This research was partially supported by the Bureau of the Census through Joint Statistical Agreements JSA 81-28 and 83-1.

STRATEGIES FOR MULTIVARIATE RANDOMIZATION ANALYSES AND APPLICATIONS TO HEALTH SCIENCES DATA

by

Ingrid Ann Amara

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1446

August 1983

STRATEGIES FOR MULTIVARIATE RANDOMIZATION ANALYSES
AND APPLICATIONS TO HEALTH SCIENCES DATA

by

Ingrid Ann Amara

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 in the School of Public Health.

Chapel Hill

1982

Approved by:

Adviser

Reader

Reader

ABSTRACT

INGRID ANN AMARA. Strategies for Multivariate Randomization Analyses and Applications to Health Sciences Data. (Under the direction of GARY G. KOCH.)

A general problem of statistical interest is the formulation of tests for the hypothesis that one or more response variables are distributed at random with respect to a set of sub-populations (e.g., treatments) within one or more strata based upon control variables (e.g., blocks, clinics, pretreatment status and/or demographic status). More specifically, if $y_{hij\ell}$ denotes the observed value for the j-th response variable for the ℓ-th subject in the i-th sub-population of the h-th stratum where $h = 1,2,\ldots,q$, $i = 1,2,\ldots,s$, $j = 1,2,\ldots,d$ and $\ell = 1,2,\ldots,n_{hi}$, then this type of hypothesis can be expressed as:

H_o: For each of the strata, $h = 1,2,\ldots,q$, there is no relationship between the response variables and the sub-populations in the sense of equally likely realizations for the

$$\prod_{h=1}^{q} \{n_h! / \prod_{i=1}^{s} n_{hi}!\}$$

possible stratified and exhaustive, random partitions of the $n_h = \sum_{i=1}^{s} n_{hi}$ data vectors $\underset{\sim}{y}_{hi\ell} = (y_{hi1\ell},\ldots,y_{hid\ell})'$ for the subjects in the h-th stratum into successive random samples of sizes $= n_{h1}, n_{h2},\ldots,n_{hs}$ for s sub-populations.

From the finite population randomization model implied by H_o, several different types of statistical tests can be formulated without any external assumptions concerning underlying distributions. These

include univariate and multivariate rank analysis of variance statistics, rank analysis of covariance statistics, Spearman rank correlation test statistics, and contingency table chi-square test statistics. All of these methods together with certain more general counterparts are discussed in terms of a common underlying methodology. This conceptual framework is then used as a guide for the straightforward computation of these types of non-parametric statistical tests via a standard set of operations.

The randomization methods are applied to health sciences data such as clinical trials for which patients are randomly assigned to treatment groups and observational studies for which randomization can be viewed as a hypothesis of interest. These examples illustrate multivariate and covariate capabilities. In examples involving clinical trials, weighted least squares and/or maximum likelihood modeling procedures are used to describe the variation across sub-populations and strata and to provide predicted values when such methods are applicable. Also, the assumptions and capabilities of these latter methods are contrasted with those of randomization methods.

ACKNOWLEDGMENTS

With great appreciation and pleasure I acknowledge my committee chairman, Professor Gary G. Koch, for involving me in his exposition of randomization statistics and for providing invaluable guidance, assistance and encouragement throughout the course of our work together. Particularly, he has given freely of his insight and knowledge clarifying the difficult statistical issues in medical studies; I have enjoyed his friendship immensely. I also thank each of my other committee members: Dr. Ronald Helms, Dr. Craig Turnbull, Dr. Michele Ibrahim and Dr. Kerry Lee for their cooperation and support. Special thanks go to Dr. Lee for his careful reading and helpful comments on every chapter.

Much of the applied statistical methodology in this dissertation was motivated from the special problems facing medical researchers. Their interest and commitment provided the essential inspiration. In particular, I'd like to thank Dr. Ross Simpson who generously contributed his complex cardiology data and helpful discussions to Chapter III; I'd like to thank Dr. Michael Ryan for his brilliantly cogent questions and observations that precipitated my writing.

