ANALYSIS OF CATEGORICAL DATA FROM LONGITUDINAL STUDIES OF SUBJECTS WITH POSSIBLY CLUSTERED STRUCTURES

by

Eliana Heiser de Freitas Marques

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1844T

January 1988

ANALYSIS OF CATEGORICAL DATA FROM LONGITUDINAL STUDIES OF SUBJECTS WITH POSSIBLY CLUSTERED STRUCTURES

by

Elianã Heiser de Freitas Marques

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill, 1987

Approved by:

Gary G. Koch

Advisor

Reader

Reader

ABSTRACT

ELIANA HEISER DE F. MARQUES. Analysis of Categorical Data from Longitudinal Studies of Subjects with Possibly Clustered Structures. (Under the direction of GARY G. KOCH).

Many references in the health sciences dealing with categorical data from studies with repeated measurements and possibly clustering present analyses that do not attempt to capture either of these special aspects of the data. Sometimes there may be analyses taking into consideration the longitudinal feature of the data; but for the most part, little attention has been given to any clustering of the attribute data.

Studies which involve a data structure with clustering should be analyzed first by viewing the cluster as the basic unit of analysis. Only if correlation among cluster sub-units is found negligible can cluster sub-units be appropriately viewed as the basic analytical units. In this way, care is taken so that estimates of variances are not underestimated (or overestimated).

Applications of existing methods of analysis for longitudinal categorical data are discussed, and extensions of available methods for analysis of clustered attribute data from a two-stage nested design are developed for these non standard type datasets. The settings of interest for the extensions are the two-stage nested design with one treatment applied within each cluster, and two treatments applied within each cluster. Two designs are considered; the completely balanced design and the partially balanced design. Examples from health sciences studies are used for illustration.

The methodology for the analysis of clustered attribute univariate data from a two-stage nested design is extended to the case of an outcome measure evaluated under two conditions (or time points) and generating two binary responses on each subject within a cluster for the case of one treatment per cluster and the case of two treatments per cluster. Quantities of interest in such studies are the mean value of an observed

response and correlation coefficients which reflect strength of clustering within a condition or condition (time) correlation. The proposed extended methodology focuses on parameter estimators within each condition (time) and the correlation coefficients between these conditions. Extensions of this methodology to a set of r binary responses on each subject within a cluster are noted.

ACKNOWLEDGMENTS

I wish to thank two professors, Gary Koch and Dennis Gillings, for their guidance, understanding and support during the period of my graduate study at UNC. I feel deeply fortunate to have had the opportunity to work closely with Gary Koch, my dissertation advisor. He has generously shared with me his ideas, comments and criticisms, and has patiently instructed me on statistical matters. His concern for his students made him an excellent advisor to me as well as a valuable friend. I also would like to express my appreciation to Dennis Gillings who from the beginning believed in my doctoral project and helped and encouraged me to pursue my goal.

My very warm thanks go to Mauro for his constant loving encouragement and companionable support. He was an excellent listener and healer of my dissertation aches. He and I 'dissertated' long hours together! Also, very special thanks go to my mother and my sister who instilled in me the energy to be strong in the present and to work persistently toward the future.

I am grateful for the support of my trailer friends in Biostatistics who made my life more bearable during this endeavor.

I would like to acknowledge the members of my committee, Drs E. C. Davis, C. M. Suchindran, C. D. Turnbull and C. Becker for their efforts. Thanks are also due to Ans Janssens for her invaluable help in producing the typescript.

Finally, a sentence to Gabriel who, even though very little, with his smile gave me a wealth of enthusiasm and made me realize that I could finish this work more quickly than I had believed possible.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION AND REVIEW OF SOME STUDIES INVOLVING REPEATED MEASUREMENTS AND CLUSTERED ATTRIBUTE DATA

1.1 Introduction

In the health and social sciences, researchers often encounter categorical data with added complexities from a profile of questions in a nested hierarchy and/or cross-classification. A common feature of these studies is a non-standard data structure with repeated measurements (subjects of investigation observed under two or more conditions), which may have some degree of clustering (correlated categorical data from distinct observational units). For example, Kempthorne (1982) analyzes a psychiatric diagnosis study previously discussed by Landis and Koch (1977), where 30 subjects are each separately classified as neurotic or not by six psychiatrists. Each subject is evaluated six times, and subjects are viewed as clusters.

Some specific examples illustrating one or more of these distinguishing features come from studies with a series of categorical responses directed at providers or users of social services. These studies may involve different groups of clusters of subjects, who are asked questions about their choices among a fixed set of alternative actions under several scenarios. Moreover, these scenarios may be cross-classified according to two or more study factors. This is the case, for example, of a study of day care center criteria for sending sick children home, with members of the staff at centers reporting on the actions taken under different situations related to a perceived ill child. In terms of the data structure, there are different features to be identified with relation to this study. Day care centers are stratified by level (certified center/non-certified center) which can be referred to as strata. There are a number of conditions (questions) for which subjects (staff) are observed and for which day care center may be a source of clustering; if besides staff, parents are also asked to answer the same set of questions posed to staff, about criteria to send sick children home, then staff and parents are designated as subpopulations; other dimensions to be considered, generally also identified as factors, are age of children (age<2, ages between 2 and 5), presence/absence of fever and presence/absence of specific symptoms or signs of illness. The answers 'do nothing', 'tell parent at the end of the day', 'call parent to tell them' and 'call parent for immediate pickup', for combinations of the study factors just mentioned above, generate the response variables that will describe variation among staff and/or parents for different levels of day care centers.

Another group of studies that need special consideration for its relatively non-standard data structure comes from the evaluation of dental materials. Such clinical studies involve patients who receive one or more materials in one or more teeth at one or more sites. For these types of studies, patient (or the mouth), should be viewed as a cluster. Technique, operator and restoration material are factors. Differently from other studies, here only a subset of possible treatments is assigned to any particular patient. Various characteristics are measured with respect to each restoration. They are, for example, retention of color, sensitivity, marginal discoloration, number of caries, etc. The status of each site for each tooth of each person is classified into ordinal categories at one or more follow-up times which brings in a multiple time structure. Other aspects of these studies which require attention are the role of multiple operators and the role of multiple evaluators who perform the follow-up classifications.

Other studies may present a complex discrete data structure encompassing repeated measurements and clustering attributes. For example, in a diary science clinical trial for treatment of mastitis in dairy cows discussed by Koch et al. (1978), the pre-treatment and post-treatment disease status of each of a cow's four udder quarters is assessed. In this study, the repeated measurements feature is identified from the before and after treatments, and the measurements obtained from the cow's four udder quarters, imply that the cow's udder is viewed as a cluster. The only factor here is the treatment which is randomly assigned within herds such that every treatment is assigned at least once within herd. In the original dataset, there is information on the two major types of bacteria causing the udder's infection. This additional component of the repeated measurements structure characterizes bacteria A and/or bacteria B as being present/absent in a quarter. One might consider if a larger number of herds and of cows were available, clustering features at different stages in the study, by identifying bacteria type within quarter, quarter within cow, cow within herd; this last step might present further difficulties since cows within the same herd receive different treatments.

Another interesting study which has been in fashion for sometime is the multi-center, multi-visit clinical trial in which patients are assigned at random to two or more treatments. A profile of repeated measurements initially and at each of one or more follow-up visits is recorded for each patient. Such follow-up produces a profile of possibly correlated measurements for each patient at the several points in time. The data structure for this clinical trial is pictured as a simpler version than what could be described for the previous examples. It is a rectangular array that consists of multiple visits of each of several subjects. A variation here is that some subjects do not appear at all visits, so one aspect of this structure is missing visits. Because these clinical studies as well as the dental material studies described earlier encompass a situation where the response variables are examined along a specific dimension like time or intensity of a particular stimulus, they are also called longitudinal studies. When the subjects in such studies come from multiple treatment centers, there can be clustering due to common aspects of the management and assessment of subjects from the same center.

Analyses of the previously described types of studies in the health and/or social sciences have often been undertaken ignoring the clustered aspects of the data and the potential impact of the repeated measurements structure. As an illustration, in studies involving day care centers and their staff (e.g., incidence of illnesses in day care centers), usually the staff member is regarded as the unit of analysis rather than day care center. Such analyses actually do not examine day care center, but staff characteristics. So, in this situation, whenever differences are found, it is not clear if it is the nature of staff or the nature of day care centers that accounts for the encountered difference. A similar situation occurs with clinical dental studies where patients are treated for more than one tooth and sometimes more than one site per tooth. In many of these studies, the unit of analysis is the tooth; however if mouth was the unit of analysis, it would account for possible correlation between teeth within the same mouth. These choices of units of analysis have been a limitation of the standard methods used to analyse such studies. For example, to ignore a possible clustering effect of day care center when staff is used as the unit of analysis implies an assumption of no intra cluster correlation; i.e. within cluster variation characterizes all the variation in the dataset. However, if correlation among 'sub-units' (staff, teeth) in a cluster (day care center, mouth) is strong, it is important to account for the clustered structure of the data in the analysis, i. e., it is important to adjust for cluster to cluster variability so that estimates of variance are not under estimated or over estimated. So, an analysis of clustered data structures should be based first on viewing the cluster as the unit of analysis. If it is possible to demonstrate that there is negligible correlation among cluster sub-units, i.e., no clustering effect, then an analysis based on viewing the cluster sub-units as the basic units of study is justified.

For settings like the ones described above, where researchers may be presented with particular statistical questions which are not routine, it is important to investigate alternative statistical methods of analysis. Also, since decision making based on the use of statistical tools almost always involves collection of data, the way in which data are collected becomes extremely important. This involves giving consideration to issues of structure of the design and types of measurements to be obtained.

1.1.1 Considerations of Study Design.

Study design determines the nature of the randomization process that will generate the data. This is important for any type of study because it identifies the basis for inference (or lack of thereof) from observed data to some target population. Randomization is employed to help average out the effect of many extraneous variables which may be present in an experiment, and by use of certain special designs, the effects of many important factors can be studied in the experiment as well as the interrelationships between these factors. For example, in clinical dental material studies, it involves using a randomization process to allocate sites, teeth and patients to combinations of materials, techniques and clinicians. In clinical trials, patients are allocated to two or more treatments. This allocation may be an unrestricted randomization scheme or one in which blocks of patients are assigned to different treatments at random so as to equalize the number of patients in each treatment group within relevant strata(e.g.,clinics).

1.1.2 Aspects of Measurements.

Measurement structure is concerned with whether the data are expressed on a continuous, discrete enumeration, ordinal, binary, or nominal scale. In some studies it also has to do with the method of follow-up; whether it has a single or multi-visit structure, and also with the potential issue of censoring data from subjects discontinuing a study before final follow-up. All these aspects are important because the measurement scale of response variables expresses the nature of information available for statistical analysis. Illustrative of different aspects of measurements are day care center study data obtained from subjects (staff) answering a questionnaire with a set of fixed alternative choices and clinical dental study dentists which give scores to their direct clinical evaluation of different restorations placed with different techniques and different materials. Data obtained from both of these examples are primarily categorical and based on the cross-classification of response and explanatory variables.

Essentially, all the information involving these often occurring studies gives rise to datasets with a fairly complex pattern of data on each subject. Difficulties may increase if data do not correspond to a fixed pattern of response in the same way across all subjects, i.e., if it is incomplete, in which case methods to be applied become more complicated or need to be worked further. Statistical questions of interest generated from these studies are, for example, comparison of treatments in a clustered data structure, within and between subject covariation with clustered data, issues of variation over time in a longitudinal data structure with clustering, evaluation of agreement of clinicians and/or observers etc. Strategies for the analysis of such datasets include available methods and refinements or extensions of such methods. For these, special computing techniques and/or simplifying assumptions may be needed.

1.2 Background on Day Care Center Studies

The number of children enrolled in day care facilities in the United States is large and rapidly expanding. Presently over eleven million children spend a minimum of ten hours per week at a day care center (Fosburg,1981). Demand for day care centers will increase in the future. Information regarding the epidemiology, transmission and control of infections in child care centers, even though incomplete, is accumulating and researchers in these areas have been analysing these data in order to generate suggestions for more effective disease control practices in day care centers.

Doyle (1978), studied the incidence of illness in early group and family day care using the child as the unit of analysis. Her paper examined rates of reported illness in chidren under three years of age enrolled in a group day care center and compared these rates with those in home-reared children and children in family day care homes. Data on seven categories of illnesses were collected by bimonthly interviews with parents during two consecutive winters. The statistical analysis developed for the data was based on analysis of variance for five of the seven symptoms; Because of skewed distributions and heterogeneity of variance, the remaining two variables were reduced to nominal data and chi-square and Fisher tests were used as appropriate.

It is worthwhile noting that Doyle's study was done comparing twelve children from each type of day care and that all data was obtained from parental reports. These small sample sizes and the questionable validity of parental report are two limitations of this study. On the other hand, previous studies done on this subject were as much or more questionable than Doyle's. For example, Loda et al. (1972), relied on care-taker reports of illness verified by throat and nasal cultures. Yet, their study lacking an adequate control group, compared rates for day care children with rates in the general population from the Cleveland Family Study. Strangert (1976), included control groups but the method of data collection was different for the center children compared to family day care and home-reared children. Only center children were examined and reported routinely by a pediatrician or care-taker; the other children were reported by parents.

Lemp et al. (1984), reported the first randomized prospective study examining specific characteristics of day care center staff and their association with day care centers diarrhea. This was an eight month period study in Texas, where sixty day care centers were randomly selected from a list of 736 licensed day care centers. Data on diarrheal illnesses was collected by weekly contacts by phone to determine number of children and staff who experienced diarrhea during the preceding week. Self-administered questionnaires were mailed to day care centers' staff and directors to obtain data on characteristics of the centers and their staff. Two units of analysis were used; center and staff. The statistical analysis done was restricted to calculation of incidence of diarrhea and correlation of this incidence with characteristics of the center and staff.

Sullivan et al. (1984), reported on a longitudinal study of occurrence of diarrheal disease in day care centers. Their study was based on sixty randomly selected day care centers in Houston, Texas and the method used to obtain information on episodes of diarrhea was every week telephone calls. Data was also obtained from directors concerning management, facility and procedual characteristics of each center. Analysis of variance and stepwise regression between incidence of diarrhea and selected day care center characteristics were the statistical tools used to analyse the data, with day care center as the unit of analysis.

Bartlett et al. (1985) examined diarrheal illness among infants and todlers in day care centers, comparing it with day care homes and households. This was a prospective two year study of diarrheal disease, with twenty-two randomly selected infant-todler day care centers, thirty (random) day care homes with 0-36 month old children and a house (random) sample with 102 houses with 0-36 month old children. They compared rates of diarrhea in the three settings using five serial biweekly family based surveys during the period of highest diarrheal rates, with children as the unit of analysis. Their statistical treatment of the data involved chi-square tests and a stepwise loglinear model including occurrence of diarrhea, type of child care setting, presence of 0-36 month sibling and presence of school age sibling in the family of the study child.

The special article on public health considerations of infectious diseases in child day care center, in the Journal of Pediatrics (1984), brings about in its conclusions a very important issue not only to the public health researcher but also to the statistician helping with the analysis of these types of data. It says that even though more carefully planned studies have been published more recently, there are still limitations on data on which to base control strategies for day care centers. This argument is indeed very important because well designed studies granting thorough statistical analyses will add very much to the present understanding of this subject, possibly permitting formulation of definite guidelines in the future, for all day care centers in the United States .

1.3 Background on Clinical Dental Material Studies

Evaluation of the clinical performance of dental restorative materials has been an integral part of dental practice for a very long time. The difference between early studies and the ones done in the last ten to fifteen years lies in the methods used. Many of the early studies were based predominantly on pure subjective judgement and were often very extensive. In some situations, there were examininations of several thousand clinical restorations of unknown history (Allan, D.N.(1969)). Though informative, such studies did little to provide guidance for future developments. By now, methodology has become progressively more sophisticated. On the other hand, the rate at which clinical evaluation techniques have developed lately in dentistry has been so rapid that there has been little assessment as to whether the methods used are indeed effective.

Statistical analysis underlying almost all analyses of clinical dental studies has been at a descriptive level for summarization of results; and only very recently, there has been consideration of study design at the early stage of planning these dental experiments.

Mahler and Marantz (1980), in their study of clinical assessment of dental amalgam restorations, spend a good part of their paper discussing factors influencing marginal fractures of restorations. They discuss many factors which may influence the incidence and extent of marginal fractures, pointing out that these factors should be considered in any experimental design for these types of studies. These factors are: nature of alloy, the ability of operator, the oral environment of the patient, the type and position of the tooth, and class and size of the restoration. Of all factors influencing marginal fracture, the nature of the alloy has proved to be one of the most significant and has been the subject of many studies. The operator effect has been recognized in the last ten years as important. Even though it has been assumed that after suitable instruction, differences in operators would not be a factor affecting the outcome of a clinical investigation, operator differences have been shown to affect marginal fracture performance in supposedly well controlled studies (Osborne and Gale, 1974; Letzel et al., 1978; Mahler and Marantz, 1979b; Goldberg et al., 1979). The patient factor, or more specifically, the combination of oral environment and occlusal loading within each patient has also been shown to have a significant effect on marginal fractures (Forsten and Kallio, 1976; Letzel et al.,1978; Goldberg et al., 1979). The questions of influence of tooth type and position, and restoration class and size on marginal fracture have been addressed more recently.

More recent literature in clinical dental material studies show that most of the factors discussed above have been taken into account. Operators place approximately equal numbers of restorations of each alloy. The patient factor is considered in the experimental design along with regard to both size of patient sample and distribution of alloys within each patient. Tooth type and position, restoration class and size have been balanced over the variables being studied. However, one very important aspect that has not been mentioned and/or studied in these more recent clinical dental material studies has to do with the oral environment of each patient and the correlation structure that may exist involving teeth and/or sites within the mouths of the patients. Consideration of this clustering aspect of the data would enhance statistical sensitivity of the experimental design to be used and along with the factors discussed above would enable better evaluation for these clinical studies.

The nature of the type of data obtained from social service areas like day care center studies and clinical dental material studies, and their different complexities arouses the interest of the applied statistician. Examination of these data from an underlying statistical viewpoint, with more comprehensive statistical analyses, may provide knowledge on the analysis of the data which may have not been explored until this time or that may have been underestimated because it has not been brought together in this fashion before.

1.4 Dairy Cows Clinical Trial Data

In a clinical trial for treatment of mastitis described by Heald et al. (1977) and discussed by Koch et al. (1978) and Kempthorne (1982), two drugs (penicillin and novobiocin) were used for treatment of infection in dairy cows. The objectives of the clinical trial were to optimize the doses of each of these two drugs for treatment of different bacteria and to determine if a combination of penicillin and novobiocin was justified for treatment of both infections at drying off.

Cows from sixteen southwestern Virginia Holstein dairy farms were in the study. Technicians who visited the farms on a weekly basis assigned subjective management scores (poor, fair, good and excellent) to the herds. Cows within each herd were assigned sequentially to receive one of the following treatments:

1. No treatment (control)
2. $1x10^5$ I.U. Penicillin
3. $2x10^5$ I.U. Penicillin
4. $4x10^5$ I.U. Penicillin
5. $4x10^2$ Novobiocin
6. $1x10^5$ I.U. Penicillin + $4x10^2$ Novobiocin
7. $2x10^5$ I.U. Penicillin + $4x10^2$ Novobiocin
8. $4x10^5$ I.U. Penicillin + $4x10^2$ Novobiocin
9. $6x10^2$ Novobiocin

The presence or absence of any infection prior to treatment and after treatment for each quarter of a cow's udder constituted the data for analysis. In this study, pre-treatment and post-treatment status were determined and this together with udder quarters illustrate a repeated measurements data structure. The basic experimental unit here was the cow since the four quarters of the cow's udder were not assumed independent.

1.4.1 Dairy Cows Data Analysis

Koch et al (1978) analyzed the mastitis data described in section 1.3 by several methods for the categorical data on the response to treatment of mastitis. The focus of the analysis was three fold: 1) to assess extent of interaction among experimental factors (e.g. herd management, drug treatment) and its implications to model fitting in order to describe relationships; 2) to account for pre-treatment scores as a covariable and 3) undertake multivariate analysis with respect to the four quarters simultaneously.

Four analyses were performed all together. Weighted least squares regression was the basic method of analysis in all of them. The first analysis examined the effects of drug treatment and herd management on post-treatment with the response variable being post-treatment mean scores, based on the post-treatment number of infected quarters: 0, 1, 2, 3, 4. The second and third analyses evaluated treatment effectiveness. One, a bivariate analysis of pre-treatment and post-treatment numbers of infected quarters was used to assess simultaneously the pre-treatment and post-treatment mean numbers of infected quarters and hence, their corresponding differences. The other directly analyzed the estimated mean pre-treatment versus post-treatment difference score with scores ranging -4, -3, -2, -1, 0, 1, 2, 3, 4. The last methodological strategy undertook multivariate analysis with respect to the four quarters simultaneously. The response variable of interest was the udder's quarter difference; and so there were four response variables. Three possible outcomes were associated with each quarter depending on if i) quarter was infected before treatment and not infected after (score=1); ii) quarter had same infection status after treatment as before (score=0); iii) quarter was not infected before treatment but infected after (score=1). Since there were nine sub-populations corresponding to the treatments and there were 81 possible multivariate profiles (3^4=81), this multivariate analysis of quarter difference score involved a 9x81 contingency table for the cross classification of treatment versus the multivariate quarter difference profiles. It was of interest in this analysis to investigate if there were differences among treatments for the four quarters , both separately and simultaneously. The pre-treatment versus post-treatment quarter difference mean score for infected quarters was the response of interest. The basic findings of the four analyses agreed with each other. From the first analysis resulted differences among treatments, no herd management score effect and no differences among the active treatments. When pre-treatment versus post-treatment mean number of infected quarters was analyzed through a bivariate analysis (second analysis), results showed that there were no differences among pre-treatment mean scores but significant differences among post-treatment mean scores; differences among post-treatment mean scores for the active (non-control) treatments were non significant. The result of no differences among pre-treatment mean scores was expected because of the nature of the design. So, in a spirit analogous to covariance analysis, an analysis was done reflecting equality of pre-treatment means through standardizing the pre-treatment scores to a common baseline, to study post-treatment comparisons. Here, the results indicated differences among post-treatment mean scores for the active treatments and that for control. Also, there was no pre-treatment versus post-treatment difference for the no treatment (control) group. Results of this third analysis were essentially the same as for the second analysis. For this alternative approach to the bivariate analysis, mean difference scores were different for the treatments although this was not true for the active treatments. In addition, the estimated mean difference score for the control treatment was not significantly different from zero. For the last analysis, results for the questions pertaining to quarter differences within treatments and treatment differences within quarters indicated that treatment differences were most apparent for quarter two and that quarter differences were most apparent for treatments six and seven; quarter differences for the control-treatment were non-significant. Comparison of active treatments within each quarter were also provided resulting in that active treatments within each quarter were non-significant but the simultaneous test of active treatments within quarters showed a borderline significance, differently from the result of the third analysis. In view of this situation, two analyses were performed, one where the simultaneous test was regarded non significant and another regarding it as significant. In the first case, the extent of improvement between pre-treatment and post-treatment for the respective four quarters for cows of any of the active (no-control) treatments was greater for the two quarters on the left than their two counterparts on the right and it was greater for the front two quarters than their two counterparts at the rear. Moreover, quarter location differences (left vs right, front vs rear) were both significant and the interaction between them was not significant. An overall estimate of improvement was obtained by adding the separate estimates for the four quarters together. It essentially yielded the same overall estimate as the previous analyses so in this sense, this multivariate analysis rendered the same conclusion concerning treatment differences as the simpler analyses described previously for the total numbers of infected quarters at pre-treatment and post-treatment, but at the same time provided additional interesting information about the differences in improvement for the four separate quarters. Alternatively, if the simultaneous test statistic for within quarter equality of active treatments was interpreted as significant, it became necessary to identify which of the four quarters was responsible for this significance. Quarter two (left rear) appeared to be a strong possibility because it had a wider range of difference scores for the active treatments than the other three quarters. Moreover, treatment nine seemed to be less effective for this quarter than the other active treatments which again seemed to be relatively similar with respect to extent of improvement. When an analysis was done putting all these considerations together, even though the model used provided a satisfactory framework for characterizing the variaton among quarter difference scores, the results were more difficult to interpret than for the analysis regarding the simultaneous test statistic as non-significant, because there did not seem to be any sensible reason for treatment nine to act in relation to quarter two as it did. In other words, the relatively small (estimated) improvement of quarter two for cows given treatment nine seemed to be either a consequence of data collection or transcription error or simply a random event. Since this latter interpretation was considered more reasonable, the analysis regarding the simultaneous test statistic as non-significant was judged as more preferable than this one. The conclusion from these analyses was that the various active treatments under study were homogeneous and superior to no treatment.

Kempthorne (1982), used the dairy cows mastitis data to illustrate the analysis of clustered attribute data in the framework of a two-stage nested design structure. The methodology for analysing clustered attribute data is aimed at assessment of the extent of clustering in the data and its effect on the estimation of the mean response. In this analysis, the focus of estimation is directed at two parameters, the mean response (π) and the intraclass correlation coefficient (ρ). Here, as before, the cow rather than the quarter of udder is the basic analysis unit.

Since the pre-treatment and post-treatment disease status of each of the cow's four udder quarters is assessed at given times, this analysis illustrates the applicability of the methodology for clustering to data having a repeated measurements structure with the cow's udder being viewed as the cluster and its quarters as the sub-units of the cluster.

For simplicity, the nine treatments were combined into four groups according to the two drugs, novobiocin and penicillin. The new sub-groupings were control, penicillin, novobiocin and a combination of novobiocin and penicillin. The contingency table for the cross-classification of the four treatments with the bivariate pre-treatment versus post-treatment joint distribution for numbers of infected quarters is shown in Table 1.1.

Estimation of the mean response and intraclass correlation coefficient was undertaken with weighted least squares analysis where pre-treatment and post-treatment logits and ln(1-ρ) were the elements of the response vector. The mean response for pre-treatment and post-treatment control obtained was 0.453 (s.e. = 0.021)

Table 1.1 Tabulation of the Bivariate Pre-treatment vs. Post-treatment Joint Distribution of Number of Infected Quarters for Four Treatments.

Treatment	Pre-treatment Number of Infected Quarters	Post-treatment Number of Infected Quarters 0	1	2	3	4
	0	2	2	3	1	0
	1	2	2	0	1	1
Control	2	0	1	2	0	0
	3	0	2	1	1	0
	4	0	0	0	5	0
	0	15	12	1	0	1
	1	13	7	1	0	0
Penicillin	2	13	6	6	0	0
	3	4	3	5	2	0
	4	9	6	3	2	2
	0	12	0	0	0	2
	1	6	8	2	0	0
Novobiocin	2	6	1	0	0	0
	3	8	2	0	0	0
	4	3	2	1	3	1
	0	17	2	2	1	0
	1	8	5	2	0	0
Combination	2	5	1	2	4	0
(Penicillin and	3	5	1	2	0	0
Novobiacin)	4	11	5	1	4	2

and for active post-treatment was 0.196 (s.e.=0.016). This indicated that the proportion of infected quarters was reduced by 43% under active treatment. The intraclass correlation pre-treatment was 0.404 (s.e.=0.035) and post-treatment was 0.218 (s.e.=0.038), i.e., post-treatment intra-class correlation was approximately half of the size of intra-class correlation pre-treatment.

Analysis of the same data assuming no correlation among udder quarters, (i.e., ignoring the clustering effect of the data) was also performed for comparison between estimates obtained from the two approaches. Table 1.2 summarizes the data when the udder quarter is viewed as the unit of analysis. A weighted least squares analysis was used with pre-treatment and post-treatment logits for the four drugs as the vector of response. The estimated mean response for pre-treatment and post-treatment control was 0.492 (s.e.=0.015) while the mean response for active post-treatment was 0.197 (s.e.=0.013). It is interesting to note how the standard errors differ when udder and udder quarter are viewed as units of analysis. When udder quarter is the unit of analysis, standard errors of the mean responses are substantially underestimated. So, the analysis taking into consideration the clustering of the data, using udder instead of udder quarter as the unit of study seems appropriate.

Table 1.2 Data on Number of Infected Quarters.

	Pre-treatment:	no		yes	
Drug	Post-treatment:	no	yes	no	yes
Control		40	24	17	23
Penicillin		193	39	161	51
Novobiocin		101	15	87	25
Combination		123	21	130	46

1.5 Summary and Outline of the Research

There haven't been many methods applied in the setting of multivariate repeated measurement studies involving combinations of categorical response measures, cross-classification of explanatory variables and possibly clustering. Weighted least squares methodology which is appropriate for a variety of analysis situations for categorical data, will be the focus of the methodological work developed in this dissertation to study different aspects of the analysis of data generated from these non-standard type datasets. Aspects addressed will be on calling attention to analyses which deal with available methods and/or refinements so far not explored as well as developing extensions of presently available methods when questions are identified for which available methods don't seem appropriate.

Methodology to analyse multiple time structure studies and/or clustered attribute data is presented by chapter. Chapter II presents an overview of categorical data analysis for repeated measures and linear models for means while Chapter III reviews the methodology to analyse clustered attribute data from a nested design with one binary response per subject, for the cases of one and two treatments per cluster. In Chapter IV, WLS methods will be used to describe variation among responses by staff members about how to handle perceived ill chidren in a day care center study, when different types of centers and different symptoms/illnesses of children are considered. The day care center study provides an opportunity to illustrate the methodology described in Chapter II, with an added refinement of creating summary measures for centers and by doing so, taking into account the clustered structure of the data. Chapter V extends the theory for clustered attribute data from a two stage nested design for one binary response per subject reviewed in Chapter III, to the case where subjects under investigation are observed under two conditions generating two binary responses per subject. This chapter is concerned with the situation of one treatment per cluster. Two studies, the day care center study and a clinical dental material study, illustrate this extended methodology. In Chapter VI, the one treatment two binary responses per subject case studied in Chapter V is extended to the situation of two treatments per cluster. The day care center study serves as an example for this more complex situation with an additional dimension incorporated to it, which is the answers given by the mothers of the children in those centers. Chapter VII presents a generalization to r conditions (r binary responses per subject) or alternatively to r-time points for the cases of one and two treatments per cluster. In Chapters IV to VII, the WLS function vector of interest takes a range of different forms depending on the data structure, on the parameters of interest to be estimated and on the hypotheses to be addressed. The model fits to these function vectors provide a description of the function estimates, of the variation among them and a framework to test hypotheses of interest.

In conclusion, in this work, performance of the combination of the methodology for repeated measurements and methods for categorical data analysis is evaluated to address statistical questions of interest generated from non-standard datasets. Also, methodology is developed to deal with the clustered aspect of the data for the situation of r-time points or r binary conditions per subject.

CHAPTER I I

OVERVIEW OF CATEGORICAL DATA ANALYSIS

2.1 Introduction

Until recently there were no general approaches to statistical analysis for categorical data. Emphasis was on the analysis of two-way tables; only a few books dealt with categorical data analysis (Maxwell, 1961; Cox, 1970; Fleiss, 1973) and none focused on the analysis of the relationship between categorical response variables and subpopulations (i.e., independent random samples from combinations of independent variables) based on demographic, diagnostic and/or other categorical factors in the context of a statistical model. Nowadays, for settings in health science research there are several methods and computing procedures for analytical frameworks concerning comparisons among subpopulations. The Weighted Least Squares (WLS) approach is one of these methods for categorical data. It is often an appropriate means of fitting linear models since the homogeneous variance assumptions of the usual least squares strategies are not required. The estimates being modeled might be those resulting from a cross-classification of variables from a multidimensional contingency table.

In order to narrow down this review to methods of interest for the work to be developed in the following chapters, attention will be given to the following topics: Weighted Least Squares and some aspects of Weighted Least Squares analysis for longitudinal categorical data. The notation and style in the following Sections are similar to those used by Koch et al. (1985), Koch et al. (1986).

2.2 Weighted Least Squares Methodology

The WLS methodology by Grizzle, Starmer and Koch (GSK, 1969) for discrete data represented by categories, allows one to describe the variation among a set of estimates produced from the data, in the context of a general linear model applied to categorical data. These estimates, or summary measures, include a variety of functions such as proportions, means, ratios etc. For example, in a day care center study of criteria to send sick children home, the estimates under study might be the proportion of staff choosing to call parent for child's immediate pickup; or in a clinical dental material study of retention of color by teeth restored with either material X or Z, the proportions of teeth retaining color for the different materials may be the measures of interest. To perform a WLS analysis it is necessary to have a set of summary statistics as a vector and a consistent estimator of the covariance matrix of this vector. The GSK strategy involves three stages: 1) construction of appropriate functions of responses; 2) fitting of regression models to these functions with WLS estimation; and 3) testing hypotheses about linear combinations of the model parameters.

