UNIVERSITY OF CAPE TOWN

DEPARTMENT OF STATISTICAL SCIENCES

TECHNIQUES FOR HANDLING CLUSTERED BINARY DATA

By Monique Hanslo

A mini thesis prepared under the supervision of Associate Professor June Juritz in partial fulfilment of the requirements for the degree of Master of Science in Mathematical Statistics

2002

"Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception and should be discouraged."

J. Cornfield (1978)

To René

You believed that I could do it.

ACKNOWLEDGEMENTS

I would like to thank a number of people who have in some way contributed to the writing of this thesis. Thank you first and foremost to my supervisor, Assoc Prof June Juritz. I have learned so much from you over the past years. Thank you for your guidance, inspiration, patience and for sharing your knowledge with me. It is greatly appreciated.

Thank you to the HODs and staff of the Statistical Sciences department who have assisted in any way. Thank you for your advice (statistical and otherwise), patience and understanding. I am grateful for all that you've done. Thank you also to Prof Les Underhill for funding toward my Masters degree.

Thank you to Dr Mike Sladden from the University of Tasmania, Australia, and Carl Lombard from the Medical Research Council for providing data for analysis.

To my parents and brothers, you've been great through it all.

And last but not at all least - Mario, Joanne, Edwina and René – you've all been amazing lifelines.

Thank you.

CONTENTS	Page

LIST OF TABLES	Page

SYNOPSIS

Over the past few decades there has been increasing interest in clustered studies and hence much research has gone into the analysis of data arising from these studies. It is erroneous to treat clustered data, where observations within a cluster are correlated with each other, as one would treat independent data. It has been found that point estimates are not as greatly affected by clustering as are the standard deviations of the estimates. But as a consequence, confidence intervals and hypothesis testing are severely affected. Therefore one has to approach the analysis of clustered data with caution. Methods that specifically deal with correlated data have been developed.

Analysis may be further complicated when the outcome variable of interest is binary rather than continuous. Methods for estimation of proportions, their variances, calculation of confidence intervals and a variety of techniques for testing the homogeneity of proportions have been developed over the years (Donner and Klar, 1993; Donner, 1989, and Rao and Scott, 1992). The methods developed within the context of experimental design generally involve incorporating the effect of clustering in the analysis. This cluster effect is quantified by the intracluster correlation and needs to be taken into account when estimating proportions, comparing proportions and in sample size calculations.

In the context of observational studies, the effect of clustering is expressed by the design effect which is the inflation in the variance of an estimate that is due to selecting a cluster sample rather than an independent sample. Another important aspect of the analysis of complex sample data that is often neglected is sampling weights. One needs to recognise that each individual may not have the same probability of being selected. These weights adjust for this fact (Little *et al*, 1997).

Methods for modelling correlated binary data have also been discussed quite extensively. Among the many models which have been proposed for analyzing binary clustered data are two approaches which have been studied and compared: the

population-averaged and cluster-specific approach. The population-averaged model focuses on estimating the effect of a set of covariates on the marginal expectation of the response. One example of the population-averaged approach for parameter estimation is known as generalized estimating equations, proposed by Liang and Zeger (1986). It involves assuming that elements within a cluster are independent and then imposing a correlation structure on the set of responses. This is a useful application in longitudinal studies where a subject is regarded as a cluster. Then the parameters describe how the population-averaged response rather than a specific subject's response depends on the covariates of interest. On the other hand, cluster-specific models introduce cluster to cluster variability in the model by including random effects terms, which are specific to the cluster, as linear predictors in the regression model (Neuhaus *et al*, 1991). Unlike the special case of correlated Guassian responses, the parameters for the cluster specific model obtained for binary data describe different effects on the responses compared to that obtained from the population-averaged model. For longitudinal data, the parameters of a cluster-specific model describe how a specific individuals probability of a response depends on the covariates. The decision to use either of these modelling methods depends on the questions of interest. Cluster-specific models are useful for studying the effects of cluster-varying covariates and when an individual's response rather than an average population's response is the focus. The population-averaged model is useful when interest lies in how the average response across clusters changes with covariates. A criticism of this approach is that there may be no individual with the characteristics of the population-averaged model.

CHAPTER 1

APPROACHES TO MODELLING CLUSTERED BINARY DATA

1.1 Analyzing correlated binary data

(1) Naïve approaches

(2) Response feature models

(3) Conditional specification of models

(4) Transitional models

(5) Marginal models

(6) Cluster-specific models

(7) Likelihood ratio test approach

1.2 Structure of thesis

CHAPTER 1
APPROACHES TO MODELLING CLUSTERED BINARY DATA

1.1 Analyzing correlated binary data

Over the past few years a number of methods have been developed to deal with correlated binary data. A binary variable takes on one of two possible values as a response. These values may fall naturally into two categories, eg. presence/absence, yes/no, or may be formed by dichotomising a continuous variable eg. age above or below 40 years. It is often the case that these binary measurements arise in clusters. Clustering in experimental data frequently arise in two ways. It may be due to

(i) repeated measurements taken on a specific individual
eg. measurements taken on left and right eyes, longitudinal studies which involve obtaining repeated measurements over time

(ii) or it may be a result of sampling or applying interventions to groups of individuals such as families, schools or even entire communities rather than separate individuals.

In both situations clustering of observations or individuals produce experimental units that exhibit correlation within subjects or clusters. This correlation needs to be accounted for in the analysis as well as sample size calculations.

The analysis of correlated binary data depends on both the study design and goals of the study. Studies may be classified as observational or experimental. The main difference between these two study designs is that the investigator actively applies some intervention or treatment in experimental studies but not in observational studies. Methods of analysis will be examined for both study designs. Techniques for obtaining simple descriptive statistics as well as more complex regression techniques will be discussed.

There are a number of regression type approaches that may be used to model the relationship between a number of covariates and a correlated binary variable. Many of these approaches rely on generalized linear models, specifically logistic regression or probit models. Here the correlation within a cluster is used to characterize the joint distribution of responses within the cluster. The modelling techniques most often used fall into at least one of the following categories.

(1) Naïve approaches

Naïve approaches to the analysis of clustered binary data simply ignore the correlation between subunits. The advantage of this approach is that standard computer packages may be used to fit models. Consider ophthalmologic data as an example. Measurements are often taken on both eyes of an individual. The measurements of the two eyes for a specific individual can be expected to be more similar than two measurements taken on eyes of different individuals. The naïve approach involves either analyzing one eye chosen at random for each individual, analyzing only left or right eyes for all individuals, or analyzing all data by ignoring the correlation between left and right eyes. The advantage of this naïve approach is that standard statistical methods may be used for analysis, resulting in regression estimators that are still consistent. There are disadvantages to this approach. Firstly, ignoring the existing correlation between individuals in a cluster results in incorrect standard errors and hence incorrect confidence intervals and p-values. A second disadvantage is that not all of the available information is utilized if only one eye per individual is chosen. If all data is used parameter estimates are consistent but standard errors will be incorrect.

(2) Response feature models

This strategy involves modelling by making use of independent univariate methods. The multivariate response information is summarized using a single measure for each cluster and then standard modelling techniques, such as

logistic regression, is applied. In some situations this may be achieved without major loss. For example, consider the ophthalmologic example given above. The investigator might decide to only use data from the "worst" or the "best" eye. If correlations between eyes are high then this method is appropriate since little will be gained by using information from both eyes. The decision to use particular data from a cluster needs to be made before the data is collected and should not be based on the data obtained from the study. A problem with the response feature approach is the question on how to deal with within-cluster covariates. In selecting data from a single individual/observation within a cluster one might actually be selecting the covariate on the basis of the response (Glynn and Rosner, 1992). Secondly, by collapsing responses in a cluster into one measure we lose some information.

(3) Conditional specification of models

By making use of conditionally specified models the joint distribution of the data is derived using conditional distributions at the cluster level. In the binary case, the probability of a positive response for an individual in a cluster is modelled conditionally on the responses of other individuals in the same cluster, i.e for individual j in cluster i, $\Pr(y_{ij}|y_{im}, m \neq j)$. As a consequence interpretation must be made in terms of conditional probabilities rather than joint or marginal probabilities. A reason for conditioning is that in the presence of correlation conditional distributions are conceptually simpler than multi-dimensional joint distributions. Examples of conditional models include the general log-linear model (Zhao and Prentice, 1990 and Fitzmaurice *et al*, 1993), the auto-logistic model (Besag, 1974) and response conditional models (Rosner ,1984; Connolly and Liang, 1988 and Qu *et al*, 1992).

(4) Transitional models

These models are used when responses within a cluster are ordered eg. longitudinal data collected over time. The transitional model is a special case of conditional specified modelling where instead of modelling full conditional

probabilities (conditional on all events in a cluster i.e. in the past and in the future), interest lies in conditioning on past responses only. For more information on this model see Zeger and Liang (1992) and Cox (1958).

(5) Marginal models

Correlation under this model is simply viewed as being a nuisance parameter. Then marginal probabilities are modelled in order to examine the relationship between the response and covariates. Examples of marginal models are the quadratic exponential model and the generalized estimating equation approach. We will discuss the latter technique that accounts for the effect of clustering by a simple extension to quasi-likelihood. The generalized estimating equations approach is appropriate if one needs to relax distributional assumptions. Parameter estimates are obtained by solving the multivariate analogue of the quasi-score function described by Wedderburn in 1974. This function incorporates the correlation by weighting the score function. This is achieved by including a correlation matrix in the set of equations to be solved. The properties of consistency and asymptotic normality of parameter estimates is maintained. Interpretation of resulting parameter estimates should be made at the population level i.e. parameter estimates have population-averaged interpretations because the effect of the explanatory variable is obtained by averaging across clusters.

(6) Cluster-specific models

This model includes parameters that vary by cluster. It produces results that measure the effect of the covariates on heterogeneous individuals, and hence includes cluster-specific parameters which describe the correlation within a cluster. Since the number of parameters to be estimated increases as the number of clusters increase, a popular approach to the analysis involves viewing the cluster-specific parameters as following some distribution. An example of a cluster-specific model is the random-effects or mixed-effects model. Here it is assumed that the cluster-specific parameters are an

independent sample from some distribution. The model can then be written in terms of fixed effects and random (cluster) effects.

(7) Likelihood ratio test approach

In this approach individual responses are assumed to follow some parametric distribution and is therefore a fully parametric approach. The choice of likelihood ratio test will depend on the distribution chosen. The distribution most often chosen is the beta-binomial distribution, an extension to the binomial distribution. Under this model it is assumed that the jth member in the ith cluster ($j=1,\ldots,n_i$; $i=1,\ldots,k$) has probability π_i of success and this probability follows a beta distribution with parameters a and b (both > 0). The responses for different cluster members, conditional on π_i, are independent. Then the resulting marginal distribution of y_i, the total number of subjects with the attribute in cluster i, follows a beta-binomial distribution. Griffiths (1973), Crowder (1978), Williams (1982) and Brooks (1984) are some authors who discuss this model in detail.

1.2 Structure of thesis

This thesis discusses clustered data in the context of both experimental and observational studies. Chapter 2 proceeds by discussing the cluster randomized experiment and its design. The basic analysis of binary cluster randomized trial data is discussed – deriving simple proportion estimates, standard deviations, confidence intervals and hypothesis testing. The methods discussed are illustrated with a hypertension example from Bass *et al* (1986). Chapter 3 goes on to examine the design and analysis of observational studies. This chapter focuses on survey sampling and more specifically, the one-stage cluster sample. Advice on study design and methods for analysis – proportion estimation and logistic regression – are highlighted. Methods are illustrated using data that deals with rectal bleeding in Tasmania, Australia. Chapters 4

and 5 discuss, in detail, modelling of proportions for clustered data. Chapter 4 looks at a population-averaged approach to modelling known as the generalized estimating equations approach. Estimation, interpretation, advantages and disadvantages are mentioned. An example is performed using the rectal bleeding data. Chapter 5 examines a cluster-specific model, the logistic-mixed effects model. The lymphocyte proliferation assay example in this chapter is taken from Betensky *et al* (2001). The results obtained using this model are compared to those using a population-averaged approach and standard logistic regression. Finally, a few additional comments are made in Chapter 6.

CHAPTER 2

DESIGN AND ANALYSIS OF CLUSTER RANDOMIZED TRIALS

2.5 Sample size calculations

2.5.1 Completely randomized design

(1) Estimating the number of individuals in the treatment group

(2) Estimating the number of clusters

(3) Estimating the number of clusters and the number of individuals per cluster

(a) Feng and Grizzle's formula

(b) Lee and Dubin's formula

2.5.2 Stratified cluster randomization design

CHAPTER 2
DESIGN AND ANALYSIS OF CLUSTER RANDOMIZED TRIALS

2.1 Introduction

Intervention studies are experiments in which an investigator randomly assigns an intervention to subjects who partake in a study. In order to determine the effect of the treatment the subjects are followed prospectively. Confounding factors that have been identified may be controlled in an experimental study. If random allocation of treatment is performed in a large sample then confounding due to unobserved variables may be eliminated.

A cluster randomized trial is one in which groups (clusters) of individuals rather than single individuals are randomly allocated to a specific treatment. So the unit of randomization is the cluster but the unit of intervention is the individual. As an example consider a trial on smoking prevention in adolescents where a number of orthodontic offices were randomly assigned to experimental conditions (exposure to anti-tobacco materials, prescriptions containing anti-tobacco messages, and tobacco-free environment within the office) or control conditions where no anti-tobacco instruction had been administered (Slymen & Hovell, 1997). So instead of single individuals receiving anti-tobacco instruction, offices of individuals were exposed to the intervention. The justification here for cluster randomization was to avoid treatment contamination, to reduce costs, and for administrative convenience. The resulting analysis should take into account that individuals at a specific orthodontic office cannot be considered to be independent of each other.

There are a number of possible reasons for between cluster variation that arises in these sort of trials. Individuals within a specific cluster may be affected by important factors in a similar manner. In addition, frequently individuals

within a cluster interact and their responses are more similar than responses from individuals in other clusters.

Cluster randomization may be employed for the following reasons. It is desirable or feasible from an administrative, economical and operational point of view, it avoids treatment contamination and could also serve to enhance subject compliance and co-operation of investigators. It may also be due to ethical reasons. So even though use of randomization on units larger than individual units tend to reduce the power associated with testing for an intervention effect (Donner and Klar, 1996), cluster randomization is often employed because it may not always be possible to randomize individuals.

Despite the importance of taking cluster randomization into account during both the design and analysis stage of an experiment, many investigators have failed to do this. Donner *et al* (1990) and Simpson *et al* (1995) reviewed a number of cluster randomized trials conducted during the periods 1979-1989 and 1990-1993 respectively. They found that many of these trials offered no justification for employing cluster randomization rather than individual randomization, only a quarter of these trials took the cluster design into account in the sample size and power calculations, and only half of the studies incorporated the between cluster variation in the analysis. Over the past few years investigators have become more aware of the need to use appropriate statistical methods when faced with a trial of this sort (Bland and Kerry, 1997). However, many are currently using standard statistical methods and as a consequence obtaining erroneous results.

The analysis of data arising from cluster randomized trials proceeds by recognising that even though the cluster (for example, the orthodontic office in the given example) is the unit of randomization, we often wish to draw inferences at an individual level. Because individuals within a cluster cannot

be regarded as independent, standard statistical methods for sample size estimation and analyses are not appropriate.

There are two sources of variation present in a cluster randomized trial: variation between subjects in a cluster and variation between clusters. Both of these have to be taken into account in the analysis. The presence of within-cluster homogeneity, which is quantified by the intracluster correlation coefficient, ρ, leads to an increase in the size of the standard errors and hence

(i) a reduction in the effective sample size, the extent of which depends on the intracluster correlation and average cluster size

(ii) underpowered tests and spurious statistical significance (p-values biased downwards) if standard methods for sample size estimation and statistical analyses are used. This worsens as the average cluster size and intracluster correlation increases (Donner, 1998).

The larger and fewer the clusters, the greater the effects.

2.2 Design of cluster randomized trials

There are a number of issues that arise when planning a cluster trial. Firstly, the investigator may be faced with making the decision to either randomize or not. There are a number of advantages when clusters are randomly allocated to treatment. Randomization eliminates selection bias, balances cluster-level baseline characteristics and justifies the use of statistical distributional theory. However, non-randomized trials are appropriate when random allocation is viewed as being unethical, to avoid possible contamination due to geographical situation of individuals, and when clusters are too large, resulting in too few clusters in the study.

The number and choice of intervention groups to be studied should be carefully considered. The choice of control group is another important aspect of the cluster randomized trial. It is often the case that control groups receive the

program they would have normally received in the absence of the intervention. There might be a tendency for added attention to be given to those in the intervention group compared to those in the control group. This could cause a change in the responses of those in the intervention group compared to what would have been observed if they had not received added attention. Therefore significant intervention effects might be partially due to the added attention given to those in the intervention group. Ways of reducing this effect include using more than one type of control group and viewing the control as being equally important as the intervention.

In cluster randomization trials eligibility criteria need to be set at both the cluster level and individual level. When setting these criteria the following issues should be kept in mind: the desired power, avoidance of treatment contamination and the expected intervention effect should guide the choice of the randomization unit (Murray, 1998). The degree of control over cluster size is important as the more control one has, the greater the administrative flexibility and the ability to achieve the desired power.

Finally, when designing the trial a choice has to be made as to whether the trial will be cross-sectional or a cohort study. Cross-sectional studies involve taking a sample of subjects before and after intervention and may result in a more representative sample of the population than the cohort study. It is most useful when inferences need to be made at the cluster level. With cohort designs the same subjects are followed over time. This design is most useful when investigating the effect of the intervention on changing the health or the behaviour of individuals.

The most frequently adopted cluster randomization designs are the completely randomized design, matched pair design and the stratified cluster randomization study. The completely randomized design and stratified cluster design will be discussed in the context of sample size calculation.

2.3 Analysis of cluster randomized trials

Very often in statistical and epidemiologic investigations we have to deal with clustered binary data. Various approaches to handle the analysis of clustered binary or proportional data, in the completely randomized design context, will be reviewed. For issues concerning the design and analysis of paired cluster randomized trials refer to Donner (1987), Donner and Hauck (1989), Gail *et al* (1992) and Martin *et al* (1993). Donner and Donald (1987) and Donner (1992) discuss stratified cluster randomization designs.

We will focus on the case of two treatment interventions with respect to a binary outcome variable in a cluster randomized experiment. So we focus on the special case where clusters are randomized to one of two treatments. We will assume that responses from different clusters are independent.

2.3.1 Estimating proportions

The aim of many investigations is to make inferences about proportions of individuals in different treatment groups who exhibit some characteristic of interest. We first consider estimation of proportions in a cluster randomized trial.

Assume there are k clusters of which k_l are randomized to treatment l, l=1, 2.
So k_l is the number of clusters in the lth treatment group.
Let n_{li} denote the size of the ith cluster in the lth treatment group, i=1,...,k_l; l=1,2.

Let y_{lij} = 1 if the jth response in the ith cluster of the lth treatment group is positive

= 0 if the response is negative.

Then $y_{li} = \sum_{j=1}^{n_{li}} y_{lij}$ is the number of positive responses (eg. diseased, affected) in the *i*th cluster of group *l* i.e. the number of individuals who exhibit the event of interest in cluster *i* of the *l*th treatment group.

Also assume that $y_{lij} \sim$ Bernoulli. Then $E(y_{lij}) = \pi_l$ and $\text{var}(y_{lij}) = \pi_l(1-\pi_l)$. And therefore the variance for the number of positive responses in the *i*th cluster of group *l* is

$$\text{var}(y_{li}) = \text{var}\left(\sum_{j=1}^{n_{li}} y_{lij}\right) + \sum_{i=1}^{k_l}\sum_{j\neq j'}^{n_{li}} \text{cov}(y_{lij}, y_{lij'}).$$

Because observations are not independent the covariance term is non-zero. In order to determine the covariance, we assume that the correlation between two observations in the *i*th cluster is given by some value, ρ, the intracluster correlation. Then by definition

$$\rho = \frac{\text{cov}(y_{lij}, y_{lij'})}{\sqrt{\text{var}(y_{lij})\text{var}(y_{lij'})}} = \frac{\text{cov}(y_{lij}, y_{lij'})}{\text{var}(y_{lij})}$$

since the variances are equal.

And therefore $\text{cov}(y_{lij}, y_{lij'}) = \rho\,\text{var}(y_{lij})$.

As a consequence the variance of y_{li} is

$$\begin{aligned}\text{var}(y_{li}) &= \text{var}\left(\sum_{j=1}^{n_{li}} y_{lij}\right) + \sum_{i=1}^{k_l}\sum_{j\neq j'}^{n_{li}} \rho\,\text{var}(y_{lij}) \\ &= n_{li}\pi_l(1-\pi_l) + n_{li}(n_{li}-1)\rho\pi_l(1-\pi_l) \\ &= n_{li}\pi_l(1-\pi_l)[1+(n_{li}-1)\rho]. \qquad (2.1)\end{aligned}$$

So we see that the effect of dependence within clusters is an increase in the variance of y_{li} by a factor of $[1+(n_{li}-1)\rho]$. We will assume that ρ is constant across all groups of interest.

In addition denote

$y_l = \sum_{i=1}^{k_l} y_{li}$, the total number of individuals affected in the lth treatment group,

$n_l = \sum_{i=1}^{k_l} n_{li}$, the total number of individuals in the lth treatment group,

and $n = \sum_{l=1}^{2} n_l$, the total number of individuals in the trial.

Then an estimate of the true proportion of individuals affected in the lth treatment group, $\hat{\pi}_l$, is the overall sample proportion of the lth treatment group as computed over all clusters in that group, so

$$\hat{\pi}_l = \frac{y_l}{n_l} = \frac{\sum_{i=1}^{k_l} y_{li}}{\sum_{i=1}^{k_l} n_{li}} \qquad l = 1, 2.$$

The variance follows from (2.1) and is given by

$$\text{var}(\hat{\pi}_l) = \frac{\hat{D}_l \hat{\pi}_l (1-\hat{\pi}_l)}{n_l}$$

where $\hat{D}_l$ is the variance correction factor,

$$\hat{D}_l = \frac{\sum_{i=1}^{k_l} n_{li}[1+(n_{li}-1)\hat{\rho}]}{\sum_{i=1}^{k_l} n_{li}} = 1+(c_l-1)\hat{\rho}$$

with

$$c_l = \sum_{i=1}^{k_l} \frac{n_{li}^2}{n_l}.$$

The value $\hat{D}_l$ is also known as the variance inflation factor and can be thought of as the estimated clustering effect in the lth treatment group. It is the inflation in the variance, associated with the success rate in each group, that is due to clustering.

An alternative estimator of the proportion of positive responses in the treatment group is, $\hat{\bar{\pi}}_l$, the average of the response rates in each cluster

$$\hat{\bar{\pi}}_l = \frac{1}{k_l} \sum_{i=1}^{k_l} \hat{\pi}_{li} \qquad (2.2)$$

where

$\hat{\pi}_{li} = \frac{y_{li}}{n_{li}}$ is the cluster-specific success rate.

This estimator becomes problematic when dealing with small samples because $\bar{\pi}_l$ does not weight $\hat{\pi}_{li}$ by cluster size.

An estimate of the overall population proportion, assuming proportions are equal in both groups, is given by

$$\hat{\pi} = \frac{\sum_{l=1}^{2} y_l}{\sum_{l=1}^{2} n_l}.$$

Then even though $\hat{\pi}$ does not have the usual binomial distribution due to the presence of intracluster correlation, we have $E(\hat{\pi}) = \pi$. The variance of $\hat{\pi}$ is given by

$$\text{var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}\left[1 + (\overline{n} - 1)\rho\right]$$

where $\overline{n} = \frac{\sum_{l=1}^{2} n_l}{k}$, the average cluster size.

2.3.2 A measure of the effect of clustering - Estimating the intracluster correlation coefficient

Of particular interest may be the effect of clustering. A measure of this effect is given by ρ, the intracluster correlation coefficient. It is postulated that the correlation between the responses of any two individuals within a cluster has the same value, ρ. It is also assumed that individuals between clusters are independent. Donner and Klar (1994) provide an analysis of variance (ANOVA) estimator of the intracluster correlation coefficient. This is achieved by deriving the between-cluster and within-cluster mean squares from a one-way analysis of variance of the binary response variable.

Denote $MSC = \sum_{l=1}^{2} \sum_{i=1}^{k_l} \frac{n_{li}(\hat{\pi}_{li} - \hat{\pi}_l)^2}{k-2}$,

$$MSE = \sum_{l=1}^{2} \sum_{i=1}^{k_l} \frac{n_{li}\hat{\pi}_{li}(1-\hat{\pi}_{li})}{n-k}$$

and $n_0 = \left[n - \sum_{l=1}^{2} c_l \right] \Big/ (k-2)$.

Then a consistent estimator of ρ is an extension of the intracluster correlation described by Fleiss (1981):

$$\hat{\rho} = \frac{MSC - MSE}{MSC + (n_0 - 1)MSE}.$$

The values *MSC* and *MSE* are mean square errors that measure variation in the outcome between and within clusters from a standard one-way ANOVA, respectively. This estimator was originally proposed for continuous variables but various authors subsequently suggested using it for binary outcomes (Fleiss, 1981). If responses of individuals between clusters are similar to responses of individuals within clusters then $MSC \approx MSE$ and hence $\hat{\rho} \approx 0$, there is no clustering effect. On the other hand, if all responses within a cluster are the same (i.e. equal to zero or one) then MSE=0 and $\hat{\rho} = 1$ indicating a strong dependency within clusters. Even when the intracluster correlation is small, combined with large cluster sizes the correlation affects power and invalidates standard statistical procedures (Donner and Klar, 1996).