I wish to express deep appreciation to Mrs. Gabrielle Hinnant for her expert typing of this manuscript and to Ms. Lori Turnbull for organizing and typing my tables. I am indebted to Lori's support in working through drafts of my chapters and her invaluable and expert assistance in the preparation of my final oral.

I also acknowledge Dr. Ronald Helms as my academic advisor for providing me opportunities in data management and supporting the direction of my program. I was fortunate in receiving the financial assistance of a Public Health Service stipend from the Bureau of Health Manpower. The research in this dissertation was also supported in part by the United States Bureau of the Census.

My warm thanks go to Joan Kempthorne, my dissertating colleague and close friend. Together we planned and worked toward achieving our doctoral and personal goals; we both had a share of Gary G. Koch.

Lastly, I extend my deepest appreciation and warmest thoughts to my husband Anthony and daughter Danielle who have borne much of my frustration and distance this past year while providing encouragement; their immeasurable love and faith have been essential to my inner peace.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND REVIEW OF LITERATURE

1.1 Background

A general problem of statistical interest is the formulation of tests for the hypothesis that one or more response variables are distributed at random with respect to a set of subpopulations (e.g., treatments) within one or more strata based upon control variables (e.g., blocks, clinics, pretreatment status and/or demographic status). One useful approach to this problem is randomization hypothesis testing [see Kempthorne (1955)], which involves formulating a sampling framework within which one or more response variables are expected to vary randomly when there is no association with the subpopulations. In clinical trials, the randomization framework corresponds to the experimental design accordinto to which patients are considered randomly assigned to subpopulations. In situations where data are observed without the benefit of an experimental design, randomization can be viewed as a hypothesis of interest for contrasting a set of subpopulations after accounting for the association between the response variables and the strata [see Koch, Gillings and Stokes (1980)]. Both of these frameworks are directed at a local population corresponding to the specific set of subjects under study. For this reason, these frameworks do not directly involve assumptions concerning underlying statistical distributions (e.g., normal, multinomial, etc.) or stochastic processes. Moreover, the randomization framework is analogous to finite population sampling in the sense that the actual values of the response variables for the respective subjects are considered fixed (under the hypothesis of no association) and their assigned subpopulations are considered random.

Examples of randomization hypothesis testing applications are as follows:

1. A multicenter clinical trial is undertaken to compare two treatments. For this purpose, a set of appropriately qualified clinics are invited to participate on a judgmental basis. Within each clinic, patients are randomly assigned to two treatments. Data pertaining to various aspects of the medical status of each patient are recorded both before treatment (e.g., baseline) and at weekly visits during treatment. Here, clinics comprise the strata (see section 2 of Chapter VII).
2. An intracardiac conduction study is concerned with the association between atrial abnormality and conduction impedance of premature beats initiated at specific levels of electrical stimulus thresholds. Data pertaining to the relative change in conduction time are obtained for multiple premature beats which are stimulated at successively earlier points involving higher electrical thresholds within the cardiac cycle. Here, the hypothesis corresponds to an assessment of whether the partition of responses (median change in conduction time relative to intervals of threshold change per patient) vary at random between diagnostic groups (normal and abnormal) in the sense of allocation of n_1 and n_2 patients from the combined sample (see Chapter III).

3. A nationwide clinical study with a historical control is concerned with the association between treatment (antidote versus supportive therapy) and hepatotoxicity in patients who ingested a suicidal-intent overdose of a medical product. Data pertaining to liver function values of serum bilirubin, SGOT ratio and prothrombin time ratio are obtained over a 72-hour period. A cross-classification of blood toxicity (moderate versus severe) and time of treatment (early versus late) is used to define strata within which the partition of responses (maximum liver function values per patient) between treatment groups is viewed as resulting from successive sets of random allocations of n_{h_1} and n_{h_2} patients from the combined samples $\{n_h\}$ (see Chapter 6).