Let $h=1, 2, \ldots, s$ index a set of s sub-populations as defined in terms of some set of explanatory variables and $l=1, 2, \ldots, q$ index a set of q categories which correspond to response profiles associated with the response variable of interest. Suppose that samples of size n_h are independently obtained by a process which is conceptually equivalent to simple random sampling from the respective s sub-populations. The data for such a situation can be summarized in an $(s \times q)$ contingency table as shown in Table 2.1, where n_{hl} denotes the frequency of response category l in the sample from the h-th sub-population. If $\underset{\sim}{n}_h = (n_{h1}, n_{h2}, \ldots, n_{hq})'$ denotes the (qx1) vector of the sample frequencies for the h-th subpopulation then the previous considerations concerning sampling framework imply that $\underset{\sim}{n}_h$ follows the multinomial distribution with parameters n_h and $\underset{\sim}{\zeta}_h = (\zeta_{h1}, \zeta_{h2}, \ldots, \zeta_{hq})'$ where ζ_{hl} represents the probability that a randomly selected subject from the h-th sub-population is classified in the l-th response category.

Table 2.1: An s × q Contingency Table

	Response profile categories				
Sub-population	1	2	...	q	Total
1	n_{11}	n_{12}	...	n_{1q}	n_1
2	n_{21}	n_{22}	...	n_{2q}	n_2
⋮	⋮	⋮	⋮	⋮	⋮
s	n_{s1}	n_{s2}	...	n_{sq}	n_s

The likelihood function is then given by

$$\Phi = \prod_{h=1}^{s} \{ n_h ! \prod_{l=1}^{q} [(\zeta_{hl}{}^{n_{hl}})/n_{hl} !] \} \tag{2.2.1}$$

with the ζ_{hl}'s satisfying the constraints $\sum_{l=1}^{q} \zeta_{hl} = 1$, for h=1, 2, ... ,s since each subject is classified into one of the q response categories.

The sample proportions $p_{hl} = n_{hl}/n_h$ are unbiased estimators for the corresponding parameters ζ_{hl}. The covariance matrix of the p_{hl} has the following structure

$$\mathrm{Var}\,(p_{hl}) = \zeta_{hl}\,(1 - \zeta_{hl})\,/\,n_h \tag{2.2.2}$$

$$\mathrm{Cov}\,(p_{hl}, p_{hl'}) = -\,\zeta_{hl}\,\zeta_{hl'}\,/\,n_h \qquad l \neq l' \tag{2.2.3}$$

$$\mathrm{Cov}\,(\,p_{hl}, p_{h'l'}\,) = 0 \qquad h \neq h'\ . \tag{2.2.4}$$

Using matrix notation to summarize these statements it follows that

$$\underset{\sim}{E}(\underset{\sim}{p}) = \underset{\sim}{\zeta}\ ,\ \underset{\sim}{Var}(\underset{\sim}{p}) = \begin{bmatrix} \underset{\sim}{V}_1(\underset{\sim}{\zeta}_1) & 0 & \cdots & \cdots \\ 0 & \underset{\sim}{V}_2(\underset{\sim}{\zeta}_2) & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \underset{\sim}{V}_s(\underset{\sim}{\zeta}_s) \end{bmatrix} = \underset{\sim}{V}(\underset{\sim}{\zeta}) \tag{2.2.5}$$

where

$$\underset{\sim}{V}_h(\underset{\sim}{\zeta}_h) = \{[\underset{\sim}{D}_{\underset{\sim}{\zeta}_h} - \underset{\sim}{\zeta}_h\underset{\sim}{\zeta}_h'] / \underset{\sim}{n}_h\} \tag{2.2.6}$$

is the covariance matrix for the h-th sub-population, with $\underset{\sim}{\zeta} = (\underset{\sim}{\zeta}_1', \underset{\sim}{\zeta}_2', \ldots, \underset{\sim}{\zeta}_s')'$ denoting the compound vector of parameters for all s sub-populations and similarly, $\underset{\sim}{p} = (\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s')'$ denoting the sample proportion vector. A consistent estimator for $\underset{\sim}{V}(\underset{\sim}{\zeta})$ is $\underset{\sim}{V}(\underset{\sim}{p})$, where $\underset{\sim}{p}$ has been substituted for $\underset{\sim}{\zeta}$.

A general class of linear models for describing the variation in $\underset{\sim}{F}$, the vector of functions of response data, can be expressed as

$$\underset{\sim}{E}_A(\underset{\sim}{F}) = \underset{\sim}{E}_A(\underset{\sim}{F}(\underset{\sim}{p})) = \underset{\sim}{F}(\underset{\sim}{\zeta}) = \underset{\sim}{X}\underset{\sim}{\beta} \tag{2.2.7}$$

where $\underset{\sim}{X}$ is the specification matrix, $\underset{\sim}{\beta}$ is the vector of parameters to be estimated and $\underset{\sim}{E}_A$ denotes asymptotic expectation.

Most often, $\underset{\sim}{F}$ is any vector of functions obtained from combinations of linear, logarithmic, or exponential transformations of the response data. $\underset{\sim}{V}_{\underset{\sim}{F}}$ can be obtained using linear Taylor series methods as described by Koch, Imrey, Singer, Atkinson and Stokes (1985).

The WLS estimated model parameters $\underset{\sim}{b}$ and covariance matrix $\underset{\sim}{V}_{\underset{\sim}{b}}$ are

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} \tag{2.2.8}$$

and

$$\underset{\sim}{V}_{\underset{\sim}{b}} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1} \quad . \tag{2.2.9}$$

The Wald statistic for evaluating the goodness of fit of the model is

$$Q_W = (\underset{\sim}{F}-\underset{\sim}{X}\underset{\sim}{b})'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}(\underset{\sim}{F}-\underset{\sim}{X}\underset{\sim}{b}) \quad . \tag{2.2.10}$$

When the sub-populations are moderately large (i.e., $n_h \geq 20$), Q_W has an approximate chi-square distribution with degrees of freedom equal to the difference between the dimensions of $\underset{\sim}{F}$ and $\underset{\sim}{b}$.

If model (2.2.7) fits the data as determined from the goodness of fit statistic, it is of interest to study which factors are related to the response variable, i.e., to perform tests of linear hypotheses pertaining to the parameter vector $\underset{\sim}{\beta}$.

Specifically, hypotheses of the form

$$H_0 : \underset{\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0} \tag{2.2.11}$$

where $\underset{\sim}{C}$ is a known $(c \times t)$ matrix of full rank $c \leq t$ and $\underset{\sim}{0}$ is a $(c \times 1)$ vector of 0's, may be tested with the Wald statistic

$$\underset{\sim}{Q}_c = \underset{\sim}{b}'\underset{\sim}{C}'(\underset{\sim}{C}\, \underset{\sim}{V}_{\underset{\sim}{b}}\, \underset{\sim}{C}')^{-1}\underset{\sim}{C}\underset{\sim}{b} \tag{2.2.12}$$

which has approximately a chi-square distribution with c degrees of freedom in large samples under H_0. In this framework, the test statistic $\underset{\sim}{Q}_c$ reflects the amount by which the goodness of fit Wald statistic in (2.2.10) would increase if model (2.2.7) were reduced by substitutions based on additional constraints implied by (2.2.11). Thus, these methods permit the total variation within $\underset{\sim}{F}(\underset{\sim}{\zeta})$ to be partitioned into specific sources and hence represents a statistically valid analysis of variance for the corresponding estimator functions $\underset{\sim}{F}$.

Predicted values for $\underset{\sim}{F}(\underset{\sim}{\zeta})$ based on model (2.2.7) can be obtained from

$$\hat{\underset{\sim}{F}} = \underset{\sim}{X}\hat{\underset{\sim}{\beta}} = \underset{\sim}{X}(\underset{\sim}{X}'\, \underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\, \underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} \quad . \tag{2.2.13}$$

They are useful to calculate in order to facilitate model interpretation. A consistent estimator for their covariance matrix is given by $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{X}\underset{\sim}{V}_{\underset{\sim}{b}}\underset{\sim}{X}'$.

A large range of problems in categorical data analysis can be examined using the GSK methodology. The function $\underset{\sim}{F}$ in (2.2.7) may be quite simple, for example, a proportion, or complex, for example, a rank correlation coefficient between two response variables. These functions can be expressed in terms of sequential applications of any of the following matrix operations:

i) linear transformations of the type

$$\underset{\sim}{F}_1(\underset{\sim}{p}) = \underset{\sim}{A}_1 \underset{\sim}{p} = \underset{\sim}{a}_1 \qquad (2.2.14)$$

where A_1 is a matrix of known constants coefficients;

ii) logarithmic transformations of the type

$$\underset{\sim}{F}_2(\underset{\sim}{p}) = \underset{\sim}{\ln}(\underset{\sim}{p}) = \underset{\sim}{a}_2 \qquad (2.2.15)$$

where $\underset{\sim}{\ln}$ transforms a vector to the corresponding vector of natural logarithms;

iii) exponential transformations of the type

$$\underset{\sim}{F}_3(\underset{\sim}{p}) = \underset{\sim}{\exp}\,(\underset{\sim}{p}) = \underset{\sim}{a}_3 \qquad (2.2.16)$$

where $\underset{\sim}{\exp}$ transforms a vector to the corresponding vector of exponential functions, i.e., of anti-logarithms.

Linear functions of the form $\underset{\sim}{A}_1\underset{\sim}{p}$ can be used to generate marginal frequencies in repeated measurements data; log linear functions of the form $\underset{\sim}{A}_2\ \underset{\sim}{\ln}\ (\underset{\sim}{A}_1\underset{\sim}{p})$ can be used to generate logits; and compound functions such as $(1-\ \underset{\sim}{\exp}\ \underset{\sim}{A}_3\ \underset{\sim}{\ln}\ (\underset{\sim}{C} + \underset{\sim}{A}_2\ \underset{\sim}{A}_1\underset{\sim}{p})$ where $\underset{\sim}{C}$ is a vector of constants, produce estimates such as intraclass correlation coefficients for clustered data.

2.3 Some Aspects of Longitudinal Categorical Data

An essential feature of a broad range of statistical investigations classified as repeated measurement studies is that each subject (or experimental unit) is observed under two or more conditions. Some classes of repeated measurement studies are split-plot experiments and change over designs. Split-plot experiments use two or more stages of randomization. In the first stage, subjects are allocated to treatments or randomly selected from strata; in the subsequent stages, conditions are allocated within subjects. Change over designs are arrangements in which different treatments are applied to the same subjects in different periods, i.e. response of each subject is observed at specified locations in time and space and thus, conditions correspond to combinations of location and treatment. Another important class which involves studies like the ones described in Sections 1.1 and 1.3 is that of longitudinal studies. In longitudinal studies subjects being observed are systematically linked to conditions. The usual dimension for such linkage is time, but it also can be locations in space or different components of a concept, item or process (Koch et al, 1987). Some examples of such studies are 1) In a study of clinical dental materials, different types of restorations are assigned at random to patients' teeth and plaque retention in these restorations is evaluated every six months for three years; 2) In a political survey, each subject is asked about the degree of trust in three political institutions (the Presidency, the Senate, and Supreme Court). Subjects are drawn from different demographic groups. The three institutions constitute the conditions (Koch et al, 1987); 3) In a day care center study, staff members are asked questions about how to handle perceived ill children under different situations as such: children may belong to different age groups (age<2, 2<= age <5); may have symptoms/signs of illness (runny nose, conjunctivitis) and may or may not have fever. The set of questions resulting from the cross-classification of these factors constitute the conditions while the dimension for linkage between subjects and conditions is not time but different components of a process; 4) The dairy cows mastitis study presented in Chapter I (Section 1.3) also serves as an illustration of a longitudinal study. There cows are examined before and after treatment with examinations characterizing the two conditions and the linkage between cows and conditions being the timing(before/after).

When dealing with longitudinal studies of categorical data, WLS methods are useful for describing variation of mean scores across groups and conditions and other questions of interest, through a regression model. Since the sample mean is the summary statistic of interest for the analyses to be discussed later, linear models for means is presented next.

2.3.1. Linear models for means.

Let $h=1,2, \ldots, s$ index a set of distinct sub-populations from which samples with size n_h have been independently obtained by a process conceptually equivalent to simple random sampling. Let $i=1,2, \ldots, n_h$ index the subjects in the sample from the h-th sub-population and let $\underset{\sim}{y}_{hi}=(y_{hi1}, y_{hi2}, \ldots, y_{hir})'$ denote the vector of responses according to k conditions (longitudinal categorical data) $k=1,2, \ldots, r$ for which the k-th condition for the i-th subject has possible outcomes $l=1,2, \ldots, L$. For reference purposes, consider the data structure shown in Table 2.2.

An illustration of data with this structure comes from a day care center study where day care centers are stratified in certified and not certified groups. Staff members in each day care center answer specific questions about how to handle perceived ill children. Each question generates a binary response, with staff member choosing to call parent to send sick child home or not. Here the two groups representing the two types of day care centers may be viewed as two distinct sub-populations, with subjects in each sample being staff, and conditions are the questions posed to staff; each condition has two possible outcomes. It is of interest to compare staff actions from the two types of day care centers with respect to answers given to the different types of questions posed to them.

If $\underset{\sim}{a}_k = (a_{k1}, a_{k2}, \ldots, a_{kL})$ denotes a set of finite scores corresponding to the respective outcomes of the k-th response, then the sample mean for the k-th response among subjects in the sample from the h-th sub-population can be expressed as

$$\bar{y}_{hk} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hik} = (\underset{\sim}{a}_k' \underset{\sim}{n}_h^{(k)} / n_h) = \underset{\sim}{a}_k' \underset{\sim}{p}_h^{(k)} \qquad (2.3.1)$$

Table 2.2: Data Array for Longitudinal Studies

		Categorical Variables					
Sub-population	Subject	1	2	...	k	...	r
1	1	y_{111}	y_{112}	...	y_{11k}	...	y_{11r}
1	2	y_{121}	y_{122}	...	y_{12k}	...	y_{12r}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	n_1	y_{1n_11}	y_{1n_12}	...	y_{1n_1k}	...	y_{1n_1r}
2	1	y_{211}	y_{212}	...	y_{21k}	...	y_{21r}
2	2	y_{221}	y_{222}	...	y_{22k}	...	y_{22r}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	n_2	y_{2n_21}	y_{2n_22}	...	y_{2n_2k}	...	y_{2n_2r}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s	1	y_{s11}	y_{s12}	...	y_{s1k}	...	y_{21r}
s	2	y_{s21}	y_{s22}	...	y_{s2k}	...	y_{s2r}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s	n_s	y_{sn_s1}	y_{sn_s2}	...	y_{sn_sk}	...	y_{sn_sr}

here $\underset{\sim}{n}_h^{(k)} = (n_{h1}^{(k)}, \ldots, n_{hL}^{(k)})'$ denotes the vector of frequencies $n_{hl}^{(k)}$ for the l-th outcome of the k-th response variable in the sample from the h-th sub-population and $\underset{\sim}{p}_h^{(k)}=(n_{hl}^{(k)}/n_h^{(k)})$ denotes the corresponding vector of sample proportions.

Since the scores $\{a_{kl}\}$ are finite, the elements of the covariance matrix $V\{\underset{\sim}{y}_{hi}\}$ exist; it follows from the Central Limit Theory that the vector $\bar{\underset{\sim}{y}}_h = (\bar{y}_{h1}, \cdots, \bar{y}_{hk})$ has approximately a multivariate normal distribution for sufficiently large samples (e.g. $n_h \geq 20$) with

$$E(\bar{y}_h) = \underset{\sim}{\mu}_h, \quad V(\bar{y}_h) = \frac{1}{n_h} \underset{\sim}{\Sigma}_h \tag{2.3.2}$$

where $\underset{\sim}{\mu}_h = (\mu_{h1}, \ldots, \mu_{hr})'$ is the vector of means for the r response variables in the entire h-th sub-population, i.e., $\mu_{hk} = \underset{\sim}{a}_k' \, \zeta_h^{(k)}$ where $\underset{\sim}{\zeta}_h^{(k)} = \underset{\sim}{E}(\underset{\sim}{p}_h^{(k)})$ is the vector of probabilities $\zeta_{hl}^{(k)}$ for the l-th outcome of the k-th response for a randomly selected subject from the h-th sub-population. Also, it is assumed that they are defined so that $\underset{\sim}{\Sigma}_h$ is non-singular.

The covariance matrix $V(\bar{y}_h)$ for the $\{\bar{y}_h\}$ can be consistently estimated by

$$\hat{V}(\bar{y}_h) = \frac{1}{n_h^2} \sum_{i=1}^{n_h} (\underset{\sim}{y}_{hi} - \bar{\underset{\sim}{y}}_h)(\underset{\sim}{y}_{hi} - \bar{\underset{\sim}{y}}_h)' \; ; \tag{2.3.3}$$

(for the representation of $V(\bar{y}_h)$ in terms of proportions see Koch et al,1985).

Results in (2.3.1-2.3.3) indicate conceptual similarities between the sample means $\{\bar{\underset{\sim}{y}}_h\}$ and the functions of multinomially distributed counts $\{n_{hl}\}$ presented in Section 2.2, thus providing motivation for the use of WLS methods to fit linear models which describe variation among the $\{\mu_{hk}\}$. The formulation of WLS methods through means rather than the proportion vector $\underset{\sim}{p}$ has the advantage that justifying asymptotic properties follow from more broadly applicable conditions. These are sufficient sample sizes $\{\underset{\sim}{n}_h\}$ for the approximate normality of the $\{\bar{\underset{\sim}{y}}_h\}$ and consistency for $\underset{\sim}{V}_{\bar{\underset{\sim}{y}}}$, where $\bar{\underset{\sim}{y}} = (\bar{\underset{\sim}{y}}'_1, \bar{\underset{\sim}{y}}'_2, \ldots, \bar{\underset{\sim}{y}}_s)$ is the compound vector of sample means, rather than approximate normality of $\underset{\sim}{p}$.

In Chapter IV, a description of the technical background which pertains to the use of linear models for means along with examples will be presented. There y_{hik} will again represent the response of the i-th subject for the k-th condition in the h-th group. For each group, the subjects will be assumed to be representative of a corresponding infinite sub-population in the sense of simple random sampling.

Sample means to be calculated from the different studies to be analyzed with WLS methodology for illustration in later chapters, will be produced with the help of the computer program MISCAT(Stanish et al, 1978). This program generates an array of statistics, including mean scores, corresponding standard errors, estimates of the linear model, residuals, predicted values of the mean, standard errors and test statistics. The software was developed as an extension of the categorical data analysis program GENCAT(Landis et al, 1976), which is a special-purpose program that uses weighted least squares estimation in the analysis of multidimensional contingency tables. Some of the advantages of MISCAT over GENCAT are: 1) It operates on case record data as oposed to contingency tables and so enables direct construction of proportions and means for scored categories; 2) when some data are missing, calculation of the mean and covariance matrix are modified according to the methodology presented by Stanish, Gillings, and Koch (1978) so as to reflect the inherent structure of the missing data. The methodology behind this computer program is asymptotic and thus appropriate only if sample sizes within each sub-population are large (e.g., at least 15).

The day care center study described previously provides an opportunity to illustrate the application of WLS methods to address statistical questions of interest for longitudinal studies. Chapter IV will be concerned with exploring different aspects of analyses of such data. Linear model descriptions of the proportions of staff choosing to call parent for immediate pickup of a perceived ill child will be considered for a cross-classification of such variables such as age, level of temperature, symptoms/signs of illness of child and type of day care center.

CHAPTER III

CLUSTERED ATTRIBUTE DATA FROM NESTED DESIGNS

3.1 Introduction

The word cluster, or clustering, is usually associated with sample survey methods; however nowadays, studies in a variety of contexts other than these, involve data structures which may have some degree of clustering. For example, in the study of mastitis treatment for cows discussed in Section 1.4, the pre-treatment and post-treatment disease status of each cow's four udder quarters is assessed. The cow's udder can be viewed as the cluster while its quarters are the sub-units of the cluster. Alternatively, in a rat fertility study, couples might mate three times over a period of months and each mating outcome noted as successful or not. Here, the matings are the sub-units of the clusters or couples (Kempthorne 1982). Still, another illustration of data with a clustered structure is given by the day care center study mentioned in Section 1.1. There, day care staff members answer questions about a fixed set of alternative actions regarding how to handle perceived ill children. If day care center is considered as the unit of analysis, methods to account for a possible correlation structure of multiple staff within a day care center can be applied with staff as sub-units of analysis.

In studies like the ones described, a parameter of primary interest is π, the overall mean response associated with a set of clustered attribute data. For example, in the day care center study, it may be of interest to compare proportions of staff from the two different levels of day care centers (certified/non certified) who choose to call parent to pick up the child given the presence of specific symptoms or signs of illness. Another parameter of interest in this setting is ρ, the intraclass correlation among cluster subunits.

Overall estimates of the mean response π for combining the cluster specific information may be obtained through approaches based on one of the following strategies: 1) direct formulation of different types of means and their variances; 2) maximum likelihood methods, assuming an underlying probability distribution; 3) weighted least squares methods assuming a sampling framework. The assumptions required to apply each of these three methods are

method (1): estimates used in the direct estimation procedure come from a random sample of clusters obtained from a defined population of clusters .

method (2): sample size should be sufficiently large so that the estimated parameter vector relative to maximum of the likelihood function has an approximate normal distribution.

method (3): i) random selection either explicitly via the research design (involving random selection of subjects from a large study population) or implicitly via conceptual representation from stratified simple random sample assumptions.
ii) Sufficiently large sample size for $\underset{\sim}{F}$ (vector of functions of probabilities in the analysis) to have an approximate multivariate normal distribution.
iii) Sufficiently large sample size and available information to obtain a consistent estimate $\underset{\sim}{V}_{\underset{\sim}{F}}$ for the covariance matrix of $\underset{\sim}{F}$.

Weighted least squares and maximum likelihood are methods that rely on large sample size to ensure approximate normality. For maximum likelihood methods, sample size should be sufficiently large for the model structure to imply that the maximum likelihood estimates of its parameters have an approximate multivariate normal distribution. The large sample requirements for weighted least squares methods refer to sample size being sufficiently large so that the functions to be analyzed (proportion, mean score etc.) have an approximate normal distribution. A comprehensive discussion of these issues is given by Koch, Gillings and Stokes (1980).

When dealing with clustered categorical data, the methodology developed has been mostly focused on the two stage nested design. Previous work has concentrated primarily on the estimation of the overall mean response π using individual cluster means. More recently, Kempthorne (1982) focused on two parameter estimators within each cluster: the cluster mean and the joint probability that two distinct observations within a cluster both have the same attribute. By considering both of these quantities initially, the parameters of primary interest were effectively the cluster mean and the intraclass correlation coefficient. A summary of the theoretical framework underlying this type of analysis follows. It presents methodology for analysing data with one binary response in the setting of a two-stage nested design. The notation and style used are based on Kempthorne (1982).

3.2 The Two Stage Nested Design with One Treatment per Cluster

In the simplest sampling setting, a certain property that is to be measured is decided and is recorded for every unit that comes into the sample. This property of the units is referred to as characteristic or attribute.

Suppose now that each unit in the population can be divided into a number of smaller units, or sub-units; i.e., consider a finite population of N clusters with D observations in each cluster. Let $y_{\imath\jmath}$ be a typical population member ($\imath$=1, 2, ... ,N; $\jmath$=1, 2, ... ,D). Assume that n clusters are randomly selected without replacement from the population of N clusters and that d_i observations are randomly sampled without replacement from the population of D observations within the i-th cluster. Refer to a typical sample element as y_{ij} (i=1, 2, ..., n; j=1, 2, ... ,d_i). Note that styled subscripts refer to observations in the population and roman subscripts refer to those obtained from the two-stage sampling process. It is of interest in this framework to consider the situation when $D \rightarrow \infty$ and $N \rightarrow \infty$. In this case, sampling with replacement and sampling without replacement are equivalent. This is convenient because the mathematics of sampling with replacement is easier to work with than the mathematics of sampling without replacement i.e., formulas for the variances and estimated variances of estimates made from the sample are often simpler when sampling is with replacement than without replacement. So, the theory corresponding to sampling without replacement from a finite population is examined first for the particular situation when D and N tend to infinity, where D is considered infinite either if actual size of cluster is very large, or conceptually if range of potential observations in a cluster is large. Later, sampling with replacement is assumed based on the equivalence of the two sampling procedures for large D and N so that variances of estimators of interest can be shown.

Following Kempthorne (1982), let

$$y_{ij} = \begin{cases} 1 & \text{if the j-th observation in the i-th cluster has the attribute of interest} \\ 0 & \text{otherwise} \end{cases}$$

The model of interest is characterized by

$E(y_{ij})=\pi=$ Pr(an observation in a selected cluster has the attribute)

$E(y_{ij}\ y_{ij'})=\lambda=$ Pr(two distinct observations in the same cluster both have the same attribute)

$E(y_{ij}\ y_{ij'}\ y_{ij''})=\xi=$ Pr(three distinct observations in the same cluster all have the same attribute)

$E(y_{ij}\ y_{ij'}\ y_{ij''}\ y_{ij'''}) = \eta =$ Pr(four distinct observations in the same cluster all have the same attribute)

Then, the i-th cluster in the population has mean

$$\pi_i = \frac{1}{D}\sum_{j=1}^{D} y_{ij} \ ; \tag{3.2.1}$$

the population mean is

$$\pi = \frac{1}{N}\sum_{i=1}^{N} \pi_i \ . \tag{3.2.2}$$

The probability that two distinct observations from the i-th cluster both have the same attribute is given by

$$\lambda_i = \frac{1}{D(D-1)} \sum_{j \neq}^{D} \sum_{j'}^{D} y_{ij} y_{ij'} = \frac{(\sum_j y_{ij})^2 - (\sum_j y_{ij}^2)}{D(D-1)} =$$

$$\frac{(D\pi_i)^2 - D\pi_i}{D(D-1)} = \pi_i \left[\pi_i - \frac{1-\pi_i}{D-1} \right], \tag{3.2.3}$$

and the overall probability that two distinct observations from the same cluster both have the attribute is

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} \lambda_i \; . \tag{3.2.4}$$

As the number of observations D tend to infinity, these parameters have simple representations in terms of the π_i's, the mean responses for the clusters in the population. For example, for λ_i,

$$\lim_{D \to \infty} \lambda_i = \lim_{D \to \infty} \left\{ \pi_i \left(\pi_i - \frac{1-\pi_i}{D-1} \right) \right\} = \pi_i^2 \tag{3.2.5}$$

so, the overall parameter λ can be expressed as

$$\lambda = \lim_{D \to \infty} \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_i \right) = \frac{1}{N} \sum_{i=1}^{N} \lim_{D \to \infty} \lambda_i = \frac{1}{N} \sum_{i=1}^{N} \pi_i^2 . \tag{3.2.6}$$

When N tends to infinity the overall mean response $\pi = \frac{1}{N} \sum_{i=1}^{N} \pi_i$ may be viewed as the average over an infinite number of small intervals $(\pi, \pi + d\pi)$ between 0 and 1 inclusive so that

$$\pi = \frac{1}{N} \sum_{i=1}^{N} \pi_i = \int_0^1 p \, f(\, p \,) \, dp \tag{3.2.7}$$

while

$$\lambda = \frac{1}{N}\sum_{i=1}^{N} \pi_i^2 = \int_0^1 p^2\, f(\, p\,)\, dp \qquad (3.2.8)$$

i.e., as $D \rightarrow \infty$ and $N \rightarrow \infty$ the basic parameters have integral representation in terms of f(p), the density of π_i .

Another parameter of interest is the intraclass correlation coefficient ρ. It reflects the strength of the clustering for a set of data. It is a function of the parameters π and λ.

$$\rho = \frac{\text{Cov}(y_{ij},y_{ij'})}{\text{Var}(y_{ij})} = \frac{\lambda-\pi^2}{\pi(1-\pi)} = \text{intraclass correlation coefficient.} \qquad (3.2.9)$$

Note that when $D \rightarrow \infty$

$$(\lambda - \pi^2) = \left[\frac{1}{N}\sum_{i=1}^{N} \pi_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} \pi_i\right)^2\right] = \frac{1}{N}\sum_{i=1}^{N} (\pi_i - \pi)^2 \geq 0 \qquad (3.2.10)$$

so the numerator of 3.2.9 is ≥ 0 and ρ is non-negative.

Two designs are examined next in the context of two stage random sampling with binary responses; the completely balanced and the partially balanced design.

3.2.1 The Completely Balanced Design

Consider a population of N clusters, each cluster with D observations. First select a sample of n clusters (primary units) from N, then select a sample of size d of sub-units from each chosen primary unit. Even though sampling is done without replacement, since D and N are considered to be infinite, sampling with and without replacement are viewed as equivalent to each other and so sampling with replacement may be assumed.

Unbiased estimators of the population quantities are given by

$$\hat{\pi} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij}}{nd} = \frac{\sum_{i=1}^{n} \bar{y}_{i.}}{n} = \bar{y}_{..} \tag{3.2.11}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n}\sum_{j\neq j'=1}^{d}\sum^{d} y_{ij}\, y_{ij'}}{nd(d-1)} = \frac{\sum_i y_{i.}\,(y_{i.}-1)}{nd(d-1)} = \frac{\sum_i y_{i.}^2 - y_{..}}{nd(d-1)} \tag{3.2.12}$$

$$\hat{\xi} = \frac{\sum_{i=1}^{n}\sum_{j\neq}^{d}\sum_{j'\neq}^{d}\sum_{j''}^{d} y_{ij}y_{ij'}y_{ij''}}{nd(d-1)} = \frac{\sum_i y_{i.}(y_{i.}-1)(y_{i.}-2)}{nd(d-1)(d-2)} = \frac{\sum_i y_{i.}^3 - 3y_{i.}^2 + 2y_{..}}{nd(d-1)(d-2)} \tag{3.2.13}$$

$$\hat{\eta} = \frac{\sum_{i=1}^{n}\sum_{j\neq j'\neq j''\neq j'''}^{d}\sum^{d}\sum^{d}\sum^{d} y_{ij}y_{ij'}y_{ij''}y_{ij'''}}{nd(d-1)(d-2)(d-3)} = \frac{\sum_i y_{i.}(y_{i.}-1)(y_{i.}-2)(y_{i.}-3)}{nd(d-1)(d-2)(d-3)}$$

$$= \frac{\sum_i y_{i.}^4 - 6\sum_i y_{i.}^3 + 11\sum_i y_{i.}^2 - 6y_{..}}{nd(d-1)(d-2)(d-3)} \tag{3.2.14}$$

The analysis of variance table for the two-stage sampling process is presented in Table 3.1.

Table 3.1 Analysis of Variance: two- stage sampling process

source of variation	degrees of freedom	sum of squares	expected mean squares
mean	1	$nd\bar{y}_{..}^2$	$\sigma_e^2 + d\sigma_a^2 + nd\mu^2$
between clusters	$n-1$	$d\sum_i(\bar{y}_{i.}-\bar{y}_{..})^2$	$\sigma_e^2 + d\sigma_a^2$
within clusters	$n(d-1)$	$\sum_i\sum_j(y_{ij}-\bar{y}_{i.})^2$	σ_e^2
total(corrected for the mean)	$nd-1$	$\sum_i\sum_j(y_{ij}-\bar{y}_{..})^2$	

Note that as a consequence of the two-stage random sampling process, the y_{ij} observations may be written in terms of π and the π_i

$$y_{ij} = \pi + (\pi_i - \pi) + (y_{ij} - \pi_i) \quad , i= 1, 2 , \ldots ,n ; j=1, 2, \ldots , d \qquad (3.2.15)$$

where $(\pi_i - \pi)$ is the cluster effect and $(y_{ij} - \pi_i)$ is the sub-sampling residual; both random effects. Now, if $\mu=\pi$, $a_i = \pi_i - \pi$ and $e_{ij}=y_{ij} - \pi_i$, then (3.2.15) may be written as

$$y_{ij} = \mu + a_i + e_{ij}, \quad i=1, 2, \ldots ,n ; j=1, 2, \ldots ,d \quad . \qquad (3.2.16)$$

Observe that in terms of population parameters, according to a sampling process from an infinite population, $E(\pi_i-\pi) = 0$, $E(y_{ij}-\pi_i) = 0$, $Var(\pi_i-\pi) = E(\pi_i-\pi)^2= =E(a_i^2)= \sigma_a{}^2$ and $Var(y_{ij} - \pi_i) = E\,(y_{ij} - \pi_i)^2= E(e_{ij}^2) = \sigma_e{}^2$ where

$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N} (\pi_i - \pi)^2 = \lambda - \pi^2 \quad , \qquad (3.2.17)$$

$$\sigma_e^2 = \frac{1}{ND}\sum_{i=1}^{N} \pi_i(1-\pi_i) = \pi - \lambda \, . \qquad (3.2.18)$$

Also,

$$Cov\{\pi_i-\pi, y_{ij}-\pi_i\}=E\{(\pi_i-\pi)(y_{ij}-\pi_i)\} = E\{a_ie_{ij}\} = 0. \qquad (3.2.19)$$

Thus, the analysis of variance for the two-stage sampling process (Table 3.1) is similar in spirit to the two-stage nested random effects ANOVA where a_i and e_i are independent and identically distributed random variables, for which $E\{a_i\} = E\{e_{ij}\} =0$, $Var\{a_i\}= \sigma_a^2$. $Var\,\{e_{ij}\} = \sigma_e^2$ and $Cov\{a_i, e_{ij}\}=0$.