Ridout *et al* (1999) discuss 20 estimates of the intracluster correlation that have been proposed and perform a detailed simulation study on these estimates. They point out the result from Prentice (1986) that states a lower bound constraint for ρ:

$$\rho \geq \frac{-1}{(n_{\max} - 1)} + \frac{\tau(1-\tau)}{n_{\max}(1-n_{\max})\pi(1-\pi)}$$

where π is the probability of success, $n_{\max}$ is the largest cluster size and $\tau = n_{\max}\pi - \text{int}(n_{\max}\pi)$, int(·) being the integer part. From their simulation study they find that an intraclass correlation estimate suggested by Williams (1982) performed best. This correlation estimate is calculated using an iterative scheme which involves weighting probabilities of success within each cluster

by weights proportional to $\frac{n_i}{1+(n_i-1)\hat{\rho}}$ where n_i is the size of the ith cluster and $\hat{\rho}$ is a current estimate of ρ. See Ridout *et al* (1999) for a more detailed discussion on estimating the intraclass correlation for binary data.

The estimator of the intraclass correlation follows a normal distribution when the number of clusters is large (Shoukri and Martin, 1992). But the exact sample distribution of $\hat{\rho}$ is difficult to obtain. Hill and Smith (1977) proposed that as $k \to \infty$, $\sqrt{k}(\hat{\rho}-\rho)$ is asymptotically unbiased and normally distributed with variance given by

$$\text{var}(\hat{\rho}) \approx \frac{2(1-\rho)^2[1+(\bar{n}-1)\rho]^2}{\bar{n}(\bar{n}-1)k}.$$

Mak (1988) showed that this was inadequate in that it underestimates the variance, and provided an approximation for the asymptotic variance of $\hat{\rho}$.

This is

$$\begin{aligned}\text{var}(\hat{\rho}) &\approx k^{-1}\left[H_1^2V_{11}+2H_1H_2V_{12}+H_2^2V_{22}\right] \\ &= k^{-1}\phi(\rho,\pi)\end{aligned}$$

where

$$\phi(\rho,\pi)=\left[H_1^2V_{11}+2H_1H_2V_{12}+H_2^2V_{22}\right]$$

with

$$H_1=\left[(\bar{n}-1)\pi(1-\pi)\right]^{-1},$$
$$H_2=-\left[\bar{n}(\bar{n}-1)\pi(1-\pi)\right]^{-1}\left[1+(\bar{n}-1)(2\pi+\rho(1-2\pi))\right],$$

$$V_{11} = \frac{1}{k}\sum_{l=1}^{2}\sum_{i=1}^{k_l}\left[E\left(\pi_{li}^4 n_{li}^2\right) - \left\{E\left(\pi_{li}^2 n_{li}\right)\right\}^2\right],$$

$$V_{12} = \frac{1}{k}\sum_{l=1}^{2}\sum_{i=1}^{k_l}\left[E\left(\pi_{li}^3 n_{li}^2\right) - E\left(\pi_{li}^2 n_{li}\right)E\left(\pi_{li} n_{li}\right)\right]$$

and

$$V_{22} = \frac{1}{k}\sum_{l=1}^{2}\sum_{i=1}^{k_l}\left[E\left(\pi_{li}^2 n_{li}^2\right) - \left\{E\left(\pi_{li} n_{li}\right)\right\}^2\right].$$

We replace ρ, π, and π_{li} (l=1,2; i=1,...,k_l,) by consistent estimates to obtain a consistent estimator of the asymptotic variance of $\hat{\rho}$.

2.3.3 Confidence intervals for proportions

We now turn to the construction of confidence intervals for proportions in cluster randomized trials. Donner and Klar (1993) recognise that clustering has to be taken into account in the construction of confidence intervals. Their method of deriving confidence limits is discussed below.

The confidence interval for a single proportion for the lth treatment group, π_l, under the cluster design is

$$\hat{\pi}_l \pm z_{\alpha/2} S\hat{E}\left(\hat{\pi}_l\right)$$

$$= \hat{\pi}_l \pm z_{\alpha/2}\left[\frac{\hat{D}_l \hat{\pi}_l\left(1-\hat{\pi}_l\right)}{n_l}\right]^{1/2}$$

where $z_{\alpha/2}$ is the $(1-\alpha)100\%$ two-sided critical value of the standard normal distribution and $\hat{D}_l$ is as defined in section 2.3.1.

Now consider the treatment effect estimated by the difference between proportions in two treatment groups, $(\hat{\pi}_1 - \hat{\pi}_2)$. The standard error of this quantity is estimated by

$$S\hat{E}(\hat{\pi}_1 - \hat{\pi}_2) = \left[\sum_{l=1}^{2} \frac{\hat{D}_l \hat{\pi}_l (1 - \hat{\pi}_l)}{n_l}\right]^{1/2}.$$

An approximate two-sided $(1-\alpha)100\%$ confidence interval for the true difference $(\pi_1 - \pi_2)$ is then given by

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2} S\hat{E}(\hat{\pi}_1 - \hat{\pi}_2)$$

$$= (\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\left[\sum_{l=1}^{2} \frac{\hat{D}_l \hat{\pi}_l (1 - \hat{\pi}_l)}{n_l}\right]^{1/2}.$$

The confidence interval for the overall proportion π under the cluster design is

$$\hat{\pi} \pm z_{\alpha/2} S\hat{E}(\hat{\pi})$$

$$= \hat{\pi} \pm z_{\alpha/2}\left[\frac{\hat{\pi}(1-\hat{\pi})}{n}[1 + (\bar{n} - 1)\hat{\rho}]\right]^{1/2}.$$

If analyses were to ignore the clustering of individuals resulting confidence intervals would be too narrow.

2.3.4 Testing the homogeneity of proportions in cluster randomized trials

Under the assumption of statistical independence of individuals we can test whether two group proportions are homogenous by testing the hypothesis $H_0: \pi_1 = \pi_2$ using the standard Pearson chi-square test. The test statistic of interest is

$$\chi_P^2 = \sum_{i=1}^{2} \frac{(y_i - n_i\hat{\pi})^2}{n_i\hat{\pi}(1-\hat{\pi})}$$

which has an asymptotic chi-square distribution with one degree of freedom.

Assume now that clusters of individuals are randomly assigned to treatments so that the assumption of independence is violated. In the presence of clustering χ_P^2 no longer follows a chi-square distribution. The magnitude of the bias associated with χ_P^2 increases as both the intracluster correlation and average cluster size increases (Donner and Klar, 1994).

A number of approaches have been developed to deal with the problem of comparing proportions in a cluster randomized study. We examine some of these.

(1) Two-sample t-test

The two-sample t-test involves applying the standard two-sample t-test to cluster-specific proportions. Donner and Klar (1994) point out reasons for dissatisfaction with this approach.

(i) The assumptions that the cluster-specific proportions are normally distributed with equal variances is not strictly satisfied, especially so if there is large variation in cluster size, and therefore the test may not be valid. However, research simulations have shown that the t-test remains robust despite violations of these assumptions.

(ii) In addition, the t-test ignores variation in cluster size (this problem can be solved by developing a weighted t-test discussed in the next section).

(iii) Finally, it does not allow one to easily make inferences with respect to odds ratios (which is often of interest).

(2) Weighted two-sample t-test

As pointed out in the previous section, a weighted t-test takes cluster size into account and is therefore preferable if the cluster sizes are unbalanced. It also incorporates the degree of intracluster correlation. This is achieved by incorporating weights in the test statistic. The appropriate weight for the ith cluster in the lth group is the cluster size divided by the variance inflation factor (Campbell, 1999),

$$w_{li} = \frac{n_{li}}{1+(n_{li}-1)\hat{\rho}} \qquad l=1,2 \; ; \; i=1,2,\ldots,k_l$$

Define an estimate of the proportion in the lth treatment group. This is simply the weighted version of the proportion estimate defined before by (2.2).

$$\hat{\bar{\pi}}_{lw} = \frac{\sum_{i=1}^{k_l} w_{li}\hat{\pi}_{li}}{\sum_{i=1}^{k_l} w_{li}}$$

where $\hat{\pi}_{li}$ are the cluster-specific proportions in group l.

Also define

$$w_l = \sum_{i=1}^{k_l} w_{li} \qquad l=1,2 \,,$$

the sum of the weights in the lth treatment group and

$$w_l^{\,2} = \sum_{i=1}^{k_l} w_{li}^2 \qquad l=1,2,$$

the sum of the squares of the weights in the *l*th treatment group.

In order to test for homogeneity of proportions in the two treatment groups we test H_0: $\bar{\pi}_{1w} = \bar{\pi}_{2w}$.

The estimated variance of $\left(\hat{\bar{\pi}}_{1w} - \hat{\bar{\pi}}_{2w}\right)$ is given by

$$s_w^2\left(\frac{w_1^2}{(w_1)^2} + \frac{w_2^2}{(w_2)^2}\right)$$

where

$$s_w^2 = \frac{\sum_{l=1}^{2}\sum_{i=1}^{k_l} w_{li}\left(\hat{\pi}_{li} - \hat{\bar{\pi}}_{lw}\right)^2}{\sum_{l=1}^{2}\sum_{i=1}^{k_l} w_{li}}.$$

Then the appropriate test statistic with $k_1 + k_2 - 2$ degrees of freedom is

$$t_w = \frac{\hat{\bar{\pi}}_{1w} - \hat{\bar{\pi}}_{2w}}{s_w\sqrt{\frac{w_1^2}{(w_1)^2} + \frac{w_2^2}{(w_2)^2}}}.$$

The resulting statistic follows a t-distribution. This weighted t-test is more powerful than the standard t-test when there is considerable variation in cluster size (Campbell, 1999).

(3) Non-parametric approaches

Another option is to analyze the data making no assumptions about the distribution of the cluster-specific proportions by using a non-parametric approach. We briefly look at two non-parametric tests.

(a) Wilcoxon rank-sum test

This test relies on the ranks of observed proportions. This procedure involves pooling the two groups' observed proportions and then ranking them. If the two groups are homogenous then the sum of the ranks should be the same for both groups. Hence the test statistic is the sum of the ranks of one of the groups. Valid *p*-values may be obtained from standard tables but the test lacks power. The downfall of this method is a loss in precision since both variation in cluster size and actual magnitude of proportions are ignored. Another problem is that it is impossible to achieve statistical significance at the 5% level, even if there is a treatment effect, if there are fewer than four clusters per group.

(b) Fisher's two-sample permutation test

This exact test of significance incorporates the magnitude of the cluster-specific proportions and is calculated using randomization theory. The resulting test statistic for this approach is calculated by looking at the number of ways in which the cluster-specific proportions could be arranged between treatment groups. While maintaining the same number of clusters per group, the test statistic which is the difference in average proportions, is calculated for each permutation of the data. Then the two-tailed significance level of the test is the proportion of test statistics of the permuted data that are at least as large as the absolute values of the test statistic found using the observed data. A limitation of this test is that it is not always possible to obtain statistically significant results when there are few clusters per treatment group. As with the

Wilcoxon test, at least four clusters in each group are needed to obtain significance (Donner and Klar, 1994).

(4) Adjusted chi-square procedures

A number of authors have developed tests that take the clustering effect into account by making simple adjustments to the usual Pearson chi-square statistic. Each yields a chi-square statistic with one degree of freedom. Shoukri and Pause (1999) suggest two adjustments, one proposed by Donner (1989) which is based on a direct adjustment of the Pearson chi-square statistic, and the other by Rao and Scott (1992) which is based on ratio estimate theory. We discuss these adjustments.

(a) Donner's adjustment

In Donner (1989) an adjustment of the chi-square statistic that involves computing clustering correction factors for each intervention group is proposed. Suppose that there exists a dependence in clusters measured by the intraclass correlation, ρ, which can be regarded as homogenous across treatment groups. Donner (1989) presents a group-specific adjustment approach by adjusting the standard Pearson chi-square test statistic. This is achieved by weighting the test statistic by an estimate of the clustering correction factor presented before, $\hat{D}_l$, for each group l.

Hence Donner's chi-squared statistic is

$$\chi^2_D = \sum_{l=1}^{2} \frac{(y_l - n_l\hat{\pi})^2}{\hat{D}_l n_l \hat{\pi}(1-\hat{\pi})}$$

with

$\hat{D}_l = 1 + (c_l - 1)\hat{\rho}$.

An underlying assumption is that D_1 and D_2 are homogenous. This assumption holds if clusters had been randomly assigned to treatment groups. This assumption may not hold for observational studies, especially if the average cluster sizes differ in the two treatment groups. Note that if $D_l = 1$, that is $\rho = 0$, then Donner's chi-square statistic, χ^2_D, simply reduces to the standard Pearson chi-square statistic , χ^2_P.

Donner's adjustment presents some problems. Donner proposed that χ^2_D follows a chi-square distribution with one degree of freedom. However, Rao and Scott (1992) point out that this only holds for the special cases where
(i) the population inflation factors are equal, that is, $D_1 = D_2$ for l=1,2 or
(ii) $n_{li} = \bar{n}$ for all (l,i), l=1,2; i=1,…,k_l.

Donner's adjustment relies on the assumption that $\hat{D}_1$ and $\hat{D}_2$ estimate the same population design effect i.e. are not significantly different. And close examination of D_l (l=1,2) reveals that even for small $\hat{\rho}$, a large average cluster size can lead to a large inflation in variance due to clustering. We therefore move on to discuss an alternative adjustment proposed by Rao and Scott (1992).

(b) Rao and Scott's adjustment

Shoukri and Pause (1999) also discuss this robust adjustment developed by Rao and Scott (1992). Here they regard $\hat{\pi}_l$, l=1,2, as ratios rather than proportions. This method assumes no model for the intraclass correlation and in order to determine the chi-squared test statistic it is not necessary to obtain an estimate of ρ as was the case with Donner's adjustment.

Recall that $\hat{\pi}_l$ is the estimated overall proportion of successes in group l. The estimated pure binomial variance is given by

$$\hat{\text{var}}_B(\hat{\pi}_l) = \frac{\hat{\pi}_l(1-\hat{\pi}_l)}{n_l}.$$

Rao and Scott (1992) proceed by writing $\hat{\pi}_l$ as a ratio of two sample means

$$\hat{\pi}_l = \frac{\bar{y}_l}{\bar{n}_l}$$

where $\bar{y}_l = \frac{1}{k_l}\sum_{i=1}^{k_l} y_{li}$ and $\bar{n}_l = \frac{1}{k_l}\sum_{i=1}^{k_l} n_{li}$.

The value $\hat{\pi}_l$ is the ratio of two means, and for large n_l and large k_l, a consistent estimator of the variance of $\hat{\pi}_l$ is the estimated ratio variance,

$$v_l = \hat{\text{var}}_R(\hat{\pi}_l) = \frac{k_l}{(k_l-1)}\frac{1}{n_l^2}\sum_{i=1}^{k_l}(y_{li} - n_{li}\hat{\pi}_l)^2.$$

It is then possible to examine the effect of clustering on the variance by computing estimated design effects in each group. This is obtained by taking the ratio of v_l, the estimate of the variance of $\hat{\pi}_l$ under the cluster design, to the estimated variance of $\hat{\pi}_l$ under pure binomial conditions. This is given by

$$d_l = \frac{\hat{\text{var}}_R(\hat{\pi}_l)}{\hat{\text{var}}_B(\hat{\pi}_l)} = \frac{n_l v_l}{\hat{\pi}_l(1-\hat{\pi}_l)}.$$

The value d_l is the inflation in the variance of the estimated proportion of the lth treatment group due to clustering. It is similar to the variance correction

factor D_l but is derived in the context of ratio estimation. In survey sampling d_l is commonly known as the design effect and will be discussed in Chapter 3. It can be used to determine an effective sample size, $\tilde{n}_l$, by setting $\tilde{n}_l = \frac{n_l}{d_l}$ as discussed by Kish (1965).

The design effect, d_l, is used to compute the effective number of successes

$$\tilde{y}_l = \frac{y_l}{d_l}$$

and hence the effective proportion, assuming it is the same in both groups, is

$$\tilde{\pi} = \frac{\sum_{l=1}^{2} \tilde{y}_l}{\sum_{l=1}^{2} \tilde{n}_l}.$$

Rao and Scott (1992) found that one could obtain asymptotically correct results for cluster design studies by replacing aggregate data (y_l, n_l) by $(\tilde{y}_l, \tilde{n}_l)$ and then treating the transformed variable $\tilde{y}_l$ as a binomial variable with parameters $(\tilde{n}_l, \tilde{\pi}_l)$.

Then in order to test the homogeneity hypothesis $H_0: \pi_1 = \pi_2$, we simply replace (y_l, n_l) by $(\tilde{y}_l, \tilde{n}_l)$ in the standard chi-square statistic and hence obtain Rao and Scott's adjusted chi-square statistic

$$\chi^2_{RS} = \sum_{l=1}^{2} \frac{(\tilde{y}_l - \tilde{n}_l \tilde{\pi})^2}{\tilde{n}_l \tilde{\pi}(1-\tilde{\pi})}.$$

Under H_0, χ^2_{RS} asymptotically follows a chi-square distribution with one degree of freedom. In order to ensure that the statistic follows a chi-square distribution with one degree of freedom the number of clusters in each treatment group must be large (Rao and Scott, 1992).

This method is not restricted to special cases and does not make any assumption on the dependence structure between individuals in a cluster as Donner's adjustment does. This method requires a large number of clusters per group and is best suited for comparisons in observational studies.

Rao and Scott (1992) also proposed a second estimate that is simply a variation of χ^2_{RS}. This estimate depends on a pooled estimate of a design effect which is assumed to be constant. The statistic is

$$\chi^2_{RSP} = \frac{\chi^2_P}{d}$$

where

$$d = \sum_{l=1}^{2} \frac{\left(1 - {}^{n_l}\!/_{n}\right)\hat{\pi}_l(1-\hat{\pi}_l)d_l}{\hat{\pi}(1-\hat{\pi})}.$$

2.4 Illustration of the analysis of a cluster randomized trial - London hypertension study

To illustrate the use of the techniques described in this chapter we consider a simplified version of the data from a 1978 London Hypertension study (Bass *et al*, 1986) also used by Donner and Klar (1993). We focus on a subgroup of 5772 female patients over the age of 45 years and the effect of blood pressure

screening and management on 5-year mortality risks. Even though the original study was a matched-pair design, it has been rearranged to illustrate the methods discussed.

A total of 34 practices were included in the study, 17 assigned to a treatment and 17 acting as controls. So the unit of assignment is a general practice but the units of observation are individual females. The total number of females included in the cluster randomized study, the number who died in each practice and the mortality rate for each of the 34 general practices are displayed in Table 2.1. This data will be used to illustrate the calculation of proportions, confidence intervals and significance testing for a cluster randomized trial.

Practice (i)	**Treated Group**			**Control Group**		
	Dead y_{Ti}	Alive $n_{Ti} - y_{Ti}$	Mortality rates $\hat{\pi}_{Ti}$	Dead y_{Ci}	Alive $n_{Ci} - y_{Ci}$	Mortality rates $\hat{\pi}_{Ci}$
1	3	98	0.0297	2	108	0.0182
2	3	50	0.0566	3	25	0.1071
3	2	211	0.0094	0	80	0.0000
4	8	287	0.0271	3	237	0.0125
5	1	163	0.0061	4	122	0.0317
6	1	223	0.0045	12	376	0.0309
7	7	311	0.0220	8	156	0.0488
8	3	169	0.0174	9	187	0.0459
9	4	285	0.0138	6	208	0.0280
10	2	89	0.0220	2	50	0.0385
11	2	126	0.0156	5	167	0.0291
12	1	16	0.0588	3	48	0.0588
13	7	97	0.0673	3	115	0.0254
14	6	88	0.0638	4	115	0.0336
15	6	277	0.0212	11	226	0.0464
16	8	207	0.0372	10	241	0.0398
17	4	233	0.0169	7	221	0.0307
Total	68	2930	$\hat{\pi}_T = 0.023$	92	2682	$\hat{\pi}_C = 0.033$

Table 2.1 Females dead and alive in the treated and control groups in each of the 34 practices in the London hypertension study

2.4.1 Results of analysis

The formulae for MSE, MSC and n_o in section 2.3.2 yielded the following: a value of MSE=0.027, MSC=0.039 and $n_0 = 162.164$. Therefore the intracluster correlation was calculated as $\hat{\rho} = 0.0027$. Even though the correlation is small it has an effect of increasing the variance of proportion estimates. This effect is characterized by a variance inflation given by $\hat{D}_T$=1.605 in the treatment group and $\hat{D}_C$=1.577 in the control group (formula provided in section 2.3.1). These values indicate that due to clustering of patients within practices, the variance of the mortality rates are over one and a half times larger than if observations were independent. So even though $\hat{\rho}$ is quite small, since cluster sizes are quite large the design effects are fairly large as well.

Then an estimate of the mortality rate of females in the treatment group is $\hat{\pi}_T$=0.023 with standard deviation 0.0035 and in the control group it is $\hat{\pi}_C$=0.033 with standard deviation 0.0043. The average cluster size of $\bar{n}$=169.76 was used to determine the variance inflation for the overall group. This was given by $1+[(\bar{n}-1)\rho]$=1.456. And therefore an estimate of the overall mortality rate of females in the combined treatment groups is $\hat{\pi}$=0.028, standard deviation 0.0026. These estimates and their 95% confidence intervals along with the corresponding estimates assuming independence are provided in Table 2.2. We observe that the point estimates of the proportions do not change but clustering of females in general practices causes an increase in the standard deviations and hence slight widening of confidence intervals.

	Assuming Independence				Cluster Trial			
	Estimate	Standard deviation	95% confidence	interval	Estimate	Standard deviation	95% confidence	interval
$\bar{\pi}_T$	0.023	0.0027	0.018	0.028	0.023	0.0035	0.016	0.030
$\bar{\pi}_C$	0.034	0.0034	0.026	0.040	0.033	0.0043	0.025	0.041
$\bar{\pi}$	0.028	0.0022	0.024	0.032	0.028	0.0026	0.023	0.033

Table 2.2 Mortality rates, standard deviations and 95% confidence intervals for the treated group, control group and overall sample of females

It appears as if the treatment group is superior to the control with respect to mortality. Its mortality rate is 2.3% compared to a mortality rate of 3.3% in the control group. Interest centres on comparing the mortality rates of the two treatment groups i.e. testing H_0: $\pi_T = \pi_C$. We consider a number of ways in which one could achieve this.

(1) Standard Pearson chi-square test

The first inappropriate method is to use the standard Pearson chi-square statistic to test the effect of the intervention on mortality. Applying the Pearson chi-square statistic to test H_0: $\pi_T = \pi_C$ yields a value of $\chi_P^2 = 5.875$ indicating that there is a significant difference in mortality rates between the two treatment groups (p=0.015).

(2) Two-sample t-test

The second approach is to apply a two-sample t-test to compare the average values of the mortality rates in the two groups. For the data in Table 2.1 the mean mortality rates in the treatment group is 0.029 and in the control group it is 0.037 with standard deviations given by 0.021 and 0.023 respectively. A test for the homogeneity of variances revealed that the variances could be assumed to be equal (F=1.246, p>0.2). The resulting value of the test statistic to compare proportions with 32 degrees of freedom is t=1.071 (p>0.2). This test reveals that there is no difference in mortality between the two groups of

interest. Note that this is a contradiction to the result obtained using the chi-square test. However, because the assumptions of the t-test are not totally satisfied and owing to a number of theoretical objections as discussed before, even though the t-test is robust to violations of its underlying assumptions, this method is not a particularly attractive one to investigators.

(3) Weighted two-sample t-test

The weighted two-sample t-test was performed on the hypertension data. The average weighted mortality rates were calculated as $\hat{\pi}_{Cw} = 0.033$ for the control group and $\hat{\pi}_{Tw} = 0.024$ for the treated group. The estimated standard error of the difference in mortality rates was 0.006. The weighted test statistic with 32 degrees of freedom was $t_w = 1.65$ (p=0.109) indicating no significant difference in mortality rates in the two groups. Since there is variation in cluster size this test is more powerful than the standard t-test.

(4) Wilcoxon rank-sum test

This non-parametric test was used to compare the proportions of the treated and control groups. Having pooled and ranked the proportions from the two groups we found that the sum of ranks for the treated group was 255.5 and for the control group it was 339.5. The test revealed that there was no difference between the two groups with respect to mortality (U=102.5, p=0.148).

(5) Donner's chi-square test

The adjusted chi-squared statistic proposed by Donner (1989) can be used for this example. This statistic depends on the clustering correction factors which in turn depend on the intracluster correlation. The estimated variance inflations for each group were calculated to be $\hat{D}_T = 1.605$ and $\hat{D}_C = 1.577$. Donner's adjusted chi-square statistic is $\chi^2_D = 3.695$ (p=0.055). Note that while the standard chi-square statistic was clearly significant at the 5% and even the 2.5% level, χ^2_D is only borderline significant at the 5% level.

(6) Rao and Scott's chi-square test

Rao and Scott's adjustment regards the mortality rate $\hat{\pi}_l$, l=T or C, as a ratio rather than a proportion. The estimates of π_T and π_C are as calculated initially in section 2.4.1, $\hat{\pi}_T = 0.023$ and $\hat{\pi}_C = 0.033$. The estimated design effects using this method are d_T=3.540 and $d_C = 1.204$. We use these estimates to adjust the standard chi-square statistic. We obtain $\tilde{n}_T$=846.927, $\tilde{y}_T$=19.210 and $\tilde{n}_C$=2304.385, $\tilde{y}_C$=76.425 and $\tilde{\pi}$=0.030. This yields a test value of χ^2_{RS}=2.313 (p=0.128). Once again this statistic is not significant at the 5% level.