For univariate hypotheses of no association, Kruskal and Wallis (1953) considered a nonparametric procedure which involved analysis of variance computations on ranks. The Kruskal-Wallis statistic has an approximate chi-square distribution in large samples; it can be computed as the ratio of the among-groups sum of squares to the total sum of squares. For the special case of two groups, the Kruskal-Wallis statistic can be simplified to the procedures described by Mann and Whitney (1947) and by Wilcoxon (1945). For stratified analyses involving one observation per treatment within strata, Friedman (1937) discussed a two-way analysis of variance on ranks statistic. An extension to more than one observation per treatment has been given by van Elteren (1960). A multivariate extension of the Kruskal-Wallis statistic is discussed in Chatterjee and Sen (1966) and later in Puri and Sen (1971). Such statistics can be viewed as equivalent to a parametric trace criterion MANOVA statistic with respect to the product of among groups sums of products matrix and the inverse of the total sum of products matrix based on rank transformed data. Koch (1969) has applied the rank MANOVA methods in analyses concerning split plot experiments. A multivariate extension of the Friedman statistic is presented by Gerig (1969). Relatedly, a covariance extension of the univariate rank analysis of variance procedure is presented by Quade (1967). This procedure involves an analysis of variance on the residuals of a preceding regression which adjusts the ranks of the response variable for the variation in ranks of the independent variable (covariable) in accordance with their correlation structure. Puri and Sen (1969) discuss the generalization of Quade's procedure and provide both exact and asymptotic based randomization tests.

For categorical responses in the two group case, Cochran (1954) essentially presented the hypothesis of no average partial association relative to sets of (2 x 2) tables for strata based on control variables. Hopkins and Gross (1971) presented a generalization of Cochran's procedure for sets of (s x r) tables by constructing the set of $\binom{s}{2}$ pairwise comparisons within (r - 1) of the r levels of the response profile. Both of these procedures assume an infinite population-based binomial model for each table used in obtaining mean differences weighted across sets of tables. Alternatively, Mantel and Haenszel (1959) showed that for sets of (2 x r) tables, expected values and covariance for (r - 1) pivot cells could be computed per table under the multiple hypergeometric probability model for finite populations. This approach was appropriate for small samples within the strata and was motivated by randomization considerations. With a quadratic form test statistic based on the across table sums of {observed-expected} values and covariance quantities, the asymptotic properties depend on the across table sample size. Koch and Reinfurt (1974) applied an extension of the procedure to sets of (s x r) tables involving (s - 1) (r - 1) pivot cells. Mantel (1963) applied scores to ordered response profiles and produced a mean score test statistic. Furthermore, he showed a score correlation-type test statistic to be pertinent for the case of both ordered responses and ordered subpopulations (e.g., factor levels).

Equivalency of the generalized Mantel-Haenszel mean score and correlation procedures to other categorical or nonparametric procedures has been documented by several authors [Birch (1965), Landis, Heyman and Koch (1980), Koch, Amara, Stokes and Gillings (1980) and Koch and Bhapkar (1982)]. More specifically, the Mantel-Haenszel procedures can be applied to any scaling of the data such that specific scoring applications yield equivalent results to other tests. These scores include actual values, integer, binary, rank, modified ridit and logrank scalings discussed in further detail in section 5 of Chapter II. For the mean score case, Cochran's Q test for matched proportions can be constructed for binary responses and the restrictions of one observation per subpopulation per block. Furthermore, if the responses are ranks with ties assigned mid-rank values, then this situation of one observation per subpopulation per block produces the Friedman chi-square statistic. In this regard, for rank scores, Cochran's Q test is the same as Friedman's chi-square test. McNemar's statistic for matched-pair samples is the two subpopulation and single stratum case of Cochran's Q test. For correlation situations in which both responses and subpopulations are assigned rank scores with mid-ranks assigned to ties, the generalized Mantel-Haenszel procedure for the single stratum produces the Spearman rank correlation coefficient test statistic. Relatedly, a return to the $(s - 1)(r - 1)$ degrees of freedom test for contingency tables can be obtained via the multivariate extension of the generalized Mantel-Haenszel mean score procedure. The multivariate response profile corresponds to $d = (r - 1)$ binary responses identifying the jth $= 1,\ldots,d$ category to which an observation belongs. In the single stratum case, the multivariate mean score procedure produces the Pearson chi-square statistic for the hypothesis of no associations in the $\{s \times [d + 1]\}$ contingency table except for a multiplicative factor $\{n/[n - 1]\}$.