In order to estimate the variance components from Table 3.1, the expected mean squares are equated to the observed mean squares;

$$\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{ij}-\bar{y}_{i.})^2}{n(d-1)} \qquad (3.2.20)$$

and

$$\hat{\sigma}_e^2 + d\hat{\sigma}_a^2 = s_a^2 = \frac{d\sum_{i=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2}{n-1} \quad . \tag{3.2.21}$$

From (3.2.18) it follows that

$$E(s_e^2) = \pi - \lambda \qquad \text{hence} \qquad (\widehat{\pi-\lambda}) = s_e^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{ij}-\bar{y}_{i.})^2}{n(d-1)} \quad ; \tag{3.2.22}$$

and it is straightforward to show that

$$s_a^2 = \frac{n}{n-1}(\hat{\pi} + (d-1)\hat{\lambda} - d\hat{\pi}^2) \quad . \tag{3.2.23}$$

So, if the number of clusters n is large (which is the case of interest here), the parameter estimators $\hat{\pi}$ and $\hat{\lambda}$ (or equivalently $\bar{y}_{..} = \hat{\pi}$ and $s_e^2 = \widehat{\pi-\lambda}$) contain essentially all of the information needed to estimate the variance components of the two-stage nested design.

The covariance matrix of the vector of parameter estimates $\bar{y}_{..}$ and $s_e^2 = s^2$ from a sampling with replacement perspective can be written as the following sum of matrices:

$$\mathrm{Var}\begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} = \frac{1}{n}\begin{bmatrix} (\lambda-\pi^2) & (\lambda-\pi^2)-(\xi-\pi\lambda) \\ (\lambda-\pi^2)-(\xi-\pi\lambda) & (\lambda-\pi^2)-2(\xi-\pi\lambda)+(\eta-\lambda^2) \end{bmatrix} +$$

$$+ \frac{1}{nd}\begin{bmatrix} (\pi-\lambda) & (\pi-\lambda)-2(\pi-\xi) \\ (\pi-\lambda)-2(\lambda-\xi) & (\pi-\lambda)-4(\pi-\xi)+4(\xi-\eta) \end{bmatrix} +$$

$$+ \frac{2}{nd(d-1)}\begin{bmatrix} 0 & 0 \\ 0 & \lambda-2\xi+\eta \end{bmatrix} \quad . \tag{3.2.24}$$

Now, if the observations can be grouped in an appropriate compact way, expressions for the parameter estimators $\bar{y}_{..}$ and s^2 can be formulated by using a MISCAT computation approach through means. For each i, let $F_{i1}=\bar{y}_{i.}$ and $F_{i2}=s_i^2=\sum_{j=1}^{d} \frac{(y_{ij}-\bar{y}_i)^2}{(d-1)}$ then

$$\underset{\sim}{F} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \bar{y}_{i.} \\ s_i^2 \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} F_{i1} \\ F_{i2} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_i \quad ; \tag{3.2.25}$$

also

$$V(\underset{\sim}{F}) = \frac{1}{n^2}\sum_{i=1}^{n} (\underset{\sim}{F}_i-\underset{\sim}{F})(\underset{\sim}{F}_i-\underset{\sim}{F})' \triangleq \mathrm{Var}\begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} \tag{3.2.26}$$

is a consistent estimate for the covariance matrix of $\underset{\sim}{F}$.

Alternatively, data grouping can be done in a frequency table fashion. In this regard, a compact structure for the data can be presented as

j	0	1	2	...	d
number of clusters with j units having the attribute	f_0	f_1	f_2	...	f_d

where the n clusters are sub-divided into (d+1) groups based on how many units they have possessing the attribute of interest.

Note that

$$\sum_{j=0}^{d} f_j = n$$

and

$$\sum_{i=1}^{n} y_{i.}^k = \sum_{j=0}^{d} j^k f_j \qquad \text{for} \qquad k=1,2,3,4 .$$

This grouped data formulation shows the parameter estimators $\hat{\pi}$ and $\hat{\lambda}$ as

$$\hat{\pi} = \frac{\sum_{j=0}^{d} j\, f_j}{nd} = \sum_{j=0}^{d} (\tfrac{j}{d})\, p_j \ ,$$

$$\hat{\lambda} = \frac{\sum_{j=0}^{d} j(j-1)\, f_j}{nd(d-1)} = \sum_{j=0}^{d} \frac{j(j-1)}{d(d-1)}\, p_j \ ,$$

and

$$s^2 = \hat{\pi - \lambda} = \frac{\sum_{j=0}^{d} j(d-j)\, f_j}{nd(d-1)} = \sum_{j=0}^{d} \frac{j(d-j)}{d(d-1)}\, p_j \ .$$

where $p_j = f_j/n$ is the proportion of clusters for which j units have attribute.

Using the GSK methodology described in Chapter II, the proportions (f_j/n) can be written as a vector of response functions:

$$\begin{bmatrix} f_0/n \\ f_1/n \\ \vdots \\ f_d/n \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_d \end{bmatrix} \tag{3.2.27}$$

To perform the WLS analysis it is assumed that the observed data $\{f_j\}$ coming from this one sub-population follows a multinomial distribution with parameters n and $\underset{\sim}{\zeta} = (\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_d)'$; then the sample distribution $\underset{\sim}{p} = (p_0, p_1, \ldots, p_d)$ has (from Chapter II, (2.2.5)), $E(\underset{\sim}{p}) = \underset{\sim}{\zeta}$, and $\mathrm{Var}\{\underset{\sim}{p}\} = \{D_{\zeta} - \underset{\sim}{\zeta}\underset{\sim}{\zeta}'\}/n$. A consistent estimator for the covariance matrix of $\underset{\sim}{\zeta}$ is $\mathrm{Var}(\underset{\sim}{p}) = \{D_{p} - \underset{\sim}{p}\underset{\sim}{p}'\}/n$.

Thus,

$$\begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} = \underset{\sim}{A}_1 \underset{\sim}{p} =$$

$$= \begin{bmatrix} 0 & \frac{1}{d} & \frac{2}{d} & \frac{3}{d} & \cdots & \frac{d-1}{d} & 1 \\ 0 & \frac{1}{d} & \frac{2(d-2)}{d(d-1)} & \frac{3(d-3)}{d(d-1)} & \cdots & \frac{1}{d} & 0 \end{bmatrix} \begin{bmatrix} f_0/n \\ f_1/n \\ \vdots \\ f_d/n \end{bmatrix} = \underset{\sim}{F}. \qquad (3.2.28)$$

The estimated covariance matrix of the estimated $\underset{\sim}{F}$ is given by

$$\mathrm{Var}\begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} = \underset{\sim}{A}_1\, \underset{\sim}{V}(\underset{\sim}{p})\, \underset{\sim}{A}_1' = \underset{\sim}{V}(\underset{\sim}{F}) \quad . \qquad (3.2.29)$$

The intraclass correlation coefficient

$$\hat{\rho} = 1 - \frac{\widehat{\pi - \lambda}}{\hat{\pi}(1-\hat{\pi})}$$

is estimated by a compound function, through a series of matrix operations:

$$\hat{\rho} = 1 - \underset{\sim}{\exp}\, \underset{\sim}{A}_3\, \underset{\sim}{\ln} \left[\underset{\sim}{C} + \underset{\sim}{A}_2 \underset{\sim}{A}_1 \underset{\sim}{F}\right] \qquad (3.2.30)$$

where $(\underset{\sim}{C}+\underset{\sim}{A}_2\underset{\sim}{A}_1\underset{\sim}{F})$ produces $\hat{\pi}$, $(1-\hat{\pi})$ and $s^2(=(\widehat{\pi-\lambda}))$; and $\underset{\sim}{A}_3\ln(\underset{\sim}{C}+\underset{\sim}{A}_2\underset{\sim}{A}_1\underset{\sim}{F})$ generates the fraction $\ln(\widehat{\pi-\lambda}/\hat{\pi}(1-\hat{\pi}))$ with

$$\underset{\sim}{A}_2 = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}, \underset{\sim}{A}_3 = \begin{bmatrix} -1 & -1 & 1 \end{bmatrix} \text{ and } \underset{\sim}{C} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} . \qquad (3.2.31)$$

The estimated variance of $\hat{\rho}$ based on linear Taylor series, is given by

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{\rho}) = \underset{\sim}{D}_{a_2} \underset{\sim}{A}_3 \underset{\sim}{D}_{a_1}^{-1} \underset{\sim}{A}_2\, \underset{\sim}{A}_1 \underset{\sim}{V}(\underset{\sim}{p})\, \underset{\sim}{A}_1' \underset{\sim}{A}_2' \underset{\sim}{D}_{a_1}^{-1}\, \underset{\sim}{A}_3' \underset{\sim}{D}_{a_2} \qquad (3.2.32)$$

where

$$\underset{\sim}{a}_1 = \underset{\sim}{C} + \underset{\sim}{A}_2 \underset{\sim}{A}_1 \underset{\sim}{F} \quad \text{and} \quad \underset{\sim}{a}_2 = \exp\{\underset{\sim}{A}_3 \ln(\underset{\sim}{C} + \underset{\sim}{A}_2 \underset{\sim}{A}_1 \underset{\sim}{F})\} .$$

An illustration of the GSK methodology to analyse categorical data from a two-stage nested design with one treatment per cluster is given by Kempthorne and Koch (1983). This data was adapted from Fleiss and discussed by Landis and Koch (1977). It is a study where 30 subjects were each classified separately by six psychiatrists as neurotic or not. Here each cluster is of size 6. The frequency of subjects diagnosed as neurotic by the psychiatrists (j=0, 1, ..., 6) is presented in Table 3.2.

Table 3.2 Frequency of Subjects Diagnosed as Neurotic by Psychiatrists

Number of diagnoses of neurosis (j)	Number of Subjects	Proportion of Subjects
0	14	14/30
1	4	4/30
2	0	0
3	2	2/30
4	6	4/30
5	3	5/30
6	1	6/30

Estimates of the overall mean and within cluster variance are obtained by

$$\begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} = \begin{bmatrix} \hat{\pi} \\ \pi \hat{-} \lambda \end{bmatrix} = \underset{\sim}{A}_1 \underset{\sim}{p} =$$

$$= \begin{bmatrix} 0 & 1/6 & 2/6 & 3/6 & 4/6 & 5/6 & 1 \\ 0 & 5/30 & 8/30 & 9/30 & 8/30 & 5/30 & 0 \end{bmatrix} \begin{bmatrix} 14/30 \\ 4/30 \\ 0 \\ 2/30 \\ 6/30 \\ 3/30 \\ 1/30 \end{bmatrix} = \begin{bmatrix} 0.305 \\ 0.112 \end{bmatrix} = \hat{\underset{\sim}{F}}$$

The covariance matrix of $\hat{F}$ is obtained by using (3.2.29) :

$$\hat{\mathrm{V}}\mathrm{ar}\begin{bmatrix} \bar{y}_{..} \\ s^2 \end{bmatrix} = \begin{bmatrix} 0.00396 & 0.00096 \\ 0.00096 & 0.00047 \end{bmatrix}$$

and the standard errors for $\bar{y}_{..}$ and s^2 are 0.063 and 0.022 respectively. If clustering were ignored $\mathrm{Var}(\bar{y}_{..})=0.00118$ and the s.e.=0.034. Thus, the variance ignoring clustering is about (1/3) the variance accounting for clustering.

The intraclass correlation is estimated using (3.2.30) :

$$\hat{\rho} = 0.471 \ ,$$

with standard error of 0.073.

Note that when clustering is taken into account, $\mathrm{Var}\{\bar{y}_{..}\}= =\frac{\sigma_a^2}{n}+\frac{\sigma_e^2}{nd}=\frac{\lambda-\pi^2}{n}+\frac{(\pi-\lambda)}{nd}$. Now if $\sigma_a^2=0$, or $\lambda=\pi^2$, then $\mathrm{Var}\{\bar{y}_{..}\}= =(\sigma_e^2/nd)=(\pi-\lambda)/nd=(\pi-\pi^2)/nd=\pi(1-\pi)/nd$ and the design effect due to clustering is

$$\frac{\left(\frac{\lambda-\pi^2}{n}\right)+\left(\frac{(\pi-\lambda)}{nd}\right)}{\frac{\pi(1-\pi)}{n}} = \frac{d(\lambda-\pi^2)+(\pi-\lambda)}{\pi(1-\pi)} =$$

$$= \frac{\pi(1-\pi)+(d-1)(\lambda-\pi^2)}{\pi(1-\pi)} = \frac{0.00396}{0.00118}=\{1+(d-1)\rho\} \ = \ 3.36 \ .$$

It is of interest to test if $\rho=0$. Since $\frac{\hat{\rho}}{\mathrm{s.e.}(\hat{\rho})}>1.96$, one can interpret ρ as significantly different from zero; thus some adjustment for clustering is needed. However, because $\hat{\rho}$ is defined as a relatively complex ratio, concern may arise about the use of a normal distribution approximation. So, $\ln(\hat{\lambda}/\hat{\pi}^2)/\mathrm{s.e.}(\ln(\hat{\lambda}/\hat{\pi}^2)$ is also evaluated, since it is linear on log-scale and thereby may be more realistically normal. It follows that $\ln(\hat{\lambda}/\hat{\pi}^2)=0.73$ and $\mathrm{s.e.}(\ln(\hat{\lambda}/\hat{\pi}^2))=0.184$. Thus $\ln(\hat{\lambda}/\hat{\pi}^2)/\mathrm{s.e.}(\ln(\hat{\lambda}/\hat{\pi}^2))>1.96$, and thereby reinforces the conclusion for an analysis that adjusts for clustering.

3.2.2 The Partially Balanced Design

In some experimental studies there may be different groups of clusters, each group having different numbers of observations, i.e., there may be a group of clusters each having 5 observations, another group which has 7 observations in each cluster and etc. This is a situation where the design is constructed so that there are modules with different cluster sizes. Each module is balanced so, there is only one set of unbiased estimators.

Suppose the m-th module has n_m clusters each of size d_m . Let

$$y_{ijm} = \begin{cases} 1 & \text{if the j-th observation in the i-th cluster} \\ & \text{of the m-th module has the attribute} \\ 0 & \text{otherwise} \end{cases}$$

$(m = 1, 2, \ldots ,M; i = 1, 2, \ldots ,n_m ; j = 1, 2, \ldots ,d_m)$.

The set of estimators for the m-th module have the same forms as expressions (3.2.11), (3.2.12), (3.2.13) and (3.2.14) with addition of an 'm' subscript on y_{ij} , n , and d. Here, similarly to the completely balanced case, the parameter estimators $\bar{y}_m$ $(=\bar{y}_{\cdot\cdot m}$ $=\hat{\pi}_m)$ and s_m^2 $(=(\hat{\pi}-\hat{\lambda})_m)$ contain essentially all the informatiom necessary for estimation of the variance components.

Grouping the data as for the completely balanced case, provides a convenient way for evaluation of the covariance structure of $\bar{y}_m$ and s_m^2. since both estimators may be expressed as linear functions of the frequencies $f_{j,m}$ of the number of clusters in the m-th module possessing j units with the attribute.

The vector of parameter estimators may be written as

$$\begin{bmatrix} \bar{y}_m \\ s_m^2 \end{bmatrix} = \underset{\sim}{A}_{1,m} \underset{\sim}{p}_m = \underset{\sim}{F}_m \qquad (3.2.33)$$

where

$$\underset{\sim}{A}_{1,m} = \begin{bmatrix} 0 & \frac{1}{d_m} & \frac{2}{d_m} & \frac{3}{d_m} & \cdots & \frac{d_m-1}{d_m} & 1 \\ 0 & \frac{1}{d_m} & \frac{2(d_m-2)}{d_m(d_m-1)} & \frac{3(d_m-3)}{d_m(d_m-1)} & \cdots & \frac{1}{d_m} & 0 \end{bmatrix}$$

and $\underset{\sim}{p}'_m = (p_{0,m}, p_{1,m}, \ldots, p_{d,m})$. An unbiased estimator of the covariance matrix of $\underset{\sim}{F}_m$ is

$$\underset{\sim}{V}_m = \hat{\underset{\sim}{Var}} \begin{bmatrix} \bar{y}_m \\ s_m^2 \end{bmatrix} = \underset{\sim}{A}_{1,m}\, \underset{\sim}{V}\, (\underset{\sim}{p}_m)\, \underset{\sim}{A}'_{1,m} \tag{3.2.34}$$

where $\underset{\sim}{V}(\underset{\sim}{p}_m) = [(\frac{1}{n_m})(\underset{\sim}{D}_{\underset{\sim}{p}_m} - \underset{\sim}{p}_m\, \underset{\sim}{p}'_m)]$ is a consistent estimator of the covariance matrix of $\underset{\sim}{F}_m$.

These modular estimators can be combined by WLS methods in GSK to produce overall estimators of the parameters of interest, provided each module fulfills the requirement of sufficiently large n_m(say $n_m \geq 15$). Thus, $\underset{\sim}{F}'$ = $=[\bar{y}_1., s_1^2, \bar{y}_2., s_2^2, \ldots, \bar{y}_M, s_M^2]$ is the (2M x 1) vector of modular estimators of π and $(\pi - \lambda)$. Variation among the elements of $\underset{\sim}{F}$ can be investigated by fitting linear regression models by WLS and goodness of fit of these models may be tested as described in Section 2.1.

Here, the model of interest is

$$\underset{\sim}{F} = \underset{\sim}{A}\, (\underset{\sim}{\zeta}) = \underset{\sim}{X}\underset{\sim}{\beta} \tag{3.2.35}$$

where

$$\underset{(2m\text{x}2)}{\underset{\sim}{X}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ \cdots & \cdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \quad \underset{\sim}{\beta} = \begin{bmatrix} \pi \\ \pi - \lambda \end{bmatrix}.$$

Estimates $\hat{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$ are obtained from

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\, \underset{\sim}{V}_{\underset{\sim}{F}}^{-1} \underset{\sim}{X})^{-1}\, \underset{\sim}{X}'\, \underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\, \underset{\sim}{F}$$

with covariance

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = (X'\boldsymbol{V}_{\boldsymbol{F}}^{-1}X)^{-1}$$

where

$$\boldsymbol{V}_{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{V}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{V}_2 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \boldsymbol{V}_M \end{bmatrix} \tag{3.2.36}$$

where each $\boldsymbol{V}$ of $\boldsymbol{V}_{\boldsymbol{F}}$ is defined as in (3.2.34).

If the number of clusters in a module is small (say, $n_m<10$) then the elements of $\boldsymbol{p}_m$ are not based on sufficiently large sample sizes for WLS and the estimates obtained by applying GSK methodology do not necessarily have the optimal properties which follow from asymptotic arguments. In this case, the mean of the $\boldsymbol{F}_m$ could be used as an estimator of $\boldsymbol{\beta}$ when the model was considered reasonable.

A dose response study discussed in Kempthorne and Koch (1983) will be used to illustrate the modular case situation. Consider a study in which it is of interest to estimate the mean proportion of successful matings among pairs of rats. Suppose there are four modules of mating pairs such that there are two distinct cluster sizes. Table 3.3 shows the data arising from this study.

In previous examination of these data (Koch and Kempthorne, 1981) it was determined that the doses of 0 and 50 were equivalent so distinction between them can be ignored.

Table 3.3 Number of Mating Pairs with 'j' Successiful Matings

Module	Dose	Number of mating per mating pair	Number of successful mating j: 0	1	2	Number of mating pairs
1	0	1	2	28	0	30
2	0	2	2	4	24	30
3	50	1	1	19	0	20
4	50	2	1	5	13	19

Thus, model (3.2.35) for the modular estimates of π and $(\pi-\lambda)$ has the form :

$$\underset{\sim}{F} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ (\pi-\lambda)_2 \\ \pi_3 \\ \pi_4 \\ (\pi-\lambda)_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \pi \\ \pi-\lambda \end{bmatrix} = \underset{\sim}{X}\,\underset{\sim}{\beta} \ ,$$

noting that only π can be estimated from modules with cluster size 1. The fit of the model was reasonable (Q=4.04, 4 d.f., p=0.40). Estimates of π and ρ were obtained by application of a series of linear, exponential and logarithmic functions to the estimated parameter vector $\hat{\underset{\sim}{\beta}}$ obtained from the WLS procedure;

$$\hat{\pi}=0.906\pm0.026 \quad , \quad \hat{\rho}=0.230\pm0.270.$$

If clustering were ignored then s.e. for $\hat{\pi}$ would be 0.029. Since $\frac{\hat{\rho}}{\text{s.e.}(\hat{\rho})}$ and $\ln(\hat{\lambda}/\hat{\pi}^2)/\text{s.e.}(\ln(\hat{\lambda}/\hat{\pi}^2))$ are both <1.96, ρ can be interpreted as not significantly different from zero and for this study, an analysis considering clustering would not be necessary.

3.3 The Two-stage Nested Design with Two Treatments per Cluster

In Section 3.2.1, the theoretical framework for the two-stage nested design with one treatment (sub-population) per cluster was presented. The theory discussed there is now extended for the case of two treatments (sub-populations) per cluster along the lines discussed in Kempthorne and Koch (1983). There are three possible situations under this design. They are described as 1) after randomly drawing a fixed number of observations from each cluster, they are randomly allocated to each treatment; 2) a fixed number of observations are randomly drawn from sub-population 1 in each cluster and a fixed (usually equal) number of observations are randomly drawn from sub-population 2 in each cluster; 3) a fixed number of observations are first randomly drawn from each cluster and then classified as falling into sub-population 1 and 2. The number of observations classified into each sub-population under this third situation might not be equal. Due to this unbalanced nature, only settings 1 and 2 will be studied here.

The structure of the data in each of these settings involves one type of psu (the cluster) and two types of ssu's (treatments or sub-populations). For example, in a sample survey, men and women (sub-populations) in neighbohoods (clusters) may be asked if they have liberal political views or not. Alternatively, two types of interviewing methods (treatments) may be applied to sets of subjects (units) in neighborhoods (clusters). Each of the two settings of interest involves random selection of clusters. Consider setting (1) where there is a finite population of clusters. In each of these clusters there are D potential observations for treatment 1 and D potential observations for treatment 2 so that there are 2D potential observations for the 2 treatments in each of the N clusters. Suppose a random sample without replacement is obtained from this population of N clusters. From each sampled cluster, 2d observational units are drawn at random and then randomly allocated to the two treatments so that d receive treatment 1 and d receive treatment 2. Now, with regard to setting (2), let each population within a cluster have D observations so that there are 2D observations within each of the N clusters. Let n clusters be randomly sampled without replacement from the population of N clusters. Further suppose that d observations are drawn at random from the D observations of sub-population 1 in each sampled cluster and d observations are drawn at random from the D observations of sub-population 2 in each sampled cluster. So, there are 2d observations in each of the sampled clusters. The sampling framework of primary interest in both of these settings is when $D \rightarrow \infty$ and

$N \to \infty$, in which case, sampling with replacement and sampling without replacement are equivalent.

Let

$$y_{hij} = \begin{cases} 1 & \text{if the j-th observation receiving treatment h in the} \\ & \quad \text{i-th cluster has the attribute under study} \\ 0 & \text{otherwise} \end{cases}$$

where $h=1,2$; $i=1,2,\ldots,n$; $j=1,2,\ldots,d$. In setting (2), h would denote the sub-population rather than treatment. The model of interest is characterized as:

$E(y_{hij}) = \pi_h =$ Pr(an observation receiving treatment h in a randomly selected cluster has the attribute),

$E(y_{hij}y_{hij'}) = \lambda_h =$ Pr(two distinct observations receiving treatment h in a randomly selected cluster both have the attribute),

$E(y_{1ij}y_{2ij'}) = \psi =$ Pr(two observations receiving different treatments in a randomly selected cluster both have the attribute),

The finite population formulas for these parameters are not necessarily the same for settings (1) and (2); but when $D \to \infty$, their formulations are the same and may be expressed in terms of π_{1i} and π_{2i}, the mean responses associated with treatments 1 and 2 in the clusters of the population.

Also of interest are the within cluster treatment correlation coefficients (ρ_h) and the intraclass cross-correlation τ which is the correlation among different subunits of the same cluster receiving different treatments:

$$\rho_h = \frac{\mathrm{Cov}(y_{hij}, y_{hij'})}{\mathrm{Var}(y_{hij})} = \frac{\lambda_h - \pi_h^2}{\pi_h(1-\pi_h)}, \; h=1,2 \qquad (3.2.37)$$

$$\tau = \frac{\mathrm{Cov}\,(y_{1ij}, y_{2ij'})}{\sqrt{\mathrm{Var}(y_{1ij})\mathrm{Var}(y_{2ij})}} = \frac{\psi - \pi_1\pi_2}{\sqrt{\pi_1(1-\pi_1)(1-\pi_2)}}. \qquad (3.2.38)$$

These parameters are measures of the strength of clustering and thus are of interest in the study of clustered data structures.

Various inequalities among the population parameters are of interest in establishing bounds. They are important with respect to estimation of certain covariance matrices. Some quantities of interest may be shown to be non-negative directly, like in expression (3.2.10) in Section 3.2, while other inequalities are established by Cauchy-Schwarz inequality. Bounds concerning the parameters of interest are studied by Kempthorne (1982).

<u>The Completely Balanced Case: Two Treatments (or Sub-populations) per Cluster</u>

Consider a two-stage random sample of n clusters containing 2d observations where the population of clusters is of size N, the population of (potential) observations is of size D, and both N and D tend to infinity. There are d subjets randomly allocated to each treatment within each cluster. This case of the two-stage nested design may be viewed as a generalized randomized block design where two treatments are assigned within each block. Unbiased sample estimators for the basic parameters in this balanced case are

$$\hat{\pi}_h = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hij}}{nd} = \frac{\sum_{i=1}^{n} \bar{y}_{hi\cdot}}{n} \quad h=1,2 \ , \tag{3.2.39}$$

$$\hat{\lambda}_h = \frac{\sum_{i=1}^{n}\sum_{j\neq j'=1}^{d}\sum^{d} y_{hij}\, y_{hij'}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{hi\cdot}^2 - y_{h\cdot\cdot}}{nd(d-1)} \quad h=1,2 \ , \tag{3.2.40}$$

and

$$\hat{\psi} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{1ij}\, y_{2ij'}}{nd^2} = \frac{\sum_{i} \bar{y}_{1i\cdot}\, \bar{y}_{2i\cdot}}{n} \ . \tag{3.2.41}$$

As for the one treatment per cluster situation, the observations y_{hij} may be written as a consequence of the two-stage random sampling process, in terms of a set of components:

$$y_{hij} = \pi + (\pi_h - \pi) + (\pi_{*i} - \pi) + (\pi_{hi} - \pi_h - \pi_{*i} + \pi) + (y_{hij} - \pi_{hi}), \qquad (3.2.42)$$

where $\pi_{hi} = \frac{1}{D}\sum_{j=1}^{D} y_{hij}$, $\pi_{*i} = (\pi_{1i} + \pi_{2i})/2$, $\pi_h = \frac{1}{N}\sum_{i=1}^{N} \pi_{hi}$, and $\pi = (\pi_1 + \pi_2)/2$ are the population parameters associated with y_{hij} ; $(\pi_h - \pi)$ is the fixed effect due to treatment h, $(\pi_{*i} - \pi)$ is the random effect due to cluster i, $(\pi_{hi} - \pi_h - \pi_{*i} + \pi)$ is the random effect associated with the h-th treatment and i-th cluster, and $(y_{hij} - \pi_{hi})$ is the random subsampling residual. Observe that

$$\sum_{h=1}^{2} (\pi_h - \pi) = 0$$

and

$$\sum_{h=1}^{2} (\pi_{hi} - \pi_h - \pi_{*i} + \pi) = 0 \quad \text{for all i.}$$

The sample observation y_{hij} may be presented like in (3.2.16), Section 3.2.1, as

$$y_{hij} = \mu + t_h + a_i + (ta)_{hi} + e_{hij}. \qquad (3.2.43)$$

where $\sum_{h=1}^{2} t_h = 0$ and $\sum_{h=1}^{2} (ta)_{hi} = 0$.

Under a sampling process with replacement, $E\{(ta)_{hi}\} = 0$, $E\{a_i\} = 0$, $E\{e_{hij}\} = 0$, and $Cov\{(a_i, e_{hij}\} = Cov\{(ta)_{hi}, e_{hij}\} = 0$.

Suppose appropriate conditions apply to the model parameters so that one can assume

$$E\{(\pi_{*i} - \pi)^2\} = E\{a_i^2\} = \sigma_a^2$$

$$E\{(\pi_{hi} - \pi_h - \pi_{*i} + \pi)^2\} = E\{(ta)_{hi}^2\} = \sigma_{ta}^2$$

$$E\{(y_{hij}-\pi_{hi})^2\} = E\{e_{hij}^2\} = \sigma_e^2$$

and

$$\mathrm{Cov}\{a_i, (ta)_{hi}\} = 0 \ .$$

Thus, the model for the two-stage sampling process with two treatments per cluster and the classical two-stage nested mixed effects model are analogous to each other and the analysis of variance methods can be used to obtain unbiased estimates of the variance components in the balanced case.

Functions of the population parameters of interest here are:

$$\frac{1}{2}\sum_{h=1}^{2} t_h^2 = \frac{1}{2}\sum_{h=1}^{2}(\pi_h-\pi)^2 = \left(\frac{\pi_1-\pi_2}{2}\right)^2, \tag{3.2.44}$$

$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N}(\pi_{*i} - \pi)^2 = \frac{(\lambda_1-\pi_1^2)+(\lambda_2-\pi_2^2)+2(\psi-\pi_1\pi_2)}{4}\ , \tag{3.2.45}$$

$$\sigma_{ta}^2 = \frac{1}{2N}\sum_{h=1}^{2}\sum_{i=1}^{N}(\pi_{hi}-\pi_h-\pi_{*i}+\pi)^2 = \frac{(\lambda_1-\pi_1^2)+(\lambda_2-\pi_2^2)-2(\psi-\pi_1\pi_2)}{4}, \tag{3.2.46}$$

$$\sigma_e^2 = \frac{1}{2ND}\sum_{h=1}^{2}\sum_{i=1}^{N}\sum_{j=1}^{D}(y_{hij} - \pi_{hi})^2 = \frac{(\pi_1-\lambda_1)+(\pi_2-\lambda_2)}{2}\ , \tag{3.2.47}$$

and

$$\mu^2 = \pi^2 = \left(\frac{\pi_1+\pi_2}{2}\right)^2 . \tag{3.2.48}$$

Thus,

$$(\sigma_a^2-\sigma_{ta}^2)=(\psi-\pi_1\pi_2) \tag{3.2.49}$$

and

$$(\sigma_a^2+\sigma_{ta}^2)=\frac{\{(\lambda_1-\pi_1^2)+(\lambda_2-\pi_2^2)\}}{2} \tag{3.2.50}$$

where $(\sigma_a^2-\sigma_{ta}^2) = \mathrm{Cov}\,(y_{hij}, y_{h'ij'})$ and $(\sigma_a^2+\sigma_{ta}^2) = \frac{1}{2}\sum_{h=1}^{2}\mathrm{Cov}(y_{hij}, y_{hij'})$.

It is important to note from the above expressions that the ANOVA method does not give estimates of the principal functions of the parameters such as $(\pi_h-\lambda_h)$, $(\lambda_h-\pi_h^2)$, or $(\psi-\pi_1\pi_2)$ but estimates of linear functions of these principal functions of the parameters. By using expressions (3.2.45), (3.2.46) and (3.2.47) one may derive estimates for the variance components from estimators such as (3.2.39)-(3.2.41) for the π_h, λ_h, and ψ. For example

$$s^2 = \frac{\sum_{h=1}^{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{hij}-\bar{y}_{hi\cdot})^2}{2n(d-1)} = \hat{\sigma}_e^2 = \tfrac{1}{2}\{(\pi_1\hat{-}\lambda_1)+(\pi_2\hat{-}\lambda_2)\}. \qquad (3.2.51)$$

A convenient way to estimate π_1, π_2, $(\lambda_1-\pi_1)$, $(\lambda_2-\pi_2)$, and ψ is through means of appropriate functions. This also facilitates the construction of estimates of the corresponding covariance matrix.

Let for each i, $F_{i1}=\bar{y}_{1i.}$, $F_{i2}=\bar{y}_{2i.}$, $F_{i3}=s_{1i}^2$, $F_{i4} = s_{2i}^2$ and $F_{i5} = \bar{y}_{1i\cdot}\bar{y}_{2i\cdot}$.

Then

$$\underset{\sim}{F} = \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ (\hat{\pi}_1-\hat{\lambda}_1) \\ (\hat{\pi}_2-\hat{\lambda}_2) \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \end{bmatrix} = \begin{bmatrix} \bar{y}_{1\cdot\cdot} \\ \bar{y}_{2\cdot\cdot} \\ s_1^2 \\ s_2^2 \\ \hat{\psi} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \bar{y}_{1i\cdot} \\ \bar{y}_{2i\cdot} \\ s_{1i}^2 \\ s_{2i}^2 \\ \bar{y}_{1i\cdot}\bar{y}_{2i\cdot} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} F_{i1} \\ F_{i2} \\ F_{i3} \\ F_{i4} \\ F_{i5} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_i \qquad (3.2.52)$$

and

$$V(\underset{\sim}{F}) = \frac{1}{n^2}\sum_{i=1}^{n} (\underset{\sim}{F}_i - \underset{\sim}{F})(\underset{\sim}{F}_i - \underset{\sim}{F})' \doteq \mathrm{Var}\begin{bmatrix} \bar{y}_{1\cdot\cdot} \\ \bar{y}_{2\cdot\cdot} \\ s_1^2 \\ s_2^2 \\ \hat{\psi} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \pi_1\hat{-}\lambda_1 \\ \pi_2\hat{-}\lambda_2 \\ \hat{\psi} \end{bmatrix} \qquad (3.2.53)$$

Alternatively, the observations y_{hij} can be grouped in such a way that a 2 way table can be formed with the observed number of clusters with j successes on treatment 1 and j′ successes on treatment 2 as the frequency of cell counts. Such compact structure for the data is presented in Table 3.4 where n clusters are sub-divided into (d+1)groups based on the number of units receiving treatments 1and 2 that have the attribute of interest.