2.4.2 Summary of procedures and results

Table 2.3 summarizes the techniques used and results obtained from analyzing the clustered binary experimental data concerning a hypertension study (Bass *et al*, 1986). Note that statistical significance was achieved at the 5% significance level when employing the standard statistical chi-square procedure. However, this was not so with any of the remaining methods. One is therefore able to see the way in which clustering may seriously affect results if incorrect analysis procedures are employed. In fact, using the standard chi-square test we could conclude that there is a significant difference between the two treatment groups whereas all the more appropriate tests show otherwise. The assumptions of normality and homogeneity of variances for the two-sample t-test are not strictly satisfied and therefore the weighted t-test is the more appropriate test, especially in the case of moderate to severe variation in cluster sizes. The non-parametric Wilcoxon test is a valid one but lacks statistical power. The most efficient test in this case is either Donner or Rao and Scott's adjusted chi-square test. Both account for clustering with Donner's test being suited for experimental comparisons and Rao and Scott's best for observational comparisons. Donner's test is valid when variance inflation factors of the two groups are homogenous but Rao and Scott's can be employed even if they're not homogenous. However, it requires a large number of clusters.

Procedure	Test Statistic	p-value	Comment
Standard chi-square test	$\chi_P^2 = 5.875$	p=0.015	Biased in presence of clustering
Two-sample t-test	t=1.071	p>0.2	Assumptions for t-test not satisfied
Weighted t-test	$t_w = 1.65$	p=0.109	More likely to satisfy assumptions
Wilcoxon rank sum test	U=102.5	p=0.148	Valid but lacks power
Donner's chi-square test	$\chi_D^2 = 3.695$	p=0.055	Population inflation factors must be homogeneous
Rao & Scott's chi-square test	χ_{RS}^2=2.313	p=0.128	Requires large number of clusters

Table 2.3 Comparison of procedures and results for hypertension example

2.5 Sample size calculations

Many techniques for estimating sample sizes for randomized studies have been developed over the years. These techniques have been based on statistical considerations of precision and power. Over the past few decades there has been an interest in sample size calculation in the case of cluster randomized studies (Donner *et al*, 1981; Hsieh, 1988; Shipley *et al*, 1989 and Donner, 1992).

Standard sample size calculations cannot be used when carrying out a cluster randomized trial. This is due to the fact that individuals within a cluster cannot be regarded as independent of each other and hence there is a reduction in the effective sample size. The sample size calculation must take both the size of the cluster and variability within cluster into account. Variability between clusters leads to the loss of some power and this also needs to be reflected in the sample size calculations. Both the number of individuals and number of

clusters should be considered (Kerry & Bland, 1998). The intraclass correlation generally increases the total sample size required to achieve a specified power.

There are three principal designs that are most frequently used in assigning clusters to treatments: completely randomized, stratified and matched-pair design. The methods employed in the calculation of sample sizes for the case of the completely randomized and stratified design will be discussed.

2.5.1 Completely randomized design

A number of authors have discussed sample size calculations for completely randomized design trials where the response is binary and cluster randomization has taken place (Donner *et al*, 1981; Feng and Grizzle, 1992; Shoukri and Martin, 1992, and Lee and Dubin, 1994). Each of the authors has developed sample size formulae. The appropriate formula to be used depends on the objectives of the study and the amount of information available to the researcher. We will examine different methods when the aim is to determine

(1) the number of individuals in each treatment group,

(2) the number of clusters and

(3) the number of clusters and the number of individuals per cluster.

(1) Estimating the number of individuals in the treatment group

Donner *et al* (1981) discuss the sample size requirements for a completely randomized study in which clusters are randomly assigned to treatment groups without matching or stratification. They show that the sample size required under cluster randomization is simply an adjustment of the standard sample size required under the assumptions of independence. More specifically, when clusters consist of an equal number of subjects, n^*, the sample size in a cluster

randomized study is obtained by multiplying the standard sample size formula by the design effect term $1+(n^{*}-1)\rho$. If the cluster sizes vary then n^{*} is replaced by the average anticipated cluster size, $\bar{n}=\frac{1}{k}\sum_{l=1}^{2} n_l$.

Recall that $E(\hat{\pi})=\pi$ and $\text{var}(\hat{\pi})=\frac{\pi(1-\pi)}{n}\left[1+(\bar{n}-1)\rho\right]$. An implication of this is that in order to find the necessary sample size that will provide the same power as would be obtained by randomizing independent individuals to each intervention group, we would have to multiply the usual sample size estimate of the number of individuals (required for each of the groups) by the inflation factor (Donner *et al*, 1981). Then when performing a two-sided test in the case of equal sample sizes the number of individuals required for each treatment group in order to test for homogeneity of proportions is

$$n_l=\frac{\left(z_{\alpha/2}-z_{1-\beta}\right)^2\left[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)\right]\left[1+(n^{*}-1)\rho\right]}{(\pi_1-\pi_2)^2}, \qquad l=1,2$$

where α is the probability of a type I error, β is the probability of a type II error, and $z_{\alpha/2}$ and $z_{1-\beta}$ are standard normal deviates obtained from the tables of the normal distribution for given α and β. Note that when the intracluster correlation is zero we have the sample size formula under independence of individuals. In order to apply the formula and calculate the sample sizes, replace π_1, π_2 and ρ by their estimates.

In the case of unequal sample sizes n^{*} is replaced by the cluster average $\bar{n}$. This leads to an underestimation of the required number of individuals per group. If variability in cluster sizes is small then this underestimation is only slight (Donner *et al*, 1981). If variability among cluster sizes is large Donner *et*

al (1981) suggest substituting the largest expected cluster size for n^*. This provides a more conservative estimate of group size.

Due to cluster randomization the additional parameter ρ has been incorporated in the sample size calculation. Estimates of ρ are generally unavailable. Feng and Grizzle (1992) suggest investigating the effect of variation in the estimate of ρ on the sample size by performing sample size calculations for selected values of ρ.

Example

As an example consider an epidemiological study referred to by Donner *et al* (1981) and also by Shoukri and Pause (1992). This study investigated the use of dietry sodium (treatment) and specifically looked at the binary response, hypertensive status (≥ 140/90 mmHg versus < 140/90mmHg). The unit of randomization was the spouse pair, $n^*=2$. A previous study on the familial aggregation of blood pressure in couples by Tishler *et al* (1977) showed that an estimate of ρ was 0.45. Assume that we would like to determine whether a treated group of spouse pairs tend to have a lower prevalence of hypertension than a control group. Suppose we wish to detect, with 80% power and at a 5% significance level, a significant difference between the two groups when the proportion of hypertensives in the treated group is 0.05 and in the control group is 0.15 (two-sided test). Then the required number of individuals in each of the control and treated group is

$$n_1 = \frac{(1.96+0.842)^2\left[(0.15)(0.85)+(0.05)(0.95)\right]\left[1+(2-1)(0.45)\right]}{(0.10)^2} = 199.224$$

that is, 200 individuals or 100 spouse pairs.

If no clustering were present then the number of individuals needed in each group to achieve the same power would have been

$$n_I = \frac{(1.96+0.842)^2\left[(0.15)(0.85)+(0.05)(0.95)\right]}{(0.10)^2} = 137.396$$

that is 138 individuals or 69 spouse pairs. So 62 additional individuals (31 spouse pairs) would have to be studied in a cluster randomized study to achieve the same power as an individual randomized study.

The effect of the intracluster correlation and cluster size is illustrated in the table of sample size calculations given below. Table 2.4 details the number of individuals required in each treatment group for various values of ρ and n^*.

n^*	ρ			
	0.01	0.1	0.45	0.9
2	139	152	200	262
5	143	193	385	632
20	164	399	1 313	2 487
50	205	811	3 167	6 197
100	274	1 498	6 259	12 380

Table 2.4 Number of individuals required for various values of ρ and n^*

Clearly, the number of individuals required increases as the degree of intracluster correlation and/or cluster size increases.

(2) Estimating the number of clusters

Rather than calculating the number of individuals in each treatment group, the investigator might be interested in determining the number of clusters to

include in the trial. Shoukri and Martin (1992) discuss estimation of the number of clusters and point out two problems associated with estimating the number of clusters. Firstly, as mentioned before, ρ is unknown and needs to be estimated. Secondly, if cluster sizes are variable then the distributions of statistics used in inferences (regarding model parameters) are more complex.

The aim of the study is important in estimating cluster numbers. Shoukri and Martin (1992) determine the number of clusters needed to test whether there is a clustering effect, H_0: $\rho = 0$ versus H_1: $\rho = \rho_1 > 0$.

The power of the test is obtained using the asymptotic normality of $\hat{\rho}$. The power is given by

$$P(\rho) = P\left[z_{1-\beta} > \frac{z_\alpha \left\{ \sqrt{\phi(0,\pi)/k} \right\} - \rho_1}{\sqrt{\phi(\rho_1,\pi)/k}} \right] = 1 - \beta$$

where α is the probability of a type I error, β is the probability of a type II error, z_α and $z_{1-\beta}$ are standard normal deviates obtained from the tables of the normal distribution for given α and β, and $\phi(\rho,\pi) \approx k \,\mathrm{var}(\hat{\rho})$. So the number of clusters necessary to achieve a specified power will depend on the values of ρ_1, π, the level of significance and also the distribution of the cluster sizes (Shoukri and Martin, 1992).

Using the equation given above and solving for the number of clusters we obtain

$$k = \left[\frac{z_\alpha \sqrt{\phi(0,\pi)} - z_{1-\beta} \sqrt{\phi(\rho_1,\pi)}}{\rho_1} \right]^2 .$$

The accuracy of the sample size calculation depends on the validity of the assumption concerning the asymptotic normality of $\hat{\rho}$.

Example

We once again consider the hypertension example discussed above. For the special case where $n^*=2$ we have

$$\phi(\rho,\pi)=\frac{(1-\rho)[\rho(2-\rho)+2\pi(1-\pi)(1-\rho)(1-2\rho)]}{2\pi(1-\pi)}.$$

We now wish to determine the number of spouse pairs in order to have an 80% chance at $\alpha=0.05$ of detecting a value of the intracluster correlation of 0.35 or larger (one-sided) knowing that the proportions of hypertensives is 0.15. So $\phi(0,0.15)=1$ and $\phi(0.35,0.15)=1.599$. Then

$$k=\left[\frac{1.64+0.842\sqrt{1.599}}{0.35}\right]^2=59.719$$

that is, 60 randomly selected couples should be included in the study.

(3) Estimating the number of clusters and the number of individuals per cluster

(a) Feng and Grizzle's formula

Feng and Grizzle (1992) developed formulae for calculating both the number of clusters and the number of individuals per cluster. They assume equal cluster sizes, n^*, and that the interest lies in comparing proportions. Then, to achieve a specified power the number of clusters required for a two-sided test is

$$k = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2 \left[1 + (n^* - 1)\rho\right] \left[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right]}{n^*(\pi_1 - \pi_2)^2} \qquad \text{for fixed } n^*.$$

The number of individuals per cluster is

$$n^* = \frac{\left(z_{\alpha/2} - z_{1-\beta}\right)^2 (1-\rho) \left[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right]}{(\pi_1 - \pi_2)^2 k - \left(z_{\alpha/2} - z_{1-\beta}\right)^2 \left[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\right] \rho} \qquad \text{for fixed } k.$$

It may happen that the denominator of the sample size formula for the cluster size turns out to be negative. This occurs when the number of clusters, k, is too small and as a consequence the study cannot achieve the desired power for any n^*.

Example

We return to the hypertension example to illustrate the use of the formulae given above. Recall that the proportion of people with hypertension was 0.15 in the control group and 0.05 in the treated group, the intracluster correlation being 0.45. If we would like to detect a significant difference between the two groups with 80% power and at the 5% significance level then the number of spouse pairs $(n^* = 2)$ that should be included in the study is

$$k = \frac{(1.96 + 0.842)^2 (1 + 0.45) \left[(0.15)(0.85) + (0.05)(0.95)\right]}{2(0.10)^2} = 99.612 \text{ or } 100.$$

This is equal to the value obtained using Donner's formula given in section (1) above.

Assume now that we were examining people and their hypertension status within general practices, treatment being applied at the general practice level.

Suppose also that $\rho = 0.01$ and that we have k=6 practices participating in the investigation. In order to find a significant difference between groups at the same significance level and power given above, we would need to include in the study

$$n^* = \frac{(1.96+0.842)^2(1-0.01)\ [(0.15)(0.85)+(0.05)(0.95)]}{(0.10)^2(6)-(1.96+0.842)^2[(0.15)(0.85)+(0.05)(0.95)](0.01)} = 29.404,$$

that is 30 people from each of the 6 general practices.

(b) Lee and Dubin's formula

Now, in order to make use of the formulae above we require knowledge of the cluster size or need to make the assumption that the cluster sizes are equal. In addition, we have to assume that the correlation between any two individuals within a cluster is the same. Lee and Dubin (1994) present sample size calculations for the number of clusters required in a study. This formula may be used if cluster sizes are not known in advance or are unequal, and if pairwise correlation varies with cluster. They are interested in estimating a proportion and testing whether the proportion equals a specified value or not i.e. H_0: $\pi = \pi_0$ versus H_1: $\pi = \pi_1$ where $\pi_1 \neq \pi_0$.

Let k denote the number of clusters in the study with n_i individuals in the ith cluster. We assume that n_i follows a distribution that takes on positive integer values only. Let y_{ij} denote a binary random variable where $y_{ij} = 1$ if the response of the jth individual in cluster i is a success and $y_{ij} = 0$ if it's a failure. Lee and Dubin (1994) start off by considering estimation of π under the working assumption that individuals within clusters are independent. Then an estimate of the proportion of successes $\hat{\pi}$ is

$$\hat{\pi} = \frac{\sum_{i=1}^{k} y_i}{\sum_{i=1}^{k} n_i}.$$

Under independence the estimated variance of $\hat{\pi}$ is therefore

$$\hat{\text{var}}_B(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{\sum_{i=1}^{k} n_i}.$$

If dependence exists between individuals within a cluster then the true variance of $\hat{\pi}$ is either larger or smaller than the variance of $\hat{\pi}$ using the binomial distribution. Henderson *et al* (1988) proposed the use of the ratio estimate of the proportion and its variance. This was mentioned when discussing Rao and Scott's adjustment to the chi-square statistic. The point estimate of the proportion $\hat{\pi}$ is the same as the estimate given above. The variance is different to the binomial variance and has the form

$$\hat{\text{var}}_{RS}(\hat{\pi}) = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{n_i}{n}\right)^2 \frac{(\hat{\pi}_i - \hat{\pi})^2}{(k-1)}$$

$$= \frac{1}{k\bar{n}} \frac{\sum_{i=1}^{k} y_i - 2\hat{\pi}\sum_{i=1}^{k} y_i n_i + \hat{\pi}^2 \sum_{i=1}^{k} n_i^2}{k-1}$$

where $\hat{\pi}_i$ is an estimate of the proportion of successes in the ith cluster and $\bar{n}$ is the average number of individuals per cluster.

Prior knowledge of the size of each cluster is needed in order to calculate $\hat{\text{var}}_B(\hat{\pi})$ and $\hat{\text{var}}_{RS}(\hat{\pi})$. But generally this is unknown and hence $\hat{\text{var}}_B(\hat{\pi})$ and $\hat{\text{var}}_{RS}(\hat{\pi})$ cannot be used in sample size calculation. When deriving the ratio

estimate, the same weight is assigned to each individual regardless of the size of the cluster to which they belong. It is important to realize though that owing to correlation between individuals within a cluster, a single cluster with k subunits will contribute less information than k clusters each consisting of one individual. Lee and Dubin (1994) continue by reweighting. We therefore reweight the clusters so that the estimate of π becomes

$$\hat{\pi}_W = \frac{1}{k}\sum_{i=1}^{k} w_i n_i \hat{\pi}_i$$

where w_i is the weight assigned to the ith cluster and satisfies the constraint

$$\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_i = 1,$$

and $\hat{\pi}_i = \frac{y_i}{n_i}$ is the estimated cluster-specific proportion.

If $w_i = \frac{k}{\sum_{i=1}^{k} n_i}$ then $\hat{\pi}_w$ becomes the ratio estimator discussed above. Lee and Dubin (1994) suggest the use of $w_i = \frac{1}{n_i}$ which results in $\hat{\pi}_w = \frac{1}{k}\sum_{i=1}^{k}\hat{\pi}_i$, the simple average of success rates in each cluster. Use of this weighting scheme avoids the dominance of a few large clusters in the sample.

Now, using conditional expectation arguments,

$$E(\hat{\pi}_i) = E\left(\frac{y_i}{n_i}\right) = E\left[E\left(\frac{y_i}{n_i}\middle| n_i\right)\right] = E(\pi) = \pi \qquad i=1,2,\ldots,k$$

and

$$\begin{aligned}\text{var}(\hat{\pi}_i) &= E\left[\text{var}(\hat{\pi}_i|n_i)\right] + \text{var}\left[E(\hat{\pi}_i|n_i)\right] \\ &= E\left[\text{var}(\hat{\pi}_i|n_i)\right] + \text{var}(\pi) \qquad\qquad i=1,2,\ldots,k.\end{aligned}$$

$E\left[\text{var}(\hat{\pi}_i|n_i)\right]$ is a constant because the cluster sizes are identically and independently distributed and each cluster has the same correlation structure. The second term is the variance of a constant, π, which equals zero. Furthermore, it can be shown that all the $\hat{\pi}_i s$ have the same continuous distribution function, call it $f(z)$, that takes on values between 0 and 1. Then, by the central limit theorem

$$\sqrt{n(\hat{\pi}_w - \pi)} \sim N(0, \sigma^2)$$

where σ is the standard deviation of Z.

So,

$$\text{var}(\hat{\pi}_w) = \frac{\text{var}(\hat{\pi}_i)}{k} = \frac{\sigma^2}{k}.$$

Hence, a consistent estimate of $\text{var}(\hat{\pi}_w)$ is

$$\hat{\text{var}}(\hat{\pi}_w) = \frac{1}{k}\frac{\sum_{i=1}^{k}(\hat{\pi}_i - \hat{\pi}_w)^2}{k-1}.$$

Lee and Dubin (1994), having determined a consistent estimate of $\text{var}(\hat{\pi}_w)$, go on to derive a formula for calculating the number of clusters that should be chosen so as to estimate $\hat{\pi}_w$ to within a distance D of π with probability 1-α. That is, we want

$$D = \frac{z_{\alpha/2}\sigma}{\sqrt{k}}.$$

Hence

$$k = \frac{\left(z_{\alpha/2}\sigma\right)^2}{D^2}.$$

We incorporate power into the sample size calculation and determine the number of clusters needed such that the Type I error rate is α and the power 1-β. Then the number of clusters needed is

$$k = \left[\frac{\sigma\left(z_{\alpha} - z_{1-\beta}\right)}{\pi_1 - \pi_0}\right]^2.$$

Both these formulae are valid under any correlation structure.

The value for σ is needed in the calculation and can be obtained using information from a prior or pilot study. If there is no estimate of σ available in advance then the choice of σ has to be based on other considerations.

Lee and Dubin (1994) illustrate sample size estimation using the beta-binomial model. Since the probability of success π takes on values between 0 and 1 they view it as being a random variable. They then allow π to follow a beta distribution with parameters a and b. This results in the beta-binomial distribution which has

$$E(\pi) = \frac{a}{a+b} = \mu \text{ and}$$

$$\text{var}(\pi) = \frac{ab}{(a+b)^2\gamma} = \gamma\mu(1-\mu)$$

where $\gamma = \frac{1}{a+b+1}$.

We are able to make use of this distribution to find the sample size necessary to estimate π to within say 10% of the true value. The estimates depend on the level of concordance or discordance and the success rates within the clusters. We examine some of the examples provided by Lee and Dubin (1994). These are summarized in the table below. Lee and Dubin (1994) look at the cases of

i) concordant responses, where the investigator believes that the responses in a cluster are mostly successes or failures

ii) discordant responses which imply that success rates in the cluster vary

iii) high success rate (the probability of success is high in all clusters)

iv) low success rate (the probability of failure is high in all clusters)

v) no prior information

Assume a 95% confidence level.

	Concordant	Discordant	High success rate	Low success rate	No prior information
Value of a	$a<1$	$a>1$	$a>1$	$a\leq 1$	$a=1$
Value of b	$b<1$	$b>1$	$b\leq 1$	$b>1$	$b=1$
Example	$a=b=0.5$	$a=b=1.5$	$a=1.5$ $b=0.5$	$a=0.5$ $b=1.5$	$a=b=1$
Variance σ^2	0.125	0.0625	0.0625	0.0625	0.0833
Number of clusters	49	25	25	25	33

Table 2.5 Sample size examples using Lee and Dubin's (1994) method

2.5.2 Stratified cluster randomization design

The stratified cluster randomization design is an extension of the matched-pair design in which two or more clusters per stratum are randomly assigned within strata to a treatment group. Donner (1992) generalizes a formula developed by Woolson *et al* (1986) for stratified case-control studies.

Suppose that k_h clusters of size n_h are randomly assigned in a balanced fashion to the two treatment groups in stratum h, h=1,2,...,s. The probability of success in the control group in stratum h is denoted by π_{h1}, and π_{h2} denotes the success probability characterizing the intervention group. Woolson *et al* (1986) considers the odds ratio, ω, as the effect of intervention. Under the assumption that the treatment effect is characterized by the odds ratio, ω, one can calculate π_{h2} given the value of π_{h1}. This is

$$\pi_{h2} = \frac{\pi_{h1}\omega}{1-\pi_{h1}+\pi_{h1}\omega}.$$

Woolson *et al* (1986) developed a formula that enables the investigator to determine k_h, the number of clusters in stratum h that will test H_0: $\omega = 1$ at the α level of significance and with power 1-β.

Assume that individuals instead of clusters are randomly assigned to one of the treatment groups within strata. Let t_h denote the proportion of individuals belonging to stratum h with $\sum_{h=1}^{s} t_h = 1$. Also let $\overline{\pi}_h = \frac{\pi_{h1}+\pi_{h2}}{2}$ denote the overall success rate for stratum h. Then Woolson *et al*'s formula for the total number of subjects to test a given two-sided hypothesis is

$$n_w = \frac{\left[z_{\alpha/2}T - z_{1-\beta}U\right]^2}{V^2}$$

where

$$T = \frac{1}{2}\sqrt{\sum_{h=1}^{s} t_h\left[\overline{\pi}_h(1-\overline{\pi}_h)\right]},$$

$$U = \sqrt{\frac{1}{8}\left[\sum_{h=1}^{s} t_h\left[\pi_{h1}(1-\pi_{h1}) + \pi_{h2}(1-\pi_{h2})\right]\right]} \quad \text{and}$$

$$V = \frac{1}{4}\sum_{h=1}^{s} t_h(\pi_{h1} - \pi_{h2}).$$

This formula can be modified to take cluster randomization into account. This is achieved by adjusting for the within-cluster correlation, quantified by ρ, which is assumed to be constant across strata. Then the total number of subjects needed in a stratified cluster randomized design is

$$n = \frac{\left[z_{\alpha/2}T' - z_{1-\beta}U'\right]^2}{V^2}$$

where

$$T' = \frac{1}{2}\sqrt{\sum_{h=1}^{s} t_h\left[\bar{\pi}_h(1-\bar{\pi}_h)\right]\left[1+(n_h-1)\rho\right]},$$

$$U' = \sqrt{\frac{1}{8}\left[\sum_{h=1}^{s} t_h\left[\pi_{h1}(1-\pi_{h1}) + \pi_{h2}(1-\pi_{h2})\right]\left[1+(n_h-1)\rho\right]\right]} \quad \text{and}$$

$V = \frac{1}{4}\sum_{h=1}^{s} t_h(\pi_{h1} - \pi_{h2})$ as before.

The formula for T' and U' are similar to the unclustered case but simply multiplied by an inflation factor $\left[1+(n_h-1)\rho\right]$ for each stratum h, h=1,2,…,s.

Therefore the number of clusters that need to be assigned to each treatment group in stratum h is given by

$$k_h = \frac{nt_h}{2n_h}.$$

This formula allows for

(i) an equal number of subjects in each stratum in which case $t_h = t$ or

(ii) an equal number of clusters within each stratum when $t_h = \frac{n_h}{\sum_{h=1}^{s} n_h}$. Then

$$k_h = k^* = \frac{n}{2\sum_{h=1}^{s} n_h}.$$

To make use of the formula, information on ρ and the stratum-specific success rates π_{h1} and π_{h2} (h=1,2,...,s) is needed.

If all clusters over all strata are to be the same size, n^*, the formula developed by Woolson *et al* (1986) may be used and simply multiplied by the inflation factor $[1+(n^*-1)\rho]$. However, if cluster sizes are to vary within strata then the adjustment to Woolson *et al*'s formula involves replacing n_h by the average anticipated cluster size in stratum h, $\bar{n}_h$. There is a tendency to underestimate the sample size but this underestimation is small if variation in cluster size within strata isn't too large (Donner, 1992).

Example

To illustrate stratified sample size calculation we adapt an intervention trial example presented by Donner (1992). The study aimed to compare virucidal-impregnated tissues to regular tissues with respect to the prevention of respiratory illness. Entire families of size two, three and four were randomized to the treatments to enhance compliance and avoid treatment contamination. Each of these family sizes formed a stratum. We thus have h=3 strata with cluster sizes of n_1=2, n_2=3 and n_3=4 in each of the strata. Assume that the investigator wishes to determine the number of families, k_h,

that must be assigned to each treatment group within each of the strata. Assume that the proportions of individuals in stratum h that experience respiratory illness in the control group, $\hat{\pi}_{hc}$, is as given in the table below. Suppose also that equal fractions of subjects belong to each of the three strata i.e. $t = 1/3$, the estimated odds ratio is given by $\hat{\omega} = 2$ and that the intracluster correlation is $\rho = 0.1$. Then $\hat{\pi}_{hT}$ and $\bar{\hat{\pi}}_h$ can be calculated and is given in Table 2.6 below.

h	Control respiratory illness rate $\hat{\pi}_{hC}$	Treatment respiratory illness rate $\hat{\pi}_{hT}$	Overall respiratory illness rate $\bar{\hat{\pi}}_h$	k_h (equal number of individuals per stratum)	k (equal number of clusters per stratum)
Stratum 1	0.038	0.056	0.047	581	388
Stratum 2	0.035	0.052	0.043	388	388
Stratum 3	0.042	0.062	0.052	291	388

Table 2.6 Hypothesized respiratory illness rates and resulting sample sizes for a stratified cluster sample

The number of families that must be assigned to each treatment group in each stratum at a 5% significance level (two-sided test) and with 80% power is then given by k_1=581, k_2=388 and k_3=291. The total number of subjects to be included in the trial is 6980. If the number of clusters assigned to each treatment group are to be equal then $k^* = 388$ families should be assigned to each treatment group within each stratum.

