journal of the American Statistical Association 1992;87(417):17-24.

Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. Statistics in Medicine 1988;8:1195-1201.

Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the 5'th Berkeley Symposium on Mathematical Statistics and Probability 1967;1:221-234.

Hulley SB. Symposium on CHD prevention trials: design issues in testing life style interventions. Introduction. American Journal of Epidemiology 1978;108(2): 85-86.

Humphreys K, Carr-Hill R. Area variations in health outcomes: artefact or ecology. International Journal of Epidemiology 1991;20(1):251-258.

Jennrich R, Sampson P. General mixed model analysis of variance (BMDP 3V) pages 1025-43 in Dixon WJ (Ed.). BMDP Statistical software Manual Vol2. University of California Press, Berkeley 1988.

Johnson NL, Kotz S. Distributions in Statistics. Continuous Univariate Distributions -2. John Wiley and Sons New York 1970.

Jones B, Kenward MG. Design and Analysis of Cross-Over Trials. London: Chapman and Hall 1989.

Jooste PL, Yach D, Steenkamp HJ, Botha JL, Rossouw JE. Drop-out and newcomer bias in a community cardiovascular follow-up study. International Journal of Epidemiology 1990;19(2):284-289.

Kalter H. Choice of the number of sampling units in teratology. Teratology 1974;9(3):257-8.

Kennedy Jr. WJ, Gentle JE. Statistical Computing, Marcel Dekker; New York 1980.

Kent JT. Robust properties of likelihood ratio tests. Biometrika 1982;69(1):19-27.

Kenward MG, Jones B. Alternative approaches to the analysis of binary and categorical repeated measurements. Journal of Biopharmaceutical Statistics 1992;2(2):137-170.

Kleinman JC. Proportions with extraneous variance: single and independent samples. Journal of the American Statistical Association 1973;68:46-54.

Koch GG, Imrey PB, Singer JM, Atkinson SS, Stokes ME. Analysis of Categorical Data. University of Montreal Press 1985.

Koepsell TD, Martin DC, Diehr PH, Psaty BM, Wagner EH, Perrin EB, Cheadle A. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis o. variance approach. Journal of Clinical Epidemiology 1991;44(7):701-13.

Koepsell TD, Wagner EH, Cheadle AC et al. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. Annual Review of Public Health 1992;13:31-57.

Korff M von, Koepsell T, Curry S, Dieher P. Multi-level analysis in epidemiologic research on health behaviors and outcomes. American Journal of Epidemiology 1992;135(10):1077-1082.

Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the study design. American Journal of Public Health 1991;81(9):1166-1173.

Kraemer HC. Ramifications of a population model for kappa as a coefficient of reliability. Psychometrika 1979;44(4):461-472.

Kramer MS. Clinical Epidemiology and Biostatistics. A Primer for Clinical Investigators and Decision Makers. Springer-Verlag 1988.

Kruskal WH. Historical notes on the Wilcoxon unpaired two-sample test. Journal of the American Statistical Association 1957;52:356-360.

Kupper LL, Haseman JK. The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics 1978;34:69-76.

Kupper LL, Portier C, Hogan MD, Yamamoto E. The impact of litter effects on dose-response modeling in teratology. Biometrics 1986;42:85-98.

Kuritz SJ, Landis JR, Koch GG. A general overview of Mantel-Haenszel methods: applications and recent developments. Annual Review of Public Health 1988;9:123-160.

Lancaster HO. Significance tests in discrete distributions. Journal of the American Statistical Association 1961;56:223-34.

Landis JR, Koch GG. A one-way components of variance model for categorical data. Biometrics 1977;33:671-679.

LaPrelle J, Bauman KE, Koch GG. High intercommunity variation in adolescent cigarette smoking in a 10-community field experiment. Evaluation Review 1992;16(2):115-130.

Lee PM. Bayesian Statistics: An Introduction. Oxford University Press 1989.

Leisenring W, Ryan L. The effect of missing covariates in a clustered data setting. Abstract from the Annual Meeting of the American Statistical Association Boston, 1992.

Lehmann E.L. Nonparametrics. Statistical Methods Based on Ranks. Holden-Day, In. Oakland, California 1975.

Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association 1989;84:1074-1078.

Liang KY. Odds ratio inference with dependent data. Biometrika 1985;72(3):678-682.

Liang KY, Beaty TH, Cohen BH. Application of odds ratio regression models for assessing familial aggregation from case-control studies. American Journal of Epidemiology 1986;124(4):678-683.

Liang KY, Self SG, Chang YC. Modelling marginal hazards in multivariate failure time data. Journal of the Royal Statistical Society (Series B) 1993;55(^):441-453.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22.

Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. Journal of the Royal Statistical Society (Series B) 1992;54(1):3-40.

Lilienfeld AM. Ceteris Paribus: The evolution of the clinical trial. Bulletin of the History of Medicine 1982;56:1-18.

Lindquist EF. Statistical Analysis in Educational Research. Houghton Mifflin Company Boston 1940.

Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika 1991;78(1):153-160.

Lloyd DM, Alexander HM, Callcott R, Dobson AJ, Hardes GR, O'Connell DL, Leeder SR. Cigarette smoking and drug use in schoolchildren:III- Evaluation of a smoking prevention education program. International Journal of Epidemiology 1983:12(1):51-58.

Luning KG, Sheridan W, Ytterborn KH, Gullberg U. The relationship between the number of implantations and the rate of intra-uterine death in mice. Mutation Research 1966;3:444-51.

Magnusson D, Bergman LR (Eds.). Data quality in longitudinal research. Cambridge University Press Cambridge 1990.

Mainland D. Elementary Medical Statistics. The Principles of Quantitative Medicine. WB Saunders Company, Philadelphia 1952. pp 114-5.

Mak TK. Analysing intraclass correlation for dichotomous variables. Applied Statistics 1988;37(3):344-352.

Manly BFJ. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London 1991.

Mantel N. Chi-square tests with one-degree of freedom: extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association 1963;58:690-700.

Mantel N. Ridit analysis and rclated ranking procedures - use at your own risk. American Journal of Epidemiology 1979;109(1):25-33.

Mantel N. Understanding Wald's test for exponential families. American Statistician 1987;41(2):147-148.

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 1959;22:719-48.

Martin DC, Diehr PH, Perrin EB, Koepsell TD. The effect of matching on the power of randomized community intervention studies. Statistics in Medicine 1993;12:329-338.

Marubini E, Correa ML, Milani S. Analysis of dichotomous response variables in teratology. Biometrical Journal 1988;30(8):965-74.

Mauritsen, RH. Logistic regression with random effects. University of Washington Ph.D. Thesis 1984.

McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society (Series B) 1980;42:109-142.

McCullagh P, Nelder JA. Generalized Linear Models. Chapman and Hall London 1983.

McCullagh P, Nelder JA. Generalized Linear Models 2'nd. Ed. Chapman and Hall London 1989.

McDonald CJ, Hui SL, Smith DM, Tierney WM, Cohen SJ, Weinberger M, McCabe GP. Reminders to physicians from an introspective computer medical record. A two-year randomized trial. Annals of Internal Medicine 1984;100:130-138.

McNemar, Q. Book review of Lindquist EF. Statistical Analysis in Educational Research. Psychological Bulletin 1940;37(9):746-748.

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. British Medical Journal 1948;2:769-782.

Mehta CR, Patel N, Senchaudhuri P. Exact stratified linear rank tests for ordered categorical and binary data. Journal of Computational and Graphical Statistics 1992;1(1):21-40.

Meier P. Stratification in the design of a clinical trial. Controlled Clinical Trials 1981;1:355-361.

Mickey RM, Goodwin GD. The magnitude and variability of design effect for community intervention trials. American Journal of Epidemiology 1993;137(1):9-18.

Mickey RM, Goodwin GD, Costanza MC. Estimation of the design effect in community intervention studies. Statistics in Medicine 1991;10:53-64.

Minkin S. re: On the computation of likelihood ratio and score test based confidence intervals in generalized linear models. Statistics in Medicine 1993;12:989.

Montgomery DC. Design and Analysis of Experiments 2'nd Ed. Wiley NY 1984.

Mood AM, Graybill FA, Boes DC. Introduction to the Theory of Statistics 3'rd Ed. McGraw-Hill Book Comp New York 1974.

Moore DF. Asymptotic properties of moment estimators for overdispersed counts and proportions. Biometrika 1986;73(3):583-8.

Moore DF. Method of Moments Estimation for Overdispersed Counts and Proportions. Unpublished Ph.D. Thesis Seattle 1985.

Moore DF, Tsiatis A. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. Biometrics 1991;47:383-401.

Moran PAP. An Introduction to Probability Theory. Clarendon Press. Oxford 1968.

Moulton LH, Zeger SL. Analyzing repeated measures on generalized linear models via the bootstrap. Biometrics 1989;45:381-394.

Munoz A, Rosner B, Carey V. Regression analysis in the presence of heterogenous intraclass correlations. Biometrics 1986;42:653-658.