Table 3.4: Frequencies of Clusters with j Successes on Treatment 1 and j' Successes on Treatment 2.

	j'	Treatment 2					
j		0	1	2	...	d	total
	0	f_{00}	f_{01}	f_{02}	...	f_{0d}	f_{0+}
	1	f_{10}	f_{11}	f_{12}	...	f_{1d}	f_{1+}
Treatment 1	2	f_{20}	f_{21}	f_{22}	...	f_{2d}	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	d	f_{d0}	f_{d1}	f_{d2}	...	f_{dd}	f_{d+}
	total	f_{+0}	f_{+1}	f_{+2}	...	f_{+d}	$f_{++}=n$

Note that f_{j+} = number of clusters with j successes on treatment 1 and f_{+j} = number of clusters with j successes on treatment 2.

Here

$$\sum_{i=1}^{n} y_{1i\cdot}^{k} = \sum_{j=0}^{d} j^{k} f_{j+}$$

$$\sum_{i=1}^{n} y_{2i\cdot}^{k} = \sum_{j=0}^{d} j^{k} f_{+j} \,.$$

This grouped data formulation gives

$$\bar{y}_{1\cdot\cdot} = \hat{\pi}_{1\cdot\cdot} = \frac{\sum_{j=0}^{d} jf_{j+}}{nd} = \sum_{j=0}^{d} \left(\frac{j}{d}\right) p_{j+}$$

$$\bar{y}_{2\cdot\cdot} = \hat{\pi}_{2} = \frac{\sum_{j=0}^{d} f_{+j}}{nd} = \sum_{j=0}^{d} \left(\frac{j}{d}\right) p_{+j}$$

$$s_1^2 = (\pi_1 \hat{-} \lambda_1) = \frac{\sum_{j=0}^{d} j(d-j)f_{j+}}{nd(d-1)} = \sum_{j=0}^{d} \left[\frac{j(d-j)}{d(d-1)}\right] p_{j+}$$

$$s_2^2 = (\hat{\pi_2 - \lambda_2}) = \frac{\sum_{j=0}^{d} j(d-j)f_{+j}}{nd(d-1)} = \sum_{j=0}^{d} \left[\frac{j(d-j)}{d(d-1)}\right] p_{+j}$$

and

$$\hat{\psi} = \frac{\sum_{j=0}^{d}\sum_{j'=0}^{d} jj' f_{jj'}}{nd^2} = \sum_{j=0}^{d}\sum_{j'=0}^{d} \left(\frac{jj'}{d^2}\right) p_{jj'} ,$$

where observed proportions p_{j+} and p_{+j} are shown in Table 3.5.

Table 3.5: Observed proportions of clusters with j successes on treatment 1 and j′ successes on treatment 2.

	j′	Treatment 2					
j		0	1	2	...	d	Total
	0	p_{00}	p_{01}	p_{02}	...	p_{dd}	p_{0+}
Treatment 1	1	p_{10}	p_{11}	p_{12}	...	p_{1d}	p_{1+}
	2	p_{20}	p_{21}	p_{22}	...	p_{2d}	p_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	d	p_{d0}	p_{d1}	p_{d2}	...	p_{dd}	p_{d+}
	Total	p_{+0}	p_{+1}	p_{+2}	...	p_{+d}	p_{++}

Since these five parameter estimators are linear functions of the frequency counts $f_{jj'}$, they can be expressed in matrix notation. Using the GSK methodology described in Section 2.3, the vector of response functions is $\underset{\sim}{p}=(p_{00}, p_{01}, \ldots, p_{dd})$. To perform the WLS analysis it is assumed that the observed frequencies $\{f_{jj'}\}$ follow a multinomial distribution with parameters n and $\underset{\sim}{\zeta}=(\zeta_{00}, \ldots, \zeta_{dd})$.

The vector of parameter estimators may be expressed in matrix notation as

$$\begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{s}_1^2 \\ \hat{s}_2^2 \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ (\hat{\pi_1 - \lambda_1}) \\ (\hat{\pi_2 - \lambda_2}) \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{1}' \otimes \underset{\sim}{u}' \\ \underset{\sim}{u}' \otimes \underset{\sim}{1}' \\ \underset{\sim}{1}' \otimes w' \\ \underset{\sim}{w}' \otimes \underset{\sim}{1}' \\ \underset{\sim}{u}' \otimes \underset{\sim}{w}' \end{bmatrix} \underset{\sim}{p} = \underset{\sim}{A}_1\underset{\sim}{p} = \underset{\sim}{F} \quad (3.2.54)$$

where

$$\underline{u}' = \left[0, \frac{1}{d}, \frac{2}{d}, \dots \frac{d-1}{d}, 1\right],$$

$$\underline{w}' = \left[0, \frac{1}{d}, \frac{2(d-2)}{d(d-1)}, \dots, \frac{1}{d}, 0\right]$$

and

$$\underline{1}' = [1, 1, \dots, 1]$$

are vectors of dimension (d+1). The covariance matrix of this vector of parameter estimators is estimated by

$$\mathrm{Var}\begin{bmatrix} \hat{\pi}_1 \\ \dots \\ \hat{\psi} \end{bmatrix} = \underline{A}_1 \underline{V}_1(\underline{p})\ \underline{A}_1' = \underline{V}(\underline{F}) \qquad (3.2.55)$$

The parameters of interest in this two treatment per cluster case of the two-stage nested design are π_1, π_2, ρ_1, ρ_2 and τ. These may be estimated jointly by applying a series of linear, logarithmic and exponential transformations to the estimated parameter vector $(\hat{\pi}_1, \hat{\pi}_2, s_1^2, s_2^2, \hat{\psi})'$. If the correlation coefficients are rewritten as

$$\hat{\rho}_h = \frac{\hat{\lambda}_h - \hat{\pi}_h^2}{\hat{\pi}_h(1-\hat{\pi}_h)} = 1 - \frac{\hat{\pi}_h - \hat{\lambda}_h}{\hat{\pi}_h(1-\hat{\pi}_h)} = 1 - \frac{s_h^2}{\hat{\pi}_h(1-\hat{\pi}_h)} \quad h=1,2 \qquad (3.2.56)$$

and

$$\hat{\tau} = \frac{\hat{\psi} - \hat{\pi}_1\hat{\pi}_2}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)\hat{\pi}_2(1-\hat{\pi}_2)}} = \left(\frac{\psi}{\hat{\pi}_1\hat{\pi}_2} - 1\right)\left\{\left(\frac{\hat{\pi}_1}{1-\hat{\pi}_1}\right)\left(\frac{\hat{\pi}_2}{1-\hat{\pi}_2}\right)\right\}^{\frac{1}{2}} \qquad (3.2.57)$$

it follows that

$$\begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\rho}_1 \\ \hat{\rho}_2 \\ \hat{\tau} \end{bmatrix} = \underset{\sim}{\text{exp}}\ \{\underset{\sim}{A}_5 \underset{\sim}{\text{ln}}[\underset{\sim}{A}_4\{\underset{\sim}{C}_2 + \underset{\sim}{\text{exp}}[\underset{\sim}{A}_3 \underset{\sim}{\text{ln}}(\underset{\sim}{C}_1 + \underset{\sim}{A}_2 \underset{\sim}{A}_1 \underset{\sim}{F})]\}]\} \qquad (3.2.58)$$

where

$$\underset{\sim}{A}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \underset{\sim}{C}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\underset{\sim}{A}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & -0 \end{bmatrix}, \quad \underset{\sim}{C}_2 = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

$$\underset{\sim}{A}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \text{and } \underset{\sim}{A}_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

(3.2.59)

If the design is constructed so that there are M modules with different cluster sizes, then representation of $\underset{\sim}{F}$ becomes

$$\underset{\sim}{F} = (\hat{\pi}_{1,1}, \hat{\pi}_{2,1}, s_{1,1}^2, s_{2,1}^2, \hat{\psi}_1, \ldots, \hat{\pi}_{1,M}, \hat{\pi}_{2,M}, s_{1,M}^2, s_{2,M}^2, \hat{\psi}_M)' \qquad (3.2.60)$$

where the elements of $\underset{\sim}{F}$ may be presumed to have an approximately multivariate normal distribution, provided the number of clusters $\{n_m\}$ are sufficiently large for all $m(m=1,2,\ldots,M)$. A model analogous to (3.2.35) is then fit by WLS.

The study of treatment of mastitis in dairy cows described in Section 1.3 is used to illustrate the balanced data case of a two-stage nested design with two sub-populations per cluster (Kempthorne and Koch (1983)). Each cow's pre-treatment status (diseased versus not diseased) was determined; Table 3.6 summarizes this information. Questions of interest are 1) whether there is a difference between the proportion of diseased quarters on the front of the udder (π_F) and proportion of diseased quarters on the back of the udder (π_B) and 2) whether the intraclass correlations for front and back quarters (ρ_F and ρ_B) and the intraclass cross-correlation (τ) are significantly different from one another as well as zero.

Table 3.6: Number of Cows with 0, 1, or 2 Diseased Quarters

		Number of Cows with j Diseased Front Quarters			
		0	1	2	Total
Number of Cows	0	7	6	3	16
with j' Diseased	1	81	48	25	154
Back Quarters	2	11	25	69	104
Total		99	78	97	274

This is an example of setting (2). The front and back portions of the udder constitute the two sub-populations in each cluster, i.e. , each quarter of the udder is either diseased or not diseased. Since the data are balanced, a grouped data formulation is used to examine the questions of interest. The data for this example is presented in Table 3.6. By applying expression (3.2.54) to the (9 × 1) vector of observed proportions $\underset{\sim}{p}$ obtained from Table 3.6 the vector of parameter estimates can be obtained. Table 3.7 shows these estimates and standard errors.

Table 3.7: Parameter Estimates and Standard Errors: Dairy Science Example

Parameter function	π_F	π_B	ρ_F	ρ_B	τ
Estimate	0.496	0.661	0.431	0^{+}	0.241
s.e.	(0.026)	(0.017)	(0.055)	$-^{+}$	(0.028)

$^{+}\rho_B = 0$

Results from Table 3.7 indicate that the mean number of diseased quarters on the back portion of the udder is significantly greater than the mean number of diseased quarters on the front portion of the udder. The intraclass correlation for front quarters and intraclass cross correlation coefficient are both significantly different from zero ($\frac{\hat{\rho}}{\text{s.e.}\hat{\rho}}$>1.96 and $\frac{\ln(\hat{\lambda}/\hat{\pi}^2)}{\text{s.e.}\ln(\hat{\lambda}/\hat{\pi}^2)}$> 1.96 for both correlation coefficients). The intraclass correlation coefficient for back quarters is estimated as zero since $(\hat{\lambda}_B-\hat{\pi}_B^2)<0$. The variance of π_F and of π_B, if clustering were ignored, would be $\text{Var}(\hat{\pi}_F)$=0.00091 (s.e.=0.030) and $\text{Var}(\hat{\pi}_B)$=0.00082 (s.e.=0.027). Also results for s.e.$(\pi_F-\pi_B)$ with and without accounting for clustering are 0.023 and 0.041 respectively. So, an analysis adjusting for clustering seems appropriate.

Studies like the ones described in Chapter I are suitable to analyses based on clustered attribute data methodology. The day care center study is one such example. Several ways of applying WLS methods will be illustrated with this dataset. One way will be for a two-stage nested design setting with one treatment per cluster if day care center is the primary unit and staff is the secondary unit of analysis (staff viewed as conceptually representative of a corresponding large sub-population in the sense equivalent to simple random sampling). Another way will be to take into consideration answers given by parents of day care children. With addition of this new information, the data may be viewed as a two stage nested design situation with two sub-populations, staff and parents, per cluster. Different linear model descriptions will also be pursued for this day care center data where the simplest one will deal with (within) day care means of staff's answers analyzed across day care centers (Chapter IV). A more sophisticated description will attempt to develop extensions of the methodology presented in Sections 3.2 and 3.3 where instead of observing a binary response on each unit, a set of binary responses is of interest (each staff member answers yes/no to a set of questions). Chapters V, VI and VII will address these issues, via definitions of extended models along with their parameters of interest.

CHAPTER IV

DAY CARE CENTER ANALYSIS: A REPEATED MEASUREMENTS STUDY WITH CLUSTERED ATTRIBUTE DATA

4.1 Introduction

This chapter describes applications of weighted least squares methods to analyze a series of categorical responses that pertain to cross-classified scenarios for categorical choices of actions related to the health and social service areas. WLS methods for the analysis of repeated measures of categorical data for the chosen responses from a day care center study is presented with a refinement to deal with correlated categorical data, taking into account the possible clustering of the data. Two examples dealing with day care centers are discussed. These examples evolve from a study of day care centers criteria for sending sick children home. The study deals with day care center staff who are asked questions of what to do under the situations where the child may have fever, other illness symptoms or a combination of these two factors. There, the questions of interest are:

1. How day care center staff take care of children who are perceived to be sick;
2. What specific symptoms or signs in children prompt day care center staff to act;
3. Do these actions vary with the center characteristic of being certified (i.e., allowing for partial reimbursement from the government for needy children) or non-certified.

Analysis of these multivariate data is done using the computer program MISCAT.

4.2 WLS methodology for repeated measurements

A straightforward framework for repeated measurements studies involves a set of subjects who often come from two or more subpopulations, with subjects in each group (or subpopulation) having their responses observed for a distinct set of conditions. These conditions can be based on a single factor or a cross-classification of two or more factors. For this framework, questions of interest are:

1. evaluation of group effects
2. evaluation of condition effects
3. description of the pattern of variation of response distributions across groups and conditions.

For this study, attention is focused on responses from $s=2$ groups of subjects. The response for the i-th subject in the h-th group for the k-th condition is denoted by y_{hik}; $h=1,2,\ldots,s$; $i=1,2,\ldots,n_h$; $k=1,2,\ldots,r$. Each subject's response is classified into two possible values (a binary observation) indexed by $l=0$ or $l=1$. Weighted least squares methods, as described by Grizzle, Starmer and Koch (1969), are useful for describing the variation among groups and conditions for certain summary statistics, such as proportions, means and ratios, through regression models. Here, attention is given to the extensions of these methods provided by Koch, et al. (1977) for repeated measurements studies and Koch, et al. (1986) for longitudinal data. They are directed at the description of the variation of the mean scores $\mu_{hk} = \sum_{l=0}^{1} a_l \, \phi_{hkl}$ with respect to $a_l=0$ and $a_l=1$ (binary case), across groups and conditions, where $\phi_{hkl} = \Pr(y_{hjk}=l)$. The binary response situation leads to $\mu_{hk} = \phi_{hkl}$.

An estimator for μ_{hk} is $\bar{g}_{hk} = \sum_{i=1}^{n_h} g_{hik} \,/\, n_h$ where $g_{hik} = a_l$ for $y_{hik}=l$; $\bar{g}_{hk}$ is the across subjects mean of the scores a_l ($l=0,1$) for the h-th group and the k-th condition. Let $\underset{\sim}{g}_{hi*} = (g_{hi1}, \ldots, g_{hir})'$ denote the (r x 1) vector of response scores for the i-th subject in the h-th group. When the groups of subjects from the respective subpopulations are conceptually equivalent to simple random samples and are sufficiently large (e.g., $n_h \geq r + 25$), then $\bar{\underset{\sim}{g}}_h = \sum_{i=1}^{n_h} \underset{\sim}{g}_{hi*}/n_h$ has approximately a multivariate normal distribution. A consistent estimate of the covariance matrix is given by

$$\underset{\sim}{V}_{g,h} = \frac{1}{n_h^2} \sum_{i=1}^{n_h} (\underset{\sim}{g}_{hi*} - \bar{\underset{\sim}{g}}_h)(\underset{\sim}{g}_{hi*} - \bar{\underset{\sim}{g}}_h)'. \quad (4.2.1)$$

A general class of linear models for describing the variation among the μ_{hk} for the cross-classification of the s=2 groups and d conditions can be expressed as $\mu = \underset{\sim}{X}\underset{\sim}{\beta}$ where $\underset{\sim}{\mu} = (\mu_{11}, \mu_{12}, ..., \mu_{1r}, \mu_{21}, \mu_{22}, ..., \mu_{2r})'$; $\underset{\sim}{X}$ is a known (2r x t) design matrix of full rank $t \leq 2r$ and $\underset{\sim}{\beta}$ is the unknown (t x 1) vector of model parameters.

The weighted least squares estimators for $\underset{\sim}{\beta}$ are given by

$$\underset{\sim}{b} = \left(\underset{\sim}{X}' \underset{\sim}{V}_{\bar{g}}^{-1} \underset{\sim}{X}\right)^{-1} \underset{\sim}{X}' \underset{\sim}{V}_{\bar{g}}^{-1} \bar{\underset{\sim}{g}} \quad (4.2.2)$$

where $\bar{\underset{\sim}{g}} = (\bar{\underset{\sim}{g}}_1', \bar{\underset{\sim}{g}}_2')'$ is the composite vector of means $\bar{g}_{hk}$ for all (group x condition) combinations; also $\underset{\sim}{V}_{\bar{g}}$ is the (2r x 2r) block diagonal matrix with $\underset{\sim}{V}_{g,h}$ as diagonal blocks and serves as a consistent estimate for the covariance matrix of $\bar{\underset{\sim}{g}}$.

The goodness of fit for the model can be assessed through the Wald statistic

$$Q_W = \left(\bar{\underset{\sim}{g}} - \underset{\sim}{X}\, \underset{\sim}{b}\right)' \underset{\sim}{V}_{\bar{g}}^{-1} \left(\bar{\underset{\sim}{g}} - \underset{\sim}{X}\, \underset{\sim}{b}\right) = \bar{\underset{\sim}{g}}' \underset{\sim}{W}' \left(\underset{\sim}{W}\, \underset{\sim}{V}_{\bar{g}}\, \underset{\sim}{W}'\right)^{-1} \underset{\sim}{W}\bar{\underset{\sim}{g}} \quad (4.2.3)$$

for which $\underset{\sim}{W}$ is a full rank orthocomplement to $\underset{\sim}{X}'$; the statistic Q_W has an approximate chi-square distribution with (2r−t) degrees of freedom (d.f.). For models considered appropriate, the linear hypothesis $H_0{:}\underset{\sim}{C}\ \underset{\sim}{\beta} = \underset{\sim}{0}_C$ can be tested with the statistics $Q_C = \underset{\sim}{b}'\ \underset{\sim}{C}'\ (\underset{\sim}{C}\ \underset{\sim}{V}_b \underset{\sim}{C}')^{-1}\underset{\sim}{C}\underset{\sim}{b}$ which has approximately a chi-square distribution with d.f.=c under the null hypothesis.

Finally, predicted values for μ are defined by $\hat{\underset{\sim}{\mu}} = \underset{\sim}{X}\underset{\sim}{b}$ and a consistent estimate for their covariance matrix is given by $V_{\hat{\mu}} = \underset{\sim}{X}\underset{\sim}{V}_b\underset{\sim}{X}'$ where $\underset{\sim}{V}_b$ is the estimated covariance matrix for $\underset{\sim}{b}$.

Extensions for the previously described methods can be specified for situations with s>2 subpopulations and for functions other than the mean scores μ_{hk}. For other functions of the first order marginal distributions $\{\phi_{hkl}\}$ or of higher order marginal distributions, see Koch, et al. (1986). The application of weighted least squares methods to the mean scores μ_{hk} for responses by day care center staff to questions concerning children with illness symptoms is illustrated with examples in Section 4.3

4.3 Examples

All licensed Day Care Centers located in three North Carolina counties, as of January 1985, identified from the North Carolina Day Care Licensing Board, were stratified by their "level" into two groups; certified centers (C) and uncertified centers (NC). Thirty three of 126 uncertified centers were randomly selected, and all 29 certified centers not previously used in a pretest were included in the study. All staff working in each of the selected DCC's were included as the study population. Each staff member completed a self-administrative questionnaire which asked among other questions, how staff would handle mildly ill children given the presence or absence of fever and presence or absence of specific symptoms/signs. Only a subset of the questions asked will be addressed here. The responses to questions of temperature alone, symptoms/signs alone and combinations of temperature and symptoms were: "Do nothing," "Tell parent at the end of the day," "Call parent to tell them," and "Call parent for immediate pick-up". In total, the data comprised information from 302 questionnaires. Some questions were specific to care of children in a certain age bracket. Staff members of DCC's that did not handle children of that specific age left those questions blank.

Since it is recognized that staff within the same center may tend to provide relatively similar responses, first day care center is viewed as a cluster and staff members as the subunits of a cluster; and second, subjects or staff are viewed as independent.

4.3.1 Analysis for clustered data

The study of day care center criteria to send sick children home involves a data structure with possibly some degree of clustering. If day care center is considered as the unit of analysis, WLS methods to account for the correlation structure of multiple staff within a day care center can be applied by working with (within) day care center means of staff's answers analyzed across day care centers. Accounting for the clustered structure of the data in the analyses is relevant so that applicable variances are not underestimated (or overestimated).

The estimation methods described in Section 4.2 are revised to focus on day care centers as primary, and staff members as secondary units of analysis. For this situation, the response for the j-th subject in the i-th cluster, h-th group and k-th condition is denoted by y_{hijk}, with $i=1,2, \ldots, n_h$; $h=1,2, \ldots s$; $k=1,2, \ldots, r$; and $j=1,2, \ldots, d_{hi}$. An estimator for μ_{hk}, the mean score, is given by

$$\tilde{g}_{h*k} = \frac{\sum_{i=1}^{n_h} \bar{g}_{hik}}{n_h} = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{d_{hi}} \left(g_{hijk} \big/ d_{hi}\right)}{n_h} . \tag{4.3.1}$$

where $g_{hijk}=a_l$ for $y_{hijk}=l$. Let $\bar{\underset{\sim}{g}}_{hi*}=(\bar{g}_{hi1}, \bar{g}_{hi2}, \ldots, \bar{g}_{hir})'$ denote the response vector of mean scores for the h-th group and the i-th cluster then

$$\tilde{\underset{\sim}{g}}_h = \frac{\sum_{i=1}^{n_h} \bar{\underset{\sim}{g}}_{hi*}}{n_h} \tag{4.3.2}$$

represents the vector of mean of means scores at the k-th condition for the h-th group.

If the sample sizes n_h are sufficiently large, then the $\{\tilde{\underset{\sim}{g}}_h\}$ approximately have multivariate normal distributions and consistent estimates for the corresponding covariance matrices are

$$V_{\tilde{g}h} = \frac{1}{n_h^2} \sum_{i=1}^{n_h} (\bar{g}_{hi*} - \tilde{g}_h)(\bar{g}_{hi*} - \tilde{g}_h)' . \qquad (4.3.3)$$

<u>Example 1: Choice for Immediate Pick-up When Type of Day Care Center, Age of Child and Level of Temperature are taken into account</u>

This first example is concerned with staff's choice of calling the parent for immediate pick-up for a child with diarrhea, taking into consideration both types of DCC's (C and NC), scenarios encompassing two age groups (age < 2, 2 < age < 5) and three levels of temperature (99° - 99.9°F, 100° - 100.9°F, 101° - 101.9°F).

Analysis of the proportion of staff choosing immediate pick-up is done by first transforming the responses to the indicator

$$g_{hijk} = \begin{cases} 1 & \text{if the response } y_{hijk} \text{ is call parent for pick up} \\ 0 & \text{if otherwise} \end{cases} \qquad (4.3.4)$$

then calculating the mean of means $\tilde{g}_{h*k} = \frac{\sum_{i=1}^{n_h} \bar{g}_{hik}}{n_h}$ by calculating the proportion of staff choosing immediate pick-up within a DCC and then averaging these proportions across DCC's for the 2 groups (certified (level 0) and non-certified (level 2) DCC's). There are $n_1 = 17$ certified and $n_2 = 23$ non-certified DCC's with both age groups. The (12 x 12) covariance matrix $V_{\tilde{g}}$ has the form

$$V_{\tilde{g}} = \begin{bmatrix} V_{\tilde{g},1} & 0_{6,6} \\ 0_{6,6} & V_{\tilde{g},2} \end{bmatrix} \qquad (4.3.5)$$

where $V_{\tilde{g},1}$ and $V_{\tilde{g},2}$ are the estimated covariance matrices for the two types of DCC's obtained from (4.3.3). The vector of means $\tilde{g} = (\tilde{g}_1, \tilde{g}_2)'$ for the respective (temperature x DCC level x age of child) combinations and the corresponding standard errors are given in Table 4.1.

Table 4.1: Observed and Predicted Staff's Choice of Immediate Pick-up Averaged Proportions and Standard Errors for Example 1

Level	Age	Temperature	Observed proportion	s.e.	Predicted proportion	s.e.
0	<2	99°- 99.9°F	0.162	0.061	0.127	0.026
0	<2	100°-100.9°F	0.369	0.078	0.348	0.045
0	<2	101°-101.9°F	0.646	0.078	0.630	0.062
0	2-5	99°- 99.9°F	0.071	0.045	0.039	0.023
0	2-5	100°-100.9°F	0.275	0.072	0.260	0.041
0	2-5	101°-101.9°F	0.557	0.078	0.542	0.053
2	<2	99°- 99.9°F	0.089	0.040	0.127	0.026
2	<2	100°-100.9°F	0.395	0.070	0.348	0.045
2	<2	101°-101.9°F	0.870	0.053	0.860	0.035
2	2-5	99°- 99.9°F	0.039	0.030	0.039	0.023
2	2-5	100°-100.9°F	0.265	0.068	0.260	0.041
2	2-5	101°-101.9°F	0.787	0.061	0.772	0.031

In the absence of an apriori model for the means $\tilde{g}_{h*k}$, the cell mean (or identity) model $\underset{\sim}{X}_1 = \underset{\sim}{I}_{12}$ is used to obtain a preliminary assessment of the sources of variation across the two subpopulations (levels of DCC's) and the six scenarios.

Table 4.2 shows preliminary hypotheses of interest, with their corresponding contrast matrices relative to the cell mean model $\underset{\sim}{X}_1$.

The first three entries address respectively the null hypotheses of no variation among (i) the two subpopulations, (ii) the two age groups and (iii) the three levels of temperature; test statistics for hypotheses H_2 and H_3 are significant (α=0.05). Tests for the interaction of age x DCC level, age x temperature level and age x temperature x DCC level show non-sigificant results (α=0.05). However, for the null hypothesis H_5, no DCC level x temperature level interaction, the Wald statistic of 14.93 (d.f.=2) is significant with $P<0.01$. Because of this significant interaction, a model including main effects and the DCC level x temperature interaction is fit to the data.

Table 4.2: Results of Hypothesis Tests from Cell Mean Model for Staff's Choice of Immediate Pick Up Proportions for Cluster Analysis of Example 1

Hypothesis	C Matrix	Approximate Q_C (d.f.)	p-value
H_1: No difference between the 2 DCC levels	$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$	0.83 (1)	0.36
H_2: No difference between the 2 age groups	$\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}$	12.87 (1)	<0.01
H_3: No difference among the 3 levels of temperature	$\begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \end{bmatrix}$	257.64 (2)	<0.01
H_4: No age group vs. DCC level interaction	$\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$	0.0007 (1)	1.00
H_5: No DCC level vs. temp. level interaction	$\begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix}$	14.93 (2)	<0.01
H_6: No age group vs. temp. level interaction	$\begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix}$	1.54 (2)	0.46
H_7: No age group vs. temp. level vs. DCC level	$\begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix}$	0.97 (2)	0.62

The design matrix $\underset{\sim}{X}_2$ along with $\underset{\sim}{\beta}_2$ the vector of parameters have the form

$$\underset{\sim}{X}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \quad \underset{\sim}{\beta}_2 = \begin{bmatrix} \beta_{2,1} \\ \beta_{2,2} \\ \beta_{2,3} \\ \beta_{2,4} \\ \beta_{2,5} \\ \beta_{2,6} \\ \beta_{2,7} \end{bmatrix} \tag{4.3.6}$$

where $\beta_{2,1} =$ reference cell parameter corresponding to age<2, NC DCC and temperature of 99-99.9 F.

$\beta_{2,2} =$ increment for C DCC's for temperature of 99-99.9 F.

$\beta_{2,3} =$ increment for 2<age<5.

$\beta_{2,4} =$ increment for temperature of 100-100.9 F for NC DCC's

$\beta_{2,5} =$ increment for temperature of 101-100.9 F for NC DCC's

$\beta_{2,6} =$ interaction increment for C DCC's and temperature of 100-100.9 F.

$\beta_{2,7} =$ interaction increment for C DCC's and temperature of 101-101.9 F.

The use of this model is supported by the non-significance ($\alpha=0.25$) of the Wald Statistic $Q_W=1.24$ (d.f.=5).

Statistical tests for model parameters indicate that the model $\underset{\sim}{X}_2$ can be simplified by removal of $\beta_{2,2}$ and $\beta_{2,6}$. Thus, consideration is given to the final model $\underset{\sim}{X}_3$ and its parameters $\underset{\sim}{\beta}_3$ where

$$\underset{\sim}{X}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix}, \quad \underset{\sim}{\beta}_3 = \begin{bmatrix} \beta_{3,1} \\ \beta_{3,2} \\ \beta_{3,3} \\ \beta_{3,4} \\ \beta_{3,5} \end{bmatrix} \qquad (4.3.7)$$

for which $\beta_{3,1}$, $\beta_{3,2}$, $\beta_{3,3}$, $\beta_{3,4}$, $\beta_{3,5}$ have the same interpretation as their respective counterparts $\beta_{2,1}$, $\beta_{2,3}$, $\beta_{2,4}$, $\beta_{2,5}$, $\beta_{2,7}$ for the model $\underset{\sim}{X}_2$. The use of the model $\underset{\sim}{X}_3$ is supported by the non-significance (α=0.25) of the Wald statistic Q_W=4.56 with d.f.=7 and p=0.71. The estimated model parameters are $\underset{\sim}{b}_3$ = [0.127, -0.088, 0.221, 0.503, 0.230]′ and their estimated covariance matrix is

$$\underset{\sim}{V}_{\underset{\sim}{b}_3} = \begin{bmatrix} 66.21 & -26.46 & -4.25 & 7.03 & -57.13 \\ -26.47 & 41.11 & -10.02 & -46.83 & 38.83 \\ -4.25 & -10.02 & 146.13 & 121.08 & -55.07 \\ 7.03 & -46.83 & 121.20 & 305.18 & -231.02 \\ 57.13 & 38.83 & -55.07 & -231.42 & 311.71 \end{bmatrix} \times 10^{-5}$$

Predicted values along with standard errors for this final model are presented in the last two columns of Table 4.1.

We conclude from this example that age and temperature of the perceived ill child are factors that contribute significant information towards staff's choice for immediate pick-up. Predicted percentages of pick-up for sick children of age less than two years are higher than those for children between ages of two and five years. As for temperature level, predicted percentages increase with temperature. It is interesting to note that staff's choice for immediate pick-up predicted percentages for temperature bracket of 101°-101.9° F are higher for certified centers than for uncertified centers.

Example 2: Choice for Immediate Pick-Up for Children Between Ages of 2 and 5 Years, Three Symptoms/Signs of Illnesses and Elevated or Normal Temperature

In this example, weighted least squares methods are applied to data involving only children between ages of 2 and 5 years, with one of the three symptoms crankiness, diarrhea or conjunctivitis and in the presence (100°-100.9° F) or absence (<100°F) of elevated temperature. In order to study the proportion of staff choosing immediate pick-up under this new set of conditions, we first dichotomize the responses of interest as in (4.3.4) of Example 1; then the proportion of staff choosing to call parent is calculated for each day care center and then averaged across day care centers for each condition of interest. These means of means for the (temperature x symptoms) computations by level of DCC are shown in Table 4.3 together with their corresponding estimated standard errors.

Table 4.3: Observed and predicted staff's choice of immediate pick-up average proportions and standard errors for Example 2

		Uncertified Day Care Ctr				Certified Day Care Ctr			
		Observed		Predicted		Observed		Predicted	
Fever	Symptom	Prep.	s.e.	Prep.	s.e.	Prep.	s.e.	Prep.	s.e.
No	None	-	-	-	-	-	-	-	-
No	Crankiness	0.07	0.041	0.04	0.035	0.03	0.013	0.02	0.012
No	Diarrhea	0.46	0.075	0.51	0.061	0.74	0.036	0.71	0.027
No	Conjunctivitis	0.57	0.067	0.51	0.061	0.67	0.056	0.71	0.027
Yes	None	0.30	0.069	0.37	0.057	0.47	0.061	0.49	0.051
Yes	Crankiness	0.37	0.067	0.37	0.057	0.48	0.056	0.49	0.051
Yes	Diarrhea	0.65	0.069	0.70	0.042	0.83	0.042	0.84	0.041
Yes	Conjunctivitis	0.73	0.050	0.70	0.042	0.84	0.050	0.84	0.041

Since an apriori model is not available, the cell mean (or identity) model $\underset{\sim}{X}_1 = I_7$ is used to obtain a preliminary assessment of the statistical significance for the sources of variation concerning the cross-classification of (temperature x symptoms) for each of the two levels of DCC's separately.