****

In many studies investigators choose a few large clusters rather than a large number of clusters thinking that the large cluster size will make up for the small amount of clusters that have been chosen. However, a greater power gain is achieved by increasing the number of clusters in a study rather than

increasing the average cluster size. This can be seen by looking at the variance of an observed proportion $\hat{\pi}$ for a study in which k clusters of size n^* are assigned to a treatment group. This variance is, as given before,

$$\text{var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n^* k}\left[1 + (n^* - 1)\rho\right].$$

Then we can clearly see that $\text{var}(\hat{\pi}) \to 0$ only as k increases and not n^*. And in most cluster randomized trials there is usually only minimal control over the cluster sizes. In addition, Donner (1998) points out that we can only improve the power up to a certain threshold if we were to increase the average cluster size. This threshold value depends on the value of ρ.

CHAPTER 3
DESIGN AND ANALYSIS OF CLUSTERED NON-EXPERIMENTAL STUDIES

3.4.5 Logistic regression

(1) Logistic regression ignoring cluster design and weighting

(2) Logistic regression taking weighting into account

(3) Logistic regression accounting for clustering only

(4) Logistic regression incorporating correct weighting and consideration of cluster design

3.4.6 Summary of logistic regression results

CHAPTER 3
DESIGN AND ANALYSIS OF CLUSTERED NON-EXPERIMENTAL STUDIES

3.1 Introduction

This chapter deals with a class of studies known as observational or non-experimental studies. A characteristic of this type of study is that the investigator does not assign some active intervention to a sample of individuals, as is the case with an experiment. So even though the objective of the study may be to study the causal effects of some treatment, the investigator cannot impose or withhold treatment from the subjects. Therefore the investigator is limited to taking selected observations that seem appropriate to the study. The groups that the investigator would like to compare are in fact already selected in some manner not chosen by the investigator.

3.2 Design of observational studies

A number of specialized designs fall into the broader class of observational studies. The simplest observational study is the cross-sectional study that examines the study population at a specific point in time. They are most useful for description and if sampling is necessary, large samples are encouraged to improve precision of estimates. Examples of cross-sectional studies include surveys and polls. Surveys will be examined in greater detail in the context of cluster sampling.

A second type of observational study is the cohort study. Here subjects are followed prospectively through time. In principle, they can mimic the conditions of an experiment but the investigator is not the one who assigns the intervention. Problems of confounding may arise and cohort studies may require years of observation. Surveys can be used in cohort studies as well.

Finally, the retrospective case-control study involves investigating a group of people who do have a disease or attribute of interest (cases) and a group of people who do not (controls). The cases and controls are then compared with respect to presence or absence of certain risk factors. This study is easy to perform but cannot be used to measure incidence or prevalence.

3.2.1 Survey sampling

The primary objective of most sample surveys are to estimate population parameters and their sampling variances. The validity, reliability and accuracy of these estimators depend on how well the sample was chosen and how accurately measurements were made. There are a number of ways of drawing a sample from the target population and hence a number of sampling methods have been developed. We look at some of the most common methods.

(1) Simple random sampling

Simple random sampling (SRS) has played an important role in the development of sampling theory by providing a base upon which more advanced theory could be constructed. This is the simplest form of survey sampling and is therefore used as a reference when examining other sampling methods. This method involves selecting elements to form a sample without making use of auxiliary information and in such a way that each of the various possible samples has the same chance of being drawn. However, in practice, obtaining a purely simple random sample can be quite problematic

- Simple random sampling tends to be expensive and is often not the feasible option.
- Minority subgroups may not be fairly represented by simple random sampling.
- Even though it is easy to obtain the sampling distribution with simple random sampling, other sampling methods such as stratified sampling may

give rise to sampling errors that are smaller than that produced by simple random sampling.

Other methods of sampling which prove to be more practical and useful do exist. The following sampling methods are essentially modifications of simple random sampling. These designs are substituted in place of simple random sampling because they are more economical, practical or could produce more precise results.

(2) Stratified sampling

Stratified sampling is a fairly simple and widely used technique that makes use of auxiliary information in the sampling design and analysis. This is achieved by separating the target population into a number of non-overlapping subpopulations known as strata. Populations are usually heterogeneous and in order to obtain a precise estimate many elements would need to be sampled. Stratification allows one to achieve a high level of precision with only a small sample size. The aim of stratification is to group similar elements in the same strata i.e. strata should be homogeneous with respect to the stratifying variable. Randomness enters the study when a simple random sample of individuals is chosen independently from each stratum. Other methods of sampling, such as systematic sampling, may be used to select elements from the strata. Stratification may be performed using appropriate auxiliary information like regional, demographic or socioeconomic information. The result is statistically independent strata.

- The effect of stratification may be a reduction in standard errors of resulting estimates within each stratum and hence an overall reduced standard error. This is especially true if elements within strata are homogeneous. And individuals that are similar with respect to the variation in the response are usually grouped together within a stratum. The result is a small within-stratum variance and hence a small overall population variance. This

enhanced precision is one of the main reasons for employing this sampling method.

- Administrative reasons may also motivate the use of stratification as well as the need to make full use of the auxiliary information available. In a stratified sample it is possible to obtain separate estimates for each stratum.
- Stratification ensures that minority subgroups in the population are represented in the sample.
- Cost per observation may be reduced by stratification into convenient groups.
- A disadvantage of stratification is that selection of the sample may be more time-consuming.

(3) Cluster sampling

In large-scale surveys it is often impossible to construct a list of every individual in the population and therefore simple random sampling cannot be employed. A more appropriate method in this case would be cluster sampling. Cluster sampling is the process whereby non-overlapping groups or clusters of individuals instead of single individuals are sampled. The sizes of the cluster need not be the same. Auxiliary information is used to form clusters from natural occurring groups within the target population. Then a sample of clusters is chosen from the population of clusters using a sampling method like simple random sampling. Therefore randomness is introduced into the study by the selection of clusters rather than selection of individuals directly from the population. Subgroups that are often used as clusters are hospitals in a country, classes within a school or households in a city. Then, in order to obtain a sample all that is needed is a list of clusters, and hence an easily accessible list of elements from the sampled clusters rather than a complete frame covering all population individuals.

Cluster sampling is widely used in practice due to the following reasons, but at the expense of statistical efficiency:

- It is more feasible than other sampling methods because a sampling frame at the element level is not necessary.
- Economically, cluster sampling is the most cost efficient sampling scheme especially when the target population is spread over a large region.
- Cluster sampling may also be the more feasible option in terms of administrative efficiency and is less time-consuming.
- Cluster sampling may also serve to enhance subject compliance.

In order for cluster sampling to be more efficient than simple random sampling the elements within clusters should be heterogeneous and the clusters should be homogenous. However, it is often the case that clusters are internally homogenous and therefore elements within the same cluster are not independent of each other. The correlation that exists between elements must be taken into account when obtaining estimates of standard errors and confidence intervals.

The correlation between elements in the same cluster is known as the intracluster correlation and was discussed in Chapter 2 for the experimental situation. It measures the degree to which elements within the same cluster are similar with respect to the presence or absence of the characteristic of interest. The more homogenous clusters are with respect to the characteristic, the greater will be the intracluster correlation. This leads to an increase in standard errors of estimates and hence a decrease in statistical efficiency. Even though there is a substantial loss in precision the lower cost per sampling unit compensates for this.

One-stage cluster sampling is a special case of cluster sampling. It involves choosing a simple random sample of clusters and including each element within the cluster in the sample. Subsampling may also occur within clusters once the first level of clusters has been selected. Subsampling within clusters is known as multistage sampling (selection of sampling units in two or more

stages of sampling). For example, cities may be sampled, then suburbs within cities, then households within suburbs and finally, persons within households. The units selected at the first level of sampling are known as primary sampling units (PSUs). They play a special role in multistage sampling. Variance estimates are computed using only information at the primary sampling unit level. They do not require information about secondary and beyond sampling units. In the example given above the cities are the primary sampling units.

3.3 Analysis of data arising from survey sampling

A large amount of research has been focused on the theory of sampling estimation (Cochran, 1953; Kish, 1965; Scheaffer, 1990 and Barnett, 1991). Standard methods for point estimation of sample characteristics are readily available. Methods for estimating both linear functions like means and totals, and non-linear functions like ratios, along with consistent estimators of the sample variance of these estimates have been developed.

We will consider a complex sample made up of a number of clusters. It has been found that the estimators produced by assuming an independent sample are unbiased. However, homogeneity within clusters tends to increase the variance of estimators (Kish and Frankel, 1974). Standard independence assumptions in this case will bias the variance downwards and hence produce spurious results for test statistics. Analyzing clustered sample data using procedures that ignore cluster sampling can produce incorrect standard errors, which consequently produces incorrect confidence intervals and test statistics, and hence severely misleading results. Adjustments, which take clustering into account, should be made.

Another important characteristic of survey data that arises from the data collection procedure is sampling weights. Even though observations in a sample survey are selected using a random procedure, different observations

may have different selection probabilities. Weights are equal or proportional to the inverse of the probability of being sampled. So the sampling weights effectively represent the number of individuals in the population that each sampled individual represents. If the *j*th observation has a weight of w_j, it represents w_j elements in the population from which the sample was drawn. There are a number of reasons for including weights in survey analysis. Firstly, resulting estimates are approximately unbiased when weights are incorporated. In addition, the weights affect the standard errors of the estimates (Little *et al*, 1997).

3.3.1 Estimation of proportions using clustered samples

For the purpose of this section we will only focus on the analysis of binary data from one-stage cluster sampling. Hence the estimation of the population proportion.

Suppose the population is made up of K clusters of individuals of sizes $n_1, n_2, \ldots, n_K$ with a total population size $\sum_{i=1}^{K} n_i = N$.

A one-stage cluster sample is obtained by taking a simple random sample of k clusters, with n_i individuals in the ith cluster, $i=1,2,\ldots,k$, and including all the individuals of the chosen clusters in the sample, $n = \sum_{i=1}^{k} n_i$..

(1) <u>Calculating selection probabilities and weights</u>

Because each subject in a sample may not have the same chance of being drawn under cluster sampling, weights need to be calculated for each subject in the sample. The weighting calculation is usually based on the probability selection for each subject (Lemeshow *et al*, 1998). Consider a one-stage cluster sample with probabilities

p_{1j} : the probability of selecting the *i*th primary sampling unit and

p_{2ij} : the probability of selecting the *j*th subject from the *i*th primary sampling unit, given that the *i*th primary sampling unit has been selected.

Then the total probability of choosing the *j*th subject from the *i*th primary sampling unit is given by

$$p_{ij} = p_{1j} \times p_{2ij} .$$

The weight, w_{ij}, allocated to this subject is then proportional to the reciprocal of the total selection probability

$$w_{ij} \propto \frac{1}{p_{ij}}.$$

Generally we set

$$w_{ij} = \frac{1}{p_{ij}}.$$

(2) Estimating proportions

We are interested in estimating the proportion, π, of individuals in the population that possess some characteristic of interest. Estimates of proportions for cluster samples are readily available (Barnett, 1991 and Cochran, 1953).

Let y_{ij} describe the *j*th individual in the *i*th cluster so that

$y_{ij} = 1$ if the individual possesses the characteristic of interest

$= 0$ if the individual does not possess the characteristic of interest.

The total number of individuals in the population who possess the characteristic of interest is

$$Y = \sum_{i=1}^{K} \sum_{j=1}^{n_i} y_{ij} .$$

Then an estimate of the total number of individuals that possess the characteristic of interest is given by

$$\hat{Y} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} y_{ij}$$

where w_{ij} is the weight associated with the jth individual in the ith cluster, $w_{ij} = \frac{1}{p_{ij}}$ with p_{ij} the selection probability of the jth element in the ith cluster.

Now define an additional variable x_{ij} with population total given by

$$X = \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij}$$

and estimated by

$$\hat{X} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} x_{ij} .$$

Then the population proportion of individuals who possess the characteristic is

$$\pi = \frac{Y}{X}$$

with $x_{ij} = 1$ for all i=1,2,...,k; j=1,2,...,n_i.

Note that an estimate of the population size is $\hat{N} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}$ which equals $\hat{X}$ in the case of proportion estimation.

An estimate of the proportion is

$$\hat{\pi} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}}.$$

This unbiased point estimate of π equals the estimate that would have been obtained if we assumed all observations in our sample were independent (simple random sample).

The estimated variance of $\hat{\pi}$ under cluster sampling can be obtained using the delta method (i.e. a first-order Taylor expansion). It is

$$\hat{\text{var}}_{CL}(\hat{\pi}) = \frac{k}{k-1}\frac{1}{\hat{N}^2}\sum_{i=1}^{k}\left(\sum_{j=1}^{n_i} w_{ij} y_{ij} - \hat{\pi}\sum_{j=1}^{n_i} w_{ij}\right)^2.$$

This is a weighted version of the variance of a ratio estimate discussed by Rao and Scott (1992) and derived for testing proportions in cluster randomized trials in section 2.3.4 (4).

3.3.2 Measuring the effects of cluster sampling – Design effects for proportions

Cluster sampling gives rise to dependent data that usually results in the variance of estimates being larger than expected under a simple random sample of the same size. Because the variances of estimates in particular are affected by using standard procedures under cluster sampling, an appropriate measure for the effect of clustering should take these variances into account. A measure of the effect of clustering on estimators is the design effect, also known as the variance inflation factor. The design effect is simply the ratio of the variance of a parameter, say θ, under cluster sampling and the variance under a hypothetical simple random sample of the same size,

$$\text{DEFF}(\theta) = \frac{\text{var}_{CL}(\theta)}{\text{var}_{SRS}(\theta)}.$$

One is then able to derive the design effect for the proportion estimate presented in the previous section. If a simple random sample without replacement had been obtained with the same number of elements, n, as in the cluster sample then the estimated variance of the proportion estimate $\hat{\pi}$ would be calculated using the formula

$$\hat{\text{var}}_{SRS}(\hat{\pi}) = \frac{1}{\hat{N}(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}(y_{ij} - \hat{\pi})^2 .$$

This means that the estimated design effect of the proportion estimate $\hat{\pi}$ is

$$\text{DEFF}(\hat{\pi}) = \frac{\hat{\text{var}}_{CL}(\hat{\pi})}{\hat{\text{var}}_{SRS}(\hat{\pi})}$$

$$= \frac{k(n-1)\sum_{i=1}^{k}\left(\sum_{j=1}^{n_i} w_{ij}y_{ij} - \hat{\pi}\sum_{j=1}^{n_i} w_{ij}\right)^2}{\hat{N}(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}\left(y_{ij} - \hat{\pi}\right)^2}$$

The design effect under cluster sampling usually produces values greater than unity, implying that the cluster design yields an estimate that has a higher variance than would be obtained using a simple random sample of the same size. Stratification, on the other hand, generally produces design effects less than unity.

Design effects can be used in the analysis stage for adjusting standard statistical estimates for the effects of dependence brought about by clustering. Design effects are also useful in the planning and design stage, especially in the determination of required sample sizes and as a measure of efficiency when comparing alternative designs.

3.3.3 Logistic regression for clustered samples

The estimation methods discussed before simply provided straightforward estimates of proportions. We often want to look at the relationship between these proportions and a number of independent explanatory variables. We therefore turn to analytic inference of clustered data.

There has been an increasing amount of research done on the analytical inference of complex sample data (Holt and Scott, 1981; Scott and Holt, 1982 and Nathan, 1988). These include methods like regression and analysis of variance that had originally been developed under the assumption of simple random sampling which is rarely realized in practice. We will specifically look at the case of logistic regression.

Suppose that the response of an arbitrary observation, y_j (j = 1,...,n), is binomially distributed with parameters n, the number of observations, and π_j, the probability of success. The linear logistic model under standard statistical theory relates the probability π_j to r explanatory variables, $x_j^T = (x_{1j}, x_{2j}, ..., x_{rj})$ associated with that observation in the following way

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{1-\pi_j}\right) = x_j^T \beta = \beta_0 + \beta_1 x_{1j} + ... + \beta_r x_{rj}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_r)$ is the unknown parameter vector of regression coefficients to be estimated. Then

$$\pi_j = \frac{\exp(\beta_0 + \beta_1 x_{1j} + ... + \beta_r x_{rj})}{1 + \exp(\beta_0 + \beta_1 x_{1j} + ... + \beta_r x_{rj})}.$$

Interpretation of the parameters is usually in terms of odds ratios, $\exp(\beta_s)$, s=1,2,...,r. The general method of estimation of the logistic regression parameters is the method of maximum likelihood.

An important assumption of the logistic regression model is that observations are independent. In fact, the logistic model is only valid when one can assume that a sample has been drawn from a population using simple random sampling where each element has the same probability of being selected.

Stratified sampling is accounted for in logistic regression by simply including an additional explanatory variable that indicates from which stratum an observation has been sampled. It is then also possible to determine whether there is an interaction between the strata and the remaining explanatory variables, that is, whether the effect of the explanatory variable depends on the strata.

Methods which fit logistic models when responses are correlated have been developed. The approach to modelling in complex surveys involves viewing the intracluster correlation as a nuisance effect with the aim of eliminating it from estimation and test results. This aggregated approach encompasses estimation methods like weighted least squares, pseudolikelihood (PL) estimation and the generalized estimating equations method. However, the weighted least squares model cannot be used in the case where predictor variables are continuous measurements (Lehtonen and Pahkinen, 1995). This section briefly examines the pseudolikelihood approach to modelling the relationship between a number of explanatory variables and intracluster correlated binary responses. Chapter 4 will examine the generalized estimating equations approach in detail.

For complex sample designs we cannot calculate maximum likelihood estimates due to the difficulty in obtaining appropriate likelihood functions. The approach to analyzing clustered survey data involves appropriately weighting observations, taking their unequal selection probabilities and the presence of intracluster correlation into account. Pseudolikelihood estimation is essentially a modification of maximum likelihood estimation.

Consider a binary response variable y_{ij}, for the jth (j=1,2,...,n_i , $\sum_{i=1}^{k} n_i$) individual in the ith cluster (i=1,2,...,k). Associated with each of the responses are a number of explanatory variables given by $x_{ij}^T=(x_{1ij}, x_{2ij},..., x_{rij})$. Let $\pi_i = f_i(\beta)$ be the probability of a response in the ith cluster and define the ratio estimate of the proportion of successes for the ith cluster, p_i. Also define a $n\times r$ design matrix, X^T, and a $n\times n$ diagonal matrix W with selection probability weights w_{ij} on the diagonal. Then the PL estimate for the logit model, $\hat{\beta}_{PL}$, is obtained by iteratively solving the PL estimating equations given by

$$X^T W \hat{f} = X^T W p$$

where $\hat{f} = f(\hat{\beta}_{PL}) = \left(\hat{f}_1, \hat{f}_2, ..., \hat{f}_k\right)^T$ and $p = (p_1, p_2, \ldots, p_k)^T$.

The resulting estimates of β_{PL} and $f(\beta_{PL})$ are asymptotically consistent (Roberts *et al*, 1987).

Under the simple random sample design the covariance matrix of $\hat{\beta}_{PL}$ is estimated by

$$\hat{V}_{SRS}\left(\hat{\beta}_{PL}\right) = \left(X^T WAWX\right)^{-1}$$

where A=diag $\{(\hat{f}_i(1-\hat{f}_i))/n_i\}$.

This estimate is not consistent in the case of clustered designs. A more complicated covariance estimator that is consistent under the complex design is obtained using the linearization method. It is

$$\hat{V}_{CL}\left(\hat{\beta}_{PL}\right) = \hat{V}_{SRS}\left(\hat{\beta}_{PL}\right) X^T W \hat{V}_{CL} WX \hat{V}_{SRS}\left(\hat{\beta}_{PL}\right)$$

where $\hat{V}_{CL}$ is the survey estimate of the covariance matrix of p.

Having determine the variance of $\hat{\beta}_{PL}$ one is able to obtain confidence intervals for odds ratio estimates, exp($\hat{\beta}_{PL}$). An approximate $(1-\alpha)100\%$ confidence interval for the odds ratio for the sth explanatory variable, with parameter estimate $\hat{\beta}_s$ can be calculated using

$$\exp\left(\hat{\beta}_s \pm z_{\alpha/2} S\hat{E}\left(\hat{\beta}_s\right)\right)$$

where $z_{\alpha/2}$ is the $(1-\alpha)100\%$ two-sided critical value of the standard normal distribution.

The design effect of the parameter estimates can be derived using

$$\text{DEFF}(\hat{\beta}_{PL}) = \frac{\hat{V}_{CL}(\hat{\beta}_{PL})}{\hat{V}_{SRS}(\hat{\beta}_{PL})} = X^T W \hat{V}_{CL} W X \hat{V}_{SRS}(\hat{\beta}_{PL})$$

Kish and Frankel (1974) have found by empirical observation that design effects for regression coefficients tend to be smaller than that for means.

3.4 Illustration of cluster sampling analysis - A rectal bleeding example

Rectal bleeding is an important and may be the only symptom of colorectal cancer, the most common internal malignancy in Australia (Jelfs *et al*, 1994). Because rectal bleeding occurs, before diagnosis, in over two thirds of patients with rectal cancer, and the prevalence of rectal bleeding in patients attending general practitioners (GPs) in Australia was unknown, Sladden *et al* (1999) conducted a study of rectal bleeding. The prevalence, health seeking behaviour and management of rectal bleeding in general practice patients in Northern Tasmania, Australia, was investigated. The data for this study was provided by Dr Mike Sladden from the University of Tasmania, Australia. The integrity of patient confidentiality has been honoured.

3.4.1 Data collection

Twenty GPs who worked five or more sessions a week were randomly selected from a Divisional list of 89 Northern Tasmanian GPs. In this case the GP is the

primary sampling unit. Each GP was supplied with 50 self-administered patient questionnaires that included questions on socio-demographic variables, frequency of GP attendance, looking for signs of bleeding and history of bleeding. Those with a history of bleeding were questioned in more detail about bleeding. For the most recent bleed they were asked whether they sought advice, what the bleeding implied for them and whether further investigations had been performed. Possible examinations were a rectal examination (PR), sigmoidoscopy or colonoscopy (scope - telescope examination of the bowel) or barium enema (BA enema - x-ray of the bowel). All consenting patients over 50 years of age completed the questionnaires while waiting to see the GP. Data collection at each GP ceased once the 50 questionnaires had been completed or after 3 weeks, whichever occurred first. The total number of questionnaires obtained from each GP ranged from 21 to 50 (median = 50). A total of 903 consenting patients were recruited over the study period. This data will be used to illustrate the effects of sampling weights and clustering on estimation of means, proportions, regression coefficients, and their variances.

3.4.2 Calculation of weights

In order to determine the probability weights appropriate for the design-based analysis of the data set described above, one needs to know the size of each general practice included in the study. This information was not available for the study described above and therefore, for illustrative purposes only, each practice was assigned a size. Using a random number table and assuming that the size of practices ranged from 400 to 1000 patients, a size was generated for each practice. These are provided in Table 3.1.

GP (i)	Sample size (n_i)	Practice size (N_i)	GP (i)	Sample size (n_i)	Practice size (N_i)
1	50	581	11	50	864
2	50	534	12	50	673
3	50	891	13	50	453
4	50	621	14	23	998
5	50	909	15	50	546
6	50	498	16	50	782
7	50	883	17	45	999
8	50	415	18	22	563
9	50	571	19	50	704
10	45	822	20	18	812

Table 3.1 GP size assigned using a random number table and sample size chosen from each of the 20 GPs

It is quite simple to calculate the weights bearing in mind that the weights associated with the selected subject is equal to the inverse of the probability of being sampled. The probability of selection of each individual is equal to the product of the probability of selecting

(a) the general practice (out of a total of 89)

This is given by $p_{GP} = {}^{k}/_{K} = {}^{20}/_{89}$.

and

(b) the patient in the selected general practice

For practice i this is $p_{Pi} = {}^{n_i}/_{N_i}$, i=1,...,20.

So, the resulting probability of selecting an individual from one of these clusters was

$$p_i = p_{GP} \times p_{Pi} = {}^{20n_i}/_{89N_i}.$$

Thus, it follows that the statistical weight for the observations in the ith GP is given by

$$w_i = 1/p_i = 89N_i/20n_i .$$

The resulting weights for each observation that belongs to a specific GP is given in Table 3.2. Therefore the estimated population size is

$$\hat{N} = \sum_{i=1}^{k} n_i w_i = 62830 .$$

GP (i)	Weight (w_i)	GP (i)	Weight (w_i)
1	51.709	11	76.896
2	47.526	12	59.897
3	79.299	13	40.317
4	55.269	14	193.091
5	80.901	15	48.594
6	44.322	16	69.598
7	78.587	17	98.790
8	36.935	18	113.880
9	50.819	19	62.656
10	81.287	20	200.744

Table 3.2 Weights for rectal bleeding example

The statistical package, STATA, is equipped with complex survey sampling commands and was used to analyze the data.

3.4.3 Descriptive analysis

(1) Analysis assuming independence (ignoring probability weighting and clustering)

First, consider the estimation of means and proportions for a number of variables included in the analysis. The sample was not stratified and therefore only weighting and clustering need be considered. Table 3.3 provides estimates of the means and proportions with standard errors for selected variables assuming independence of patients within GPs and ignoring weights. These estimates are incorrect and standard errors are underestimated.