Murray DM, Perry CL, Griffin G et al. Results from a statewide approach to adolescent tobacco use prevention. Preventive Medicine 1992;21:449-472.

Neuhaus JM. Estimation · ficiency and tests of covariate effects with clustered binary data. Technical Report #17. Department of Epidemiology and Biostatitics. University of California, San Francisco 1991.

Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. Statistical Methods in Medical Research 1992;1(3):249-273.

Neuhaus JM, Jewall NP. Some comments on Rosner's multiple logistic model for clustered data. Biometrics 1990;46:523-531.

Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data. University of California at San Francisco, International Statistical Review 1991;59(1):25-35.

Nutbeam D, Smith C, Catford J. Evaluation in health education. A review of progress, possibilities, and problems. Journal of Epidemiology and Community Health 1990;44: 83-89.

O'Hara Hines RJ, Lawless JF. Modelling overdispersion in toxicological mortality data grouped over time. Biometrics 1993;49:107-121.

Paik MC. Parametric variance function estimation for nonnormal repeated measurement data. Biometrics 1992;48:19-30.

Palmer AK. Statistical analysis and choice of sample units. Teratology 1974;10(3):301-2.

Paul SR, Liang KY, Self SG. On testing departure from the binomial and multinomial assumptions. Biometrics 1989;45:231-6.

Paul SR, Mantel N. Model-free analyses of litter-depletion data. The Statistician 1989;38:121-5.

Pearson E. Bayes' theorem, examined in the light of experimental sampling. Biometrika 1925;17:388-442.

Piantadosi S, Byar DP. Design of Prevention Trials. Chapter 16 (pages 441-72) in Moon TE, Micozzi MS (Eds.). Nutrition and Cancer Prevention. Investigating the Role of Micronutrients. Marcel Dekker, Inc. New York 1989.

del Pino, G. The unifying role of iterative generalized least squares in statistical algorithms. Statistical Science 1989;4(4):394-408.

Pocock SJ. Clinical Trials. A Practical Approach. John Wiley and Sons. Chichester 1987.

Pollock TM. Trials of Prophylactic Agents for the Control of Communicable Diseases. A Guide to their Organization and Evaluation. WHO Geneva 1966. Monograph Series No. 52.

Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:1033-1048.

Prentice RL, Farewell VT. Relative risk and odds ratio regression. Annual Review of Public Health 1986;7:35-58.

Proctor CH. Fitting HF Smith's empirical law to cluster variances for use in designing multi-stage sample surveys. Journal of the American Statistical Association 1985;294-300.

Prosser B. ML3: Software for Three-Level Analysis. American Statistician 1991;45(2):155-6.

Puri ML. On the combination of independent two sample tests of a general case. Review of the International Statistical Institute 1965;33(2):229-241.

Rai K, Van Ryzin J. A dose-response model for teratological experiments involving quantal responses. Biometrics 1985;41:1-9.

Rao CR. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Cambridge Philosophical Society 1948;44:50-57.

Rao CR. Linear Statistical Inference and Its Applications. Wiley NY 1975.

Rao JNK, Colin D. Fitting Dose-Response Models and Hypotheses Testing in Teratological Studies. Chapter 15 in Krewski D. and Franklin C. (Eds.) Statistics in Toxicology. Gordon and Breach Science Publishers, New York 1991.

Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics 1992;48(2):577-585.

Rawlings JO. Applied Regression Analysis: A Research Tool. Wadsworth and Brooks, Pacific Grove, California 1988.

Roberts G, Rao JNK, Kumar S. Logistic regression analysis of sample survey data. Biometrika 1987;74(1):1-12.

Robey RR, Barcikowski RS. Type I error and the number of iterations in Monte Carlo studies of robustness. British Journal of Mathematical and Statistical Psychology 1992;45:283-288.

Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. International Statistical Review 1991;58(2):227-240.

Rosner B. Multivariate methods in ophthalmology with application to other paired-data situations. Biometrics 1984;40:1025-1035.

Rosner B. Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. Biometrics 1982;38:105-114.

Rosner B, Hennekens CH. Analytic methods in matched pair epidemiological studies. International Journal of Epidemiology 1978;7(8):367-372.

Rosner B, Milton RC. Significance testing for correlated binary outcome data. Biometrics 1988;44:505-512.

Rotnitzky A, Jewall NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika 1990;77(3):485-97.

Royall R.M. Ethics and statistics in randomized clinical trials. Statistical Science 1991;6(1):52-88.

Royall RM. Model robust confidence intervals using maximum likelihood estimators. International Statistical Review 1986;54(2):221-6.