Table 4.4 shows the preliminary hypotheses of interest along with C matrices for both levels of DCC's. It can be seen that there is no statistically significant difference between elevated temperature alone and crankiness plus fever nor between the signs of diarrhea and conjunctivitis in the presence or absence of elevated temperature for either type of DCC. For uncertified DCCs there is no elevated temperature by symptom interaction. In view of these findings, the following tentative model is examined for goodness of fit, for each type of day care center separately.

$$X_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{2,1} \\ \beta_{2,2} \\ \beta_{2,3} \\ \beta_{2,4} \end{bmatrix} = X_2\beta_2 \qquad (4.3.8)$$

The parameter $\beta_{2,1}$ represents the predicted reference value for crankiness and no elevated temperature $\beta_{2,2}$ represents the effect of diarrhea and conjunctivitis under no elevated temperature, $\beta_{2,3}$ represents the effect of elevated temperature alone and crankiness in the presence of elevated temperature and $\beta_{2,4}$ represents the effect of diarrhea and conjunctivitis in the presence of elevated temperature. The use of this model is supported by the non-significance ($\alpha=0.10$) of the weighted least squares (or Wald) statistic Q_W,

$$Q_W = \begin{cases} 4.68 \text{ (d.f.=3, p=0.20) (NC)} \\ \\ 1.58 \text{ (d.f.=3, p=0.66) (C)} \end{cases} \qquad (4.3.9)$$

Table 4.4: Results of Hypothesis Tests from Cell Mean Model for Staff's Choice of Immediate Pick Up Proportions for Example 2

Hypothesis	C-Matrix	Uncertif. DCC's Q_C (d.f.)	Uncertif. DCC's p-value	Certified DCC's Q_C (d.f.)	Certified DCC's p-value
H_1: No difference between symtoms in absence of fever	$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}$	56.36 (2)	<0.01	418.85 (2)	<0.010
H_2: No difference between crankiness and conjuntivitis in the absence of fever	$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}$	56.13 (1)	<0.01	119.74 (1)	<0.010
H_3: No difference between diarrhea and conjunctivitis in the absence of fever	$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}$	2.57 (1)	0.109	1.14 (1)	0.283
H_4: No difference between symptoms in the presence of fever	$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$	39.96 (2)	<0.01	49.15 (2)	<0.010
H_5: No difference between crankiness and conjunctivitis in the presence of fever	$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$	34.93 (1)	<0.01	42.43 (1)	<0.010
H_6: No difference between diarrhea and conjunctivitis in the presence of fever	$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$	1.86 (1)	0.173	0.02 (1)	0.828
H_7: No difference between fever alone and symptoms plus fever	$\begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$	47.62 (3)	<0.01	57.98 (3)	<0.010

Table 4.4: Continued

Hypothesis	C-Matrix	Uncertif. DCC's Q_C (d.f.)	Uncertif. DCC's p-value	Certified DCC's Q_C (d.f.)	Certified DCC's p-value
H_8: No difference between fever alone and crankiness plus fever	$\begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}$	1.91 (1)	0.167	0.02 (1)	0.904
H_9: No difference between fever alone and diarrhea plus fever	$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}$	26.31 (1)	<0.01	31.95 (1)	<0.010
H_{10}: No difference between fever alone and conjunctivitis plus fever	$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$	37.55 (1)	<0.01	42.70 (1)	<0.010
H_{11}: No fever × symptom interaction	$\begin{bmatrix} 1 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{bmatrix}$	2.59 (2)	0.27	28.27 (2)	<0.010
H_{12}: No difference between symptoms in the presence and absence of fever	$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}$	14.17 (3)	<0.01	71.67 (3)	<0.010
H_{13}: No difference between crankiness in the presence and absence of fever	$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix}$	12.45 (1)	<0.01	61.23 (1)	<0.010
H_{14}: No difference between diarrhea in the presence and absence of fever	$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$	7.04 (1)	<0.01	2.80 (1)	0.09
H_{15}: No difference between conjunctivitis in the presence and absence of fever	$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}$	6.02 (1)	0.02	12.91 (1)	<0.010

The estimated model parameters and their estimated covariance matrices are

$$\underset{\sim}{b}_{(NC)} = \begin{bmatrix} 0.043 \\ 0.469 \\ 0.323 \\ 0.655 \end{bmatrix} \qquad \underset{\sim}{V}_{\underset{\sim}{b}(NC)} = \begin{bmatrix} 11.99 & -6.03 & 11.65 & -6.89 \\ -6.03 & 37.48 & 3.31 & 11.17 \\ -11.65 & 3.31 & 44.27 & 21.82 \\ -6.89 & 11.17 & 21.82 & 23.85 \end{bmatrix} \times 10^{-4}$$

$$\underset{\sim}{b}_{(C)} = \begin{bmatrix} 0.024 \\ 0.688 \\ 0.468 \\ 0.819 \end{bmatrix} \qquad \underset{\sim}{V}_{\underset{\sim}{b}(C)} = \begin{bmatrix} 1.44 & -2.00 & -1.02 & -2.12 \\ -2.00 & 9.83 & 5.94 & 8.25 \\ -1.02 & 9.83 & 26.33 & 11.98 \\ -2.12 & 8.25 & 11.98 & 19.53 \end{bmatrix} \times 10^{-4} .$$

Predicted values of the proportions for choosing immediate pick-up are obtained from $\hat{\mu}=Xb$. These quantities along with their estimated standard errors are shown in Table 4.3. There, it can be seen that the predicted percentage for staff's choice of immediate pick-up for uncertified DCC's is 70% when child presents symptoms/signs of diarrhea and conjunctivitis in the presence of elevated temperature; 51% for these symptoms but not elevated temperature; 37% when child has elevated temperature with or without crankiness and only 4% for the symptom of crankiness alone. Thus, the symptoms/signs of diarrhea and conjunctivitis with a temperature of less than 100° F are perceived by the staff as a more serious illness situation than temperature of 100°-100.9°F alone; and when elevated temperature accompanies these two symptoms, the proportion of staff personnel that opts for calling parents for immediate pick-up of the child is even higher. When predicted percentages for certified DCC's are examined, a similar pattern is observed to that of uncertified DCC's but with higher predicted percentages except for the symptom of crankiness in the absence of elevated temperature. These predicted percentages are 84% for diarrhea and conjunctivitis in the presence of elevated temperature and 71% when temperature is lower than 100°F; 49% when child has elevated temperature alone or with crankiness and only 2% for crankiness alone. So, except for crankiness alone, certified day care centers show much higher percentages for choosing to call parents for immediate pick-up of the perceived ill child than the uncertified day care centers.

While DCC appears to be the appropriate unit of analysis for these data due to the possible correlation among staff's answers in a DCC, it is interesting to note if estimates and their standard errors differ when staff is viewed as the unit of analysis. This approach essentially ignores any clustering; that is it assumes that there is no correlation among staff answers within a cluster.

4.4 Mean scores application for non-clustered data

Analysis of the proportion of staff choosing immediate pick-up for the situation of staff being the unit of analysis is done by first transforming the responses to an indicator like in (4.3.4), then calculating the means $\bar{g}_{hk}=\sum_{i=1}^{n_h}\sum_{j=1}^{d_{hi}} g_{hijk}/\sum_{i=1}^{n_h} d_{hi}$, with $d_{1\cdot}=\sum_i d_{1i}=95$ and $d_{2\cdot}=\sum_i d_{2i}=69$. The (12×12) covariance matrix $\underset{\sim}{V}_{\bar{g}}$ has a similar expression to that in (4.3.5) but with the $\underset{\sim}{V}_{\bar{g}i}$ replaced by counterparts to the $\underset{\sim}{V}_{\bar{g}i}$ in (4.2.1).

For both examples, preliminary hypotheses of interest are assessed with a cell mean model, and based on those findings tentative models similar to the ones used in section 4.3 are examined for goodness of fit. For example 1, the Q_W statistic is 5.12 (d.f.=7, p-value=0.65); and for Example 2,

$$Q_W = \begin{cases} 3.71 \text{ (d.f.=3, p=0.29)} \\ 1.67 \text{ (d.f.=3, p=0.64)} \end{cases}$$

Estimates of the observed and predicted proportions for staff's choice of Immediate Pick-Up and standard errors for Examples 1 and 2 are presented in Table 4.5.

When standard errors of the estimated observed and predicted values from Table 4.5 are compared with those in Tables 4.1 and 4.3 it is observed that in the approach where staff is the unit of analysis, the standard errors are substantially underestimated when compared to those obtained from the analyses when the correlation structure among staff within a DCC is taken into account.

Table 4.5. Estimates of the Observed and Predicted Staff's Choice of Immediate Pickup Proportions and Standard Errors

Example 1

	Level	Age	Temperature	Observed Proportion	s.e.	Final Model Predicted Proportion	s.e.
NC	0	<2	99°- 99.9°F	0.126	0.034	0.136	0.021
	0	<2	100°-100.9°F	0.347	0.049	0.351	0.034
	0	<2	101°-101.9°F	0.642	0.049	0.653	0.043
	0	2-5	99°- 99.9°F	0.074	0.027	0.060	0.019
	0	2-5	100°-100.9°F	0.242	0.044	0.275	0.032
	0	2-5	101°-101.9°F	0.579	0.051	0.577	0.040
C	2	<2	99°- 99.9°F	0.116	0.039	0.136	0.021
	2	<2	100°-100.9°F	0.420	0.059	0.351	0.036
	2	<2	101°-101.9°F	0.885	0.039	0.874	0.031
	2	2-5	99°- 99.9°F	0.058	0.028	0.060	0.019
	2	2-5	100°-100.9°F	0.290	0.055	0.275	0.032
	2	2-5	101°-101.9°F	0.812	0.047	0.799	0.031

NC: uncertified C: certified

Example 2

		Level 0				Level 2			
		Observed		Predicted		Observed		Predicted	
Fever	Symptom	Prep.	s.e.	Prep.	s.e.	Prep.	s.e.	Prep.	s.e.
No	None	-	-	-	-	-	-	-	-
No	Crankiness	0.032	0.016	0.031	0.015	0.026	0.013	0.024	0.013
No	Diarrhea	0.500	0.045	0.501	0.036	0.683	0.037	0.666	0.031
No	Conjunctivitis	0.532	0.044	0.501	0.036	0.650	0.038	0.666	0.031
Yes	None	0.294	0.041	0.324	0.036	0.455	0.040	0.476	0.034
Yes	Crankiness	0.365	0.043	0.324	0.036	0.494	0.041	0.476	0.034
Yes	Diarrhea	0.651	0.042	0.627	0.035	0.831	0.030	0.832	0.025
Yes	Conjunctivitis	0.635	0.043	0.627	0.035	0.831	0.030	0.732	0.025

4.5 Discussion

Applications of weighted least squares methods for analysis of possibly correlated categorical responses from a study of day care centers' criteria for sending perceived sick children home are evaluated through 2 examples with respect to what actions they take for situations where the child may have fever, other illness symptoms or a cross-classification of these 2 conditions.

When staff is considered the unit of analysis, i.e., when the possibility of clustering in the data is ignored, the standard errors obtained are smaller than when day care center is considered the unit of analysis with staff members as sub-units. This indicates that correlation among cluster sub-units for this study may not be negligible and that viewing sub-units (staff) as independent of one another may not give reasonable results. In the light of these conclusions, the summary of results that follows is based on considering DCC as the unit of analysis.

Day care center staff seem to act differently with regard to taking care of children perceived to be sick, depending on the age of child, temperature and level of day care center. When only these three factors are taken into consideration, age of the child seems to contribute significant information for staff's choice of calling parent or not for immediate pick-up; while levels of center and temperature interact with each other.

When three specific symptoms/signs, crankiness, diarrhea and conjunctivitis, are taken into account in the presence and absence of elevated temperature for children between the ages of 2 and 5 years, the results obtained are straightforward to interpret. Temperature of 100°F-100.9°F with no symptoms is considered a less serious symptom/sign of illness than diarrhea and conjunctivitis; elevated temperature of 100°F-100.9°F along with any of these two symptoms shows the highest percentage for calling parent for immediate pick-up. The sign of elevated temperature alone and in the presence of crankiness are similar, and finally crankiness with temperature of less than 100°F leads only to a relatively small proportion of staff to call parent for immediate pick-up.

With respect to variation in action by the two types of day care centers, the attitude of staff members under the situation where age and temperature of child are taken into account does not vary significantly with level of center except for the temperature of 101°F-101.9°F when percentage for choosing to call parents is higher for certified centers than for uncertified. When an analysis of symptoms and normal/elevated temperature is considered, except for crankiness alone, certified centers show larger percentages for choosing to call parents for immediate pick-up than uncertified centers.

The following three chapters present the theory to analyse clustered attribute data from two stage nested designs for the situation of more than one binary response per subject. Extension for the case of two binary responses and the generalization to r binary responses per subject are presented. Examples for the two binary response situation are used to illustrate that extension.

CHAPTER V

ANALYSIS OF CLUSTERED BINARY DATA EXTENDED TO TWO TIME POINTS OR CONDITIONS: ONE TREATMENT PER CLUSTER

5.1 Introduction

In Chapter III the theoretical framework for the analysis of clustered attribute univariate binary data in a two stage nested design with one treatment per cluster was reviewed. The methodology presented there is now extended to the situation where instead of observing one binary response on each sub-unit or subject within a cluster, an outcome measure is evaluated under two conditions (e.g., two time points) for a given subject generating two binary responses on each subject within a cluster.

The basic data structure in this new setting involves clusters, one treatment group per cluster and two binary responses from each subject in a cluster. For example, in a study of day care center staff's belief and knowledge about infections in DCCs, staff members may be asked if they would call the parent to pick up a child, if child had temperature and one of the symptoms, new cough or unusually cranky. Here, the basic unit of analysis is the DCC and the two binary responses are answers by staff to each of the questions related to the situations, temperature and new cough and temperature and cranky (yes, no). Alternatively, in a clinical dental material study, patients may be treated for restorations in two or more teeth. These patients are then followed over a period of two years, with an annual check up. A characteristic like plaque retention (rating of plaque presence or absence in restored teeth) may be of interest. At each follow up time, a binary response is generated from evaluation of each tooth with relation to a chosen characteristic.

5.2 Methodological Consideration

The theoretical framework of interest here is similar to that described for the univariate binary data situation where n clusters were randomly sampled without replacement from a finite population of N clusters and d_i subjects were randomly sampled without replacement from a population of D subjects within the i-th sampled cluster ($i=1,2,...,n$; $j=1,2,..., d_i$).

The sampling framework of primary interest in this setting is when $D \rightarrow \infty$ and $N \rightarrow \infty$. In this case sampling with replacement and without replacement are equivalent. For reasons discussed in Chapter III, the situation corresponding to sampling without replacement from a finite population is formulated initially and then the case where D and N tend to infinity is examined. Later on, sampling with replacement is assumed by equivalence of the two sampling procedures.

Let

$$y_{ijk} = \begin{cases} 1 & \text{if the k-th response of the j-th subject} \\ & \text{in the i-th cluster is favorable} \\ 0 & \text{otherwise} \end{cases}$$

with $i=1,2, \ldots, n$; $j=1,2, \ldots, d_i$; $k=1,2$.

The model of interest is characterized by

$E(y_{ijk}) = \pi_k =$ Pr (k-th response for a given subject in a cluster is favorable)

$E(y_{ijk}y_{ij'k}) = \lambda_k =$ Pr (k-th responses for two distinct subjects in a cluster are favorable)

$E(y_{ijk}y_{ijk'}) = \theta =$ Pr (both responses for a subject in a cluster are favorable)

$E(y_{ijk}y_{ij'k'}) = \phi =$ Pr (response k from subject j and response k′ from subject j′ in a cluster are favorable)

In this setting, the i-th cluster in the population has mean

$$\pi_{ik} = \frac{1}{D}\sum_{j=1}^{D} y_{ijk} \qquad k=1,2 \quad ,$$

and the population mean is

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} \pi_{ik} \quad k=1,2 \quad .$$

From a sampling without replacement point of view,

i) The probability that the k-th binary responses for distinct subjects in a given cluster, are both favorable is the parameter λ_k,

$$\lambda_k = \frac{1}{N}\sum_{i=1}^{N} \lambda_{ik}$$

with

$$\lambda_{ik} = \frac{1}{D(D-1)} \sum_{j \neq}^{D} \sum_{j'=1}^{D} y_{ijk}\, y_{ij'k} = \frac{\left(\sum_{j=1}^{D} y_{ijk}\right)^2 - \left(\sum_{j=1}^{D} y_{ijk}^2\right)}{D(D-1)} =$$

$$= \pi_{ik}\left[\pi_{ik} - \frac{1-\pi_{ik}}{D-1}\right] = \pi_{ik}^2 - \frac{\pi_{ik}\,(1-\pi_{ik})}{(D-1)} \,.$$

ii) The probability that both responses for the same subject in a cluster are favorable is the parameter θ,

$$\theta = \frac{1}{N}\sum_{i=1}^{N} \theta_i$$

with

$$\theta_i = \frac{1}{D}\sum_{j=1}^{D} y_{ijk}\, y_{ijk'} \quad .$$

iii) The probability that response k for subject j and response k′ for subject j′ in the same cluster are both favorable is the parameter ϕ;

$$\phi = \frac{1}{N}\sum_{i=1}^{N} \phi_i \quad .$$

with

$$\phi_i = \frac{1}{D(D-1)} \sum_{j \neq j'=1}^{D} \sum^{D} y_{ijk}\, y_{ij'k'}$$

$$= \frac{1}{D(D-1)} \left(\sum_{j=1}^{D} y_{ijk} \sum_{j'=1}^{D} y_{ij'k'} - \sum_{j=1}^{D} y_{ijk}\, y_{ijk'} \right)$$

$$= \frac{1}{D(D-1)} (D^2 \pi_{ik}\, \pi_{ik'} - D\theta_i) = \frac{D\pi_{ik}\, \pi_{ik'} - \theta_i}{(D-1)}$$

$$= \pi_{ik}\, \pi_{ik'} - \frac{\theta_i - \pi_{ik}\, \pi_{ik'}}{D-1} .$$

If for the overall probability associated with the parameters λ_{ik} the number of observations D tends to infinity, then

$$\lim_{D \to \infty} \lambda_{ik} = \lim_{D \to \infty} \left\{ [\pi_{ik}]\, [\pi_{ik} - \frac{1-\pi_{ik}}{D-1}] \right\} = \pi_{ik}^2 .$$

So, the corresponding overall parameter may be expressed as

$$\lambda_k = \lim_{D \to \infty} \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_{ik} \right) = \frac{1}{N} \sum_{i=1}^{N} \lim_{D \to \infty} \lambda_{ik} = \frac{1}{N} \sum_{i=1}^{N} \pi_{ik}^2 .$$

When both D and N tend to infinity, the same expression may be obtained. Note that the overall mean response $\pi_k = \frac{1}{N} \sum_{i=1}^{N} \pi_{ik}$ may be viewed as an average over an infinite number of small intervals $0 \leq (\pi_k, \pi_k + d\pi_k) \leq 1$ so that

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \pi_{ik} \sim \int_0^1 p_k f_k(p_k)\, d(p_k)$$

with f(p) the density of π_{ik}. Similar correspondences exist for the other parameters.

Other parameters of interest in this setting are

$$\text{Cov}\,(y_{ijk}, y_{ij'k}) = \lambda_k - \pi_k^2$$

$$\text{Cov}\,(y_{ijk}, y_{ijk'}) = \theta - \pi_k\, \pi_{k'}$$

$$\mathrm{Cov}\,(y_{ijk}, y_{ij'k'}) = \phi - \pi_k\, \pi_{k'}$$

and the correlation coefficients

$$\rho_k = \frac{\lambda_k - \pi_k^2}{\pi_k(1-\pi_k)}$$

$$\gamma = \frac{\theta - \pi_k\, \pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\, \pi_{k'}(1-\pi_{k'})}}$$

$$\nu = \frac{\phi - \pi_k\, \pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}} \quad .$$

The first correlation coefficient measures the strength of clustering in a set of data within a condition, the second reflects time correlation and the third, reflects correlation between times for different subjects in the same cluster.

5.3 The Completely Balanced Design: Two-Time Point Responses

A random sample of n clusters is obtained from a population of N clusters, a sample of d subjects is selected from D subjects within each sampled cluster, and the population of N clusters and the population of D observations both tend to infinity. Two binary responses are observed on each sampled subject.

Unbiased estimates of the population quantities are given by

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{d} y_{ijk}}{nd} = \bar{y}_{\cdot\cdot k} \tag{5.3.1}$$

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} \sum_{j \neq j'=1}^{d} \sum^{d} y_{ijk}\, y_{ij'k}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{i\cdot k}(y_{i\cdot k}-1)}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{i\cdot k}^2 - y_{\cdot\cdot k}}{nd(d-1)} \tag{5.3.2}$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ijk}\, y_{ijk'}}{nd} \tag{5.3.3}$$

$$\hat{\phi} = \frac{\sum_{i=1}^{n}\left(\sum_{j\neq j'=1}^{d}\sum^{d} y_{ijk}\, y_{ij'k'}\right)}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{i\cdot k}\, y_{i\cdot k'} - nd\hat{\theta}}{nd(d-1)} \tag{5.3.4}$$

Note that the observations y_{ijk} may be written in terms of the π_{ik} parameters as a consequence of the two stage random sampling process as:

$$\begin{bmatrix} y_{ij1} \\ y_{ij2} \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} + \begin{bmatrix} \pi_{i1}-\pi_1 \\ \pi_{i2}-\pi_2 \end{bmatrix} + \begin{bmatrix} y_{ij1}-\pi_{i1} \\ y_{ij2}-\pi_{i2} \end{bmatrix} \tag{5.3.5}$$

for $i=1,2,\ldots,n$; $j=1,2,\ldots,d$; $k=1,2$; where $(\pi_{ik}-\pi_k)$ is the random cluster effect element and $(y_{ijk}-\pi_{ik})$ is the sub-sampling residual element for response k.

If $\mu_k = \pi_k$, $a_{ik} = (\pi_{ik}-\pi_k)$ and $(e_{ijk})=(y_{ijk}-\pi_{ik})$, then (5.3.5) may be written as

$$\begin{bmatrix} y_{ij1} \\ y_{ij2} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} + \begin{bmatrix} e_{ij1} \\ e_{ij2} \end{bmatrix} . \tag{5.3.6}$$

Observe that in terms of population parameters, according to a sampling process from an infinite population,

$$E\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} = E\begin{bmatrix} \pi_{i1}-\pi_1 \\ \pi_{i2}-\pi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad E\begin{bmatrix} e_{ij1} \\ e_{ij2} \end{bmatrix} = E\begin{bmatrix} y_{ij1}-\pi_{i1} \\ y_{ij2}-\pi_{i2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} ,$$

$$\mathrm{Var}\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} \pi_{i1}-\pi_1 \\ \pi_{i2}-\pi_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a,1} & \rho_a\sigma_{a,1}\sigma_{a,2} \\ \rho_a\sigma_{a,1}\sigma_{a,2} & \sigma^2_{a,2} \end{bmatrix} ,$$

$$\mathrm{Var}\begin{bmatrix} e_{ij1} \\ e_{ij2} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} y_{ij1}-\pi_{i1} \\ y_{ij2}-\pi_{i2} \end{bmatrix} = \begin{bmatrix} \sigma^2_{e,1} & \rho_e\sigma_{e,1}\sigma_{e,2} \\ \rho_e\sigma_{e,1}\sigma_{e,2} & \sigma^2_{e,2} \end{bmatrix} .$$

$$\mathrm{Cov}\left\{\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix}, \begin{bmatrix} e_{ij1} \\ e_{ij2} \end{bmatrix}\right\} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

By defining

$$E(y_{ijk}) = \mu_k = \pi_k ,$$

$$\mathrm{Var}(y_{ijk}) = \sigma^2_{a,k} + \sigma^2_{e,k} = \pi_k(1-\pi_k) ,$$

$$\mathrm{Cov}(y_{ijk}\ y_{ij'k}) = \sigma^2_{a,k} = \lambda_k - \pi_k^2 ,$$

$$\mathrm{Cov}(y_{ijk}, y_{ijk'}) = \rho_a\sigma_{a,k}\sigma_{a,k'} + \rho_e\sigma_{e,k}\sigma_{e,k'} = \theta - \pi_k\pi_{k'} ,$$

$$\mathrm{Cov}(y_{ijk}, y_{ij'k'}) = \rho_a\sigma_{a,k}\sigma_{a,k'} = \phi - \pi_k\pi_{k'} ,$$

one can note that

$$\theta - \phi = \rho_e\sigma_{e,k}\sigma_{e,k'} ,$$

$$\pi_k - \lambda_k = \sigma^2_{e,k} ;$$

$$\rho_e = \frac{(\theta-\phi)}{\sqrt{(\pi_1-\lambda_1)(\pi_2-\lambda_2)}} ,$$

and

$$\rho_a = \frac{\phi - \pi_k\pi_{k'}}{\sqrt{(\lambda_k-\pi_k^2)(\lambda_{k'}-\pi_{k'}^2)}} .$$

The analysis of variance for the two stage sampling process for this setting has the same form for each response as that one in Table 3.3 with addition of 'k' subscript to y_{ij}, $\bar{y}_{i\cdot}$, $\bar{y}_{\cdot\cdot}$, σ_a^2 and σ_e^2. This analysis of variance is identical in form to that for the classical two-stage nested random effects ANOVA model where the random variables $\{a_{ik}\}$ and $\{e_{ijk}\}$ are mutually independent and

$$\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} \sim \mathrm{BN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{a,1}^2 & \rho_a\sigma_{a,1}\sigma_{a,2} \\ \rho_a\sigma_{a,1}\sigma_{a,2} & \sigma_{a,2}^2 \end{pmatrix}\right]$$

$$\begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix} \sim \mathrm{BN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e,1}^2 & \rho_e\sigma_{e,1}\sigma_{e,2} \\ \rho_e\sigma_{e,1}\sigma_{e,2} & \sigma_{e,2}^2 \end{pmatrix}\right]$$

where BN means 'has bivariate normal distribution'.

5.3.1 Means formulation for Estimation

If the observations y_{ijk} can be grouped in an appropriate way, expressions can be formulated for the parameter estimators.

Note that the parameter estimates $\hat{\pi}_k$ and $\hat{\theta}$ may be written as

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij1}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i1} = F_1 \quad , \tag{5.3.7}$$

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij2}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i2} = F_2 \quad , \tag{5.3.8}$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ijk}y_{ijk'}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i3} = F_3 \quad , \tag{5.3.9}$$

while

$$(\hat{\pi}_1-\hat{\lambda}_1) = s_{e,1}^2 = s_1^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{ij1}-F_{i1})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i4} = F_4 \quad , \qquad (5.3.10)$$

$$(\hat{\pi}_2-\hat{\lambda}_2) = s_{e,2}^2 = s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{ij2}-F_{i2})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i5} = F_5 \quad , \qquad (5.3.11)$$

and

$$(\hat{\theta}-\hat{\phi}) = s_{e,12} = s_{12} = \sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{d}\left(y_{ij1}-F_{i1}\right)\left(y_{ij2}-F_{i2}\right)}{n(d-1)}\right) = \frac{1}{n}\sum_{i=1}^{n} F_{i6} = F_6. \qquad (5.3.12)$$

It follows that

$$\underset{\sim}{F} = \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\theta} \\ \hat{\pi}_1-\hat{\lambda}_1 \\ \hat{\pi}_2-\hat{\lambda}_2 \\ \hat{\theta}-\hat{\phi} \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \\ F_6 \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} F_{i1} \\ F_{i2} \\ F_{i3} \\ F_{i4} \\ F_{i5} \\ F_{i6} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_i \ . \qquad (5.3.13)$$

Note that $E(s_k^2) = \pi_k - \lambda_k$ and $E(s_{kk'}) = \theta - \phi$. The data formulation presented in (5.3.7) through (5.3.12) show that the parameter estimates of interest can be obtained through means, thus the covariance matrix estimate is $V(\hat{\underset{\sim}{F}}) = \frac{1}{n^2}\sum_{i=1}^{n} (\underset{\sim}{F}_i-\underset{\sim}{F})(\underset{\sim}{F}_i-\underset{\sim}{F})'$. The means $\underset{\sim}{F}$ can be transformed to obtain estimates for other quantities of interest.

The parameters of primary interest in this one treatment two binary responses case are π_1, π_2, ρ_1, ρ_2, γ and ν. These may be estimated jointly by applying a series of linear, logarithmic and exponential transformations to the estimated parameter vector $\underset{\sim}{F} = [\hat{\pi}_1, \hat{\pi}_2, \hat{\theta}, (\hat{\pi}_1-\hat{\lambda}_1), (\hat{\pi}_2-\hat{\lambda}_2)(\hat{\theta}-\hat{\phi})]'$.

If the correlation coefficients are written as

$$\hat{\rho}_k = \frac{\hat{\lambda}_k - \hat{\pi}_k^2}{\hat{\pi}_k(1-\hat{\pi}_k)} = \left(\frac{\hat{\lambda}_k}{\hat{\pi}_k^2}-1\right)\left(\frac{\hat{\pi}_k}{1-\hat{\pi}_k}\right) = 1 - \frac{\hat{\pi}_k-\hat{\lambda}_k}{\hat{\pi}_k(1-\hat{\pi}_k)} \quad k=1,2 \ , \tag{5.3.14}$$

$$\hat{\gamma} = \frac{\hat{\theta}-\hat{\pi}_k\hat{\pi}_{k'}}{\sqrt{\hat{\pi}_k(1-\hat{\pi}_k)\ \hat{\pi}_{k'}(1-\hat{\pi}_{k'})}} = \left(\frac{\hat{\theta}}{\hat{\pi}_k\hat{\pi}_{k'}}-1\right)\left\{\left(\frac{\hat{\pi}_k}{1-\hat{\pi}_k}\right)\left(\frac{\hat{\pi}_{k'}}{1-\hat{\pi}_{k'}}\right)\right\}^{\frac{1}{2}}, \tag{5.3.15}$$

and

$$\hat{\nu} = \frac{\hat{\phi}-\hat{\pi}_k\hat{\pi}_{k'}}{\sqrt{\hat{\pi}_k(1-\hat{\pi}_k)\hat{\pi}_{k'}(1-\hat{\pi}_{k'})}} = \left(\frac{\hat{\phi}}{\hat{\pi}_k\hat{\pi}_{k'}}-1\right)\left\{\left(\frac{\hat{\pi}_k}{1-\hat{\pi}_k}\right)\left(\frac{\hat{\pi}_{k'}}{1-\hat{\pi}_{k'}}\right)\right\}^{\frac{1}{2}} \tag{5.3.16}$$

then it follows that

$$\begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\rho}_1 \\ \hat{\rho}_2 \\ \hat{\gamma} \\ \hat{\nu} \end{bmatrix} = \exp\,[\underset{\sim}{A}_4 \ln[\underset{\sim}{A}_3[\underset{\sim}{C}_2+\exp[\underset{\sim}{A}_2 \ln(\underset{\sim}{C}_1+\underset{\sim}{A}_1\underset{\sim}{F})]]]] \tag{5.3.17}$$

where

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}, \ C_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ 0 \\ -1 \\ 0 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } A_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

Expression (5.3.17) is easily calculated with the help of the computer program MISCAT, which produces estimates of the parameters along with their estimated covariance matrix. This vector of estimated parameters is the result of a sequence of operations where the matrix operations $(\underset{\sim}{C}_1 + \underset{\sim}{A}_1 \hat{\underset{\sim}{F}})$ create from $\hat{\underset{\sim}{F}}$ in (5.3.13) the parts necessary to form numerators and denominators in the expressions for the four correlation coefficients (5.3.14), (5.3.15) and (5.3.16). Taking logarithms results in $\ln(\underset{\sim}{C}_1 + \underset{\sim}{A}_1 \hat{\underset{\sim}{F}})$. By premultiplying this vector by $\underset{\sim}{A}_2$, the logarithms of the ratios involved in estimating the ρ's are formed with addition of appropriate logarithms generating desired products for numerators and denominators, and subtraction generating ratios. If now the exponential of the logarithmic function is taken, the vector of constant C_2 is added and the matrix $\underset{\sim}{A}_3$ is pre-multiplied, then all terms necessary to form the expressions for the

ρ's are generated. Premultiplication of the vector just obtained by $\underset{\sim}{A}_4$, adds logarithmic terms producing products and resulting in the logarithm of the primary functions of interest. Finally taking exponential of the logarithmic functions produces the desired estimates of the parameters π_1, π_2, ρ_1, ρ_2, γ and ν.