(2) Analysis assuming independence and incorporating weights

If one incorporates the correct probability weighting scheme but ignores clustering in the analysis, the estimates of means and proportions, and their variances change slightly. In fact, variances have increased for all variables. Point estimates are now unbiased but standard errors are still incorrect. See Table 3.4. One is able to determine the inflation in the variance, due to weighting, by examining the design effects. There is quite a large increase in the variance for the variable: Sought advice elsewhere. It is approximately 75% larger when weighting is considered compared to when it is ignored.

Variable	Simple Random Sample Design			
	Estimate (mean/ proportion)	Standard error	95% confidence	interval
Age	66.261	0.308	65.657	66.866
Males	0.438	0.017	0.405	0.470
Patient report of number of visits to GP	6.353	0.102	6.154	6.553
Patients seeing doctor about rectal bleeding	0.058	0.013	0.033	0.082
Ever looks at paper for signs of bleeding	0.765	0.014	0.738	0.793
Ever looks at pans for signs of bleeding	0.702	0.015	0.672	0.732
Have ever had rectal bleeding	0.331	0.016	0.300	0.362
Age at first bleed	48.165	0.991	46.212	50.117
Age at last bleed	58.699	0.699	57.322	60.075
FOR THE MOST RECENT BLEED				
Single pattern of bleeding (as opposed to multiple)	0.705	0.029	0.648	0.761
Paper bleeding (as opposed to pan)	0.740	0.028	0.686	0.794
Pain	0.284	0.028	0.228	0.339
Similar to previous episodes of bleeding	0.619	0.028	0.563	0.674
First ever bleed	0.147	0.021	0.107	0.188
Sought GP advice	0.454	0.030	0.395	0.514
Sought family advice	0.087	0.016	0.055	0.119
Sought advice elsewhere	0.033	0.010	0.013	0.054
Had a PR	0.365	0.028	0.310	0.419
Had a BA Enema	0.130	0.020	0.092	0.169
Had a scope	0.211	0.024	0.164	0.257

Table 3.3 Estimates, standard errors and 95% confidence intervals assuming independence and ignoring weighting

Variable	Weighted Sample Design				
	Estimate (mean/ proportion)	Standard error	95% confidence	interval	Design Effect
Age	66.137	0.335	65.479	66.795	1.197
Males	0.433	0.018	0.398	0.469	1.228
Patient report of number of visits to GP	6.402	0.112	6.183	6.621	1.209
Patients seeing doctor about rectal bleeding	0.056	0.014	0.028	0.084	1.303
Ever looks at paper for signs of bleeding	0.758	0.016	0.726	0.789	1.242
Ever looks at pans for signs of bleeding	0.695	0.017	0.662	0.729	1.241
Have ever had rectal bleeding	0.332	0.017	0.298	0.367	1.233
Age at first bleed	47.603	1.042	45.550	49.656	1.151
Age at last bleed	58.217	0.744	56.751	59.683	1.173
FOR THE MOST RECENT BLEED					
Single pattern of bleeding (as opposed to multiple)	0.707	0.032	0.644	0.770	1.248
Paper bleeding (as opposed to pan)	0.727	0.032	0.664	0.791	1.333
Pain	0.296	0.032	0.232	0.360	1.310
Similar to previous episodes of bleeding	0.621	0.031	0.560	0.682	1.219
First ever bleed	0.137	0.021	0.094	0.179	1.167
Sought GP advice	0.465	0.034	0.398	0.531	1.247
Sought family advice	0.089	0.018	0.053	0.124	1.183
Sought advice elsewhere	0.042	0.015	0.012	0.072	1.747
Had a PR	0.362	0.031	0.301	0.423	1.233
Had a BA Enema	0.140	0.023	0.095	0.186	1.339
Had a scope	0.219	0.027	0.165	0.273	1.313

Table 3.4 Estimates, standard errors and 95% confidence intervals assuming independence but taking weighting into account

(3) Analysis taking clustered design into account but ignoring weighting

If weights are ignored and only clustering is taken into account, then the estimates of means and proportions are the same as those assuming independence. The effect of clustering is clearly seen in the difference between standard error estimates under the different sampling designs. See Table 3.5. The standard error produced under the cluster design is, in most cases, larger

than that under the simple random sampling design. A consequence of the inflated variances is wider confidence intervals. The inflation in the variance that is due to the clustering of individuals within GPs is estimated by the design effect.

Variable	Cluster Design				
	Estimate (mean/ proportion)	Standard error	95% confidence	interval	Design Effect
Age	66.261	0.569	65.071	67.451	3.405
Males	0.438	0.021	0.393	0.483	1.664
Patient report of number of visits to GP	6.353	0.194	5.948	6.759	3.633
Patients seeing doctor about rectal bleeding	0.058	0.013	0.031	0.085	1.053
Ever looks at paper for signs of bleeding	0.765	0.024	0.716	0.815	2.818
Ever looks at pans for signs of bleeding	0.702	0.028	0.644	0.760	3.253
Have ever had rectal bleeding	0.331	0.022	0.285	0.377	1.994
Age at first bleed	48.165	0.942	46.194	50.136	0.902
Age at last bleed	58.699	0.780	57.065	60.332	1.245
FOR THE MOST RECENT BLEED					
Single pattern of bleeding (as opposed to multiple)	0.705	0.030	0.642	0.767	1.094
Paper bleeding (as opposed to pan)	0.740	0.025	0.688	0.792	0.805
Pain	0.284	0.030	0.221	0.346	1.137
Similar to previous episodes of bleeding	0.619	0.021	0.574	0.664	0.578
First ever bleed	0.147	0.021	0.104	0.190	0.998
Sought GP advice	0.454	0.031	0.389	0.520	1.070
Sought family advice	0.087	0.023	0.040	0.134	1.923
Sought advice elsewhere	0.033	0.011	0.011	0.055	1.018
Had a PR	0.365	0.031	0.300	0.429	1.232
Had a BA Enema	0.130	0.018	0.093	0.167	0.819
Had a scope	0.211	0.024	0.161	0.260	0.991

Table 3.5 Estimates, standard errors, 95% confidence intervals and design effects produced by taking cluster design into account but ignoring weights

(4) Correct analysis incorporating both cluster design and weighting

The results for the correct analysis compared to the incorrect analysis ignoring design and weights, is given in Table 3.6. Here both the weighting and clustering of observations have been accounted for in the correct analysis. As a result one is able to obtain confidence intervals whose true coverage is close to 95%.

The estimated mean age of these patients was approximately 66 years and the estimated average age at first bleed was 47.6 years. The percentage of males in the study was 43.3% and the percentage of females 56.7%. Approximately 33% of patients had experienced rectal bleeding before. About 6% of the patients were seeing the GP with a rectal bleeding problem.

As far as signs of rectal bleeding were concerned, patients examined both paper and pan for signs of bleeding. A larger proportion of patients examined paper (75.8%) compared to pan (69.5%).

Patients were also questioned about their most recent bleed. Many said that it was similar to previous episodes of bleeding (62%). Most experienced a single pattern of bleeding (70.7%) as opposed to a multiple pattern. Paper bleeding occurred more frequently amongst patients (in 73% of cases) than pan bleeding. Approximately 30% of these patients experienced some degree of pain with the bleed.

Quite a few patients consulted their GP (46.5%) when faced with a bleeding problem. About 9% and 4% sought advice from a family member or elsewhere respectively. For the most recent bleed, 36.2% had a rectal examination, 21.9% had a telescope examination of the bowel and 14% had an x-ray of the bowel.

Variable	Simple Random Sample Design				Weighted Cluster Design				
	Estimate (mean/ proportion)	Standard error	95% confidence	interval	Estimate (mean/ proportion)	Standard error	95% confidence	interval	Design Effect
Age	66.261	0.308	65.657	66.866	66.137	0.516	65.058	67.216	2.833
Males	0.438	0.017	0.405	0.470	0.433	0.026	0.378	0.489	2.563
Patient report of number of visits to GP	6.353	0.102	6.154	6.553	6.402	0.202	5.979	6.825	3.965
Patients seeing doctor about rectal bleeding	0.058	0.013	0.033	0.082	0.056	0.013	0.028	0.084	1.149
Ever looks at paper for signs of bleeding	0.765	0.014	0.738	0.793	0.758	0.030	0.695	0.820	4.412
Ever looks at pans for signs of bleeding	0.702	0.015	0.672	0.732	0.695	0.033	0.626	0.765	4.666
Have ever had rectal bleeding	0.331	0.016	0.300	0.362	0.332	0.024	0.283	0.382	2.263
Age at first bleed	48.165	0.996	46.212	50.117	47.603	0.862	45.799	49.407	0.787
Age at last bleed	58.699	0.703	57.322	60.075	58.217	0.638	56.881	59.553	0.862
FOR THE MOST RECENT BLEED									
Single pattern of bleeding (as opposed to multiple)	0.705	0.029	0.648	0.761	0.707	0.028	0.648	0.766	0.979
Paper bleeding (as opposed to pan)	0.740	0.028	0.686	0.794	0.727	0.035	0.655	0.800	1.539
Pain	0.284	0.029	0.228	0.339	0.296	0.041	0.211	0.381	2.075
Similar to previous episodes of bleeding	0.619	0.029	0.563	0.674	0.621	0.035	0.548	0.695	1.562
First ever bleed	0.147	0.020	0.107	0.188	0.137	0.022	0.091	0.182	1.201
Sought GP advice	0.454	0.030	0.395	0.514	0.465	0.032	0.398	0.532	1.114
Sought family advice	0.087	0.016	0.055	0.119	0.089	0.024	0.039	0.138	2.047
Sought advice elsewhere	0.033	0.010	0.013	0.054	0.042	0.014	0.013	0.070	1.375
Had a PR	0.365	0.028	0.310	0.419	0.362	0.030	0.299	0.425	1.162
Had a BA Enema	0.130	0.019	0.092	0.169	0.140	0.019	0.101	0.180	0.867
Had a scope	0.211	0.024	0.164	0.257	0.219	0.025	0.167	0.271	1.077

Table 3.6 Estimates, standard errors and 95% confidence intervals under simple random sampling and the correct weighted cluster design

3.4.4 Summary of descriptive statistics results

A summary of estimates, standard errors and design effects for each of the four analyses is given in Table 3.7. The first value in each cell corresponds to the mean/proportion estimate, the second to the standard error and the third to the design effect.

By examining the estimates obtained under independence and then under the correct weighted cluster design, we see that the standard errors are generally larger under the correct design in all but four cases: age at first bleed, age at last bleed, singular pattern of bleeding and had a BA enema. This is also reflected in the design effects that are less than 1 for each of these variables. For all of the cases the design effects are only slightly below 1 indicating that the variances under cluster design and independence are approximately the same. The slight decrease in variances is caused by small but slightly negative intracluster correlations.
The design effect is as large as 4.666 for the variable: Ever looks at pan for signs of bleeding. This means that the variance is over 4 ½ times larger under cluster sampling compared to simple random sampling.

Overall, the probability weighting causes a slight change in both the estimate and the standard error of the estimate. In all cases the standard errors have increased. Taking clustering only into account has the effect of increasing standard errors quite considerably but mean and proportion point estimates remain unchanged. Both the estimates and standard errors have changed under the correct design compared to that under independence.

Variable		Independent Design	Weighted Design	Clustered Design	Weighted Clustered Design
Age	Estimate	66.261	66.137	66.261	66.137
	Standard error	0.308	0.018	0.569	0.516
	Design effect		1.197	3.405	2.833
Males		0.438	0.433	0.438	0.433
		0.017	0.018	0.021	0.026
			1.228	1.664	2.563
Patient report of number of visits to GP		6.353	6.402	6.353	6.402
		0.102	0.112	0.194	0.202
			1.209	3.633	3.965
Patients seeing doctor about rectal bleeding		0.058	0.056	0.058	0.056
		0.013	0.014	0.013	0.013
			1.303	1.053	1.149
Ever looks at paper for signs of bleeding		0.765	0.758	0.765	0.758
		0.014	0.016	0.024	0.030
			1.242	2.818	4.412
Ever looks at pans for signs of bleeding		0.702	0.695	0.702	0.695
		0.015	0.017	0.028	0.033
			1.241	3.253	4.666
Have ever had rectal bleeding		0.331	0.332	0.331	0.332
		0.016	0.017	0.022	0.024
			1.233	1.994	2.263
Age at first bleed		48.165	47.603	48.165	47.603
		0.991	1.042	0.942	0.862
			1.151	0.902	0.787
Age at last bleed		58.699	58.217	58.699	58.217
		0.699	0.744	0.780	0.638
			1.173	1.245	0.862

Table 3.7 Comparisons of estimates, standard errors and design effects obtained using different methods of analysis

(Remaining variables continued overleaf)

Variable	Independent Design	Weighted Design	Clustered Design	Weighted Clustered Design
FOR THE MOST RECENT BLEED				
Single pattern of bleeding (as opposed to multiple)	0.705 0.029	0.707 0.032 1.248	0.705 0.030 1.094	0.707 0.028 0.979
Paper bleeding (as opposed to pan)	0.740 0.028	0.727 0.032 1.333	0.740 0.025 0.805	0.727 0.035 1.539
Pain	0.284 0.028	0.296 0.032 1.310	0.284 0.030 1.137	0.296 0.041 2.075
Similar to previous episodes of bleeding	0.619 0.028	0.621 0.031 1.219	0.619 0.021 0.578	0.621 0.035 1.562
First ever bleed	0.147 0.021	0.137 0.021 1.167	0.147 0.021 0.998	0.137 0.022 1.201
Sought GP advice	0.454 0.030	0.465 0.034 1.247	0.454 0.031 1.070	0.465 0.032 1.114
Sought family advice	0.087 0.016	0.089 0.018 1.183	0.087 0.023 1.923	0.089 0.024 2.047
Sought advice elsewhere	0.033 0.010	0.042 0.015 1.747	0.033 0.011 1.018	0.042 0.014 1.375
Had a PR	0.365 0.028	0.362 0.031 1.233	0.365 0.031 1.232	0.362 0.030 1.162
Had a BA Enema	0.130 0.020	0.140 0.023 1.339	0.130 0.018 0.819	0.140 0.019 0.867
Had a scope	0.211 0.024	0.219 0.027 1.313	0.211 0.024 0.991	0.219 0.025 1.077

Table 3.7 (continued) Comparisons of estimates, standard errors and design effects obtained using different methods of analysis

3.4.5 Logistic regression

(1) Logistic regression ignoring cluster design and weighting

In order to determine whether any of the variables jointly affect whether patients with rectal bleeding consulted a GP or not, a logistic regression was performed. In order to make valid inferences it is necessary, as with mean and proportion estimation, to take weighting and clustering into account. If these are ignored the logistic regression results assuming independence of patients within GPs are as given in Table 3.8. We are able to obtain easily interpretable odds ratios for each of the variables by simply calculating the exponent of each regression coefficient. These odds ratios are provided along with confidence intervals in Table 3.9. The variables which affect whether a patient consults a GP are possibly age (p=0.067), pattern of bleeding; paper or pan (p<0.001) and singular or multiple (p=0.047), family advice (p=0.001) and whether they had experienced a similar bleed before (p<0.001).

Sought GP advice	Independent Logistic Regression					
	Coefficient	Standard error	z	p-value	95% Confidence	Interval
Constant	0.209	1.167	0.179	0.858	-2.079	2.497
Age	0.032	0.017	1.834	0.067	-0.002	0.066
Paper bleeding pattern	-1.595	0.367	-4.347	<0.001	-2.314	-0.876
Single bleeding pattern	-0.681	0.343	-1.988	0.047	-1.352	-0.010
Family advice	2.043	0.623	3.278	0.001	0.822	3.265
Similar bleed	-1.417	0.336	-4.215	<0.001	-2.076	-0.758

Table 3.8 Coefficients, standard errors, p-values and 95% confidence intervals produced by independent logistic regression ignoring weighting

Sought GP advice	Independent Logistic Regression			
	Odds Ratio	Standard error	95% Confidence	Interval
Age	1.033	0.018	0.998	1.069
Paper bleeding pattern	0.203	0.074	0.099	0.417
Single bleeding pattern	0.506	0.173	0.259	0.991
Family advice	7.715	4.808	2.274	26.171
Similar bleed	0.242	0.082	0.125	0.469

Table 3.9 Odds ratios, standard errors and 95% confidence intervals produced by independent logistic regression ignoring weighting

(2) Logistic regression taking weighting into account

If the proper selection weighting was incorporated the results would be as seen in Table 3.10 and Table 3.11. Both coefficients and odds ratio estimates have changed. Examination of the design effects (Table 3.11) indicate that the standard errors for all variables have increased as all design effects are greater than 1. Single bleeding pattern exhibits the largest design effect indicating an increased variance that is almost 1 ½ times greater under the weighted design compared to the independence design.

Sought GP advice	Weighted Independent Logistic Regression					
	Coefficient	Standard error	*z*	*p*-value	95% Confidence	Interval
Constant	0.090	1.308	0.069	0.945	-2.486	2.667
Age	0.034	0.020	1.704	0.090	-0.005	0.073
Paper bleeding pattern	-1.711	0.417	-4.100	<0.001	-2.534	-0.889
Single bleeding pattern	-0.494	0.408	-1.210	0.227	-1.299	0.310
Family advice	1.901	0.641	2.967	0.003	0.639	3.163
Similar bleed	-1.347	0.369	-3.646	<0.001	-2.074	-0.619

Table 3.10 Coefficients, standard errors, *p*-values and 95% confidence intervals produced by independent logistic regression but accounting for weighting

Sought GP advice	Weighted Independent Logistic Regression				
	Odds Ratio	Standard error	95% Confidence	Interval	Design effect
Age	1.034	0.020	0.995	1.075	1.276
Paper bleeding pattern	0.181	0.075	0.079	0.411	1.295
Single bleeding pattern	0.610	0.249	0.273	1.364	1.422
Family advice	6.692	4.288	1.894	23.643	1.057
Similar bleed	0.260	0.096	0.126	0.538	1.207

Table 3.11 Odds ratios, standard errors and 95% confidence intervals produced by independent logistic regression but accounting for weighting

(3) Logistic regression accounting for clustering only

Accounting for clustering only produces the following results.

Sought GP advice	Clustered Logistic Regression					
	Coefficient	Standard error	*z*	*p*-value	95% Confidence	Interval
Constant	0.209	1.049	0.199	0.844	-1.987	2.406
Age	0.032	0.016	1.966	0.064	-0.002	0.066
Paper bleeding pattern	-1.595	0.397	-4.017	0.001	-2.426	-0.764
Single bleeding pattern	-0.681	0.432	-1.575	0.132	-1.586	0.224
Family advice	2.043	0.677	3.016	0.007	0.625	3.461
Similar bleed	-1.417	0.347	-4.088	0.001	-2.143	-0.691

Table 3.12 Coefficients, standard errors, *p*-values and 95% confidence intervals produced by logistic regression that takes cluster design into account

If we compare this set of estimates with those obtained assuming independence we see that coefficient and odds ratio estimates (Table 3.13) have remained the same but standard errors have increased for all variables except age. As a consequence single

bleeding pattern is no longer significantly related to GP consultation. However, the results are still erroneous. Weighting needs to be taken into account along with clustering. All design effects, except that for age, is greater than 1. This indicates that clustering generally increases the variance of regression estimates.

Sought GP advice	Clustered Logistic Regression				
	Odds Ratio	Standard error	95% Confidence	Interval	Design effect
Age	1.033	0.017	0.998	1.068	0.844
Paper bleeding pattern	0.203	0.081	0.088	0.466	1.175
Single bleeding pattern	0.506	0.219	0.205	1.251	1.523
Family advice	7.715	5.226	1.869	31.845	1.036
Similar bleed	0.242	0.084	0.117	0.501	1.008

Table 3.13 Odds ratios, standard errors, 95% confidence intervals and design effects produced by logistic regression that takes cluster design into account

(4) Logistic regression incorporating correct weighting and consideration of cluster design

The results obtained assuming independence are erroneous. The true sampling design is incorporated in the analysis by specifying that each GP is a primary sampling unit with patients within the GPs possibly being more alike than patients between GPs. Performing the logistic regression again, taking the clustering into account and incorporating the proper weighting scheme, we obtain the correct results in Table 3.14. As a consequence of increased standard errors we find that confidence intervals are wider, except for the case of age where the standard error decreased.

Design effects are not too large. Those for age, family advice and similar bleed are quite close to 1, meaning that the inflation in the variance under cluster sampling is

not very much larger than that under independence. A value of 1.467 for the paper bleeding design effect indicates that the variance for this regression coefficient is close to one and a half times larger under cluster sampling.

Note that under the correct design we question whether age is related to GP consultation and single bleeding pattern clearly needs to be omitted from the model. We are now able to make valid interpretations based on the correct method of analysis.

The final model estimates for this set of data is given in Table 3.16. The final model is given by

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = 1.921 - 1.864 \text{ (Paper bleeding pattern)} + 1.972 \text{ (Family advice)}$$
$$- 1.282 \text{ (Similar bleed)}$$

Unlike the model under independence, age and single vs multiple bleeding pattern are no longer significantly related to GP consultation. The three variables included in the final model were paper bleeding pattern, family advice and similar bleed. Odds ratios are provided in Table 3.17. A paper bleed or a bleed similar to a previous occasions had a negative effect on a patients decision to consult a GP or not. The odds of a patient who had a paper bleed visiting their GP were 84% less compared to those who observed pan bleeding. The odd of those who had had a similar bleed previously seeing their GP were 72% less compared to the odds of someone who did not have a bleed that was similar to a previous occasion. Family advice also played a major role in GP consultation. The odds of those who received advice from a family member consulting their GP was 7 times more than those who did not receive family advice.

Sought GP advice	Independent Logistic Regression						Weighted Cluster Logistic Regression					
	Coefficient	Standard error	*z*	*p*-value	95% Confidence	Interval	Coefficient	Standard error	*z*	*p*-value	95% Confidence	Interval
Constant	0.209	1.167	0.179	0.858	-2.079	2.497	0.090	1.206	0.075	0.941	-2.433	2.613
Age	0.032	0.017	1.834	0.067	-0.002	0.066	0.034	0.018	1.850	0.080	-0.004	0.072
Paper bleeding pattern	-1.595	0.367	-4.347	<0.001	-2.314	-0.876	-1.711	0.436	-3.921	0.001	-2.625	-0.798
Single bleeding pattern	-0.681	0.343	-1.988	0.047	-1.352	-0.010	-0.494	0.525	-0.941	0.359	-1.593	0.605
Family advice	2.043	0.623	3.278	0.001	0.822	3.265	1.901	0.688	2.764	0.012	0.461	3.341
Similar bleed	-1.417	0.336	-4.215	<0.001	-2.076	-0.758	-1.347	0.361	-3.733	0.001	-2.102	-0.592

Table 3.14 Coefficients, standard errors, *p*-values and 95% confidence intervals produced by both independent logistic regression and weighted clustered logistic regression

Sought GP advice	Independent Logistic Regression				Weighted Cluster Logistic Regression				
	Odds ratio	Standard error	95% Confidence	Interval	Odds ratio	Standard error	95% Confidence	Interval	Design effect
Age	1.033	0.018	0.998	1.069	1.034	0.019	0.996	1.074	1.087
Paper bleeding pattern	0.203	0.074	0.099	0.417	0.181	0.079	0.072	0.450	1.467
Single bleeding pattern	0.506	0.173	0.259	0.991	0.610	0.320	0.203	1.831	2.096
Family advice	7.715	4.808	2.274	26.171	6.692	4.603	1.586	28.238	1.138
Similar bleed	0.242	0.081	0.125	0.469	0.260	0.094	0.122	0.553	1.067

Table 3.15 Coefficients, standard errors, *p*-values and 95% confidence intervals produced by both independent logistic regression and weighted clustered logistic regression

Sought GP advice	Weighted Clustered Logistic Regression					
	Coefficient	Standard error	z	p-value	95% Confidence	Interval
Constant	1.921	0.359	5.353	<0.001	1.170	2.673
Paper bleeding pattern	-1.864	0.431	-4.327	<0.001	-2.766	-0.963
Family advice	1.972	0.662	2.977	0.008	0.586	3.358
Similar bleed	-1.282	0.343	-3.744	0.001	-1.999	-0.565

Table 3.16 Final model coefficients, standard errors, p-values and 95% confidence intervals produced by logistic regression that takes cluster design into account

Sought GP advice	Weighted Clustered Logistic Regression				
	Odds Ratio	Standard error	95% Confidence	Interval	Design effect
Paper bleeding pattern	0.155	0.067	0.063	0.382	1.453
Family advice	7.185	4.759	1.796	28.738	1.075
Similar bleed	0.277	0.095	0.135	0.568	1.048

Table 3.17 Final model odds ratios, standard errors, 95% confidence intervals and design effects produced by logistic regression that takes cluster design into account

3.4.6 Summary of logistic regression results

A summary of estimates, variances, p-values and design effects for each of the four logistic regression analyses is given in Table 3.18. The first value in each cell corresponds to the mean/proportion estimate, the second to the standard error, the third to the p-value and the fourth to the design effect.

Similar inferences as those for the descriptive analysis can be made. Correct weighting affects both parameter estimates and standard errors while clustering affects standard errors only. We are clearly able to see the effect that the correct weighting and clustering procedure has on the results. The two variables, age and single bleeding pattern, are no longer significantly related to whether an individual consults a GP or not.