Salonen JT, Kottke TE, Jacobs Jr. DR, Hannan PJ. Analysis of community-based cardiovascular disease prevention studies - Evaluation issues in the North Karelia Project and the Minnesota Heart Health Program. International Journal of Epidemiology 1986;15(2):176-82.

SAS Institute Inc. SAS/STAT User's Guide, Version 6, 4'th Ed., Volume 1, Carey, NC. SAS Institute Inc 1989.

Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics Bulletin 1946;2(6):110-114.

Scheffe H. The Analysis of Variance. John Wiley and Sons New York 1959.

Schwartz D, Flamant R, Lellouch J. Clinical Trials. Academic Press, London 1980.

Scott AJ, Holt D. The effect of two stage sampling on ordinary least squares methods. Journal of the American Statistical Association 1982;77(380):848-854.

Searle SR. Matrix Algebra for Statistics. Wiley NY 1982.

Searle SR. Mixed models and unbalanced data: wherefrom, whereat and whereto? Communications in Statistics - Theory and Methods 1988;17(4):935-68.

Searle SR, Casella G, McCulloch CE. Variance Components. Wiley NY 1992.

Segal MR, Neuhaus JM. Robust inference for multivariate survival data. Statistics in Medicine 1993;12:1019-1031.

Sharples KJ. Regression analysis of correlated binary data. Ph.D. Thesis, University of Vashington, 1989.

Sharples KJ, Breslow N. Regression analysis of correlated binary data: some small sample results for the estimating equation approach. Journal of Statistical Computing and Simulation 1992;42:1-20.

Sherwin R. Controlled trials of the diet-heart hypothesis: some comments on the experimental unit. American Journal of Epidemiology 1978;108(2):92-99.

Shipley MJ, Smith PG, Dramaix M. Calculation of power for matched pair studies when randomization is by group International Journal of Epidemiology 1989;18(2):457-61.

Shirley EAC. Applications of ranking methods to multiple comparison procedures and factorial experiments. Applied Statistics 1987;36(2):205-13.

Shirley EAC, Hickling R. An evaluation of some statistical methods for analysing numbers of abnormalities found amongst litters in teratology studies. Biometrics 1981;37:819-829.

Simon R. Composite randomization designs for clinical trials. Biometrics 1981;37:723-731.

Skellam JG. A probability distribution from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society (Series B) 1948;10:257 261.

Skrabaneck P. Why is preventive medicine exempted from ethical constraints? Journal of Medical Ethics 1990;16:187-90.

Smith HF. An empirical law describing heterogeneity in the yields of agricultural crops. The Journal of Agricultural Science 1938;28:1-23.

Smith PG, Pike MC. Generalisations of two tests for the detection of household aggregation of disease. Biometrics 1976;32:817-828.

Smith PJ. A comparative study of Cochran-Mantel-Haenszel methods and generalized estimating equations for making inference on common odds in K 2x2 tables. Unpublished 1993.

Sommer A, Djunaedi E, Loeden AA, Tarwotjo I, West KP, Tilden R, Mele L and the ACEH Study Group. Impact of vitamin A supplementation on childhood mortality. A randomized control trial. Lancet 1986;1169-1172.

Speed FM, Hocking RR, Hackney OP. Methods of analysis of linear models with unbalanced data. Journal of the American Statistical Association 1978;73:105-112.

SPIDA. Statistical Package for Interactive Data Analysis. The Statistical Laboratory, Macquarie University, Australia 1992.

Spitzer WO, Feinstein AR, Sackett DL. What is a health care trial? Journal of the American Medical Association 1975,233(2):161-3.

Stanish WM, Taylor N. Estimation of the intraclass correlation coefficient for the analysis of covariance model. American Statistician 1983;37(3):221-224.

Stanton BF, Clemens JD. An educational intervention for altering water-sanitation behaviors to reduce childhood diarrhea in urban Bangledesh. American Journal of Epidemiology 1987;125(2):292-301.

StatXact. StatXact: Statistical Software for Exact Nonparametric Inference, Version 2. Cytel Software, Cambridge, Mass 1991.

Stigler SM. The History of Statistics. The Measurement of Uncertainty before 1900. Harvard University Press Cambridge 1986.

Stiratelli R, Laird N, Ware H. Random-effects models for serial observations with binary response. Biometrics 1984;40:961-971.

Strasser T, Jeanneret O, Raymond L. Ethical Aspects of Prevention Trials Chapter 15 in Ethical Dilemmas in Health Promotion, Doxiadis S. (Ed.). John Wiley & Sons Ltd. 1987.

Surgeon General. The Health Consequences of Using Smokeless Tobacco. A Report of the Advisory Committee to the Surgeon General. US Department of Health and Human Services, Bethesda, Maryland 1986.