Alternatively, statistical tests for the correlation coefficient parameters being zero are performed by evaluating

$$\frac{\ln\left(\frac{\hat{\lambda}_k}{\hat{\pi}_k^2}\right)}{\text{s.e. } \ln\left(\frac{\hat{\lambda}_k}{\hat{\pi}_k^2}\right)}, \tag{5.3.18}$$

$$\frac{\ln\left(\frac{\hat{\theta}}{\hat{\pi}_k \hat{\pi}_{k'}}\right)}{\text{s.e. } \ln\left(\frac{\hat{\theta}}{\hat{\pi}_k \hat{\pi}_{k'}}\right)}, \tag{5.3.19}$$

and

$$\frac{\ln\left(\frac{\hat{\phi}}{\hat{\pi}_k \hat{\pi}_{k'}}\right)}{\text{s.e. } \ln\left(\frac{\hat{\phi}}{\hat{\pi}_k \hat{\pi}_{k'}}\right)}, \tag{5.3.20}$$

using a normal distribution approximation. Note that the numerators of these expressions and their estimated covariance matrix are obtained by using the following matrix operations on the vector (5.3.13):

$$\begin{bmatrix} \ln\left(\dfrac{\hat{\lambda}_1}{\hat{\pi}_1^2}\right) \\ \ln\left(\dfrac{\hat{\lambda}_2}{\hat{\pi}_2^2}\right) \\ \ln\left(\dfrac{\hat{\theta}}{\hat{\pi}_1\hat{\pi}_2}\right) \\ \ln\left(\dfrac{\hat{\phi}}{\hat{\pi}_1\hat{\pi}_2}\right) \end{bmatrix} = \underset{\sim}{A}_2\ln(\underset{\sim}{A}_1\underset{\sim}{F})$$

where

$$\underset{\sim}{A}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \text{ and } \underset{\sim}{A}_2 = \begin{bmatrix} -2 & 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

If it would be of interest to estimate $\ln(\theta/\phi)$ along with its s.e., one last line (0 0 0 0 1-1) should be added to matrix $\underset{\sim}{A}_2$.

5.4. <u>The Partially Balanced Design: Two Time-Point Responses</u>

As it was described in Chapter III Section 3.2.2, in some experimental situations modules may be formed with clusters of different sizes grouped accordingly. As before, subjects within a cluster generate two binary responses each. In this modular case, each module is balanced so there is only one set of estimates for each module.

Suppose the m-th module has n_m clusters, each of size d_m. Let

$$y_{ijkm} = \begin{cases} 1 & \text{if the k-th response for subject j in the i-th} \\ & \text{cluster of the m-th module is favorable} \\ 0 & \text{otherwise} \end{cases}$$

where $i=1,2, \ldots, n_m$; $j=1,2,\ldots, d_m$; $k=1,2$; $m=1,2,\ldots, M$.

The set of estimators for the m-th module has the same form as expressions (5.3.1) to (5.3.4) with addition of an 'm' subscript on y_{ijk}, n and d. Grouping the data to form means as was done in Section 5.3, provides a convenient way for evaluation of the basic parameters for this case as well as their covariance matrix. These modular estimators can be combined to generate overall estimators of the parameters by using the WLS approach, provided each module fulfills the requirement of sufficiently large n_m (say $n_m \geq 15$).

In this way,

$$\underset{\sim}{F}_m = [\hat{\pi}_{1m}=F_{1m}, \hat{\pi}_{2m}=F_{2m}, \hat{\theta}_m=F_{3m}, s^2_{1m}=F_{4m}, s^2_{2m}=F_{5m}, s_{12m}=F_{6m}]'$$

is a (6×1) vector of modular estimates and variation among the elements of $\underset{\sim}{F}_{(6M\times1)}$

$$\underset{\sim}{F}_{(6M\times1)} = [\ \underset{\sim}{F}_1, \underset{\sim}{F}_2 \cdots \underset{\sim}{F}_m \cdots \underset{\sim}{F}_M]'$$

can be studied by using the GSK methodology with the model

$$E(\underset{\sim}{F}) = \underset{\sim}{X}\,\underset{\sim}{\beta}$$

where

$$\underset{\sim}{X}_{(6M\times6)} = \begin{bmatrix} I_6 \\ I_6 \\ \vdots \\ I_6 \end{bmatrix} \text{ and } \underset{\sim}{\beta} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \theta \\ \pi_1-\lambda_1 \\ \pi_2-\lambda_2 \\ \theta-\phi \end{bmatrix}.$$

The estimate $\hat{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$ is obtained from $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'V_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'V_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}$ with

$\underset{\sim}{V}_{\hat{\beta}}=(\underset{\sim}{X}'\underset{\sim}{V}_{F}^{-1}\underset{\sim}{X})^{-1}$. Here, $\underset{\sim}{V}_{F}$ is a block diagonal matrix with diagonal blocks $\underset{\sim}{V}_m$ where $\underset{\sim}{V}_m= \underset{\sim}{V}(\hat{\underset{\sim}{F}}_m)=\frac{1}{n_m^2}\sum_{i=1}^{n}(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)'$.

If the number of clusters in a module is small ($n_m \leq 15$) then estimates derived by the GSK methodology do not necessarily have the optimal properties which follow from asymptotic arguments and so, alternative approaches such as direct estimates like mean of means may be used, i.e. with $\underset{\sim}{G}=\frac{1}{M}\sum_{m=1}^{M}\underset{\sim}{F}_m$ and $V_{\underset{\sim}{G}}=\frac{1}{M^2}\sum_{m=1}^{M}\underset{\sim}{V}_m$.

Examples are presented in the next Section to illustrate the extensions of estimation strategies developed in this chapter, and to provide a description of the possible correlations within clusters.

5.5 Examples

In this Section, the methodology for clustered attribute binary data with two responses per subject is illustrated by way of examples for the balanced case and modular case.

5.5.1 Day Care Center Study

The methodology developed in Section 5.3 can be used to examine data from the day care center study. The data are adapted to be compatible with what is required by the estimation strategies being illustrated. In this study, staff members answer questions about decision of calling parent to take a perceived ill child home if child has certain symptoms or signs of illness. Of interest are the two situations: i) child has new cough; ii) child is unusually cranky. These questions involve children between ages 2 and 5 and running a temperature of 100°F to 101°F. The answers to the two questions generate the binary responses from each subject within each cluster. There are 18 day care centers with 5 staff members in each center. The 5 staff members were randomly selected from staffs of at least this size.

Estimates for π_1, π_2, θ, $(\pi_1-\lambda_1)$, $(\pi_2-\lambda_2)$ and $(\theta-\phi)$ along with standard errors, can be obtained through the categorical data analysis program MISCAT which provides for the direct calculation of means.

So,

$$\underset{\sim}{F} = \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\theta} \\ (\hat{\pi}_1 - \hat{\lambda}_1) \\ (\hat{\pi}_2 - \hat{\lambda}_2) \\ (\hat{\theta} - \hat{\phi}) \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \\ F_6 \end{bmatrix} = \begin{bmatrix} 0.39 \\ 0.41 \\ 0.34 \\ 0.17 \\ 0.18 \\ 0.11 \end{bmatrix},$$

and s.e.($\underset{\sim}{F}$) = (0.076, 0.075, 0.071, 0.026, 0.028, 0.026)′. If clustering had been ignored, s.e. ($\hat{\pi}_1$)=0.051 and s.e.($\hat{\pi}_2$)=0.052.

The hypothesis of no difference between proportions of staff answering favorably to call parent for immediate pick up of sick child under the two situations of new cough ($\hat{\pi}_1$) and unusually cranky ($\hat{\pi}_2$) was tested using $H_0{:}\underset{\sim}{C}\underset{\sim}{\beta}=\underset{\sim}{0}$, $\underset{\sim}{C}=[1 \;\; -1 \;\; 0 \;\; 0 \;\; 0 \;\; 0]$. The result was non significant.

The parameters of primary interest in this one treatment two binary responses example are π_1, π_2, ρ_1, ρ_2, γ and ν. These may be estimated by using (5.3.17). Estimates of the four correlation coefficients along with standard errors are presented in Table 5.1.

A test of whether these quantities are null, obtained by evaluating expressions (5.3.18), (5.3.19) and (5.3.20) and using a normal distribution approximation, indicated that they are significantly different from zero (α=0.05). Also, the intraclass correlation coefficients ρ_1, ρ_2 and ν are not significantly different from each other but are different from the (condition) correlation coefficient γ (α=0.05). It is interesting to note that the result obtained with this subset of the data for the estimated proportion of staff calling

Table 5.1: Correlation Coefficients for Day Care Center Study, Completely Balanced Case

Parameter function	Estimate	(s.e.)
$\rho_1 = \rho_{\left(\text{cough}\right)}$	0.299	(0.104)
$\rho_2 = \rho_{\left(\text{cranky}\right)}$	0.266	(0.105)
$\gamma = \gamma_{\left(\substack{\text{cough, cranky} \\ \text{same subject}}\right)}$	0.769	(0.078)
$\nu = \nu_{\left(\substack{\text{diff. illnesses} \\ \text{diff. subjects}}\right)}$	0.306	(0.090)

for parent to pick up a child with elevated temperature and unusually cranky is in the range of those obtained in Chapter IV. However, the extended methodology presented in this chapter deals with estimates of cluster specific information, namely the vector (5.3.13) with its elements expressed as means of the observed data. This provides a method to estimate the different intraclass correlation coefficients as well as the condition correlation coefficient along with standard errors through a straightforward compound function vector (5.3.17).

5.5.2 Clinical Dental Material Study

In a twelve month clinical study of performance of restoration materials, patients were treated for one or more teeth and examined at follow-up times at six and twelve months. At each of these points in time, the 178 placed restorations were evaluated for color matching ability. The criterion for rating the restorations is defined as:

Alfa -	restoration matches adjacent tooth structure in color and translucency
Bravo -	mismatch in color and/or translucency is not outside normal range of tooth color and translucency
Charlie -	mismatch in color and/or translucency is outside the normal range

No restorations received "Charlie" ratings (clinically unacceptable).

Three materials were used in the study. Previous data analyses revealed that the materials were equivalent and so distinction between them could be ignored. It is of interest to take into consideration the oral environment of each patient, studying the correlation structure that may exist involving teeth within the same mouth. To perform this analysis, a subset of the data was used with 21 patients or clusters, each with 3 restored teeth.

Estimates of the proportions of teeth rated Bravo at six and twelve months along with the four correlation coefficient estimates and standard errors for this completely balanced case are presented in Table 5.2.

Table 5.2: Estimates of Proportions and Correlation Coefficients for Dental Study

Parameter function	Estimate	(s.e.)
$\pi_1=\pi_{(\text{6 months})}$	0.15	(0.041)
$\pi_2=\pi_{(\text{12 months})}$	0.30	(0.069)
$\rho_1=\rho_{\binom{\text{6 months}}{\text{diff. subjects}}}$	0.02	(0.112)
$\rho_2=\rho_{\binom{\text{12 months}}{\text{diff. subjects}}}$	0.25	(0.121)
$\gamma=\gamma_{\binom{\text{6,12 months}}{\text{same subject}}}$	0.46	(0.121)
$\nu=\nu_{\binom{\text{diff. times}}{\text{diff. subjects}}}$	0.24	(0.076)

Table 5.2 shows that the proportion of teeth rated Bravo doubled at 12 months. Tests of the correlation coefficients being null were performed. The correlation coefficients ρ_2, γ and ν are all different from zero ($\alpha=0.05$) with the intraclass correlation coefficients ρ_2 and ν different from the time correlation coefficient γ. The intraclass correlation coefficient at 6 months, ρ_1, is not significantly different from zero.

5.5.3 Example for the Partially Balanced Case

Data from the day care center study are adapted to illustrate how partially balanced clustered attribute data can be analyzed with the extended proposed methodology. Consider the situation where there are three modules of 17, 20 and 13 clusters each, with clusters in each module of sizes 2, 5 and 8 respectively. The two questions posed to staff members in the Example in Section 5.5.1 for the completely balanced case will be used to generate the two binary responses in this modular situation. The number of clusters in each module is considered reasonably good for the GSK methodology to be applied, so that overall estimates of the parameters of interest can be obtained.

The following model was fit to the vector of module estimates of π_1, π_2, θ, $(\pi_1-\lambda_1)$, $(\pi_2-\pi_2)$, and $\theta-\phi$.

$$\underset{(18\times 1)}{\underset{\sim}{F}} = \begin{bmatrix} \hat{\pi}_{1(1)} \\ \hat{\pi}_{2(1)} \\ \hat{\theta}_{(1)} \\ (\hat{\pi}_1-\hat{\lambda}_1)_{(1)} \\ (\hat{\pi}_2-\hat{\lambda}_2)_{(1)} \\ (\hat{\theta}-\hat{\phi})_{(1)} \\ \vdots \\ \hat{\pi}_{1(3)} \\ \hat{\pi}_{2(3)} \\ \hat{\theta}_{(3)} \\ (\hat{\pi}_1-\hat{\lambda}_1)_{(3)} \\ (\hat{\pi}_2-\hat{\lambda}_2)_{(3)} \\ (\hat{\theta}-\hat{\phi})_{(3)} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{I}_6 \\ \underset{\sim}{I}_6 \\ \underset{\sim}{I}_6 \end{bmatrix} \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\theta} \\ (\hat{\pi}_1-\hat{\lambda}_1) \\ (\hat{\pi}_2-\hat{\lambda}_2) \\ (\hat{\theta}-\hat{\phi}) \end{bmatrix} = \underset{\sim}{X}\hat{\underset{\sim}{\beta}} \qquad (5.3.21)$$

The fit of the model was reasonable (Q=19.2 with 12 degrees of freedom; p=0.08). Parameter estimates for model (5.3.21) are presented in Table 5.3. By applying a series of linear, logarithmic and exponential functions as in (5.3.17) to the estimated parameter vector $\hat{\beta}$ in (5.3.21) obtained from the weighted least squares procedure, estimates of the intraclass correlation coefficients and condition correlation coefficient were obtained. These estimates are summarized in Table 5.4.

Table 5.3: Parameter estimates and standard errors for modular example.

Parameter estimator	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\theta}$	$(\hat{\pi}_1-\hat{\lambda}_1)$	$(\hat{\pi}_2-\hat{\lambda}_2)$	$(\hat{\theta}-\hat{\phi})$
Module 1	0.559	0.412	0.353	0.206	0.176	0.088
	(0.092)	(0.095)	(0.091)	(0.060)	(0.058)	(0.046)
Module 2	0.394	0.419	0.350	0.174	0.181	0.106
	(0.075)	(0.075)	(0.071)	(0.027)	(0.028)	(0.025)
Module 3	0.541	0.484	0.461	0.224	0.201	0.155
	(0.070)	(0.082)	(0.087)	(0.026)	(0.025)	(0.024)
$\hat{\beta}$	0.535	0.537	0.486	0.203	0.201	0.139
	(0.034)	(0.031)	(0.036)	(0.013)	(0.015)	(0.016)

Table 5.4: Correlation Coefficient Parameter Estimates for the Modular Example

Overall parameter	Weighted Least Square Estimate	s.e.
$\rho_1=\rho\binom{\text{cough}}{\text{diff. children}}$	0.184	0.050
$\rho_2=\rho\binom{\text{cranky}}{\text{diff. children}}$	0.192	0.048
$\gamma=\gamma\binom{\text{cough, cranky}}{\text{diff. children}}$	0.800	0.034
$\nu=\nu\binom{\text{diff. illnesses}}{\text{diff. children}}$	0.242	0.051

The proportion of staff calling for parent to pick up child unusually cranky is approximately the same as for the situation of a child with new cough. The intraclass correlation coefficients are all different from zero (α=0.05). Also $\hat{\rho}_1$ and $\hat{\rho}_2$ are different from $\hat{\nu}$(α=0.05). The condition correlation coefficient is different from zero and different from the three intraclass correlation coefficients. Findings for this modular case agree with those in Section 5.5.1, although standard errors for this example are smaller due to increased sample size.

CHAPTER VI

ANALYSIS OF CLUSTERED BINARY DATA EXTENDED TO BIVARIATE RESPONSES FOR TWO TREATMENTS PER CLUSTER

6.1 Introduction

In Chapter V discussion centered on the two stage nested design with one treatment (or sub-population) per cluster and two conditions for each subject within a cluster. The methodology presented there is now extended to the case of two treatments (or sub-populations) per cluster.

The basic data structure in this setting involves clusters, two treatment (sub-population) groups and two binary responses for each subject within a cluster. Of interest here is the situation where a fixed number of observations (subjects) are randomly drawn from sub-population 1 in each cluster and a fixed (usually equal) number of observations are randomly drawn from sub-population 2. For example, in a sample survey, men and women (sub-populations) in neighborhoods (clusters) may be asked to choose in an opinion poll at time t_1, and time t_2 (six months later), between two candidates for public office. Alternatively, in the day care center study discussed in Chapter IV, two questions about how to handle a perceived ill child are posed to staff members and also to mothers of the day care centers' children. In this example, there are two sub-populations, staff and mothers, day care centers are the clusters and subjects, staff and mother, contribute with two binary responses each. Here, it is of interest to develop measures of within cluster association and then apply these measures to staff and parents responses in order to compare them.

6.2. Methodological Considerations

The framework for this problem of two sub-populations per cluster is similar to that of one treatment per cluster. However, new parameters need to be defined because of more than one treatment within a cluster.

Let

$$y_{hijk} = \begin{cases} 1 & \text{if k-th response of the j-th subject in the i-th} \\ & \text{cluster and h-th sub-population is favorable} \\ 0 & \text{otherwise} \end{cases}$$

where $h=1,2$; $i=1,2, \ldots, n$; $j=1,2, \ldots, d_i$; $k=1,2$.

The model for this framework is formulated from the following definitions:

$E(y_{hijk}) = \pi_{hk}$ = Pr (the k-th response of a subject in sub-population h in a randomly selected cluster is favorable),

$E(y_{hijk}\, y_{hij'k}) = \lambda_{hk}$ = Pr (the k-th responses for two different subjects in sub-population h in a randomly selected cluster, are both favorable),

$E(y_{hijk} y_{hijk'}) = \theta_h$ = Pr (both responses for the same subject in sub-population h in a randomly selected cluster are favorable),

$E(y_{hijk} y_{hij'k'}) = \phi_h$ = Pr (response k of subject j and response k′ of subject j′ in sub-population h in a randomly selected cluster are favorable),

$E(y_{hijk} y_{h'ij'k}) = \psi_k$ = Pr (the k-th responses for subjects j and j′ receiving treatments h and h′ respectively are both favorable),

and

$E(y_{hijk} y_{h'ij'k'}) = \psi_{hk,h'k'}$ = Pr (response k for subject j in sub-population h, and response k′ for subject j′ in sub-population h′ are both favorable).

In this setting, the mean response associated with treatment (or sub-population) h for the i-th cluster in the population is

$$\pi_{hik} = \frac{1}{D} \sum_{j=1}^{D} y_{hijk} \quad h=1,2;\ k=1,2.$$

and the overall mean response associated with treatment (or sub-population) h is

$$\pi_{hk} = \frac{1}{N}\sum_{i=1}^{N} \pi_{hik} \qquad h=1,2;\ k=1,2.$$

Other parameters of interest are

$$\mathrm{Cov}(y_{hijk}, y_{hij'k}) = \lambda_{hk} - \pi_{hk}^2$$

$$\mathrm{Cov}(y_{hijk}, y_{hijk'}) = \theta_h - \pi_{hk}\pi_{hk'}$$

$$\mathrm{Cov}(y_{hijk}, y_{hij'k'}) = \phi_h - \pi_{hk}\pi_{hk'}$$

$$\mathrm{Cov}(y_{hijk}, y_{h'ij'k}) = \psi_k - \pi_{hk}\pi_{h'k}$$

$$\mathrm{Cov}(y_{hijk}, y_{h'ij'k'}) = \psi_{hk,h'k'} - \pi_{hk}\pi_{h'k'}$$

and the correlation coefficients

$$\rho_{hk} = \frac{\lambda_{hk} - \pi_{hk}^2}{\pi_{hk}(1-\pi_{hk})}$$

$$\gamma_h = \frac{\theta_h - \pi_{hk}\pi_{hk'}}{\sqrt{\pi_{hk}(1-\pi_{hk})\pi_{hk'}(1-\pi_{hk'})}}$$

$$\nu_h = \frac{\phi_h - \pi_{hk}\pi_{hk'}}{\sqrt{\pi_{hk}(1-\pi_{hk})\pi_{hk'}(1-\pi_{hk'})}}$$

$$\tau_k = \frac{\psi_k - \pi_{hk}\pi_{h'k}}{\sqrt{\pi_{hk}(1-\pi_{hk})\pi_{h'k}(1-\pi_{h'k})}}$$

$$\omega_{(hk,h'k')} = \frac{\psi_{hk,h'k'} - \pi_{hk}\pi_{h'k'}}{\sqrt{\pi_{hk}(1-\pi_{hk})\pi_{h'k'}(1-\pi_{h'k'})}} .$$

These correlation coefficients describe aspects of clustering that are of interest in situations with the two stage nested design, two treatments per cluster, two binary responses per subject.

6.3 The Completely Balanced Design: Two-Time Point Responses

Consider a two-stage random sample with n clusters, two sub-populations per cluster, d subjects with two outcomes per subject within a sampled cluster. Discussion here focuses on the case where the cluster and subject population sizes N and D are viewed as infinite. In the balanced case, the unbiased sample estimators of the population parameters are

$$\hat{\pi}_{hk} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hijk}}{nd} = \bar{y}_{h\cdot\cdot k} , \qquad (6.2.1)$$

$$\hat{\lambda}_{hk} = \frac{\sum_{i=1}^{n}\sum_{j\neq j'=1}^{d}\sum^{d} y_{hijk}y_{hij'k}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{hi\cdot k}^{2} - y_{h\cdot\cdot k}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{hi\cdot k}^{2} - nd\hat{\pi}_{hk}}{nd(d-1)} , \qquad (6.2.2)$$

$$\hat{\theta}_{h} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hijk}y_{hijk'}}{nd} , \qquad (6.2.3)$$

$$\hat{\phi}_{h} = \frac{\left(\sum_{i=1}^{n}\sum_{j\neq j'=1}^{d}\sum^{d} y_{hijk}y_{hij'k'}\right)}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{hi\cdot k}\, y_{hi\cdot k'} - nd\theta_{h}}{nd(d-1)} , \qquad (6.2.4)$$

$$\hat{\psi}_{k} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{hijk}y_{h'ij'k}}{nd^{2}} = \frac{\sum_{i=1}^{n}\bar{y}_{hi\cdot k}\,\bar{y}_{h'i\cdot k}}{n} , \qquad (6.2.5)$$

and

$$\hat{\psi}_{hk,h'k'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{hijk}y_{h'ij'k'}}{nd^2} = \frac{\sum_{i=1}^{n}\bar{y}_{hi\cdot k}\bar{y}_{h'i\cdot k'}}{n}. \tag{6.2.6}$$

As a consequence of the two stage random sampling process, the observations y_{hijk} may be written in terms of the π_{hk} parameters as

$$\begin{bmatrix} y_{hij1} \\ y_{hij2} \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} + \begin{bmatrix} \pi_{h1}-\pi_1 \\ \pi_{h2}-\pi_2 \end{bmatrix} + \begin{bmatrix} \pi_{*i1}-\pi_1 \\ \pi_{*i2}-\pi_2 \end{bmatrix} + \begin{bmatrix} \pi_{hi1}-\pi_{h1}-\pi_{*i1}+\pi_1 \\ \pi_{hi2}-\pi_{h2}-\pi_{*i2}+\pi_2 \end{bmatrix} + \begin{bmatrix} y_{hij1}-\pi_{hi1} \\ y_{hij2}-\pi_{hi2} \end{bmatrix} \tag{6.2.7}$$

$i=1,2,\ldots,n$; $j=1,2,\ldots,d$; $h=1,2$; $k=1,2$. Here, $\pi_k=(\pi_{1k}+\pi_{2k})/2$, π_{hk}, $\pi_{*ik}=(\pi_{1ik}+\pi_{2ik})/2$ and π_{hik} are the population parameters associated with the sampled observations y_{hijk}. The $(\pi_{hk}-\pi_k)$, $(\pi_{*ik}-\pi_k)$, $(\pi_{hik}-\pi_{hk}-\pi_{*ik}+\pi_k)$ and $(y_{hijk}-\pi_{hik})$ in (6.2.7) are respectively, a fixed element due to treatment h, a random effect element for the i-th cluster, a random effect element associated with the h-th treatment and i-th cluster, and the sampling residual element, all for response k. Also, $\sum_{h=1}^{2}(\pi_{hk}-\pi_k)=0$ and $\sum_{h=1}^{2}(\pi_{hik}-\pi_{hk}-\pi_{*ik}+\pi_k)=0$.

Expression (6.2.7) may be rewritten in terms of an analysis of variance model as:

$$\begin{bmatrix} y_{hij1} \\ y_{hij2} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} t_{h1} \\ t_{h2} \end{bmatrix} + \begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} + \begin{bmatrix} (ta)_{hi1} \\ (ta)_{hi2} \end{bmatrix} + \begin{bmatrix} e_{hij1} \\ e_{hij2} \end{bmatrix} \tag{6.2.8}$$

where $\sum_{h=1}^{2} t_{hk}=0$ and $\sum_{h=1}^{2} (ta)_{hik}=0$.

From the definitions of the elements in the model (6.2.8), it follows that

$$E\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} = E\begin{bmatrix} \pi_{*i1}-\pi_1 \\ \pi_{*i2}-\pi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad E\begin{bmatrix} e_{hij1} \\ e_{hij2} \end{bmatrix} = E\begin{bmatrix} y_{hij1}-\pi_{hi1} \\ y_{hij2}-\pi_{hi2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$E\begin{bmatrix} (ta)_{hi1} \\ (ta)_{hi2} \end{bmatrix} = E\begin{bmatrix} \pi_{hi1}-\pi_{h1}-\pi_{*i1}+\pi_1 \\ \pi_{hi2}-\pi_{h2}-\pi_{*i2}+\pi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and

$$\mathrm{Cov}\left\{\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix}, \begin{bmatrix} e_{hij1} \\ e_{hij2} \end{bmatrix}\right\} = \mathrm{Cov}\left\{\begin{bmatrix} (ta)_{hi1} \\ (ta)_{hi2} \end{bmatrix}, \begin{bmatrix} e_{hij1} \\ e_{hij2} \end{bmatrix}\right\} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Suppose appropriate conditons apply to model parameters so that one can assume

$$\mathrm{Var}\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} \pi_{*i1}-\pi_1 \\ \pi_{*i2}-\pi_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a,1} & \rho_a\sigma_{a,1}\sigma_{a,2} \\ \rho_a\sigma_{a,1}\sigma_{a,2} & \sigma^2_{a,2} \end{bmatrix},$$

$$\mathrm{Var}\begin{bmatrix} e_{hij1} \\ e_{hij2} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} y_{hij1}-\pi_{hi1} \\ y_{hij2}-\pi_{hi2} \end{bmatrix} = \begin{bmatrix} \sigma^2_{e,1} & \rho_e\sigma_{e,1}\sigma_{e,2} \\ \rho_e\sigma_{e,1}\sigma_{e,2} & \sigma^2_{e,2} \end{bmatrix},$$

$$\mathrm{Var}\begin{bmatrix} (ta)_{hi1} \\ (ta)_{hi2} \end{bmatrix} = \mathrm{Var}\begin{bmatrix} \pi_{hi1}-\pi_{h1}-\pi_{*i1}+\pi_1 \\ \pi_{hi2}-\pi_{h2}-\pi_{*i2}+\pi_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{ta,1} & \rho_{ta}\sigma_{ta,1}\sigma_{ta,2} \\ \rho_{ta}\sigma_{ta,1}\sigma_{ta,2} & \sigma^2_{ta,2} \end{bmatrix}.$$

and

$$\mathrm{Cov}\left\{\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix}, \begin{bmatrix} (ta)_{hi1} \\ (ta)_{hi2} \end{bmatrix}\right\} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It then follows from the structure of the analysis of variance model (6.2.8) that

$$E(y_{hijk}) = \mu_k ,$$

$$Var(y_{hijk}) = \sigma^2_{a,k} + \sigma^2_{ta,k} + \sigma^2_{e,k} = \pi_{hk}(1-\pi_{hk}) ,$$

$$Cov(y_{hijk}, y_{hij'k}) = \sigma^2_{a,k} + \sigma^2_{ta,k} = \lambda_{hk} - \pi^2_{hk} ,$$

$$Cov(y_{hijk}, y_{hijk'}) = \rho_a\sigma_{a,k}\sigma_{a,k'} + \rho_{ta}\sigma_{ta,k}\sigma_{ta,k'} + \rho_e\sigma_{e,k}\sigma_{e,k'} = \theta_h - \pi_{hk}\pi_{hk'} ,$$

$$Cov(y_{hijk}, y_{hij'k'}) = \rho_a\sigma_{a,k}\sigma_{a,k'} + \rho_{ta}\sigma_{ta,k}\sigma_{ta,k'} = \phi_h - \pi_{hk}\pi_{hk'} ,$$

$$Cov(y_{hijk}, y_{h'ij'k}) = \sigma^2_{a,k} - \sigma^2_{ta,k} = \psi_k - \pi_{hk}\pi_{h'k} \text{ since } (ta)_{h'ik} = -(ta)_{hik} ,$$

$$Cov(y_{hijk}, y_{h'ij'k'}) = \rho_a\sigma_{a,k}\sigma_{a,k'} - \rho_{ta}\sigma_{ta,k}\sigma_{ta,k'} = \psi_{hk,h'k'} - \pi_{hk}\pi_{h'k'} .$$

One can also note that

$$\theta_h - \phi_h = \rho_e\sigma_{e,k}\sigma_{e,k'} ,$$

$$\pi_{hk} - \lambda_{hk} = \sigma^2_{ek} ,$$

and

$$\rho_e = \frac{\theta_h - \phi_h}{\sqrt{(\pi_{hk} - \lambda_{hk})(\pi_{hk'} - \lambda_{hk'})}} .$$

The analysis of variance for the k-th response from this two stage sampling process with two treatments per cluster is presented in Table 6.1.

Table 6.1: Analysis of Variance for the Two-Stage Nested Design with Two Treatments per Cluster

Source of Variation	Degrees of Freedom	Sum of Square	Expected Mean Square
mean	1	$2n\bar{y}^2_{\cdots k}$	$\sigma^2_{e,k}+2d\sigma^2_{a,k}+2nd\mu^2_k$
between treatments	1	$nd\sum_{h=1}^{2}(\bar{y}_{h\cdot\cdot k}-\bar{y}_{\cdots k})^2$	$\sigma^2_{e,k}+d\sigma^2_{ta,k}+nd\sum_{h=1}^{2}t^2_{hk}$
among clusters	$n-1$	$2d\sum_{i=1}^{n}(\bar{y}_{\cdot i\cdot k}-\bar{y}_{\cdots k})^2$	$\sigma^2_{e,k}+2d\sigma^2_{a,k}$
treatment by cluster (interaction)	$n-1$	$d\sum_{h=1}^{2}\sum_{i=1}^{n}(\bar{y}_{hi\cdot k}-\bar{y}_{h\cdots k}+ -\bar{y}_{\cdot i\cdot k}+\bar{y}_{\cdots k})^2$	$\sigma^2_{e,k}+d\sigma^2_{ta,k}$
subject within treatment by cluster (residual)	$2n(d-1)$	$\sum_{h=1}^{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{hijk}-\bar{y}_{hi\cdot k})^2$	$\sigma^2_{e,k}$
total (corrected)	$2nd-1$	$\sum_{h=1}^{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{hijk}-\bar{y}_{\cdots k})^2$	

This analysis of variance is similar to that for the classical two-stage nested model with random and fixed effects, (model 6.2.2), where $\{a_{ik}\}$, $\{(ta)_{hik}\}$ and $\{e_{hijk}\}$ are mutually independent and

$$\begin{bmatrix} a_{i1} \\ a_{i2} \end{bmatrix} \sim BN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{a,1} & \rho_a\sigma_{a,1}\sigma_{a,2} \\ \rho_a\sigma_{a,1}\sigma_{a,2} & \sigma^2_{a,2} \end{pmatrix}\right]$$

$$\begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix} \sim BN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e,1} & \rho_e\sigma_{e,1}\sigma_{e,2} \\ \rho_e\sigma_{e,1}\sigma_{e,2} & \sigma^2_{e,2} \end{pmatrix}\right]$$

$$\begin{bmatrix} (ta)_{h1} \\ (ta)_{h2} \end{bmatrix} \sim BN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{ta,1} & \rho_{ta}\sigma_{ta,1}\sigma_{ta,2} \\ \rho_{ta}\sigma_{ta,1}\sigma_{ta,2} & \sigma^2_{ta,2} \end{pmatrix}\right]$$

where $\sim$ BN means 'has bivariate normal distribution'.

In this model, the $\{(ta)_{hik}\}$ are subject to the constraint $\sum_h (ta)_{hik}=0$ and the fixed treatment effects $\{t_{hk}\}$ have the constraint $\sum_h t_{hk}=0$.

6.3.1 Means Formulation for Estimation

If the observations y_{hijk} are grouped in an appropriate way, then the parameters of interest can be estimated using GSK methodology to analyse categorical data.