Variable		Independent Design	Weighted Design	Clustered Design	Weighted Clustered Design
Age	Parameter estimate,	0.032	0.034	0.032	0.034
	Standard error,	0.017	0.020	0.016	0.018
	p-value,	0.067	0.090	0.064	0.080
	Design effect		1.276	0.844	1.087
Paper bleeding pattern		-1.595	-1.711	-1.595	-1.711
		0.367	0.417	0.397	0.436
		<0.001	<0.001	0.001	0.001
			1.295	1.175	1.467
Single bleeding pattern		-0.681	-0.494	-0.681	-0.494
		0.343	0.408	0.432	0.525
		0.047	0.227	0.132	0.359
			1.422	1.523	2.096
Family advice		2.043	1.901	2.043	1.901
		0.623	0.641	0.677	0.688
		0.001	0.003	0.007	0.012
			1.057	1.036	1.138
Similar bleed		-1.417	-1.347	-1.417	-1.347
		0.336	0.369	0.347	0.361
		<0.001	<0.001	0.001	0.001
			1.207	1.008	1.067

Table 3.18 Comparisons of logistic regression estimates, standard errors, *p*-values and design effects obtained using different methods of analysis

The largest design effect under the correctly weighted and clustered design is for single bleeding pattern where inflation in the variance of the regression estimate is

over two times larger in the weighted cluster design compared to that under independence. However, this variable needs to be excluded from the model as it is no longer significantly related to GP consultation. This is an effect of the increased variance. All the design effects for the remaining coefficients are not very much larger than 1 indicating not much of an increase in the variance. Note that this example supports the empirical findings by Kish and Frankel (1974). The design effects for regression coefficients are generally smaller than the design effects for means and proportions (see section 3.4.4 for design effects for proportions).

CHAPTER 4
POPULATION-AVERAGED MODELS

CHAPTER 4
POPULATION-AVERAGED MODELS

4.1 Introduction

One of the approaches to handling clustered binary data involves not specifying a joint distribution for the responses but considering the marginal model where a model is constructed for the marginal expectation of the responses for each observation within a cluster. Marginal models are a class of statistical models that allows one to model the regression relationship between a response and a number of covariates in the presence of clustering. Using this approach, marginal effects are averaged over all clusters and therefore this is also referred to as the population-averaged approach. The marginal response is modelled as a function of the covariates without explicitly accounting for subject-to-subject heterogeneity.

The marginal expectation for individual j in cluster i, $\mu_{ij} = E(y_{ij})$, is the focus. We assume that the marginal expectation is related to a $r \times 1$ covariate vector x_{ij}^T by

$$g(\mu_{ij}) = x_{ij}^T \beta$$

for some known link function $g(\cdot)$, and the marginal variance is a function of the marginal mean

$$\text{var}(y_{ij}) = \upsilon(\mu_{ij})\varphi$$

where υ is a known function and φ is the over-dispersion parameter which accounts for the variance of y_{ij} not explained by $\upsilon(\mu_{ij})$.

The marginal regression coefficients have the same interpretation as coefficients from a cross-sectional analysis i.e. they have population-averaged interpretations. Here the β parameter describes the way in which the average response in the population changes with a change in the explanatory variable.

We discuss the marginal approach known as generalized estimating equations (GEE) introduced by Zeger and Liang (1986). This approach can be used when the response has a distribution in the exponential family.

4.2 The generalized estimating equations (GEE) method

This marginal approach to analyzing correlated binary data is an extension of generalized linear models (McCullagh and Nelder, 1983). The non-likelihood approach arose in the context of longitudinal studies and allows one to estimate regression parameters without specifying the entire likelihood. The method is a semi-parametric one in that it only requires specification of the first two moments of the marginal distribution of the repeated outcome i.e. the mean and variance of the responses.

The generalized estimating equation approach is closely related to quasi-likelihood proposed by Wedderburn (1974) and was adapted by Zeger and Liang (1986). Quasi-likelihood is a method that only requires specification of the relationships between the mean and covariates and between the mean and the variance. So unlike a likelihood approach where the actual form of the distribution is specified, only the mean-covariance structure need be specified. Quasi-likelihood then involves solving score equations that are likelihood-type functions. The generalized estimating equations extension to the generalized linear model takes the correlation inherent in longitudinal studies into account. Statistical methods for this sort of clustered data when responses are approximately Gaussian are well developed eg. Laird and Ware (1982). The

analysis of repeated measurements complicated by correlation is made even more complex when responses are non-Gaussian. Until recently less attention had been focused on the analysis of non-Gaussian response data.

Liang and Zeger (1986) derive the GEEs by firstly assuming that the marginal distribution of the responses can be put in the form of a generalized linear model. They then propose a so-called working model under the working assumption that observations within a cluster are independent of each other. This independence working model is then extended to explicitly account for correlation, resulting in the GEEs. The estimating equations produce consistent and asymptotically Gaussian regression parameter estimates, as well as consistent estimates of variances, under the mild assumptions concerning the actual correlation structure within clusters. An added feature of the GEE approach is that the regression parameter estimates remain consistent and asymptotically normal even if the correlation structure is misspecified (Liang and Zeger, 1986). It is therefore a good approach to use when the main interest lies in the estimation of regression coefficients and the correlation is seen as nuisance.

4.2.1 Independence estimating equations

Liang and Zeger (1986) fit the generalized linear model to clustered data. When dealing with repeated measures data we consider data y_{ij}, the jth measurement on the ith subject, with the subject being viewed as a cluster. The marginal density of y_{ij} can be written in the form

$$f(y_{ij}) = \exp\left[\{y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(y_{ij}, \phi)\right]$$

where

$$\theta_{ij} = g(\eta_{ij}),$$

$\eta_{ij} = x_{ij}^T \beta$ and

a and b are functions of known form.

The function η_{ij} is the linear predictor and β is a $r \times 1$ vector of parameters that measures the effects of the covariates, $x_{ij}^T = (x_{1ij}, x_{2ij}, ..., x_{rij})$. The link function $g(\cdot)$ is a monotonically increasing and differentiable function linking the linear predictor to the parameters of the generalized linear model. The mean value of y_{ij} is $E(y_{ij}) = \mu_{ij} = a'(\theta_{ij})$ and the variance is $\text{var}(y_{ij}) = a''(\theta_{ij})$. The main objective is to estimate β using information on the mean and variance.

In the case of binary data and using the logistic regression model we make use of the logit link function,

$$\log it(\pi_{ij}) = x_{ij}^T \beta$$

and

$$\theta_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right),$$

$$a(\theta_{ij}) = \log(1-\pi_{ij}),$$

$$b(y_{ij}) = \log\binom{1}{y_{ij}}$$

and $\phi = 1$.

Then

$E(y_{ij}) = \mu_{ij} = \pi_{ij}$ and $\text{var}(y_{ij}) = \pi_{ij}(1-\pi_{ij})$.

If we assume that individuals within and between clusters are independent then we are able to derive maximum likelihood estimates of β_I from the score equations (McCullagh and Nelder, 1983). These equations can be written as the sum over all clusters of a matrix product with three factors:

$$\sum_{i=1}^{k} D_i^T A_i^{-1} S_i = 0 \qquad (4.1)$$

where

(i) D is a vector of partial derivatives, $D_i = \frac{\partial \mu_i}{\partial \beta_I} = \frac{\partial \pi_i}{\partial \beta_I}$ for binary data, $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{in_i})$.

(ii) $A_i = diag\{\text{var}(y_{ij})\}$ is a $n_i \times n_i$ diagonal matrix representing the variance of y_{ij}. Since we are assuming that individuals within a cluster are independent the (j,j)th element is equal to $\pi_{ij}(1-\pi_{ij})$.

(iii) $S_i = y_i - \mu_i$ is a $n_i \times 1$ residual vector for the ith cluster. In the dichotomous case $S_i = y_i - \pi_i$ where $y_i = (y_{i1}, \ldots, y_{in_i})^T$ is the $n_i \times 1$ vector of binary outcome values (0,1) for the ith cluster.

The parameter estimator under independence, $\hat{\beta}_I$, is the maximum likelihood solution of the equations (4.1) given above. The variance of $\hat{\beta}_I$ can be consistently estimated by

$$\hat{\text{var}}(\hat{\beta}_I) = \left(\sum_{i=1}^{k} D_i^T A_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{k} D_i^T A_i^{-1} S_i S_i^T A_i^{-1} D_i\right) \left(\sum_{i=1}^{k} D_i^T A_i^{-1} D_i\right)^{-1}. \qquad (4.2)$$

Both estimates of β_I and its variance are consistent if the regression model for $E(y_{ij})$ is correctly specified (Liang and Zeger, 1986). If the correlation between observations is quite large $\hat{\beta}_I$ may not have high efficiency. Liang and Zeger (1986) therefore propose generalized estimating equations that lead to estimators with higher efficiency by accounting for the correlation structure.

4.2.2 Generalized estimating equations

The set of estimating equations given by (4.1) can be extended to account for clusters of dependent data. To improve efficiency of the estimation process a working correlation matrix to account for dependence is introduced. Let $R(\alpha)$ be a $n_i \times n_i$ symmetric working correlation matrix for the y_i's, fully specified by a $t \times 1$ vector of unknown parameters, α. Assume that $R(\alpha)$ is the same for all clusters. $R(\alpha)$ is known as a working correlation matrix because it is often the case that it is incorrectly specified. However, we would like to obtain consistent estimators that have consistent variance estimates even when $R(\alpha)$ is incorrect. By generalizing the quasi-likelihood approach to account for correlations between subunits or repeated observations, we can define the working covariance matrix for cluster *i*

$$V_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi \qquad (4.3)$$

This is $\text{cov}(y_i)$ if $R(\alpha)$ is the true correlation matrix. This approach is quite important because often we are not able to specify the correct correlation structure.

Then the GEE estimate of β_{GEE} is based on the k clusters and is essentially an extension of the quasi-likelihood equations. For given α it is obtained by replacing the A_i matrix in (4.1) by V_i and obtaining the solution $\hat{\beta}_{GEE}$ to the general estimating equations

$$\sum_{i=1}^{k} D_i^T V_i^{-1} S_i = 0 .$$

The estimating equations can be written as

$$\sum_{i=1}^{k} U_i(\alpha, \beta_{GEE}) = 0$$

where $U_i(\alpha, \beta_{GEE}) = D_i^T V_i^{-1} S_i$.

These are similar to the quasi-likelihood equations but with V_i being a function of both the correlation and regression parameters, α and β_{GEE}, whereas with quasi-likelihood equations V_i is simply a function of β. Note that if $R(\alpha)$ is the identity matrix then the GEEs are identical to the independence equations for binomial data.

Liang and Zeger (1986) compute $\hat{\beta}_{GEE}$ by iterating between solving for the regression coefficients, β_{GEE}, using a modified version of Fisher scoring, and moment estimation in solving for the correlation and scale parameters, α and ϕ. Having determined an estimate of β_{GEE}, we calculate standardized residuals $\hat{r}_{ij} = (y_{ij} - \hat{\mu}_{ij}) / \left[\hat{\text{var}}(y_{ij})\right]^{1/2}$. This in turn is used to re-estimate α and ϕ. The above two steps are reiterated until convergence occurs.

For the following discussions we will refer to β_{GEE} as β.

4.2.3 Estimation of α and β

The estimating equations depend on both α and β but can be expressed in terms of β only. This is achieved by calculating a $k^{1/2}$-consistent estimator of α, $\hat{\alpha}(\beta,\phi)$. This correlation parameter is estimated by making use of current Pearson residuals at a given iteration

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{[\hat{\text{var}}(y_{ij})]^{1/2}}$$

$$= \frac{y_{ij} - \hat{\pi}_{ij}}{\sqrt{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}} \text{ for binary data.}$$

We borrow strength over the k clusters to obtain a consistent estimate of α. This estimate depends on the correlation structure $R(\alpha)$. A general form for the estimate of α is

$$\hat{\alpha}_{uv} = \sum_{i=1}^{k} \frac{\hat{r}_{iu}\hat{r}_{iv}}{N-r}$$

where $N = \sum_{i=1}^{k} n_i$.

The resulting estimate is asymptotically efficient as those obtained if α were known (Diggle *et al*, 1994). Specific choices of $R(\alpha)$ along with the estimates of α will be discussed.

Then ϕ in $\hat{\alpha}(\beta,\phi)$ is replaced by its $k^{1/2}$-consistent estimator $\hat{\phi}(\beta)$. In order to obtain this estimate we once again make use of the Pearson residuals,

$$\hat{\phi}^{-1} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\hat{r}_{ij}^2}{N-r}$$

Then the estimate of β for a given $R(\alpha)$, is the solution to

$$\sum_{i=1}^{k} U_i\left[\hat{\alpha}\left\{\beta, \hat{\phi}(\beta)\right\}, \beta\right] = 0 \ .$$

Once estimates for the correlation parameter and scale parameter are obtained, the following iterative procedure, a modification of Fisher's scoring method, can be used to calculate β:

$$\beta^{(m+1)} = \beta^{(m)} + \left\{\sum_{i=1}^{k} D_i^T\left(\beta^{(m)}\right) \tilde{V}_i^{-1}\left(\beta^{(m)}\right) D_i^T\left(\beta^{(m)}\right)\right\}^{-1} \left\{\sum_{i=1}^{k} D_i^T\left(\beta^{(m)}\right) \tilde{V}_i^{-1}\left(\beta^{(m)}\right) S_i\left(\beta^{(m)}\right)\right\}$$

where

$$\tilde{V}_i(\beta) = V_i\left[\hat{\alpha}\left\{\beta, \hat{\phi}(\beta)\right\}, \beta\right]$$

Here β describes how the average population response rather than a specific individual's response depends on the covariates. If we define $D = \left(D_1^T, ..., D_k^T\right)^T$ and $S = \left(S_1^T, ..., S_k^T\right)$, and let $\tilde{V}$ be a $nk \times nk$ block diagonal matrix with $\tilde{V}_i$s as diagonal elements we can define a modified dependent variable

$$Z = D\beta - S \ .$$

Then the iterative procedure described above is equivalent to performing an iteratively reweighted linear regression of Z on D with weight $\tilde{V}^{-1}$ (Liang and Zeger, 1986).

4.3 Properties of $\hat{\beta}$

The GEE equations are designed to produce consistent regression coefficients when the link function has been correctly specified and even under minimal assumptions about the dependence between subjects in a cluster (Zeger and Liang, 1986). The estimate $\hat{\beta}$ is a consistent estimate of β if the relationship between μ_i and β is correctly specified (Shoukri and Pause, 1999). Because $D_i^T V_i^{-1}$ does not depend on y, the estimating equations converge to 0 and have consistent roots provided that $E(y_i - \pi_i) = 0$ (Zeger and Liang, 1986).

If the working correlation matrix is approximately correct then the asymptotic efficiency of $\hat{\beta}$ is expected to be close to unity (Prentice, 1988). However, if one has incomplete follow-up data or if there is a high correlation between measurements then efficiency drops (Stukel, 1993). The estimate $\hat{\beta}$ for estimated α, is nearly as efficient as the maximum likelihood estimates of β, provided that V_i has been reasonably approximated (Liang and Zeger, 1986). In fact, Fitzmaurice *et al* (1993) point out that GEE is the maximum likelihood score equation in the case of both multivariate normal and binary data, provided that V_i is correctly specified.

We examine the consistency results as presented by Liang and Zeger (1986). They show that under mild regularity conditions, as the number of clusters becomes very large i.e. as $k \to \infty$, $\hat{\beta}$ is a consistent estimator of β, and that $\sqrt{k}(\hat{\beta} - \beta)$ is asymptotically Gaussian with a mean value of 0 and robust variance estimate (the sandwich estimator):

$$\hat{V}_{GEE} = \lim_{k \to \infty} k \left[\sum_{i=1}^{k} D_i^T V_i^{-1} D_i \right]^{-1} \left[\sum_{i=1}^{k} D_i^T V_i^{-1} \operatorname{cov}(y_i) V_i^{-1} D_i \right] \left[\sum_{i=1}^{k} D_i^T V_i^{-1} D_i \right]^{-1}$$

$$= \lim_{k\to\infty} k(V_1^{-1}V_0V_1^{-1})$$

where

$$V_0 = \sum_{i=1}^{k} D_i^T V_i^{-1} \operatorname{cov}(y_i) V_i^{-1} D_i \text{ and } V_1 = \sum_{i=1}^{k} D_i^T V_i^{-1} D_i \ .$$

The covariance $\operatorname{cov}(y_i)$ is the actual and not the assumed covariance of y_i. This covariance matrix is consistent when the mean and marginal variance is correctly specified, even when $\operatorname{cov}(y_i) \neq V_i$. An estimate of $\operatorname{cov}(y_i)$ is $\hat{\operatorname{cov}}(y_i) = (y_i - \hat{\mu}_i)(y_i - \hat{\mu})^T$. Note that this asymptotic covariance matrix of $\hat{\beta}$ does not depend on the choice of α and ϕ.

Pendergast *et al* (1996) also provide a naïve estimate of the variance of $\hat{\beta}$. If the form of the variance given by (4.3) is correct then the variance of $\hat{\beta}$ is simply $\left(\sum_{i=1}^{k} X_i^T \hat{V}_i^{-1} X_i\right)^{-1}$ where $\hat{V}_i^{-1}$ depends on the estimates $\hat{\beta}$, $\hat{\alpha}$ and $\hat{\phi}$. If the variance function is misspecified then the sandwich estimate is the better estimate.

4.4 Advantages and disadvantages of GEE

A major advantage of the GEE approach is that the working correlation matrix does not have to be correctly specified in order to derive the consistent and asymptotically Gaussian estimate $\hat{\beta}$ and the consistent $\hat{V}_{GEE}$. The only requirement is that α and ϕ be estimated consistently (Zeger and Liang, 1986). Hence confidence intervals for β and other statistical methods are asymptotically valid even when $R(\alpha)$ is misspecified. This is particularly useful when the main interest lies in modelling the mean using an

approximation of the covariance. However, choosing $R(\alpha)$ closer to the true correlation increases efficiency in the estimation process. Diggle *et al* (1994) suggest a method of checking the robustness of inferences concerning β. He recommends fitting a final model using different covariance structures and then comparing the resulting parameter estimates and their standard errors. A larger difference in these estimates implies that the covariance model needs to be reconsidered. Another advantage of the GEE method is that it can be extended so that clusters do not have to share the same correlation matrix i.e. $R(\alpha)$ can vary between clusters. A further strength of this procedure is that it can be easily extended to adjust for both cluster-level and individual-level covariates (Donner and Klar, 1994).

This approach does have limitations as well. A major disadvantage is that the estimating equations have no probability distribution and hence no likelihood function can be constructed. As a consequence Lindsey (1993) points out that interpretation of the model as a representation of a physical mechanism that could have produced the data is destroyed. In addition, because models have no likelihood function or deviance, comparison of models are difficult. Another criticism of this approach is that there may be no individual in the population with the characteristics as described by the population-averaged model (Lindsey and Lambert, 1998). It may indicate that a treatment is superior on average when it might be poorer for a specific individual. Both Donner and Klar (1994) and Lindsey and Lambert (1998) advise using this method for the analysis of observational studies rather than experimental studies.

4.5 Correlation structures

Kenward and Smith (1995) suggest a variety of ways in which the correlation matrix $R(\alpha)$ could be chosen. It could be fixed, based on previous analyses or it could be calculated using the data itself. Below are various correlation structures that could be used.

(1) Independence structure

This is the simplest correlation structure. Here we are making the assumption that each individual in a cluster is uncorrelated with any other individual in that cluster. Therefore $corr(y_{ij}, y_{ij'}) = 1$ when $j = j'$ and is zero otherwise.

$$R(\alpha) = I = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Therefore the resulting GEEs are the independence estimating equations

(2) Uniform or exchangeable structure

We specifically consider the case when $t = 1$ and assume that $corr(y_{ij}, y_{ij'}) = \alpha$ for all $j \neq j'$. Every observation within a cluster is equally correlated with every other observation in that cluster. This is known as the exchangeable correlation structure obtained from a random effects model with a random level for each subject (Laird and Ware, 1982). We have

$$R(\alpha) = \begin{bmatrix} 1 & & & \\ \alpha & & & \\ \cdots & \cdots & \cdots & \cdots \\ \alpha & \alpha & \ldots & 1 \end{bmatrix}$$

Then, given ϕ, an estimate of the correlation parameter is

$$\hat{\alpha} = \phi \sum_{i=1}^{k} \sum_{j>j'}^{n_i} \hat{r}_{ij} \hat{r}_{ij'} \Big/ \sum_{i=1}^{k} \tfrac{1}{2} n_i (n_i - 1) - r$$

where r is the number of regression parameters.

An estimator of ϕ is not necessary for calculating $\hat{\beta}$ and $\hat{V}_{GEE}$.

(3) Autoregressive structure

The autoregressive correlation structure indicates that two observations taken close in time (or space) have a tendency to be more highly correlated than two observations that are further apart. As an example consider the correlation structure of a first order autoregressive process. Here $corr(y_{ij}, y_{ij'}) = \alpha^{|j-j'|}$. The correlation matrix is given by

$$R(\alpha) = \begin{bmatrix} 1 & & & & & \\ \alpha & 1 & & & & \\ \alpha^2 & \alpha & 1 & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ \alpha^{n_i-1} & \cdots & \cdots & \cdots & \alpha & 1 \end{bmatrix}$$

(4) *m*-dependence structure

Let $\alpha = (\alpha_1, \ldots, \alpha_{n_i-1})^T$ with $\alpha_j = corr(Y_{ij}, Y_{i,j+1})$, $j = 1,2,\ldots,n_i - 1$.

An estimator of α_j, given β and ϕ, is

$$\hat{\alpha}_j = \phi \sum_{i=1}^{k} \frac{\hat{r}_{ij}\hat{r}_{i,j+1}}{k-r}.$$

We have a one-dependent model if $R(\alpha)$ is tridiagonal with $R_{j,j+1} = \alpha_j$:

$$R(\alpha) = \begin{bmatrix} 1 & & & & \\ \alpha_1 & 1 & & & \\ \alpha_2 & \alpha_1 & 1 & & \\ 0 & \alpha_2 & \alpha_1 & 1 & \\ 0 & 0 & \alpha_2 & \alpha_1 & 1 \end{bmatrix}.$$

We do not need an estimator of ϕ to calculate $\hat{\beta}$ and $\hat{V}_{GEE}$.

In the special case when t =1 and $\alpha_j = \alpha$, $j = 1,2,...,n_i$, α can be estimated by

$$\hat{\alpha} = \sum_{j=1}^{n_i-1} \hat{\alpha}_j / n_i - 1.$$

This model can be extended to *m*-dependence.

(5) Unstructured correlation

In this case no assumption is made about the correlation between a pair of observations. Therefore $corr(y_{ij}, y_{ij'}) = 1$ when $j = j'$ and takes on any value between –1 and +1 for values of $j \neq j'$.

(6) User fixed structure

Here the correlation coefficients are not estimated using the data but are fixed by the user prior to the data analysis. So $corr(y_{ij}, y_{ij'}) = 1$ when $j = j'$ and takes on any value between –1 and +1 for values of $j \neq j'$ (this value being fixed before analysis).

4.6 Design effects of GEE parameter estimates

Scott and Holt (1982) present a method for deriving design effects for linear regression estimators. Neuhaus and Segal (1993) provide an extension to this method and produce design effects for regression coefficients when the response is binary. We first examine some results on the effect of cluster designs on linear regression analysis using the analytic approach developed by Scott and Holt (1982).

Scott and Holt (1982) examine the effect of intracluster correlation on linear regression. They assume that observations from the same cluster are correlated and that the covariance matrix of Y has the form

$$\text{cov}(Y) = \sigma^2 V$$

with V an exchangeable block diagonal matrix given by, $V = \bigoplus_{i=1}^{k} V_i$. V_i is an $n_i \times n_i$ matrix for the ith cluster and given by

$$V_i = \begin{bmatrix} 1 & \rho & .... & \rho \\ \rho & 1 & .... & \rho \\ .... & & & .... \\ \rho & \rho & .... & 1 \end{bmatrix}.$$

Ordinary least squares estimates for the linear regression model is given by

$$\hat{\beta}_{OLS} = \left(X^T X\right)^{-1} X^T Y.$$

This is an unbiased estimate of

$$\hat{\beta}_{GLS} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y,$$

the generalized least squares estimate with weights imposed by the variance structure and given by V^{-1}. The ordinary least squares estimate is generally used even though there is a slight loss in efficiency.

The variance of $\hat{\beta}_{OLS}$ under the cluster design is

$$\text{var}_{CL}\left(\hat{\beta}_{OLS}\right) = \sigma^2 \left(X^T X\right)^{-1} \left(X^T V X\right) \left(X^T X\right)^{-1} \qquad (4.4)$$

while the covariance matrix of $\hat{\beta}_{OLS}$ under the assumption of independence is

$$\text{var}_I\left(\hat{\beta}_{OLS}\right) = \sigma^2 \left(X^T X\right)^{-1}.$$

Therefore we can write $\text{var}_{CL}\left(\hat{\beta}_{OLS}\right) = \text{var}_I\left(\hat{\beta}_{OLS}\right) D$

where $D = \dfrac{\text{var}_{CL}\left(\hat{\beta}_{OLS}\right)}{\text{var}_I\left(\hat{\beta}_{OLS}\right)} = \left(X^T V X\right)\left(X^T X\right)^{-1}$.

By comparing $\text{var}_{CL}\left(\hat{\beta}_{OLS}\right)$ to $\text{var}_I(\hat{\beta}_{OLS})$ we see that the effect of intracluster correlation is to increase the variance of the regression estimates obtained under the assumption of independence by a factor of D. The diagonal elements of D are simply the design effects of the regression coefficients. So D is the inflation factor that accounts for clustered design by correcting standard variance results and taking the correlation into account. In the special case when all individuals within clusters are equally correlated, the intraclass correlation being ρ,

$$D = I + (\tilde{N} - 1)\rho$$

where

$$\tilde{N} = \left(\sum_{i=1}^{k} n_i X_{Bi}^T X_{Bi} \right) \left(X^T X\right)^{-1}$$

with X_{Bi} representing the $n_i \times r$ matrix with every element in the sth column (s=1,2,...,r) equal to the average value of the sth covariate over the ith cluster (Scott and Holt, 1982). Note than if $\rho = 0$ then $D = I$ and we're dealing with standard logistic regression.