Syme SL. Life style intervention in clinic-based trials. American Journal of Epidemiology 1978;108(2):87-91.

Tabar L, Gad A, Lomberg LH, Ljungquist U, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomized trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985;829-832.

Tarone RE. Testing the goodness of fit of the binomial distribution. Biometrika 1979;66(3):585-90.

Thompson WD, Walter SD. A reappraisal of the kappa coefficient. Journal of Clinical Epidemiology 1988;41(10):949-958.

Thornquist MD, Anderson GL. Small sample properties of generalized estimating equations in group-randomized designs with Gaussian response. Unpub 1992.

Turpeinen O. Diet and coronary events. Journal of the American Dietetics Association 1968; 52(3):209-13.

Vaeth M. On the use of Wald's test in exponential families. International Statistical Review 1985;53(2):199-214.

Vollset SE. Confidence intervals for a binomial proportion. Statistics in Medicine 1993;12:809-824.

Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 1943;54:426-482.

Weerasekera DR, Bennett S. Adjustments to the Mantel-Haenszel test for data from stratified multistage surveys. Statistics in Medicine 1992;11:603-616.

Weil CS. Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. Food and Cosmetics Toxicology 1970;8:177-82.

Weil CS. Choice of the number of sampling units in teratology. Teratology 1974;10(3):301.

Welch WJ. Construction of permutation tests. Journal of the American Statistical Association 1990;85:693-8.

Wichman BA, Hill ID. An efficient and portable pseudo-random number generator. Applied Statistics 1982;31:188-190.

Wickens TD. Analysis of contingency tables with between-subjects variability. Psychological Bulletin 1993;113(1):191-204.

Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945;1:80-83.

Williams D, Burke G, Hendley JO. Ascariasis: a family disease. Journal of Pediatrics 1974;84:853-4.

Williams DA. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometri‿s 1975;31:949-952.

Williams DA. Extra-binomial variation in logistic linear models. Applied Statistics 1982;31(2):144-148.

Williams, DA. Dose-response models for teratological experiments. Biometrics 1987;43:1013-6.

Williams DA. Estimation bias using the beta-binomial distribution in teratology. Biometrics 1988a;44:305-309.

Williams DA. Extra-binomial variation in toxicology. Proceedings of the 14'th International Biometric Conference, Namur, Belgium, July 1988b.

Williams, DA. The reliability of tests of hypotheses when overdispersed logistic-linear models are fitted by maximum quasi-likelihood. Biometrical Journal 1991;33:259-270.

Williams PT, Fortmann SP, Farquhar JW, Varady A, Mellen S. A comparison of statistical methods for evaluating risk factor changes in community studies: an example from the Stanford Three-Community Study. Journal of Chronic Diseases 1981;34:565-571.

Wilson DM, Taylor W, Gilbert JR, Best JA, Lindsay EA, Wilms DG, Singer J. A randomized trial of a family physician intervention for smoking cessation. Journal of the American Medical Association 1988;260(11):1570-1574.

Wolfinger R, Tobias R, Sall J. Mixed models: a future direction. SUGI 16. SAS Users Group Int. Sixteenth Annual Conference. New Orleans, Louisiana Feb 17-20, 1991.

Woltei KM. Appendix E. Computer Software for Variance Estimation. Introduction to Variance Estimation. Springer-Verlag, New York 1985.

Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. Journal of the American Statistical Association 1985;80(391):513-24.

Woolf B. On estimating the relation between blood group and disease. Annals of Human Genetics 1955;19:251-253.

Woolson RF, Bean JA, Rojas PB. Sample size for case-control studies using Cochran's statistic. Biometrics 1986;42:927-32.

World Health Organization European Collaborative Group. An international controlled trial in the multifactorial prevention of coronary heart disease. International Journal of Epidemiology 1974;3(3):219-224.

World Health Organization European Collaborative Group. European collaborative trial of multifactorial prevention of coronary heart disease: final report on the 6-year results. Lancet 1986;869-872.

Zeger SL. Discussion of papers on the analysis of repeated categorical response. Statistics in Medicine 1988;7:161-168.

Zeger KY, Liang S, Albert PS. re: The interpretation of a regression coefficient. Biometrics 1991;47:1595-1596.

Zhao XP, Prentice RL, Self SG. Multivariate mean parameter estimation by using a partly exponential model. Journal of the Royal Statistical Society (Series B) 1992;54(3):805-811.

Zucker D, Wittes J. Testing the effect of treatment in experiments with correlated binary outcomes. Biometrics 1992;48:695-710.