Note that the parameter estimators $\hat{\pi}_{hk}$ and $\hat{\theta}_h$ may be written as

$$\hat{\pi}_{11} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{1ij1}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i1} = F_1 \qquad (6.2.9)$$

$$\hat{\pi}_{12} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{1ij2}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i2} = F_2 \qquad (6.2.10)$$

$$\hat{\pi}_{21} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{2ij1}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i3} = F_3 \qquad (6.2.11)$$

$$\hat{\pi}_{22} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{2ij2}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i4} = F_4 \qquad (6.2.12)$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{1ij1}\, y_{1ij2}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i5} = F_5 \qquad (6.2.13)$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{2ij1}\, y_{2ij2}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i6} = F_6 \tag{6.2.14}$$

Also, one can form

$$\hat{\pi}_{11}-\hat{\lambda}_{11} = s_{11}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{1ij1}-\bar{y}_{1i\cdot 1})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i7} = F_7 \tag{6.2.15}$$

$$\hat{\pi}_{12}-\hat{\lambda}_{12} = s_{12}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{1ij2}-\bar{y}_{1i\cdot 2})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i8} = F_8 \tag{6.2.16}$$

$$\hat{\pi}_{21}-\hat{\lambda}_{21} = s_{21}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{2ij1}-\bar{y}_{2i\cdot 1})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i9} = F_9 \tag{6.2.17}$$

$$\hat{\pi}_{22}-\hat{\lambda}_{22} = s_{22}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{2ij2}-\bar{y}_{2i\cdot 2})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i,10} = F_{10} \tag{6.2.18}$$

$$\hat{\theta}_1-\hat{\phi}_1 = s_{1,12} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{d}\left(y_{1ij1}-\bar{y}_{1i\cdot 1}\right)\left(y_{1ij2}-\bar{y}_{1i\cdot 2}\right)}{d-1}\right\} = \frac{1}{n}\sum_{i=1}^{n} F_{i,11} = F_{11} \tag{6.2.19}$$

$$\hat{\theta}_2-\hat{\phi}_2 = s_{2,12} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{d}\left(y_{2ij1}-\bar{y}_{2i\cdot 1}\right)\left(y_{2ij2}-\bar{y}_{2i\cdot 2}\right)}{d-1}\right\} = \frac{1}{n}\sum_{i=1}^{n} F_{i,12} = F_{12} \tag{6.2.20}$$

$$\hat{\psi}_1 = \frac{\sum_{i=1}^{n} \bar{y}_{1i\cdot 1}\bar{y}_{2i\cdot 1}}{n} = \frac{1}{n}\sum_{i=1}^{n} F_{i,13} = F_{13} \tag{6.2.21}$$

$$\hat{\psi}_2 = \frac{\sum_{i=1}^{n} \bar{y}_{1i\cdot 2}\bar{y}_{2i\cdot 2}}{n} = \frac{1}{n}\sum_{i=1}^{n} F_{i,14} = F_{14} \tag{6.2.22}$$

$$\hat{\psi}_{11,22} = \frac{\sum_{i=1}^{n} \bar{y}_{1i\cdot 1}\, \bar{y}_{2i\cdot 2}}{n} = \frac{1}{n}\sum_{i=1}^{n} F_{i,15} = F_{15} \tag{6.2.23}$$

$$\hat{\psi}_{12,21} = \frac{\sum_{i=1}^{n} \bar{y}_{1i\cdot 2}\bar{y}_{2i\cdot 1}}{n} = \frac{1}{n}\sum_{i=1}^{n} F_{i,16} = F_{16}\ . \tag{6.2.24}$$

It follows that

$$\underset{\sim}{F} = \begin{bmatrix} \hat{\pi}_{11} \\ \hat{\pi}_{12} \\ \hat{\pi}_{21} \\ \hat{\pi}_{22} \\ \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\pi}_{11}-\hat{\lambda}_{11} \\ \hat{\pi}_{12}-\hat{\lambda}_{12} \\ \hat{\pi}_{21}-\hat{\lambda}_{21} \\ \hat{\pi}_{22}-\hat{\lambda}_{22} \\ \hat{\theta}_1-\hat{\phi}_1 \\ \hat{\theta}_2-\hat{\phi}_2 \\ \hat{\psi}_1 \\ \hat{\psi}_2 \\ \hat{\psi}_{11,22} \\ \hat{\psi}_{12,21} \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \\ F_6 \\ F_7 \\ F_8 \\ F_9 \\ F_{10} \\ F_{11} \\ F_{12} \\ F_{13} \\ F_{14} \\ F_{15} \\ F_{16} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} F_{i1} \\ F_{i2} \\ F_{i3} \\ F_{i4} \\ F_{i5} \\ F_{i6} \\ F_{i7} \\ F_{i8} \\ F_{i9} \\ F_{i,10} \\ F_{i,11} \\ F_{i,12} \\ F_{i,13} \\ F_{i,14} \\ F_{i,15} \\ F_{i,16} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_i\,. \tag{6.2.25}$$

The data formulation presented in (6.2.9) to (6.2.25) shows that the parameter estimates of interest can be obtained through means; the corresponding covariance matrix estimate is $\underset{\sim}{V}_{\hat{\underset{\sim}{F}}} = \frac{1}{n^2}\sum_{i=1}^{n} (\underset{\sim}{F}_i - \underset{\sim}{F})(\underset{\sim}{F}_i - \underset{\sim}{F})'$. The means $\underset{\sim}{F}$ can be transformed to obtain estimates for other quantities of interest.

The parameters of primary interest in the two treatment per cluster, two binary responses per subject case are π_{11}, π_{12}, π_{21}, π_{22}, ρ_{hk}, γ_h, ν_h, τ_k, and $\omega_{(hk,h'k')}$. These parameters may be estimated jointly by applying a series of linear, logarithmic and exponential transformations to the estimated parameter vector (6.2.25).

If the correlation coefficients ρ_{hk}, γ_h, ν_h; τ_k and $\omega_{(hk,h'k')}$ are written as

$$\hat{\rho}_{hk} = \frac{\hat{\lambda}_{hk} - \hat{\pi}_{hk}^2}{\hat{\pi}_{hk}(1-\hat{\pi}_{hk})} = 1 - \frac{\hat{\pi}_{hk} - \hat{\lambda}_{hk}}{\hat{\pi}_{hk}(1-\hat{\pi}_{hk})}$$

$$\hat{\gamma}_h = \frac{\hat{\theta}_h - \hat{\pi}_{hk}\hat{\pi}_{hk'}}{\sqrt{\hat{\pi}_{hk}(1-\hat{\pi}_{hk})\hat{\pi}_{hk'}(1-\hat{\pi}_{hk'})}} = \left(\frac{\hat{\theta}_h}{\hat{\pi}_{hk}\hat{\pi}_{hk'}} - 1\right)\left\{\left(\frac{\hat{\pi}_{hk}}{1-\hat{\pi}_{hk}}\right)\left(\frac{\hat{\pi}_{hk'}}{1-\hat{\pi}_{hk'}}\right)\right\}^{\frac{1}{2}}$$

$$\hat{\nu}_h = \frac{\hat{\phi}_h - \hat{\pi}_{hk}\hat{\pi}_{hk'}}{\sqrt{\hat{\pi}_{hk}(1-\hat{\pi}_{hk})\hat{\pi}_{hk'}(1-\hat{\pi}_{hk'})}} =$$

$$= \left(\frac{\hat{\phi}_h}{\hat{\pi}_{hk}\hat{\pi}_{hk'}} - 1\right)\left\{\left(\frac{\hat{\pi}_{hk}}{1-\hat{\pi}_{hk}}\right)\left(\frac{\hat{\pi}_{hk'}}{1-\hat{\pi}_{hk'}}\right)\right\}^{\frac{1}{2}}$$

$$\hat{\tau}_k = \frac{\hat{\psi}_h - \hat{\pi}_{hk}\hat{\pi}_{h'k}}{\sqrt{\hat{\pi}_{hk}(1-\hat{\pi}_{hk})\hat{\pi}_{h'k}(1-\hat{\pi}_{h'k})}} =$$

$$= \left(\frac{\hat{\psi}_h}{\hat{\pi}_{hk}\hat{\pi}_{h'k}} - 1\right)\left\{\left(\frac{\hat{\pi}_{hk}}{1-\hat{\pi}_{hk}}\right)\left(\frac{\hat{\pi}_{h'k}}{1-\hat{\pi}_{h'k}}\right)\right\}^{\frac{1}{2}}$$

$$\hat{\omega}_{hk,h'k'} = \frac{\hat{\psi}_{hk,h'k'} - \hat{\pi}_{hk}\hat{\pi}_{h'k'}}{\sqrt{\hat{\pi}_{hk}(1-\hat{\pi}_{hk})\hat{\pi}_{h'k'}(1-\hat{\pi}_{h'k'})}} =$$

$$= \left(\frac{\hat{\psi}_{hk,h'k'}}{\hat{\pi}_{hk}\hat{\pi}_{h'k'}} - 1\right)\left\{\left(\frac{\hat{\pi}_{hk}}{1-\hat{\pi}_{hk}}\right)\left(\frac{\hat{\pi}_{h'k'}}{1-\hat{\pi}_{h'k'}}\right)\right\}^{\frac{1}{2}}$$

then it follows that

$$\begin{bmatrix} \hat{\pi}_{11} \\ \hat{\pi}_{12} \\ \hat{\pi}_{21} \\ \hat{\pi}_{22} \\ \rho_{11} \\ \rho_{12} \\ \rho_{21} \\ \rho_{22} \\ \gamma_1 \\ \gamma_2 \\ \nu_1 \\ \nu_2 \\ \tau_1 \\ \tau_2 \\ \omega_{11,22} \\ \omega_{12,21} \end{bmatrix} = \exp(A_5 \ln(A_4(C_2 + \exp(A_3 \ln(C_1 + A_2 \exp(A_1 \ln F)))))) \qquad (6.2.26)$$

where

$$\underset{20 \times 16}{\underset{\sim}{A}_1} = \begin{bmatrix} \multicolumn{2}{c}{\underset{\sim}{I}_{16}} \\ \begin{matrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{matrix} & \underset{\sim}{0}_{4,12} \end{bmatrix},$$

$$\underset{20 \times 20}{A_2} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1
\end{bmatrix}$$

$$C_1 = [0 \; 1 \; 0 \; 1 \; 0 \; 1 \; 0 \; 1 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0]',$$

$$\underset{20\times 20}{A_3} = \begin{bmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
-1&-1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&-1&-1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&-1&-1&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&-1&-1&0&0&0&0&0&1&0&0&0&0&0&0\\
-1&0&-1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&-1&0&-1&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&0&0\\
-1&0&-1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&-1&0&-1&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&\frac{1}{2}&-\frac{1}{2}&\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&0&0\\
-\frac{1}{2}&-\frac{1}{2}&0&0&-\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&-\frac{1}{2}&-\frac{1}{2}&0&0&-\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&1&0&0\\
-\frac{1}{2}&-\frac{1}{2}&0&0&0&0&-\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&-\frac{1}{2}&0&0&0&0&0&0&0&0&0&0&0&0&0&1
\end{bmatrix}$$

$$C_2 = [0\ 0\ 0\ 0\ -1\ -1\ -1\ -1\ -1\ -1\ 0\ 0\ -1\ -1\ 0\ 0\ 0\ 0\ 0\ 0]',$$

$$\underset{20\times 20}{A_4} = \begin{bmatrix} I_4 & & \\ & -I_4 & \\ & & I_{12} \end{bmatrix},$$

and

$$\underset{16 \times 20}{A_5} = \begin{bmatrix} \underset{\sim}{I}_8 & & & \\ & \begin{matrix}1010\\0101\end{matrix} & & \\ & & \begin{matrix}1010\\0101\end{matrix} & \\ & & & \underset{\sim}{I}_4 \end{bmatrix}.$$

These computations are easily performed using the computer program MISCAT.

Alternatively statistical tests for the correlation coefficient parameters being zero are performed by evaluating

$$\frac{\ln\left(\dfrac{\hat{\lambda}_{hk}}{\hat{\pi}^2_{hk}}\right)}{\text{s.e. }\ln\left(\dfrac{\hat{\lambda}_{hk}}{\hat{\pi}^2_{hk}}\right)}, \tag{6.2.27}$$

$$\frac{\ln\left(\dfrac{\hat{\theta}_h}{\hat{\pi}_{hk}\hat{\pi}_{hk'}}\right)}{\text{s.e. }\ln\left(\dfrac{\hat{\theta}_h}{\hat{\pi}_{hk}\hat{\pi}_{hk'}}\right)}, \tag{6.2.28}$$

$$\frac{\ln\left(\dfrac{\hat{\phi}_h}{\hat{\pi}_{hk}\hat{\pi}_{hk'}}\right)}{\text{s.e. }\ln\left(\dfrac{\hat{\phi}_h}{\hat{\pi}_{hk}\hat{\pi}_{hk'}}\right)}, \tag{6.2.29}$$

$$\frac{\ln\left(\dfrac{\hat{\psi}_h}{\hat{\pi}_{hk}\hat{\pi}_{h'k}}\right)}{\text{s.e. }\ln\left(\dfrac{\hat{\psi}_h}{\hat{\pi}_{hk}\hat{\pi}_{h'k}}\right)}, \tag{6.2.30}$$

$$\frac{\ln\left(\frac{\hat{\psi}_{hk,h'k'}}{\hat{\pi}_{hk}\hat{\pi}_{h'k'}}\right)}{\text{s.e. }\ln\left(\frac{\hat{\psi}_{hk,h'k'}}{\hat{\pi}_{hk}\hat{\pi}_{h'k'}}\right)}\,. \tag{6.2.31}$$

using a normal distribution approximation. The numerators of these expressions and their estimated coavariance matrix are obtained by using the matrix operations $L=A_2\ln(A_1F)$ on vector $\underset{\sim}{F}$ in (6.2.25), where L is the vector whose elements are the numerators described in (6.2.27) to (6.2.31) and

$$\underset{16\times 16}{\underset{\sim}{A}_1} = \begin{bmatrix} \underset{\sim}{B}_1 & \underset{\sim}{0}_{4,6} & \underset{\sim}{0}_{4,4} \\ \underset{\sim}{I}_6 & -\underset{\sim}{I}_6 & \underset{\sim}{0}_{6,4} \\ \underset{\sim}{B}_2 & \underset{\sim}{0}_{2,6} & \underset{\sim}{0}_{2,4} \\ \underset{\sim}{0}_{4,6} & \underset{\sim}{0}_{4,6} & \underset{\sim}{I}_4 \end{bmatrix} \text{ where } \underset{\sim}{B}_1 = [\underset{\sim}{I}_4 \quad \underset{\sim}{0}_{4,2}] \text{ and } \underset{\sim}{B}_2 = [\underset{\sim}{0}_{2,4} \quad \underset{\sim}{I}_2]\,,$$

$$\underset{12\times 16}{\underset{\sim}{A}_2} = \begin{bmatrix} -2\underset{\sim}{I}_4 & \underset{\sim}{I}_4 & \underset{\sim}{0}_{4,4} & \underset{\sim}{0}_{4,4} \\ \begin{array}{cccc} -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 \end{array} & \underset{\sim}{0}_{4,4} & \underset{\sim}{I}_4 & \underset{\sim}{0}_{4,4} \\ \begin{array}{cccc} -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ -1 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 \end{array} & \underset{\sim}{0}_{4,4} & \underset{\sim}{0}_{4,4} & \underset{\sim}{I}_4 \end{bmatrix}.$$

6.4 The Modular Case: Two-Time Point Responses

When the structure of the data is such that modules of different cluster sizes can be formed, then the means estimation approach just presented can be used for each module, which makes estimation of parameters as well as evaluation of their covariance structure convenient and somewhat simplified.

If the m-th module has n_m clusters, each of size $2d_m$ where d_m observations are from sub-population 1 and d_m observations are from sub-population 2, then

$$y_{hijkm} = \begin{cases} 1 & \text{if k-th response in m-th module for the j-th observation in} \\ & \text{the i-th cluster and h-th sub-population, has the attribute} \\ 0 & \text{otherwise} \end{cases}$$

$(m=1,2,\ldots,M;\ h=1,2;\ i=1,2,\ldots,\ n_m;\ j=1,2,\ldots,\ d_m;\ k=1,2)$.

Estimators for the m-th module are identical to those in expressions (6.2.9) up to (6.2.24) with the addition of an 'm' subscript to n, d and y_{hijk}'s. So, the vector of parameter estimates for the m-th module may be written as

$$\underset{(16\times 1)}{\underset{\sim}{F}_m} = [\hat{\pi}_{hk,m},\ \hat{\theta}_{h,m},\ \hat{\pi}_{hk,m}-\hat{\lambda}_{hk,m},\ \hat{\theta}_{h,m}-\hat{\phi}_{h,m},\ \hat{\psi}_{h,m},\ \hat{\psi}_{(hk,h'k'),m}]' \quad (6.2.32)$$

The GSK methodology can be used to obtain overall parameter estimators and standard errors of these estimators by fitting the linear model

$$E(\underset{\sim}{F}) = \underset{\sim}{X}\underset{\sim}{\beta}$$

where

$$\underset{\sim}{F}_{(16M\times 1)} = [\underset{\sim}{F}_1',\ \ldots,\ \underset{\sim}{F}_m' \ \ldots,\ \underset{\sim}{F}_M']' \ ,$$

$\underset{(16M\times 16M)}{\underset{\sim}{X}} = \underset{(16\times 16)}{\underset{\sim}{I}} \otimes \underset{\sim}{1}_{M\times 1}$ and $\underset{\sim}{\beta} = (\pi_{hk},\ \theta_h,\ \pi_{hk}-\lambda_{hk},\ \theta_h-\phi_h,\ \psi_h,\ \psi_{hk,h'k'})'$. The estimate of $\underset{\sim}{\beta}$ is obtained as for the one treatment situation by computing $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}$ with $\underset{\sim}{V}(\hat{\underset{\sim}{\beta}}) = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}$ where $\underset{\sim}{V}_{\underset{\sim}{F}}$ is block diagonal with block diagonal elements $\underset{\sim}{V}_{F_m} = \frac{1}{n_m^2}\sum_{i=1}^{n_m}(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)'$.

Estimates derived by GSK methodology will have optimal properties only if the number of clusters in a module is reasonably large ($n_m \geq 15$); otherwise other methods such as direct estimates like mean of means may be employed.

6.5 Example

In this section a completely balanced case example with the two sub-populations per cluster and two binary responses per subject is presented.

The day care center dataset with an additional component of it, mothers of the day care children, as the other sub-population, will be used to illustrate the implementation of this methodology. Among other questions answered by the mothers in a self administered questionnaire, similar to the one filled out by the staff members of the day care centers, some had to do with how they wanted day care staff to act (call parent to pick up child or not) if child was perceived as having fever and/or symptoms/signs of illness. Some of the questions of interest in this two sub-population setting are: i) whether parents have a correlation structure in relation to day care centers; and, ii) if answer is yes, is this correlation larger or smaller than that for staff. Answers to these questions are obtained by applying the measures for within cluster response similarity developed in this chapter to staff and parent data and then comparing the ones for parents with the ones for staff.

In this example, data on parents and staff are adopted to be compatible with what is required by the estimation strategy being illustrated. There are 19 clusters and 8 subjects per cluster, with 4 in each sub-population. Each subject contributes with two binary responses (yes, no) to the question of calling or not the parent to pick up a child if he/she is i) with sore throat or ii) unusually cranky. Since the data are balanced, the means estimator formulation can be used to examine the questions of interest.

By applying the compound vector (6.2.26) with its series of linear logarithmic and exponential operators to vectors (6.2.25), π's and ρ's are jointly estimated. The weighted least squares estimates of these parameters as well as their standard deviations are summarized in Table 6.2.

Table 6.2: Parameter estimates and standard errors for the two sub-populations balanced data example

Focus of estimation	Parameter function	WLS estimate (s.e.)
π's	mean response	
	staff/sore throat (π_{11})	0.64 (0.066)
	staff/cranky (π_{12})	0.36 (0.071)
	mother/sore throat (π_{21})	0.51 (0.051)
	mother/cranky (π_{22})	0.34 (0.049)
ρ's	intraclass correlation	
	staff/sore throat (ρ_{11})	0.11 (0.084)*
	staff/cranky (ρ_{12})	0.20 (0.146)**
	mothers/sore throat (ρ_{21})	0.02 (0.093)
	mothers/cranky (ρ_{22})	0.08 (0.072)
	staff/sore throat cranky (γ_1)	0.44 (0.077)**
	mothers/sore throat cranky (γ_2)	0.59 (0.077)**
	staff/subject k, sore throat	
	subject k′, cranky (ν_1)	0.14 (0.074)*
	mothers/subject k sore throat	0.01 (0.070)
	subject k′, cranky (ν_2)	
	mothers,staff/sore throat (τ_1)	0.05 (0.065)
	mothers, staff/cranky (τ_2)	0.01 (0.062)
	mothers/sore throat ($\omega_{11,22}$)	0.01 (0.062)
	staff/cranky	
	mothers/cranky ($\omega_{12,21}$)	0.009 (0.051)
	staff/sore throat	

**Significant at 0.05

* Significant at 0.1

Results from Table 6.2 show mean responses for staff calling parent to pick up a child significantly higher (α=0.05) if child has a sore throat than if he/she is unusually cranky. Similar results hold for mothers answers of what they would want staff to do under these two situations. Proportion of staff calling for child with sore throat is not signficantly different from mothers proportion of what they wanted staff to do under that particular situation. For cranky children, proportions are very close for the two groups, staff and parents. In both cases, child with a sore throat or unusually cranky, staffs proportion are a little higher (even though not significantly) than those for mothers. If clustering had been ignored, s.e. $(\hat{\pi}_{11}) = 0.055$, s.e. $(\hat{\pi}_{12}) = 0.055$, s.e. $(\hat{\pi}_{21}) = 0.057$ and s.e. $(\hat{\pi}_{22}) = 0.054$. The time correlation coefficient for staff's responses to the two questions (0.44) and for mothers responses of what they wanted staff to do under the two situations (0.59) are significantly different from zero butchild is significantly different from zero (α=0.05) while for sore throat is of borderline not from each other. Intraclass correlation for staff's answers to the question of a cranky significance. Mothers answers intraclass correlation coefficient for the two questions are not different from zero. The correlation coefficient for staff involving different subjects and different questions is significantly different from zero (α=0.05). The correlation coefficients involving different sub-populations are not significantly different from zero.

CHAPTER VII

ANALYSIS OF CLUSTERED BINARY DATA FOR R TIME POINTS

7.1 Introduction

Chapters V and VI presented the theoretical framework for the analysis of clustered attribute data where two conditions generating one binary response each, were observed on each subject. The methods presented there are readily extended to the situation where a set of r binary responses is observed on each subject within a cluster. Some examples of studies with r binary responses are: 1) a clinical dental material study where patients with two or more treated teeth are examined every six months for a period of three years for one characteristic of interest such as plaque retention in treated teeth (Yes, No); 2) a day care center study where a set of questions about sick child care is posed to staff members. It is of interest to learn how staff handle the situations of a child having a temperature of 100°F-101°F and one of the following symptoms: i) new runny nose, ii) ear pain, iii) new cough, iv) diarrhea, v) conjunctivitis, vi) unusually cranky. The response of interest is that of calling parent for immediate pick up of perceived ill child or not.

The extension from two to r (>2) binary responses has some complexity due to the increase in number of parameters to be estimated. The one and two treatments per cluster situations will be examined. The finite population parameters are formulated by assuming a sampling without replacement framework. As in previous chapters, the infinite case is of particular interest, that is, the case where N and D tend to infinity. In this infinite case, sampling with and without replacement are equivalent.

7.2 One Treatment per Cluster: Balanced Case

Consider a finite population of N clusters, each containing D observations (or subjects). Let n clusters be randomly selected without replacement from N clusters and d observations randomly selected from D observations in the respective n sampled clusters.

Let

$$y_{ijk} = \begin{cases} 1 & \text{if k-th response for the j-th subject} \\ & \quad \text{in the i-th cluster is favorable} \\ 0 & \text{otherwise} \end{cases}$$

where $i=1,2,\ldots,n$; $j=1,2,\ldots,d$ and $k=1,2,\ldots,r$.

The following parameter definitions characterize the basic model of interest

$E(y_{ijk}) = \pi_k$ = Pr (k-th response for j-th subject in i-th cluster is favorable)

$E(y_{ijk}\ y_{ij'k}) = \lambda_k$ = Pr (k-th binary responses for two distinct subjects in a cluster are favorable)

$E(y_{ijk}\ y_{ijk'}) = \theta_{kk'}$ = Pr (that responses k and k′ of a subject in a cluster are both favorable). There are C_2^k such θ's ($1 \le k < k' \le r$).

$E(y_{ijk}\ y_{ij'k'}) = \phi_{kk'}$ = Pr (response k from subject j and response k′ from subject j′ are favorable). There are C_2^k such ϕ's ($1 \le k < k' \le r$).

So, there are for the respective clusters, when π's, λ's, θ's and ϕ's are combined, $r(r+1)$ parameters to be estimated.

Additional parameters of interest are the within cluster correlation coefficients.

Let

$$\mathrm{Cov}(y_{ijk}, y_{ij'k}) = \lambda_k - \pi_k^2 ,$$

$$\mathrm{Cov}(y_{ijk}, y_{ijk'}) = \theta_{kk'} - \pi_k \pi_{k'} \quad 1 \le k < k' \le r \quad ,$$

and

$$\mathrm{Cov}(y_{ijk}, y_{ij'k'}) = \phi_{kk'} - \pi_k \pi_{k'} \quad 1 \le k < k' \le r \quad .$$

Then,

$$\rho_k = \frac{\lambda_k - \pi_k^2}{\pi_k(1-\pi_k)} \quad ,$$

$$\gamma_{kk'} = \frac{\theta_{kk'} - \pi_k \pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}} \qquad 1 \leq k < k' \leq r \quad ,$$

and

$$\nu_{kk'} = \frac{\phi_{kk'} - \pi_k \pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}} \qquad 1 \leq k < k' \leq r \quad .$$

Unbiased estimators of the population quantities are

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ijk}}{nd} = \bar{y}_{\cdot\cdot k} \quad , \tag{7.2.1}$$

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n}\sum_{j \neq j'=1}^{d}\sum^{d} y_{ijk}\, y_{ij'k}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{i\cdot k}^2 - y_{\cdot\cdot k}}{nd(d-1)} \quad , \tag{7.2.2}$$

$$\hat{\theta}_{kk'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ijk}\, y_{ijk'}}{nd} \qquad 1 \leq k < k' \leq r \quad , \tag{7.2.3}$$

and

$$\hat{\phi}_{kk'} = \frac{\sum_{i=1}^{n}\left(\sum_{j \neq j'=1}^{d}\sum^{d} y_{ijk}\, y_{ij'k'}\right)}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{i\cdot k}\, y_{i\cdot k'} - nd\,\hat{\theta}_{kk'}}{nd(d-1)} \; 1 \leq k < k' \leq r \quad . \tag{7.2.4}$$

Also, if

$$s_k^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{ijk} - \bar{y}_{i\cdot k})^2}{n(d-1)} \tag{7.2.5}$$

and

$$s_{kk'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{ijk} - \bar{y}_{i\cdot k})(y_{ijk'} - \bar{y}_{i\cdot k'})}{n(d-1)} \; , \tag{7.2.6}$$

then, $s_k^2=(\hat{\pi}_k-\hat{\lambda}_k)$ and $s_{kk'} = (\hat{\theta}_{kk'}-\hat{\phi}_{kk'})$.

The parameter estimators $\hat{\pi}_k$ and $\hat{\theta}_{kk'}$ may be written as

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij1}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i11} = F_{11} \tag{7.2.7}$$

$$\vdots$$

$$\hat{\pi}_r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ijr}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i1r} = F_{1r} \tag{7.2.8}$$

$$\hat{\theta}_{12} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij1}\, y_{ij2}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i21} = F_{21} \tag{7.2.9}$$

$$\vdots$$

$$\hat{\theta}_{(r-1),r} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij(r-1)}y_{ijr}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i2,r(r-1)} = F_{2,r(r-1)} \tag{7.2.10}$$

Also one can form

$$(\hat{\pi}_1-\hat{\lambda}_1) = s_1^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{ij1}-F_{i11})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i31} = F_{31} \tag{7.2.11}$$

$$\vdots$$

$$(\hat{\pi}_r-\hat{\lambda}_r) = s_r^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{ijr}-F_{i1r})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i3r} = F_{3r} \tag{7.2.12}$$

$$(\hat{\theta}_{12}-\hat{\phi}_{12}) = s_{12} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{1}{d-1}\sum_{j=1}^{d} \Big(y_{ij1}-F_{i11}\Big)\Big(y_{ij2}-F_{i12}\Big)\right\} = \frac{1}{n}\sum_{i=1}^{n} F_{i41} = F_{41} \tag{7.2.13}$$

$$\vdots$$

$$(\hat{\theta}_{(r-1),r}-\hat{\phi}_{(r-1),r}) = s_{(r-1),r} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{1}{d-1}\sum_{j=1}^{d} \Big(y_{ij(r-1)}-F_{i1(r-1)}\Big)\Big(y_{ijr}-F_{i1r}\Big)\right\} =$$

$$= \frac{1}{n}\sum_{i=1}^{n} F_{i4,r(r-1)} = F_{4,r(r-1)}. \tag{7.2.14}$$

It then follows that

$$
\underset{\sim}{F} = \begin{bmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_r \\ \hat{\theta}_{12} \\ \vdots \\ \hat{\theta}_{(r-1),r} \\ (\hat{\pi}_1-\hat{\lambda}_1) \\ \vdots \\ (\hat{\pi}_r-\hat{\lambda}_r) \\ (\hat{\theta}_{12}-\hat{\phi}_{12}) \\ \vdots \\ (\hat{\theta}_{(r-1),r}-\hat{\phi}_{(r-1),r}) \end{bmatrix} = \begin{bmatrix} F_{11} \\ \vdots \\ F_{1r} \\ F_{21} \\ \vdots \\ F_{2,r(r-1)} \\ F_{31} \\ \vdots \\ F_{3r} \\ F_{41} \\ \vdots \\ F_{4,r(r-1)} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} F_{i11} \\ \vdots \\ F_{i1r} \\ F_{i21} \\ \vdots \\ F_{i2,r(r-1)} \\ F_{i31} \\ \vdots \\ F_{i3r} \\ F_{i41} \\ \vdots \\ F_{i4,r(r-1)} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_i \,. \qquad (7.2.15)
$$

The data formulation presented in (7.2.5) – (7.2.12) shows that the parameter estimates can be obtained as functions of means of suitably defined quantities. The corresponding covariance matrix estimate is $\underset{\sim}{V}_{\hat{F}} = \frac{1}{n^2}\sum_{i=1}^{n} (\underset{\sim}{F}_i-\underset{\sim}{F})(\underset{\sim}{F}_i-\underset{\sim}{F})'$ with $\underset{\sim}{F}$ and $\underset{\sim}{F}_i$ from (7.2.13).

The parameters of primary interest are π's and ρ's. Estimates of these $r(r+1)$ parameters along with their estimated covariance matrix can be obtained as it was done in Chapter V, by applying a series of appropriate linear, logarithmic and exponential transformations to the estimated vector $\hat{\underset{\sim}{F}}$ (7.1.13).

7.3 One Treatment per Cluster: Partially Balanced Design

Let

$$
y_{ijkm} = \begin{cases} 1 & \text{if the k-th response for subject j in the i-th cluster of m-th module is favorable} \\ 0 & \text{otherwise} \end{cases}
$$

$i=1,2,\ldots,n_m$; $j=1,2,\ldots,d_m$; $k=1,2,\ldots,r$; $m=1,2,\ldots,M$.

The estimators for the m-th module have the same form as expressions (7.2.1), (7.2.2), (7.2.3), and (7.2.4) with addition of an "m" subscript on y_{ijk}, n and d. In order to produce estimates of the parameters and the estimated covariance matrix for each module, a means formulation approach may be done as in section 7.2 (expressions (7.2.5) – (7.2.12)).

These modular estimators can be combined by using WLS analysis provided each cluster has sufficient sample size. The vector $\underset{\sim}{F}_m$ of means for the m-th module has the same structure as $\underset{\sim}{F}$ in (7.2.13) but with an "m" added to all subscripts of the elements of $\underset{\sim}{F}$.

This modular situation can be studied by using WLS methodology with the model

$$E(\underset{\sim}{F})=\underset{\sim}{X}\underset{\sim}{\beta}$$

where

$$\underset{[r(r+1)M\times r(r+1)]}{\underset{\sim}{X}} = \begin{bmatrix} I_{r(r+1)} \\ I_{r(r+1)} \\ \vdots \\ I_{r(r+1)} \end{bmatrix} \quad \text{and} \quad \underset{[r(r+1)\times 1]}{\underset{\sim}{\beta}} = \begin{bmatrix} \{\pi_k\} \\ \{\theta_{kk'}\} \\ \{\pi_k-\lambda_k\} \\ \{\theta_{kk'}-\phi_{kk'}\} \end{bmatrix}.$$

The estimate $\hat{\underset{\sim}{\beta}}$ is obtained from $\hat{\underset{\sim}{\beta}}=(\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}$ and an estimate for its covariance matrix is $\underset{\sim}{V}_{\hat{\underset{\sim}{\beta}}}=(\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}$. Here, $\underset{\sim}{V}_{\underset{\sim}{F}}$ is a block diagonal matrix with diagonal blocks $\underset{\sim}{V}_m=\underset{\sim}{V}(\underset{\sim}{F}_m) = \frac{1}{n_m^2}\sum_{i=1}^{n} (\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)'$ where $\underset{\sim}{F}_{im}$ is as $\underset{\sim}{F}_i$ in (7.2.13) with "m" added to all subscripts of the elements of $\underset{\sim}{F}_i$.