One can see that under compound symmetry D has a similar form to the well-known design effect for a sample mean or proportion, $\{1 + (\bar{n} - 1)\rho\}$ (see Chapter 2). Scott and Holt (1982) point out that D might better be termed a misspecification effect which is conditional on the observed X and represents the error in the variance and covariance estimates due to incorrectly omitting the intracluster correlation from the model.

An extension to determine the effect of clustered design on GEE parameter estimates - Design effects for the GEE model

Neuhaus and Segal (1993) extend the above results for continuous data and look at design effects for binary data under the cluster design and working independence assumption. In section 4.2.1 the asymptotic covariance matrix of $\hat{\beta}$ under the cluster design and working independence assumption was derived. This was given by (4.2)

$$\hat{\text{var}}_{CL}(\hat{\beta}) = \left(D^T A^{-1} D\right)^{-1} \left(D^T A^{-1} S S^T A^{-1} D\right) \left(D^T A^{-1} D\right)^{-1}$$

where D is a vector of partial derivatives, $D_i = \frac{\partial \mu_i}{\partial \beta} = \frac{\partial \pi_i}{\partial \beta}$ for binary data, $\pi_i = \left(\pi_{i1}, \pi_{i2}, \ldots, \pi_{in_i}\right)$ and $A_i = diag\{\text{var}(y_{ij})\}$ is a $n_i \times n_i$ diagonal matrix representing the variance of y_{ij}.

Now $D^T = X^T \Delta A$ with $\Delta = diag\left(\frac{\partial \theta_{ij}}{\partial \eta_{ij}}\right)$, and $SS^T = \text{var}(Y)$, and hence Neuhaus and Segal (1993) express the variance in terms of X, A and V. The variance of $\hat{\beta}$ can be written as

$$\hat{\text{var}}_{CL}\left(\hat{\beta}\right) = \left(X^T \Delta A \Delta X\right)^{-1}\left(X^T \Delta \,\text{var}(Y) \Delta X\right)\left(X^T \Delta A \Delta X\right)^{-1}.$$

The first matrix product in $\hat{\text{var}}_{CL}\left(\hat{\beta}\right)$, $\left(X^T \Delta A \Delta X\right)^{-1}$, is in fact $\hat{\text{var}}_I\left(\hat{\beta}\right)$, the covariance matrix of the parameter estimates under independence. The clustering effect is incorporated in the covariance matrix by taking the true correlation structure of the responses, var(Y), into account in the centre term.

By setting

$\tilde{X} = A^{1/2} \Delta X$ and

$\tilde{V} = A^{-1/2}\,\text{var}(Y) A^{-1/2}$,

Neuhaus and Segal (1993) express $\hat{\text{var}}_{CL}\left(\hat{\beta}\right)$ in a form similar to (4.4)

$$\hat{\text{var}}_{CL}\left(\hat{\beta}\right) = \left(\tilde{X}^T \tilde{X}\right)^{-1}\left(\tilde{X}^T \tilde{V} \tilde{X}\right)\left(\tilde{X}^T \tilde{X}\right)^{-1} = \hat{\text{var}}_I\left(\hat{\beta}\right)\tilde{D}$$

where

$\tilde{D} = \left(\tilde{X}^T \tilde{V} \tilde{X}\right)\left(\tilde{X}^T \tilde{X}\right)^{-1}$ which is the design effect. D consists of the design effects for the regression parameters.

It is difficult to draw conclusions with respect to design effects of generalized linear model regression coefficients as one must evaluate $\tilde{D}$ under different structures on var(Y). Neuhaus and Segal (1993) turn to approximations and look specifically at the case of a compound symmetric correlation structure imposed on var(Y), for binary data.

Having determined that the variance of regression coefficients under a cluster sample is some factor given by $\tilde{D}$ multiplied by the variance under independence, we conclude that the effect of clustering on regression coefficients is similar to its effects on means and proportions. It has a tendency to increase the variances of the covariate effects. Kish and Frankel (1974) point out that design effects tend to be larger for means and proportions than for regression parameters.

We've specifically looked at the derivation of design effects of regression coefficients under the working independence assumption. One is also able to derive design effects under different correlation structures.

4.7 Illustration of GEE analysis - A rectal bleeding example

We revert back to the rectal bleeding example considered before in Chapter 3. GEE estimation was used to estimate the regression coefficients for variables that affected GP consultation. GEE estimation was performed, using STATA, first assuming an independence correlation structure (this is equivalent to logistic regression) and then an exchangeable correlation. The coefficient estimates for each of the correlation structures, their standard errors, confidence intervals and p-values can be seen in Table 4.1 and Table 4.2.

The two approaches result in very similar estimates for the parameters and little difference between the estimated standard errors. Extensive simulations have shown that these two methods may produce very similar results when the cluster sizes are equal and small (McDonald, 1993 and Lipsitz *et al*, 1990).

Sought GP advice	Independent Correlation Structure					
	Coefficient	Standard error	*z*	*p*-value	95% Confidence	Interval
Constant	0.209	1.167	0.179	0.858	-2.079	2.497
Age	0.032	0.017	1.834	0.067	-0.002	0.066
Paper bleeding pattern	-1.595	0.367	-4.347	0.000	-2.314	-0.876
Single bleeding pattern	-0.681	0.343	-1.988	0.047	-1.352	-0.010
Family advice	2.043	0.623	3.278	0.001	0.822	3.265
Similar bleed	-1.417	0.336	-4.215	<0.001	-2.076	-0.758

Table 4.1 GEE coefficients, standard errors, *p*-values and 95% confidence intervals assuming an independent correlation structure

Sought GP advice	Exchangeable Correlation Structure					
	Coefficient	Standard error	*z*	*p*-value	95% Confidence	Interval
Constant	0.261	1.158	0.225	0.822	-2.009	2.530
Age	0.031	0.017	1.803	0.071	-0.003	0.065
Paper bleeding pattern	-1.591	0.367	-4.333	0.000	-2.310	-0.871
Single bleeding pattern	-0.701	0.343	-2.048	0.041	-1.373	-0.030
Family advice	2.134	0.630	3.389	0.001	0.900	3.368
Similar bleed	-1.434	0.338	-4.242	<0.001	-2.096	-0.771

Table 4.2 GEE coefficients, standard errors, *p*-values and 95% confidence intervals assuming an exchangeable correlation structure

Odds ratios along with standard errors and confidence intervals are provided for each of the variables included in the GEE analysis with an exchangeable correlation

structure (Table 4.3). Recall that parameter estimates have a population-averaged interpretation. So an odds ratio of 8.446 for family advice indicates that the odds of those who seek family advice seeing a GP was 8 times more than the odds of those who did not obtain family advice. Other regression coefficients have similar population-averaged interpretations. Table 4.3 also provides design effects for each of the GEE parameter estimates. Examination of the design effects reveal that there is virtually no difference between standard errors produced under independence and those obtained assuming an exchangeable correlation structure.

Sought GP advice	**Exchangeable**	**Correlation**	**Structure**		
	Odds ratio	Standard error	95% Confidence	Interval	Design Effect
Age	1.032	0.018	0.997	1.067	0.980
Paper bleeding pattern	0.204	0.075	0.099	0.418	1.001
Single bleeding pattern	0.496	0.170	0.253	0.970	1.000
Family advice	8.446	5.317	2.459	29.007	1.020
Similar bleed	0.238	0.081	0.123	0.462	1.011

Table 4.3 Odds ratios, standard errors and 95% confidence intervals assuming an exchangeable correlation structure

Table 4.4 presents the results obtained when employing the GEE approach with exchangeable correlation and robust standard errors. The robust variance estimator provides variance estimates and confidence intervals for the problematic case of a misspecified model. This alternative produces valid standard errors even if correlations are different from that hypothesized.

Sought GP advice	Exchangeable Correlation Structure					
	Coefficient	Robust Standard error	z	p-value	95% Confidence	Interval
Constant	0.261	1.059	0.246	0.806	-1.814	2.335
Age	0.031	0.016	1.903	0.057	-0.001	0.063
Paper bleeding pattern	-1.591	0.390	-4.075	<0.001	-2.356	-0.826
Single bleeding pattern	-0.701	0.438	-1.600	0.110	-1.561	0.158
Family advice	2.134	0.683	3.123	0.002	0.795	3.473
Similar bleed	-1.434	0.344	-4.169	<0.001	-2.108	-0.760

Table 4.4 GEE coefficients, robust standard errors, p-values and 95% confidence intervals assuming an exchangeable correlation structure

Coefficient estimates under the robust analysis remain the same. However, all standard errors except that for age increases considerably. In fact, this has an important impact on inference. The increase in standard errors has increased the p-value for single bleeding pattern which is now no longer significantly related to GP consultation (p=0.110). The increase in the standard errors is also reflected in the design effects. There is an increase in the design effects for paper bleeding pattern, single bleeding pattern and family advice. This increase in design effects indicates a slight inflation in the variances for the variables: paper bleeding pattern, family advice and similar bleed. There is a fairly large increase in the variance for single bleeding pattern. Under the cluster design it is 64% larger than under independence of observations. The final GEE model (excluding single bleeding pattern) is analogous to the final model obtained using correctly weighted survey logistic regression in Chapter 3 even though coefficient and variance estimates are slightly different.

Sought GP advice	Exchangeable	Correlation	Structure		
	Odds ratio	Robust Standard error	95% Confidence	Interval	Design Effect
Age	1.032	0.017	0.999	1.065	0.880
Paper bleeding pattern	0.204	0.080	0.095	0.438	1.132
Single bleeding pattern	0.496	0.217	0.210	1.171	1.638
Family advice	8.446	5.771	2.213	32.230	1.202
Similar bleed	0.238	0.082	0.121	0.468	1.011

Table 4.5 Odds ratios, robust standard errors and 95% confidence intervals assuming an exchangeable correlation structure

CHAPTER 5
CLUSTER-SPECIFIC MODELS

5.1 Introduction

5.2 Logistic linear mixed-effects model

5.3 Estimation and interpretation

5.4 Comparison of population-averaged and cluster-specific models

5.5 Advantages and disadvantages of mixed-effect models

5.6 LPA example

- 5.6.1 Standard logistic regression
- 5.6.2 Generalized estimating equations
- 5.6.3. Random effects model
- 5.6.4 Comparison of results

CHAPTER 5
CLUSTER-SPECIFIC MODELS

5.1 Introduction

Cluster-specific models differ from population-averaged models in that parameters that are specific to the cluster are included in the model. Then given a cluster's regression coefficients, the responses are assumed to be independent observations on a generalized linear model. This approach involves analyzing clustered data by explicitly modelling heterogeneity across clusters in the regression parameters. Examples of the cluster-specific (CS) approach, specifically for binomial data, are the mixed-effects logistic model eg. Stiratelli, Laird and Ware (1984), Anderson and Aitkin (1985) and Gilmour, Anderson and Rae (1985), and the conditional likelihood approach eg. Breslow and Day (1980).

Cluster-specific models are appropriate for situations in which covariates are obtained at both the individual and cluster level. They are especially useful in determining the effects of cluster-varying covariates. These cluster-varying covariates are covariates that may take on different values for every unit in the cluster, either by design or due to chance. The cluster-specific model includes the cluster-varying parameters that describe the correlation structure within the cluster. Hence the model for each cluster is allowed to differ.

Cluster-specific models will be examined in the context of longitudinal studies. Consider a longitudinal study in which each subject is measured with respect to some response. Each individual has a set of covariates. The covariates may be classified as a cluster-level (between-cluster) covariate that is fixed within a cluster or a cluster-varying (within-cluster) covariate that may vary within a cluster. Let x_{ij}^T denote the $r \times 1$ vector of covariates for the jth response of the

*i*th subject. Inference could be difficult because the number of parameters in the model grows as the number of clusters increase. To combat this, a popular approach to modelling involves viewing the cluster-specific parameters as a random sample from some underlying distribution.

In order to make use of the cluster-specific model we make the following assumptions:

(i) the conditional distribution of y_{ij} (the *j*th response in the *i*th cluster) given a vector of parameters specific to the *i*th cluster, α_i, satisfies a generalized linear model with

$$g\left[E\left(y_{ij}|\alpha_i\right)\right]=\alpha_i+x_{ij}^T\beta \qquad (5.1)$$

where

$g(\cdot)$ is the known link function.

(ii) $y_{i1}, y_{i2},\ldots, y_{in_i}$ are conditionally independent given α_i

(iii) α_i follows some distribution $f(a)$. Typically α_i is set to follow a multivariate Gaussian distribution with mean zero and covariance matrix Σ.

Interest focuses on estimation and interpretation of the parameter, β.

Betensky *et al* (2001) extends this model by separating the effects of cluster-level and cluster-varying covariates. This is achieved by dividing the covariates into two components, $x_{(f)}$ which includes covariates fixed at the cluster level, and $x_{(v)}$ which includes covariates which vary within a cluster. Then model (5.1) can be written as

$$g\left[E\left(y_{ij}|\alpha_i\right)\right]=\alpha_i+x_{(f)i}^T\beta^{(f)}+x_{(v)ij}^T\beta^{(v)},$$

where β has been separated into a cluster-level component, $\beta^{(f)}$ and a cluster-varying component, $\beta^{(v)}$, for the jth response in the ith cluster. This is then a special case of a multilevel model.

We examine an example of the cluster-specific approach for binary data, the mixed-effects logistic model. This is also known as the random effects model, hierarchical or multilevel model. The mixed model assumes that dependence between units within a cluster arise because regression coefficients vary across clusters. Therefore the resulting regression model is one that includes both fixed and random terms, the random effects drawn from some underlying distribution.

Another example of the cluster-specific approach involves the use of the conditional likelihood model. Here random effects are eliminated by computing the probability of a cluster response conditional on the cluster sum which is a sufficient statistic for α_i. The conditional approach only uses data from clusters with discordant outcomes while the mixed-effects approach makes use of both discordant and concordant cluster information. Thus, the conditional likelihood approach could be less efficient than the mixed-effects method (Neuhaus and Lesperance, 1996). We will only be looking at the mixed-effects method and the conditional approach will not be considered further.

5.2 Logistic linear mixed-effects model

Let y_{ij} denote binary responses for the jth observation for individual i, $i=1,2,\ldots,k$; $j=1,2,\ldots,n_i$. Also let $\pi_{ij}=\Pr(y_{ij}=1)$ and $g(\mu_{ij})=\log it(\pi_{ij})$.

The mixed-effects logistic model is simply an extension of the standard logistic model but with random effects terms, α_i, being allowed to vary between

clusters according to either a parametric or semiparametric mixing distribution with density $f(a)$. Parametric mixture models assume that $f(a)$ belongs to a specified family of distributions, often $N(\mu,\sigma^2)$ eg. Stiratelli, Laird and Ware (1984). Semiparametric mixture models involve jointly obtaining regression estimates and the non-parametric mixing distribution $f(a)$ eg. Lindsay and Lesperance (1995). Both parametric and semiparametric approaches use information from all clusters. Furthermore, mixed-effects models postulate that the α_i are independent and identically distributed (Neuhaus and Lesperance, 1996).

For the mixed model, within the *i*th cluster the responses y_{ij} are independent and follow a generalized linear model with parameters that can vary between clusters (Neuhaus and Kalbfleisch, 1998). Therefore given α_i, the general linear logistic mixed-effects model is of the form

$$\log it \Pr(y_{ij}=1|\alpha_i) = \alpha_i + x_{(f)ij}^T \beta^{(f)} + x_{(v)ij}^T \beta^{(v)} \qquad (5.2)$$

where $\beta^{(f)}$ and $\beta^{(v)}$ are the fixed and cluster-varying components respectively. The random effects α_i are assumed to be *iid* from a multivariate $N(0,\Sigma)$ distribution.

5.3 Estimation and interpretation

Estimation of the fixed parameters and the variances and covariances of the random effects terms can be obtained by employing the method of maximum likelihood using a Newton-Raphson or an EM algorithm.

Let π_{ij} be the probability of a response for the *j*th individual in cluster *i*. The likelihood function is given by

$$L(\beta,\alpha_i)=\prod_{i=1}^{k}\prod_{j=1}^{n_i}\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{(1-y_{ij})}$$

$$=\prod_{i=1}^{k}l_i$$

where

$$l_i=\prod_{j=1}^{n_i}\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{(1-y_{ij})}.$$

In order to obtain the likelihood under the maximum likelihood logistic regression model we integrate the likelihood function with respect to the distribution of the random effects. Once the random effects are integrated out the resulting marginal likelihood function depends on the fixed effect parameters and the parameters of the random effects distribution, specifically the parameters of the random effects covariance matrix Σ. The maximum likelihood of these parameters, β and Σ, are obtained by maximizing the marginal likelihood of the data

$$L(\beta,\Sigma)=\prod_{i=1}^{k}\int l_i\ df(a).$$

There are two problems associated with the direct maximization of the marginal likelihood. Firstly, in most cases this integral does not have a closed form so exact maximum likelihood estimates cannot be found. Either numerical or Monte Carlo integration methods must be used to calculate the likelihood. In the case of normally distributed random effects the marginal likelihood function is

$$L(\beta,\Sigma)=\prod_{i-1}^{k}\int l_i\frac{\exp\left(-\alpha_i^2/2\right)}{\sqrt{2\pi}}d\alpha_i. \qquad (5.3)$$

Then one way of numerical integration involves the use of the Gauss-Hermite formula (Collett, 1991). This involves approximating the integral by a sum i.e.

$$\int f(u)e^{-u^2}du \approx \sum_{j=1}^{m} c_j f(s_j)$$

where the values of c_j and s_j are given in standard tables eg. Abramowitz and Stegman (1972). Then the expression in (5.3) can be expressed as a summation and the values of the parameters that maximize this can be determined numerically.

Stiratelli *et al* (1984) discuss the EM algorithm to optimize an approximate restricted maximum likelihood. Secondly, the maximum likelihood estimate of the parameters of the matrix Σ does not take into account the loss in the degrees of freedom resulting from having estimated fixed effects. As a consequence the maximum likelihood of the variance components is biased in small samples. Therefore the REML (restricted maximum likelihood) procedure is generally recommended (Longford, 1993 and Drum and McCullagh, 1993).

Unlike the generalized estimating equation case where the parameter β_{GEE} refers to the unconditional logits of overall population prevalences, in the mixed-effects model the fixed parameters refer to covariate effects for specific individuals. So the cluster-specific parameter effect $\beta^{(f)}$ measures the change in the conditional logit of the probability of a positive response when $x_{(f)}$ changes by one unit while $x_{(v)}$ remains fixed. This effect is assumed to be constant over all the clusters. In a similar way the cluster-varying parameter effect $\beta^{(v)}$ measures the change in the conditional logit of the probability of a positive response when $x_{(v)}$ changes by one unit and $x_{(f)}$ remains fixed.

5.4 Comparison of population-averaged and cluster-specific models

Unlike the case of correlated Gaussian data the parameters of the cluster-specific and population-averaged models in the binary case describe different types of effects of the covariates. In the case of the population-averaged approach, the regression coefficients describe the average population's response to the changing covariates whereas in the cluster-specific approach the coefficient describes an individual's response (Zeger *et al*, 1988).

Neuhaus *et al* (1991) found that when there is a dependence between units within a cluster, cluster-varying covariate effects produced by the population-averaged model (discussed in Chapter 4) are smaller than those obtained using the cluster-specific model. Consider a single covariate, x_{ij} for the jth response of the ith individual. Let β_{PA} denote the corresponding population-averaged effect of the parameter obtained using a marginal model like the generalized estimating equation approach, and β the cluster-specific effect. The marginal model doesn't specify a unique mixed-effects model but the mixed-effects model does specify a marginal model for y_{ij} i.e.

$$P(y_{ij}=1|x_{ij}) = \int \left(1+e^{-\alpha_i-\beta x_{ij}}\right)^{-1} f(a)d\alpha .$$

To compare population-averaged and cluster-specific effects, Neuhaus *et al* (1991) derive a formula for approximating the population-averaged (PA) effect using the cluster-specific parameter value. The population-averaged effect that expresses a unit increase in the covariate in the log odds scale is defined to be

$$\beta_{PA}(x) = \log\left\{\frac{P(y=1|x+1)/P(y=0|x+1)}{P(y=1|x)/P(y=0|x)}\right\}.$$

In the PA model β_{PA} is independent of x. However, this quantity depends on x in the cluster-specific model and is given by

$$\beta_{PA}(x) = \log\left\{\frac{E\left[\left(1+e^{-\alpha-\beta(x+1)}\right)^{-1}\right]E\left[\left(1+e^{\alpha+\beta x}\right)^{-1}\right]}{E\left[\left(1+e^{\alpha+\beta(x+1)}\right)^{-1}\right]E\left[\left(1+e^{-\alpha-\beta x}\right)^{-1}\right]}\right\}. \qquad (5.4)$$

We can approximate the right-hand side of (5.4) by expanding it in a Taylor series about $\beta=0$. This produces the approximation of β_{PA}

$$\beta_{PA}(x) \approx \beta[1-\rho(0)]$$

where $\rho(0) = corr(y_{ij}, y_{ij'}|\beta=0)$, the intracluster correlation between y when there is no covariate effect. This result holds for both cluster-level and cluster-varying covariates (Betensky *et al*, 2001).

Zeger *et al* (1988) have shown that in particular if $\alpha_i \sim N(0,\lambda)$ then

$$\log it\,\Pr(y_{ij}=1) \approx \frac{x_{ij}^T\beta}{\sqrt{1+0.35\lambda}}$$

and therefore

$$\beta_{GEE} \approx \frac{\beta}{\sqrt{1+0.35\lambda}}.$$

Clearly the GEE estimates and mixed-effects estimates are approximately equal if either $\beta=0$ or $\lambda = \text{var}(b_i) = 0$.

Betensky *et al* (2001) show that in the cluster-specific model the PA effect of a cluster-level covariate (obtained using the cluster-specific model) is

approximately independent of that covariate but not independent of the cluster-varying covariates (obtained using the cluster-specific model). In a similar manner, the PA effect of a cluster-varying covariate (obtained using the cluster-specific model) is approximately independent of that covariate but not of the cluster-level covariate (obtained using the cluster-specific model). On the other hand if $\beta_{CS}^{(f)}$ and $\beta_{CS}^{(v)}$ are close to 0 for the PA model, the PA effect of any covariate is approximately independent of all the covariates (obtained using the cluster-specific model).

Betensky *et al* (2001) discuss tests of cluster-varying covariates. If correlation is not taken into account then tests of cluster-varying covariates using the cluster-specific model are more powerful than those using the population-averaged model. However, when the correlation structure is assumed to be exchangeable, then tests using the population-averaged models are as efficient as cluster-specific tests. For cluster-level covariates, Wald tests for cluster-specific and population-averaged models are equivalent.

5.5 Advantages and disadvantages of mixed-effect models

Laird and Ware (1982) point out a number of desirable features of the random effects model. One of the important advantages of mixed models is that it can be applied in the case of unbalanced data. Unbalanced data arises when repeated measurements are taken at different times for each subject or if each cluster consists of a different total number of measurements. Secondly mixed-effects model allows for the explicit modelling and analysis of both between-cluster and within-cluster variation. It is often the case that these parameters have a natural interpretation relevant to the goals of the study and therefore their estimates may be used for exploratory analysis. Finally, these models assist in the investigation of the background variables on the response.

Betensky *et al* (2001) also discuss further advantages of this method. This model provides estimates of variances and covariances which are often of interest. Mixed models can be extended to use in the multivariate case, to model variance heterogeneity and discrete response data. There is flexibility in the manner in which fixed and random parameters are modelled. Another advantage of this approach is that much more complex models, with multiple levels of clustering, overlapping clusters and random coefficients, are possible compared to the population-averaged approach. Estimation of mixed models can correct for heterogeneity shrinkage.

The two main limitations of this model are that it is computationally intensive and it is also limited by the special form that is assumed for the covariance structure.

5.6 LPA example

This example will be used to illustrate the cluster-specific mixed-effects method in comparison to the population-averaged method of GEE and standard logistic regression. Betensky *et al* (2001) conducted a study to assess the competence of the immune system of subjects infected with the type 1 human immunodeficiency virus (HIV). The lymphocyte proliferation assay (LPA) was performed on 52 subjects, 23 of whom were HIV positive. In the LPA lymphocytes proliferate when stimulated with antigens or mitogens. Results may be expressed in terms of a stimulation index (SI) that can be treated as a dichotomous variable. SI is dichotomized using a threshold value of 5 as proposed by Betensky *et al* (2001). Because not all laboratories are certified to perform the assay, there is a large demand on a few laboratories. Thus interest centres on determining whether handling method influences the SI reading, that is whether blood samples could be shipped or stored overnight as opposed to analyzing, as according to the standard protocol, fresh blood samples. The effect of HIV status (positive or negative), anticoagulants (acid citrate dextrose,

citrate cell preparation tube (CCPT) and heparin) and stimulant (pokeweed, candida, tetanus toxoid and streptokinase) on the LPA results is also important in this study. Thus HIV is a cluster-level covariate with values remaining constant over different blood samples taken from the same individual. Anticoagulants, stimulant and handling method are cluster-varying variables because their values are allowed to vary over blood samples taken from a specific individual. The assay was performed on up to 36 combinations of handling method, anticoagulant and stimulant for each individual in the study. A total of 1201 responses were obtained for the study. The data analyzed can be obtained from

http://www.blackwellpublishers.co.uk/rss/ .

5.6.1 Standard logistic regression

In order to determine which variables influence SI reading under the assumption of independence between measurements from a specific individual, a logistic regression can be performed. A standard logistic regression revealed that HIV status, stimulant and handling method were significantly related to SI reading being either below 5 or equal to or above 5.

SI		Coefficient	Standard error	z	p-value	95% Confidence	Interval
Constant		2.713	0.236	11.479	<0.0001	2.250	3.176
HIV		-1.594	0.160	-9.931	<0.0001	-1.908	-1.279
Antico-agulant	CCPT	-0.191	0.188	-1.015	0.310	-0.560	0.178
	Heparin	0.160	0.180	0.886	0.376	-0.194	0.513
Stimulant	Pokeweed	2.785	0.528	5.270	<0.0001	1.749	3.820
	Tetanus	-1.669	0.196	-8.508	<0.0001	-2.053	-1.284
	Streptokinase	-2.083	0.211	-9.889	<0.0001	-2.495	-1.670
Handling Method	Overnight	-0.249	0.186	-1.339	0.180	-0.613	0.115
	Shipped	-0.842	0.185	-4.540	<0.0001	-1.205	-0.478

Table 5.1 Coefficients, standard errors, p-values, and confidence intervals produced by standard logistic regression

The results can be viewed in terms of odds ratios given in Table 5.2.

SI		Odds Ratio	Standard error	95% Confidence	Interval
HIV		0.203	0.033	0.148	0.278
Antico-agulant	CCPT	0.826	0.156	0.571	1.195
	Heparin	1.173	0.212	0.824	1.671
Stimulant	Pokeweed	16.193	8.556	5.749	45.609
	Tetanus	0.188	0.037	0.128	0.277
	Streptokinase	0.125	0.026	0.082	0.188
Handling Method	Overnight	0.780	0.145	0.542	1.122
	Shipped	0.431	0.080	0.300	0.620

Table 5.2 Odds ratios, standard errors and confidence intervals produced by standard logistic regression

5.6.2 Generalized estimating equations

Using generalized estimating equations with an exchangeable correlation matrix we obtain results that are quite similar to the independence case except in the case of the stimulant pokeweed. The effect of this stimulant has decreased quite drastically compared to that of the independence analysis. Generally both estimates and standard errors have changed.

SI		Coefficient	Standard error	z	p-value	95% Confidence	Interval
Constant		2.817	0.293	9.611	<0.0001	2.252	3.392
HIV		-1.773	0.300	-5.906	<0.0001	-2.361	-1.184
Antico-agulant	CCPT	-0.149	0.174	-0.852	0.394	-0.491	0.193
	Heparin	0.191	0.166	1.151	0.250	-0.134	0.517
Stimulant	Pokeweed	1.963	0.386	5.090	<0.0001	1.207	2.719
	Tetanus	-1.698	0.185	-9.180	<0.0001	-2.060	-1.335
	Streptokinase	-2.279	0.206	-11.081	<0.0001	-2.682	-1.876
Handling Method	Overnight	-0.375	0.170	-2.202	0.028	-0.709	-0.041
	Shipped	-0.746	0.173	-4.306	<0.0001	-1.086	-0.407

Table 5.3 Coefficients, standard errors, p-values, and confidence intervals produced by GEE with exchangeable correlation

SI		Odds Ratio	Standard error	95% Confidence	Interval
HIV		0.170	0.051	0.094	0.306
Antico-agulant	CCPT	0.862	0.150	0.612	1.213
	Heparin	1.210	0.201	0.875	1.677
Stimulant	Pokeweed	7.121	2.749	3.343	15.165
	Tetanus	0.183	0.034	0.127	0.263
	Streptokinase	0.102	0.021	0.068	0.153
Handling Method	Overnight	0.687	0.117	0.492	0.960
	Shipped	0.474	0.082	0.338	0.666

Table 5.4 Odds ratios, standard errors and confidence intervals produced by GEE with exchangeable correlation

The GEE analysis was also performed using the robust standard error calculation. The results for this analysis are given in Table 5.3 and Table 5.4. We see that coefficient estimates are the same as in the usual GEE exchangeable case but the increase in the standard errors is generally quite large. And as a consequence confidence intervals are wider. However, all the variables except anticoagulants are still significantly related to SI reading.

SI		Coefficient	Robust standard error	z	p-value	95% Confidence	Interval
Constant		2.817	0.446	6.317	<0.001	1.943	3.691
HIV		-1.773	0.464	-3.820	<0.001	-2.682	-0.863
Antico-agulant	CCPT	-0.149	0.204	-0.730	0.465	-0.548	0.250
	Heparin	0.191	0.146	1.314	0.189	-0.094	0.476
Stimulant	Pokeweed	1.963	0.398	4.933	<0.001	1.183	2.743
	Tetanus	-1.698	0.356	-6.329	<0.001	-2.985	-1.573
	Streptokinase	-2.279	0.360	-4.771	<0.001	-2.395	-1.000
Handling Method	Overnight	-0.375	0.216	-1.739	0.082	-0.798	0.048
	Shipped	-0.746	0.209	-3.564	<0.001	-1.157	-0.336

Table 5.5 Coefficients, robust standard errors, p-values, and confidence intervals produced by GEE with exchangeable correlation

SI		Odds Ratio	Robust standard error	95% Confidence	Interval
HIV		0.170	0.079	0.068	0.422
Antico-agulant	CCPT	0.862	0.175	0.578	1.285
	Heparin	1.210	0.176	0.910	1.610
Stimulant	Pokeweed	7.121	2.834	3.264	15.535
	Tetanus	0.183	0.037	0.051	0.207
	Streptokinase	0.102	0.065	0.051	0.368
Handling Method	Overnight	0.687	0.148	0.450	1.049
	Shipped	0.474	0.099	0.051	0.715

Table 5.6 Odds ratios, robust standard errors and confidence intervals produced by GEE with exchangeable correlation

We can use the odds ratios to make interpretations with respect to the significant regression coefficients. Recall that interpretations should be made with respect to the population rather than a specific individual.

The effect of HIV was to lower the SI reading. The odds of patients in the population who were HIV positive having an SI reading over 5 was 83% the odds of those patients who were not HIV positive. The stimulant used in the LPA analysis also played a role in determining SI reading. In comparison to the candida stimulant, the pokeweed stimulant had 7 times the odds of producing readings over 5. The remaining two stimulants had a lesser chance of producing high readings compared to candida. The blood samples which were stimulated using tetanus toxoid and streptokinase had a reduced odds of about 90% and 82% respectively, of producing LPA results of over 5, compared to samples stimulated using candida. As far as handling method was concerned, in comparison to fresh samples, those samples held overnight and shipped were less likely to result in SI readings over 5. Overnight samples had close to a third of the odds and shipped samples slightly over half the odds of having high SI readings compared to the fresh blood samples.

5.6.3 Random effects model

SI		Coefficient	Standard error	z	p-value	95% Confidence	Interval
Constant		3.952	0.362	10.932	<0.0001	3.243	4.660
HIV		-1.771	0.317	-5.591	<0.0001	-2.392	-1.150
Antico-agulant	CCPT	-0.177	0.235	-0.755	0.451	-0.637	0.283
	Heparin	0.309	0.225	1.376	0.169	-0.131	0.750
Stimulant	Pokeweed	2.573	0.559	4.606	<0.0001	1.478	3.668
	Tetanus	-2.396	0.252	-9.497	<0.0001	-2.890	-1.901
	Streptokinase	-3.257	0.287	-11.349	<0.0001	-3.820	-2.695
Handling	Overnight	-0.598	0.230	-2.596	0.009	-1.049	-0.146
Method	Shipped	-1.022	0.237	-4.304	<0.0001	-1.487	-0.557
σ_U^2	(Between patients)	2.064	0.217				

Table 5.7 Coefficients, standard errors, p-values, and confidence intervals produced by mixed-effects logistic regression

We are also able to analyze the effect of the variables on SI readings using a random effects model. We set the random effect to be the individual and assume that the random effects are normally distributed. Performing a regression we obtain the results given in Table 5.7 above.

SI		Odds Ratio	Standard error	95% Confidence	Interval
HIV		0.170	0.054	0.091	0.317
Antico-agulant	CCPT	0.838	0.197	0.529	1.327
	Heparin	1.362	0.306	0.877	2.117
Stimulant	Pokeweed	13.105	7.326	4.384	39.173
	Tetanus	0.091	0.023	0.056	0.149
	Streptokinase	0.039	0.011	0.022	0.068
Handling	Overnight	0.550	0.126	0.350	0.864
Method	Shipped	0.360	0.085	0.226	0.573

Table 5.8 Odds ratios, standard errors and confidence intervals produced by mixed-effects logistic regression

The value of the intracluster correlation was calculated as 0.807 and a test of H_0: $\rho=0$ indicated that the intracluster correlation was non-zero ($\chi_1^2=193.70$, $p<0.0001$).

Once again, HIV status, stimulant and handling method all appear to be significantly related to SI reading. Now parameter interpretation is with respect to specific rather than "average" individuals. So for example, the odds ratio of 13.105 for pokeweed means that the odds of an individual's blood sample SI reading being above 5 was 13 times the odds if they had a pokeweed stimulant compared to the *same* individual given candida stimulant. On the other hand, an individual's blood sample was about 90% less likely to have a high SI reading if they were treated with a tetanus stimulant compared to the candida. Interpretations for the remaining variables may be made in a similar way.

5.6.4 Comparison of results

The parameter estimates and standard errors obtained under each of the statistical analyses are provided in Table 5.9. All of the methods developed for clustered data have produced standard errors that are larger than those under independence. Also note that the results support the findings by Neuhaus *et al* (1991): the effects of the cluster-varying covariates (anticoagulant, stimulant and handling method) produced using the GEE method are smaller than those using the cluster-specific model. When examining the effects of the covariates keep in mind that parameters produced using the GEE method and mixed-effects methods have different interpretations.

SI		Independent logistic regression	GEE exchangeable	GEE exchangeable with robust option	Mixed-effects
HIV	Estimate,	-1.594	-1.773	1.773	-1.771
	Standard error	0.160	0.300	0.464	0.317
Antico-	CCPT	-0.191	-0.149	-0.149	-0.177
		0.188	0.174	0.204	0.235
agulant	Heparin	0.160	0.191	0.191	0.309
		0.180	0.166	0.146	0.225
Stimulant	Pokeweed	2.785	1.963	1.963	2.573
		0.528	0.386	0.398	0.559
	Tetanus	-1.669	-1.698	-1.698	-2.396
		0.196	0.185	0.356	0.252
	Streptokinase	-2.083	-2.279	-2.279	-3.257
		0.211	0.206	0.360	0.287
Handling	Overnight	-0.249	-0.375	-0.375	-0.598
		0.186	0.170	0.216	0.230
Method	Shipped	-0.842	-0.746	-0.746	-1.022
		0.185	0.173	0.209	0.237

Table 5.9 Parameter estimates and standard errors obtained for each of the analyses

The GEE results and mixed-effect model results do not coincide with those obtained by Betensky *et al* (2001) using SAS. STATA calculates parameter estimates in a slightly different manner compared to SAS. It's also been found that the GEE results using SAS do not match that using STATA when cluster sizes are unbalanced. See the following link for a discussion on the differences between PROC GENMOD in SAS and XTGEE in STATA:

http://www.stata.com/support/faqa/stat/xtgeesas.html .

A possible explanation for the difference in parameter estimates under the mixed effects model is that SAS and STATA are using slightly different procedures in calculating the estimates. Betensky *et al* (2001) uses a penalized quasi-likelihood approximation of the likelihood by using the GLIMMIX

macro in SAS. STATA makes use of the Gauss-Hermite quadrature approximation mentioned in section 5.3 by applying the XTLOGIT command.

CHAPTER 6 – SOME FINAL COMMENTS

Clustered data have presented challenges to statisticians for a long time. Its effects were recognized in the pharmaceutical industry in the course of experiments on the adverse effects of drugs (Haseman and Kupper, 1979). Pregnant dams were injected with varying amounts of the drug of interest and the number of their offspring adversely affected by the drug recorded. This resulted in binomial data with y_j out of n_j exposed pups exhibiting the effect of interest in the jth dam. In many cases, when all the known explanatory variables had been fitted it was clear that the assumption of a binomial response did not explain all the variation in the data. This became known as the 'litter effect'. The variance of the response y_j was assumed to be

$\text{var}(y_j) = \pi_j(1-\pi_j)\sigma^2$ where π_j was the probability of success.

It was solved *ad hoc* by inflating the variance

$$\text{var}(y_j) = \pi_j(1-\pi_j)[1+(n_j-1)\rho]$$

where ρ was the unknown intracluster correlation that had to be estimated (Collett, 1991). If the cluster (i.e. litter) sizes were equal, $\sigma^2 = [1+(n\text{-}1)\rho]$. If cluster sizes were not equal ρ could be found by an iterative procedure given by Williams (1982). The model was then refitted using weights given by $1/(1+(n_j\text{-}1)\rho)$. This approach only allows cluster-level covariates to be compared.

Cluster-varying covariates are those whose values change with the units within the cluster. Models that allow simultaneous assessment of both the cluster-level and cluster-varying covariates are the cluster-specific (CS) and the population-

averaged (PA) models. Betensky *et al* (2001) compare the covariate effects for each of these approaches. In PA models the covariate effects can be interpreted as the effect averaged over the whole population. This makes them useful for assessing the effect of cluster-level covariates but they have been severely criticized by Lindsey and Lambert (1998) since such an effect may not apply to any of the individuals in the population. For this reason preference tends to be given to the cluster-specific approach.

In cluster-specific models the random effect is usually assumed to have a normal distribution. It can either be integrated out, by a numerical procedure such as that described by Collett (1991) and described in Chapter 5, or fitted using an adaptation of the residual maximum likelihood procedure to generalized linear models (Schall, 1991). Schall's procedure is useful when interest lies in estimating both fixed and random effects. It permits not only the effects of all the covariates to be assessed but also the distribution of the random effects. This method has been extended to hierarchical generalized linear models by Lee and Nelder (1996). The class of hierarchical generalized linear models consists of generalized linear models in which the random components come from a conjugate distribution in the exponential family. Ultimately the choice of approach used in an analysis should be decided by the aims of the particular study in question, with preference tending to be given to the cluster-specific approach.

REFERENCES

Abramowitz M. and Stegun I.A. (1972) *Handbook of mathematical functions with formulas, graphs and mathematical tables*. U.S. Government Printing Office, Washington.

Anderson D.A. and Aitkin M. (1985) Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society B*, **47**, 203-210.

Barnett, Vic (1991) *Sample survey : principles and methods* (2nd Edition). London : E. Arnold

Bass M.J., McWhinney I.R. and Donner A. (1986) Do family physicians need nurse assistants to detect and manage hypertension – a randomized trial. *Canadian Medical Association Journal*, **134**, 1247-1253.

Besag J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B*, **36**, 192-236.

Betensky R.A. and Williams P.L. (2001) A comparison for clustered binary outcomes: analysis of a designed immunology experiment. *Applied Statistics*, **50**, 43-61.

Bland J.M., and Kerry S.M. (1997) Trials randomised in clusters. *British Medical Journal*, **315**, 600.

Breslow N.E. and Day N.E. (1980) *Statistical Methods in Cancer Research, Vol 1: The analysis of case-control studies*. Scientific Publications No. 32. International Agency for Research on Cancer, Lyon.

Brooks R. J. (1984) Approximate likelihood ratio tests in the analysis of beta – binomial data. *Applied Statistics*, **33**, 285-289.

Campbell M. (1999) Design and analysis of cluster randomized trials. Notes provided at School on Modern Statistical Methods in Medical Research in Trieste, Italy.

Cochran W.G. (1953) *Sampling Techniques*. John Wiley & Sons, Inc.

Collett D. (1991) *Modelling binary data*. Chapman and Hall, London.

Connolly M.A. and Liang K.Y. (1988) Conditional logistic regression for correlated binary data. *Biometrika*, **75**, 501-506.

Cornfield J. (1978) Randomization by Group: A Formal Analysis. *American Journal of Epidemiology*, **108**, 100-102.

Cox D.R. (1958) Two further applications of a model for binary regression. *Biometrika*, **45**, 562-565.

Crowder M.J. (1978) Beta-binomial anova for proportions. *Applied Statistics*, **27**, 34-37.

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

Donner A. (1987) Statistical methodology for paired cluster designs. *American Journal of Epidemiology*, **126**, 972-979.

Donner A. (1989) Statistical methods in ophthalmology: An adjusted chi-square approach. *Biometrics*, **45**, 605-611.

Donner A. (1992) Sample size requirements for stratified cluster randomization designs. *Statistics in Medicine*, **11**, 743-750.

Donner A. (1998) Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics*, **47**, 95-113.

Donner A., Birkett N., and Buck C. (1981) Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology*, **114**, 906-914.

Donner A., Brown K.S., and Brasher P. (1990) A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. *International Journal of Epidemiology*, **19**, 795-800.

Donner A., and Donald A. (1987) Analysis of data arising from a stratified design with the cluster as unit of randomization. *Statistics in Medicine*, **6**, 43-52.

Donner A., and Hauck W. (1989) Estimation of a common odds ratio in paired cluster randomization designs. *Statistics in Medicine*, **8**, 599-607.

Donner A., and Klar N. (1993) Confidence interval construction for effect measures arising from cluster randomization trials. *Journal of Clinical Epidemiology*, **46**, 123-131.

Donner A., and Klar N. (1994) Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *American Journal of Epidemiology*, **140**, 279-289.

Donner A., and Klar N. (1996) Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, **49**, 435-439.

Drum M. and McCullagh P. (1993) REML estimation with exact covariance in the logistic mixed model. *Biometrics*, **49**, 677-689.

Feng A., and Grizzle J.E. (1992) Correlated binomial variates: Properties of estimator of intraclass correlation and its effect on sample size calculation. *Statistics in Medicine*, **11**, 1607-1614.

Fitzmaurice G.M., Laird N.M. and Rotnitzky A.G. (1993) Regression models for discrete longitudinal responses (with discussion). *Statistical Science*, **8**, 284-309.

Fleiss J.L. (1981) *Statistical methods for rates and proportions* (2nd edition). New York, John Wiley and Sons.

Gail M.H., Byar D.P, Pechacek T.F. and Corle D.K (1992) Aspects of statistical design for the community intervention trial for smoking cessation. *Controlled Clinical Trials*, **13**, 6-21.

Gilmour A.R., Anderson R.D. and Rae A.L. (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72**, 593-599.

Glynn R.J. and Rosner B. (1992) Accounting for the correlation between fellow eyes in regression analysis. *Archives of Ophthalmology*, **110**, 381-387.

Griffiths D.A. (1973) 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*, **29**, 637-648.

Haseman J.K. and Kupper L.L. (1979) Analysis of Dichotomous Response Data from Certain Toxicological Experiments. *Biometrics*, **35**, 281-293.

Henderson W.G., Moritz T., Goldman S., Copeland J., Souchek J., Zadina K., Ovitt T., Doherty J., Read R., Chesler E., Sako Y., Lancaster L., Emery R., Sharma G., Josa M., Pacold I., Montoya A., Parikh, D., Sethi G., Holt J., Kirklin J., Shabetai R., Moores W., Aldridge J., Masud Z., Demots H., Floten S., Haakenson C., Hsu Y.L., Urbanski S. and Harker L.A. (1988) The statistical analysis of graft patency data in a clinical trial of antiplatelet agents following coronary artery bypass grafting. *Controlled Clinical Trials*, **9**, 189-205.

Hill W.G. and Smith C. (1977) Alternative response to query: estimating heritability of a dichotomous trait. *Biometrics*, **33**, 232-233.

Holt D., and Scott A.J. (1981) Regression analysis using survey data. *The Statistician*, **30**, 169-178.

Hsieh F.Y. (1988) Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine*, **8**, 1124-1140.

Jelfs P., Giles G., Shugg D., Taylor R., Roder D., Fitzgerald, Ring I. and Condon J. (1994) Cancer in Australia 1986-88. *Australian Institute of Health and Welfare*. Australian Government Publishing Service.

Kenward M.G. and Smith D.M. (1995) Computing the generalized estimating equations with quadratic estimation for repeated Measurements. *Genstat Newsletter*, **32**, 49-61.

Kerry S.M., and Bland J.M. (1998) Sample size in cluster randomization. *British Medical Journal*, **316**, 549.

Kish L. (1965) *Survey Sampling*. John Wiley & Sons, Inc.

Kish L., and Frankel M.R. (1974) Inference from complex samples. *Journal of the Royal Statistical Society B*, **36**, 1-37.

Laird N.M. and Ware J.H. (1982) Random-effects models for longitudinal data, *Biometrics*, **38**, 963-974.

Lee E.W. and Dubin N. (1994) Estimation and sample size considerations for clusterered binary responses. *Statistics in Medicine*, **13**, 1241-1252.

Lee Y. and Nelder J. A. (1996) Hierarchical generalized linear models (with Discussion). *Journal of the Royal Statistical Society B*, **58**, 619-678.

Lehtonen R. and Pahkinen E. J. (1995) *Practical methods for design and analysis of complex surveys*. John Wiley & Sons, Chichester.

Lemeshow S., Letenneur L., Dartigues J., Lafont S., Orgogozo J., and Commenges D. (1998) Illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the PAQUID study. *American Journal of Epidemiology*, **148**, 298-306.

Liang K.L. and Zeger S.L. (1986) Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*, **73**, 13-22.

Lindsay B.G. and Lesperance M.L. (1995) *A review of semiparametric mixture models. Journal of Statistical Planning and Inference*, **47**, 29-39.

Lindsey J.K. (1993) *Models for Repeated Measurements*. Oxford University Press.

Lindsey J.K. and Lambert P. (1998) On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine*, **17**, 447-469.

Lipsitz S., Laird N. and Harrington D. (1990) Multivariate regression models and analyses for categorical data. *Technical Report 700*, Department of Biostatistics, John Hopkins University, Baltimore.

Little R.J.A., Lewitzky S., Heeringa S., Lepkowski J. and Kessler R.C. (1997) Assessment of weighting methodoloy for the National Comorbidity Survey. *American Journal of Epidemiology*, **146**, 439-449.

Longford N.T. (1993) *Random coefficient models*. Clarendon Press. Oxford.

Mak T.K. (1988) Analyzing intraclass correlation for dichotomous variables. Applied Statistics, 37, 344-352.

Martin D.C., Diehr P., Perrin E.B., and Koepsell T.D. (1993) The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine*, **12**, 329-338.

McCullagh P., and Nelder J.A. (1983) *Generalized Linear Models*. Chapman and Hall, London.

McDonald B. (1993) Estimating logistic regression parameters for bivariate binary data. *Journal of Royal Statistical Society B*, **55**, 391-397.

Murray, D.M. (1998) *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Nathan G. (1988) Inference based on data from complex sample designs. In: Krishnaiah P.R., and Rao C.R., eds., *Handbook of Statistics, Vol 6.* Elsevier Science Publishers B.V., 247-266.

Neuhaus J.M., Kalbfleisch J.D. and Hauck W.W. (1991) A comparison of cluster-specific and population-averaged approaches for analysing correlated binary data. *International Statistical Review*, **59**, 25-35.

Neuhaus J.M and Kalbfleisch J.D. (1998) Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54, 638-645.

Neuhaus J.M. and Lesperance M.L. (1996) Estimation efficiency in binary mixed-effects model setting. *Biometrika*, **83**, 441-446.

Neuhaus J.M. and Segal M.R. (1993) Design effects for binary regression models fitted to dependent data. *Statistics in Medicine*, **12**, 1259-1268.

Pendergast J.F., Gange S.J., Newton M.A., Lindstrom M.J., Palta M. and Fisher M. (1996) A survey of methods for analyzing binary response data. *International Statistical Review*, **64**, 89-118.

Prentice R.L. (1986) Binary regression using an extended beta-binomial distribution with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, **81**, 321-327.

Prentice R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033-1048.

Qu Y., Williams G.W., Beck G.J. and Medendorp S.V. (1992) Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics*, **48**, 1095-1102.

Rao J.N.K., and Scott A.J. (1992) A simple method for the analysis of clustered binary data. *Biometrics*, **48**, 577-585.

Ridout M.S., Demetrio C.G.B, and Firth D. (1999) Estimating intraclass correlation for binary Data. *Biometrics* **55**, 137-148.

Roberts G., Rao J.N.K., and Kumar S. (1987) Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1-12.

Rosner B. (1984) Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics*, **40**, 1025-1035.

Schall R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719-727.

Scheaffer R.L. (1990) *Elementary survey sampling* (4th Edition). Boston, Mass : PWS-Kent.

Scott A.J., and Holt D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, **77**, 848-854.

Shipley M.J., Smith P.G., and Dramaix M. (1989) Pair studies when randomization is by group. *International Journal of Epidemiology*, **18**, 457-461.

Shoukri M.M., and Martin S.W. (1992) Estimating the number of clusters for the analysis of correlated binary response variables from unbalanced data. *Statistics in Medicine*, **11**, 751-760.

Shoukri M.M. and Pause C. (1999) *Statistical Methods for Health Sciences* (2nd Edition). CRC Press LLC.

Simpson J.M., Klar N., and Donner A. (1995) Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*, **85**, 1378-1383.

Sladden M.J, Thomson A.N. and Lombard C.J. (1999) Rectal bleeding in general practice patients. *Australian Family Physician*, **28**, 750-754.

Slymen D.J., and Hovell M.F. (1997) Cluster versus individual randomization in adolescent tobacco and alcohol studies: Illustrations for design decisions. *International Journal of Epidemiology*, **26**, 765-771.

Stiratelli R., Laird N.M. and Ware J.H. (1984) Random effects models for serial observations with binary response. *Biometrics*, **40**, 961-971.

Stukel T. (1993) Comparison of methods for the analysis of longitudinal interval count data. *Statistics in Medicine*, **12**, 1339-1351.

Tishler P., Donner A., Taylor J.O. *et al* (1977) Familial aggregation of blood pressure in very young children. *CVD Epidemiology Newsletter*, **22**, 45.

Wedderburn R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Williams D.A. (1982) Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144-148.

Woolson R.F., Bean J.A. and Rojas P.B. (1986) Sample size for case-contol studies using Cochran's statistic. *Biometrics*, **42**, 927-932.

Zeger S.L. and Liang K. (1986) Longitudinal data Analysis for discrete and continuous Outcomes. *Biometrics*, **42**, 121-130.

Zeger S.L. and Liang K. (1992) An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, **11**, 1825-1839.

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

Zhao L. and Prentice R. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642-648.