7.4 <u>Two Treatments per Cluster: Balanced Case</u>

Let

$$y_{hijk} = \begin{cases} 1 & \text{if k-th response of j-th subject in i-th} \\ & \text{cluster of h-th population is favorable} \\ 0 & \text{otherwise} \end{cases}$$

where $h=1,2$; $i=1,2,\ldots,n$; $j=1,2,\ldots,d$; $k=1,2,\ldots,r$.

The following definitions characterize the basic model of interest in this two treatments situation:

$E(y_{hijk}) = \pi_{hk}$ = Pr (k-th response of a subject in subpopulation h, in a randomly selected cluster is favorable)

$E(y_{hijk}\, y_{hij'k}) = \lambda_{hk}$ = Pr (k-th responses for two different subjects in a randomly selected cluster in population h, are both favorable)

$E(y_{hijk}\, y_{hijk'}) = \theta_{h,kk'}$ = Pr (responses k and k' for a subject in a cluster are both favorable). There are C_2^r of such θ_h's or $2C_2^r$ θ's $(1 \le k < k' \le r)$.

$E(y_{hijk}\, y_{hij'k'}) = \phi_{h,kk'}$ = Pr (response k from subject j and response k' from subject j' are favorable). There are C_2^r ϕ_h's or $2C_2^r$ ϕ's $(1 \le k < k' \le r)$.

$E(y_{hijk}\, y_{h'ij'k}) = \psi_k$ = Pr (k-th responses for subject j in subpopulation h and subject j' in sub-population h' are favorable)

$E(y_{hijk}\, y_{h'ij'k'}) = \psi_{hk,h'k'}$ = Pr (response k for subject j in subpopulation h and response k' for subject j' in subpopulation h' are favorable). There are P_2^r such ψ's $(1 \le k \ne k' \le r,$ $1 \le h < h' \le 2)$.

Thus, for the respective clusters with two sub-populations per cluster and r responses per subject, there are $3r^2+2r$ parameters.

Other parameters of interest are the intraclass correlation coefficients.

Let

$$\mathrm{Cov}(y_{hijk}\, y_{hij'k}) = \lambda_{hk} - \pi_{hk}^2, \quad h=1,2;\ k=1,2,\ldots,r\ ,$$

$$\mathrm{Cov}(y_{hijk}, y_{hijk'}) = \hat{\theta}_{h,kk'} - \pi_{hk}\pi_{hk'} \quad 1 \le k < k' \le r;\ h=1,2\ ,$$

$$\mathrm{Cov}(y_{hijk}, y_{hij'k'}) = (\phi_{h,kk'} - \pi_{hk}\pi_{hk'}) \quad 1 \le k < k' \le r,$$

$$\mathrm{Cov}(y_{hijk}, y_{h'ij'k}) = (\psi_k - \pi_{hk}\pi_{h'k}) \quad 1 \le k < k' \le r;\ h=1,2\ ,$$

$$\mathrm{Cov}(y_{hijk}, y_{h'i'j'k'}) = \psi_{hk,h'k'} - \pi_{hk}\pi_{hk'}) \quad 1 \le k < k' \le r; \quad h=1,2 .$$

Then,

$$\rho_{hk} = \frac{\lambda_{hk} - \pi_{hk}^2}{\pi_{hk}(1-\pi_{hk})} ,$$

$$\gamma_{h,kk'} = \frac{\theta_{h,kk'} - \pi_{hk}\pi_{hk'}}{\sqrt{\pi_{hk}(1-\pi_{hk})\pi_{hk'}(1-\pi_{hk'})}} \quad 1 \le k < k' \le r ,$$

$$\nu_{h,kk'} = \frac{\phi_{h,kk'} - \pi_{hk}\pi_{hk'}}{\sqrt{\pi_{hk}\ (1-\pi_{hk})\pi_{hk'}(1-\pi_{hk'})}} \quad 1 \le k < k' \le r ,$$

$$\tau_k = \frac{\psi_k - \pi_{hk}\pi_{h'k}}{\sqrt{\pi_{hk}\ (1-\pi_{hk})\pi_{h'k}(1-\pi_{h'k})}} \quad 1 \le k < k' \le r ,$$

and

$$\omega_{(hk,h'k')} = \frac{\psi_{hk,h'k'} - \pi_{hk}\pi_{h'k'}}{\sqrt{\pi_{hk}\ (1-\pi_{hk})\pi_{h'k'}(1-\pi_{h'k'})}} \quad 1 \le k < k' \le r .$$

Unbiased estimators of the population parameters are:

$$\hat{\pi}_{hk} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hijk}}{nd} = \bar{y}_{h\cdot\cdot k} , \tag{7.4.1}$$

$$\hat{\lambda}_{hk} = \frac{\sum_{i=1}^{n}\sum_{j\neq j'=1}^{d}\sum^{d} y_{hijk}\, y_{hij'k}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{hi\cdot k}^2 - y_{h\cdot\cdot k}}{nd(d-1)} , \tag{7.4.2}$$

$$\hat{\theta}_{h,kk'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hijk}\, y_{hijk'}}{nd} \quad 1 \le k < k' \le r , \tag{7.4.3}$$

$$\hat{\phi}_{h,kk'} = \frac{\sum_{i=1}^{d}\sum_{j\neq}^{d}\sum_{j'=1}^{d} y_{hijk}\, y_{hij'k'}}{nd(d-1)} = \frac{\sum_{i=1}^{n} y_{hi\cdot k}y_{hi\cdot k'} - nd\theta_h}{nd(d-1)} \quad 1 \le k < k' \le r , \tag{7.4.4}$$

$$\hat{\psi}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{hijk}\, y_{h'ij'k}}{nd^2} = \frac{\sum_{i=1}^{n} \bar{y}_{hi\cdot k}\, \bar{y}_{h'i\cdot k}}{n} \quad 1 \leq k < k' \leq r\ , \tag{7.4.5}$$

$$\hat{\psi}_{hk,h'k'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{hijk}\, y_{h'ij'k'}}{nd^2} = \frac{\sum_{i=1}^{n} \bar{y}_{hi\cdot k}\, \bar{y}_{h'i\cdot k'}}{n} \quad 1 \leq k < k' \leq r\ . \tag{7.4.6}$$

The parameter estimators $\hat{\pi}_{hk}$, $\hat{\theta}_{kk'}$, $\hat{\psi}_k$, $\hat{\psi}_{hk,h'k'}$ may be written as

$$\hat{\pi}_{hk} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hijk}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i,1hk} = F_{1hk} \tag{7.4.7}$$

$$\hat{\theta}_{h,kk'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{hijk}\, y_{hijk'}}{nd} = \frac{1}{n}\sum_{i=1}^{n} F_{i,2h,kk'} = F_{2hkk'} \tag{7.4.8}$$

$$\hat{\psi}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{hijk}\, y_{h'ij'k}}{nd^2} = \frac{1}{n}\sum_{i=1}^{n} F_{i,3k} = F_{3k} \tag{7.4.9}$$

$$\hat{\psi}_{hk,h'k'} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=1}^{d} y_{hijk}\, y_{h'ij'k'}}{nd^2} = \frac{1}{n}\sum_{i=1}^{n} F_{i,4hk,h'k'} = F_{4hk,h'k'}\ . \tag{7.4.10}$$

Also one can form

$$(\hat{\pi}_{hk} - \hat{\lambda}_{hk}) = s^2_{hk} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} (y_{hijk} - F_{1hk})^2}{n(d-1)} = \frac{1}{n}\sum_{i=1}^{n} F_{i,5hk} = F_{5hk} \tag{7.4.11}$$

and

$$(\hat{\theta}_{h,kk'} - \hat{\phi}_{h,kk'}) = s_{h,kk'} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{\sum_{j=1}^{d} (y_{hijk} - \bar{y}_{hi\cdot k})(y_{hijk'} - \bar{y}_{hi\cdot k'})}{d-1} \right\} =$$

$$\frac{1}{n}\sum_{i=1}^{n} F_{i,6h,kk'} = F_{6h,kk'} \ . \tag{7.4.12}$$

It then follows that

$$F_{i1} = (F_{i,1,11}, \ldots, F_{i,1,1r}, F_{i,1,21}, \ldots, F_{i,1,2r})' \ , \tag{7.4.13}$$

$$F_{i2} = (F_{i,21,12}, \ldots, F_{i,21,(r-1),r}, F_{i,22,12}, \ldots, F_{i,22,(r-1),r})' \ , \tag{7.4.14}$$

$$F_{i3} = (F_{i,31}, \ldots, F_{i3r})' \tag{7.4.15}$$

$$F_{i4} = (F_{i,411,22}, F_{i,411,23}, \ldots, F_{i,411,r-1,r}, \ldots, F_{i,41r,21}, F_{i,41r,22}, \ldots, F_{i,41r,2(r-1)})' \ , \tag{7.4.16}$$

$$F_{i5} = (F_{i,511}, \ldots, F_{i51r}, F_{i521}, \ldots, F_{i52r})' \ , \tag{7.4.17}$$

and

$$F_{i6} = (F_{i,61,12}, \ldots, F_{i,61,(r-1),r}, F_{i62,12}, \ldots, F_{i,62,(r-1),r})' \ , \tag{7.4.18}$$

so

$$\underset{\sim}{F}_i = (\underset{\sim}{F}'_{i1}, \underset{\sim}{F}'_{i2}, \underset{\sim}{F}'_{i3}, \underset{\sim}{F}'_{i4}, \underset{\sim}{F}'_{i5}, \underset{\sim}{F}'_{i6})' \tag{7.4.19}$$

and

$$\underset{\sim}{F} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{F}_i \ . \tag{7.4.20}$$

This data formulation presented in (7.4.7) up to (7.4.20) shows that the parameter estimates $\hat{\pi}_{hk}$, $\hat{\lambda}_{hk}$, $\hat{\theta}_{kk'}$, $\hat{\phi}_{kk'}$, $\hat{\psi}_k$ and $\hat{\psi}_{hk,hk'}$ can be obtained through means, i.e., the vector $\underset{\sim}{F}$ in (7.4.20). The corresponding covariance matrix estimated is $\underset{\sim}{V}_{\underset{\sim}{F}} = \frac{1}{n^2}\sum_{i=1}^{n} (\underset{\sim}{F}_i - \underset{\sim}{F})(\underset{\sim}{F}_i - \underset{\sim}{F})'$ with $\underset{\sim}{F}_i$ from (7.4.19) and $\underset{\sim}{F}$ from (7.4.20).

There are $(3r^2+2r)$ parameters of primary interest in this two-subpopulation setting. The π's and ρ's can be estimated along with their covariance matrix by applying a series of appropriate linear, logarithmic and exponential transformations to the estimated vector $\underset{\sim}{F}$ in (7.4.20).

7.5 Two Treatments per Cluster: Partially Balanced Case

Suppose data can be grouped in such a way that modules of the same size clusters can be formed.

$$y_{hijkm} = \begin{cases} 1 & \text{if k-th response in m-th module for the j-th subject in the} \\ & \text{i-th cluster and h-th sub-population, has the attribute} \\ 0 & \text{otherwise} \end{cases}$$

($m=1,2,\ldots,M$; $h=1,2$; $i=1,2,\ldots,d_m$; $j=1,2,\ldots,r$; $k=1,2$).

The estimators for the m-th module are like those in expressions (7.4.1) up to (7.4.6) with the addition of an "m" subscript to n,d and the y_{hijk}'s.

Grouping the data in terms of means as it was done in (7.4.7) through (7.4.12), the modular estimators can be analyzed using the WLS methodology provided the number of clusters within modules is sufficiently large.

The parameter estimates for the m-th module have the same structure as $\underset{\sim}{F}$ in (7.4.20) but with an "m" added to all subscripts of the elements of $\underset{\sim}{F}$.

The WLS procedure can be applied to obtain overall parameter estimators by fitting the linear model

$$E(\underset{\sim}{F}) = \underset{\sim}{X}\underset{\sim}{\beta}$$

with $\underset{[(3r^2+2r)M\times(3r^2+2r)]}{\underset{\sim}{X}} = \underset{[(3r^2+2r)\times(3r^2+2r)]}{\underset{\sim}{I}} \otimes \underset{[M\times 1]}{\underset{\sim}{1}}$. The estimate for $\underset{\sim}{\beta}$ is

$\hat{\underset{\sim}{\beta}}=(\underset{\sim}{X}\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\ \underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}$ with $\underset{\sim}{V}(\hat{\underset{\sim}{\beta}})=(\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}$. Here $\underset{\sim}{V}_{\underset{\sim}{F}}$ is block diagonal with block diagonal elements $\underset{\sim}{V}_{\underset{\sim}{F}_m} = \frac{1}{n_m^2}\sum_{i=1}^{n_m} (\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)(\underset{\sim}{F}_{im}-\underset{\sim}{F}_m)'$ where $\underset{\sim}{F}_{im}$ has the same structure of $\underset{\sim}{F}_i$ in (7.4.18) but with an "m" added to all subscripts of the elements of $\underset{\sim}{F}_i$.

CHAPTER VIII

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

Many references in the health and/or social science areas dealing with categorical data type studies with repeated measurements and possibly clustering present analyses that do not attempt to capture either of these special aspects of the data. Sometimes there may be an 'over time' question addressed in the analysis taking into consideration the longitudinal feature of the data; but for the most part, little attention has been given to any clustering of the attribute data. This dissertation has aimed to examine these types of data from an underlying viewpoint with more comprehensive statistical analyses. Applications of existing methods of analysis for longitudinal categorical data were discussed, methodology for the analysis of clustered attribute binary data for different design settings was extended, and real data were used for illustration.

In Chapter I, specific examples of studies illustrating one or more distinguishing aspects of non-standard datasets were presented. Common features of these studies were that subjects under investigation were observed under two or more conditions and there were correlated categorical data for observations from distinct sampling units. Chapter II reviewed aspects of weighted least squares analysis for longitudinal data.

Methodology for analysis of clustered attribute data from nested designs was reviewed in some detail in Chapter III. The settings of interest were the two stage nested design with one treatment applied within each cluster and with two treatments applied within each cluster. Two designs were reviewed: the completely balanced, and the partially balanced designs. A sampling without replacement framework was used to define the mean and joint probability parameters within a cluster. A sampling with replacement framework (equivalent to sampling without replacement in an infinite population setting) could then be assumed in order to simplify parameter functions of interest for the covariance structure of the data; in this regard, the finite populations under consideration are viewed as sufficiently large to be essentially infinite. The models for the two stage random sampling process and the classical two stage nested random effects ANOVA with the distributional assumptions were noted to be analogous. The parameter estimates of interest were the mean $\hat{\pi}$, the within cluster variance $(\hat{\pi} - \hat{\lambda})$, and the intraclass correlation coefficient $\hat{\rho}$ (a function of $\hat{\pi}$ and the joint probability $\hat{\lambda}$ for pairs of observations from the same cluster). Two ways were described to estimate $\hat{\pi}$ and $(\hat{\pi}-\hat{\lambda})$ jointly; a means formulation and a frequency table approach. In the first, estimators of π and $(\pi-\lambda)$ are written as means while in the second, they are written as linear functions of frequency counts f_j. In both situations, after writing the data formulation in matrix notation, the methodology of Grizzle et al (1969) for analysing categorical data can be used to obtain estimates of the parameters and the estimated covariance matrix associated with them. The primary parameters of interest, π and ρ, were jointly estimated by applying a series of linear, logarithmic and exponential transformations to the estimated parameter vector whose elements are $\hat{\pi}$ and $(\hat{\pi}-\hat{\lambda})$. These estimation procedures were used in the completely balanced case and in the partially balanced case; for the latter, modular estimates $\hat{\pi}_m$ and $(\hat{\pi}-\hat{\lambda})_m$ were obtained, and then the methodology of Grizzle, et al (1969) was applied to obtain overall parameter estimates through weighted least squares (WLS) if the number of clusters in each module was sufficiently large for the asymptotic theory to be valid. Otherwise, alternative approaches such as direct estimates like means of means was suggested. Examples to illustrate the two design cases were reviewed and alternative tests for ρ being zero were suggested and evaluated.

In Chapter IV, availability of existing methodology was illustrated by way of examples. This chapter described WLS methods to analyze a day care center study encompassing a structure where staff of day care centers answered more than one question about how to handle a perceived ill child and there might be some degree of clustering associated with answers given by staff from a day care center. Analysis of this dataset was undertaken first by using day care center as the unit of analysis in order to account for the possible correlation structure of multiple staff members within a day care center; then it was analyzed by using staff members as the unit of analysis and hence ignoring the possibility of a clustering effect. When day care center was the unit of analysis, within day care center means of staff's answers were analyzed across day care centers. In both analyses, cross-classifications of factors (e.g. type of day care, age of child, fever condition) were taken into account and the response of interest was (yes, no) for decisions staff persons would make about sending a perceived ill child home or not. The results showed that when staff was the unit of analysis, i.e., when the possibility of clustering was ignored, standard errors obtained were smaller than when day care center was considered the unit of analysis with staff as sub-units. This indicated that correlation among cluster sub-units might not be negligible and viewing staff as independent of one another would underestimate variances.

In Chapter V, the methodology reviewed in Chapter III was extended to the case of clustered binary data with two time points or conditions, with one treatment per cluster in the two-stage nested design. Correspondence between the two stage sampling process and the classical two-stage nested random effects ANOVA was noted. In this extended setting, in addition to the mean response and within cluster variance for each condition or time point (π_1, π_2, $(\pi_1-\lambda_1)$ and $(\pi_2-\lambda_2)$), there are θ (the probability that both responses for a subject in a cluster are favorable) and the within cluster covariance for two conditions, $(\theta-\phi)$. A means formulation for the estimation of these parameters was presented for the completely balanced design. After writing this means formulation in terms of matrices, estimates of the parameters and the covariance matrix associated with them can be specified. The quantities of primary interest in this two binary responses or two conditions case are the mean response and the intraclass correlation coefficient for each condition (π_1, ρ_1, π_2, ρ_2) and the time or condition correlation coefficients; γ, for same subject and different conditions and ν, for different subjects and different conditions. Analogous to the one binary response case in Chapter III, estimates of these parameters may be obtained by applying a series of linear, logarithmic and exponential transformations to the estimated parameter vector of means obtained with the means formulation approach. Alternative statistical tests for the correlation coefficients being zero were proposed using a normal distribution approximation. These same estimation approaches were used in the partially balanced case within each module to obtain modular estimates $\hat{\pi}_m$, $(\hat{\pi}-\hat{\lambda})_m$, $\hat{\theta}_m$ and $(\hat{\theta}-\hat{\phi})_m$. In the modular case as in Chapter III, WLS analysis was applied to obtain overall parameter estimates if the number of clusters in each module was sufficiently large for the theory to hold. Otherwise, an alternative approach was suggested. Three examples illustrated this extended methodology, two for the balanced case and one for the modular case. The first balanced case example examined data from the day care center study. Estimated means and correlation coefficients were obtained along with standard errors. The four correlation coefficients, when tested being equal to zero, provided results different from zero. The second balanced case example was a clinical dental material study of performance of restorations. After all parameters of interest were estimated and correlation coefficients tested to be equal to zero, one intraclass correlation coefficient (for one of the times) was not significantly different from zero. The last example considered the situation where there were three modules for the day care center study. The four estimated correlation coefficients were significantly different from zero. Findings for this modular example agreed with those of the first example. Extensions to the two treatments per cluster case were given in Chapter VI. Analogous to Chapter V, the similarity between the analysis of variance for the k-th response from the two stage sampling process and the analyses of variance for the classical two-stage nested model with random and fixed effects was noted. An extension of the means formulation in Chapter V for estimation of the parameters of interest was presented for the completely balanced case. After writing this formulation in matrix notation, estimates of the parameters and the covariance matrix of these estimates were specified. A series of transformations, similar to the one in the previous chapter, was then applied to this vector of estimated means in order to obtain estimates of the primary parameters of interest, means and within clusters correlation coefficients along with their estimated covariance matrix. This set of estimated parameters for this two treatments per cluster situation involves four means, $\hat{\pi}_{11}$, $\hat{\pi}_{12}$, $\hat{\pi}_{21}$ and $\hat{\pi}_{22}$ (two for each treatment) and twelve within cluster correlation coefficients where four are intraclass correlation coefficients, $\hat{\rho}_{11}$, $\hat{\rho}_{12}$, $\hat{\rho}_{21}$, $\hat{\rho}_{22}$; two are time or condition correlation coefficients, $\hat{\gamma}_1$, $\hat{\gamma}_2$; there are two intraclass cross-correlation coefficients, $\hat{\tau}_1$, $\hat{\tau}_2$ (among different sub-units of the same cluster receiving different treaments); two correlation coefficients for different sub-populations and same conditions, $\hat{\nu}_1$, $\hat{\nu}_2$; and two for different sub-populations and different conditions $\hat{\omega}_{11,22}$, $\hat{\omega}_{12,21}$. Alternative tests for these within cluster correlation coefficients being zero were proposed. These same estimation approaches were used in the partially balanced case, within each module to obtain modular estimates. Weighted least squares analysis was then applied to obtain overall parameter estimates under the assumption of sufficiently large numbers of clusters in each module. An example, the day care center study illustrated this extended methodology. This study with the addition of answer given by parents, may be viewed as a two-stage nested design situation with two sub-populations, staff and parents. Estimates of the 16 parameters were obtained along with standard errors. Tests for the correlation coefficients being zero indicated that the two intraclass correlation coefficients for staff (ρ_{11}, ρ_{12}), the two condition correlation coefficients (γ_1, γ_2) and the correlation coefficient for staff with different subjects and different conditions (ν_1) were significantly different from zero.

The methodology presented in Chapter V and VI was extended in Chapter VII to the situation where a set of r binary responses is observed on each subject within a cluster. These extensions had some complexity due to the increase in the number of parameters to be estimated. Means formulation for this more general situation were presented and a series of transformations on this vector of means (as it was done in previous chapters) was suggested in order to obtain the set of parameters of primary interest, the π's and ρ's. For the completely balanced case there are r(r+1) such parameters for the one treatment per cluster case and $3r^2+2r$ parameters for the two treatment per cluster case (r=1, 2, ..., r). For the partially balanced situation there are r(r+1)M and $(3r^2+2r)M$ parameters for the M modules for the one treatment case and two treatments case respectively.

The analytical results obtained from the examples discussed throughout the chapters appear to indicate that studies which involve a data structure with possibly some degree of clustering should be analyzed first viewing the cluster as the basic unit of study. If correlation among cluster sub-units is found to be negligible, then cluster sub-units can be viewed as the basic analytical units. In this way, care is taken so that estimates of variances are not underestimated .

8.2 Future Research

A setting not addressed in this dissertation was that of unbalanced designs. Methods of obtaining overall parameter estimates for the unbalanced case for the one treatment and two treatments per cluster extensions can be explored. Note that the fully unbalanced data structure does not fit in the framework of the classical weighted least squares methodology of Grizzle, et al (1969) since each element of the vector $\underset{\sim}{F}$ may be based on a sample of size 1 (i.e., a single cluster). Thus, alternative methods are needed to obtain parameter estimates for this case.

One such alternative approach, direct formulation of an estimate of the mean response, was suggested when the number of clusters in the modular situations described for the one and two treatments per cluster case, was not sufficiently large for the asymptotic theory to hold. Even though the number of clusters in each module in the last example of Chapter V were of moderate sizes, it would be interesting to assess how a direct estimator such as a mean of means along with its standard error could compare with the one obtained with WLS procedure.

For both the completely balanced and partially balanced designs, the proposed methods presented in Chapters V, VI and VII can be extended further to the three stage nested design or other more complex situations.

BIBLIOGRAPHY

Allan, D.N. (1969). The durability of conservative restorations. *British Dentistry Journal* 128: 172-177.

Beale, E.M. and Little, R.J.A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, Series B 37, 129-146.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association* 61: 228-235.

Bartlett, A.V., Moore, M., Gary, G.W., Starko, K.M., Erben, J.J., Meredith, B.A. (1985). Diarrheal Illness among infants and toddlers in day care centers I. Epidemiology and pathogens. *Journal of Pediatrics vol* 107 No 4: 495-502.

Bartlett, A.V., Moore, M., Gary, G.W., Starko, K.M., Erben, J.J., Meredith, B.A. (1985). Diarrheal Illness among infants and toddlers in day care centers II. Comparison with day care homes and households. *Journal of Pediatrics*, Vol. 107 No 4: 503-509.

Breslow,N.E. and Day,N.E. (1981). Statistical Methods in Cancer Research. Vol I. The analysis of case-control studies. Lyon, France. *IARC Scientific Publications*, No 32.

Centers for Disease Control, U.S. Public Health Service, Department of Health and Human Services (1985). Public health considerations of infectious deseases in child day care centers: Special Article. *The Journal of Pediatrics vol* 105, *no* 5: 683-701.

Cochran, W. G. (1977). *Sampling Techniques* (third edition). John Wiley & Sons.

Davis, V. (1984). A review of weighted least squares methodology for analyzing residuals from categorical data. Master's of Science Paper, Department of Biostatistics, University of North Carolina, Chapel Hill.

Derkson, G.D., Richardson, A.S., Waldman, R. (1984). Clinical evaluation of compositive resin and amalgam posterior restorations: Three year results. *Journal of the Canadian Dentristy Association*, No 6, 478-480.

Doyle, A.B. (1976). Incidence of illness in early group and family day care. *Pediatrics* 1976, 58: 607-613.

Fosburg, S. (1981). Family day care in the United States: summary of findings. National day care home study. Cambrige MA: ABT Associates. *DHHS Publication no (OHDS)* 80-30282.

Goldberg, J., Munster, E., Spangberg E. et al. (1979). Experimental design in the clinical evaluation of amalgam restorations. *Journal of Dental Research.*

Grizzle, J.E., Starmer, C.F., Koch, G.G. (1969). The analysis of categorical data by linear models. *Biometrics* 25, 489-504.

Heald, C.W., Jones, G.M., Nickerson, S. and Bibb, T.L. (1977). Mastitis control by penicillin and novobiacin at drying off. *Canadian Veterinary Journal* 18, 171-180.

Hicks, C.R. (1973). *Fundamental concepts in the design of experiments* (second edition). Holt, Rhinehart and Winston, INC.

John, P.W.M. (1971). *Statistical Design and Analysis of Experiments.* The MacMillan Company, New York.

John, G.A. and Prescott, P. (1975). Estimating missing values in experiments. *Applied Statistics* 24, 190-192.

Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. (1982). *Epidemiologic Research.* Lifetime Learning Publications.

Koch, G.G. (1967a). A general approach to the estimation of variance components. *Technometrics* 9, 93-118.

Koch, G.G. (1967b). A procedure to estimate the population mean in random effects models. *Technometrics* 9, 577-586.

Koch, G.G. (1969). A useful lemma for proving the equality of two matrices with applications to least squares type of quadratic forms. *Journal of the American Statistical Association* 64, 969-970.

Koch, G.G. and Edwards, S. (1984). Logistic regression. In *Encyclopedia of Statistical Sciences.* 5 N. L. Johnson and S. Kotz (eds) 442-457. Wiley, NY.

Koch, G.G., Elashoff, J.D. and Amara, I.A. (1987). Repeated measurements studies, design and analysis. In *Encyclopedia of Statistical Sciences,* 8, N.L. Johnson and S. Kotz (eds.), Wiley, New York.

Koch, G.G., Gillings, D.B. and Stokes, M.E. (1980). Biostatistical Implications of design, sampling and measurement to health science data. *Annual Review of Public Health* 1, 163-225.

Koch, G.G., Amara, I.A., Stokes, M.E. and Gillings, D.B. (1980). Some views on parametric and non parametric analysis for repeated measurements and selected bibliography. *International Statistical Review* 48, 249-265.

Koch, G.G., Grizzle, J.E., Semenya, K. and Sen, P.K. (1978). Statistical methods for evaluation of mastitis treatment data. *Journal of Dairy Science* 61(6), 830-847.

Koch, G.G., Amara, I.A., Carr, G.J., Hirsch, M.G. and Marques, E.H. (1985). The interface between analysis strategies and design structure in repeated measurements studies. *Statistical Computing Section Proceedings of the American Statistical Association,* 19-28.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., Jr. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 33, 133-158.

Koch, G.G., Imrey, P.B., Singer, J.M., Atkinson, S.S., Stokes, M.E. (1985). Analysis of Categorical Data. In *Collection Seminaire de Mathematiques Superieures* 96, *G. Sabidussi.* Les Presses de L'Université de Montreal, Montreal.

Koch, G.G., Singer, J.M., Carr, G.J., Cohen, S.B. and Stokes, M.E.(1986). Some aspects of weighted least squares analysis for longitudinal categorical data. *Proceedings of the Workshop on Longitudinal Methods in Health Research.* Berlin, forthcoming.

Kempthorne, W.J.(1982). The Analysis of clustered attribute data from nested and/or classification designs using random effects models. Unpublished Ph.D. thesis, Department of Biostatistics, University of North Carolina, Chapel Hill, NC.

Kempthorne, W.J. and Koch, G.G. (1983). A general approach for the analysis of attribute data from a two-stage nested design: one and two treatments per cluster. *Contributions to Statistics: Essays in Honor of Norman L. Johnson.* Ed. P.K. Sen, North Holland Publishing Company, 259-280.

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

Landis, S. (1986). Day care center staff's criteria for sending sick children home. Master's of Public Health, Department of Epidemiology, University of North Carolina at Chapel Hill.

Landis, J.R. and Koch, G.G. (1977). A one way components of variance model for categorical data. *Biometrics* 33, 671-679.

Landis, J.R., Stanish, W.M., Freeman, J.L., and Koch, G.G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least square (GENCAT) *Computer Programs in Biomedicine*, North Holland Publishing Company, 6, 196-231.

Lemp, G.F., Woodward, W.E., Pickering, L.K., Sullivan, P.S. and Dupont, H.L. (1984). The relationship of staff to the incidence of diarrhea in day care centers. *American Journal of Epidemiology* vol 120 No 5, 750-758.

Letzel, H., Aardening, C.J.M.W., Fick, J.M., et al. (1978). Tarnish, corrosion, marginal fracture and creep of amalgam restorations: a two-year clinical study. *Operative Dentristy* 3, 82-92.

Loda, F.A., Glezen, W.P. and Clyde, W.A., Jr. (1972). Respiratory disease in group care. *Pediatrics* 49, 428-438.

Mahler, D.B. and Marantz, R.L. (1979b). The effect of the operator on the clinical performance of amalgam. *Journal of the American Dental Association* 99, 38-41.

Mahler, D.B. and Marantz, R.L.(1980). Clinical Assessments of dental amalgam restorations. *International Dental Journal* 30, 327-334.

Marques, E.H, Landis, S., and Koch, G.G. (1986). Analysis of cross-classified scenarios for categorical choices of actions. *Survey Research Methods Section Proceedings of the American Statistical Association*, pp. 697-702.

May, K.N., Wilder, A.D., Leinfelder, K.F. (1983). Burnished amalgam restorations: a two year clinical evaluation. *The Journal of Prosthetic Dentristy*,49,No 2, 193-196.

Moore, D.L. and Stewart, J.L. (1967). Prevalence of deffective restorations. *Journal of Prosthetic Dentistry*, 17, 372-378.

Osborne, J.W. and Gale, E.N. (1974). Long term follow-up of clinical evaluations of lathe-cuts versus spherical amalgam. *Journal of Dental Research*, 58, 1204-1207.

Stanish, W.M., Koch, G.G. and Landis, J.R. (1978). A computer program for multivariate ratio analysis (MISCAT). *Computer programs in biomedicine.* Elsevier/North-Holland Biomedical Press, 8, 197-207

Stanish, W.M., Gillings, D.B. and Koch, G.G. (1978). An application of multivariate ratio methods for the analysis of longitudinal clinical trial with missing data. *Biometrics* 34, No 2, 305-317.

Strangert, K. (1976). Respiratory illness in pre-school children with different forms of day care. *Pediatrics* 57, 191-195.

Sturdevant, J.R., Lunden, T.F., Sluder, T.B., Leinfelder, K.F. (1986). *Dental Materials* (submitted for publication) January 1986.

Sturdevant, J.R., Sturdevant, C.M., Taylor, D.F., Bayne, S.C., Marques, E.H. (1987). Clinical performance of low gold alloy after eight years. *Journal of Dental Research*, Vol. 66, pg 246 (abstract #1120).

Sullivan, P., Woodward, W.E., Pickering, L.K. and Dupont, H.L. (1984). A longitudinal study of the occurrence of diarrheal disease in day care centers. *American Journal of Public Health* 74, 987-991.

Ware, J.H. (1985). Linear models for the Analysis of Longitudinal studies. *The American Statistician* 39, 2.