In view of these considerations, the remaining sections present specific details of randomization statistics relative to the previously summarized background literature. More specifically, section 1.2 concerns the contingency table framework for a single response variable and several strata (tables); section 1.3 deals with the multivariate framework for a single stratum and data matrix structure; section 1.4 concerns the covariance framework for a single response variable and a single stratum; section 1.5 pertains to missing value strategies for multivariate response profiles. Finally, section 1.6 presents an outline of how the concepts in the previous sections are encompassed in a general framework for single or multiple strata, univariate or multivariate response profiles and direct or covariance adjusted test statistics relative to the set of examples discussed extensively in Chapters III, IV, V, VI and VII.

1.2 Average Partial Association for Sets of Contingency Tables

In this section, the technical aspects of randomization methodology are summarized for a single categorical response variable with r levels in a framework involving s subpopulations and q strata. The discussion is based upon the formulation of randomization methods for this case, as reviewed in Landis, Heyman and Koch (1978) and the computing strategy available with the program PARCAT, documented in Landis, Cooper, Kennedy and Koch (1979). Furthermore, the formulation can be viewed as encompassing both categorical and continuous data. For categorical data the response categories are the set of possible values for the actual response observed; for the continuous case, the response values correspond to mid-points of a set of small intervals. These intervals are viewed as having been actually observed in the sample relative to the totality of all possible such intervals, even though the latter is not specifically identified in the data structure. For example, age in five year intervals is categorical while age as a continuous variable could be thought of as age in days, which is categorical for the days actually observed.

The Landis, Heyman and Koch (1978) presentation of the contingency table framework, as it relates to the product multiple hypergeometric distribution, is as follows:

TABLE 1.1

OBSERVED CONTINGENCY TABLE FOR LEVEL h OF THE COVARIABLES

	Response Variable Categories				
Subpopulation	1	2	...	r	Total
1	n_{h11}	n_{h12}	...	n_{h1r}	$n_{h1\cdot}$
2	n_{h21}	n_{h22}	...	n_{h2r}	$n_{h2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	n_{hs1}	n_{hs2}	...	n_{hsr}	$n_{hs\cdot}$
Total	$n_{h\cdot 1}$	$n_{h\cdot 2}$	...	$n_{h\cdot r}$	$n_{h\cdot\cdot}$

Let $h = 1,2,\ldots,q$ index a set of $(s \times r)$ contingency tables which correspond to distinct levels of a stratification variable or combination of several such variables. Let $i = 1,2,\ldots,s$ index a set of subpopulations which are to be compared with respect to a particular response variable for which the outcome categories are indexed by $j = 1,2,\ldots,r$. Then, let $\underset{\sim}{n}_h = (n_{h11},\ldots,n_{h1r},\ldots,n_{hs1},\ldots,n_{hsr})$, where n_{hij} denotes the number of subjects in the sample who are jointly classified as belonging to the hth table, the ith subpopulation and the jth response category. These frequency data, corresponding to the hth level of the covariable set, can be summarized as shown in Table 1.1, where $N_{hi\cdot}$ denotes the marginal total number of subjects classified as belonging to the ith subpopulation, $N_{h\cdot j}$ denotes the marginal total number of subjects classified as belonging to the jth response category, and $N_{h\cdot\cdot}$ denotes the overall marginal total sample size in the hth table.

The basic hypothesis under investigation involves the relationship between the response variable and the subpopulations, adjusted for the stratification framework. Under the assumption that the marginal totals $\{N_{hi\cdot}\}$ and $\{N_{h\cdot j}\}$ are fixed, the overall null hypothesis of 'no partial association' can be stated as:

> H_o: For each of the separate levels of the stratification $h = 1,2,\ldots,q$, the response variable is distributed at random with respect to the subpopulations, i.e., the data in the respective rows of the hth table can be regarded as a successive set of simple random samples of sizes $\{N_{hi\cdot}\}$ from a fixed population corresponding to the marginal total distribution of the response variable $\{N_{h\cdot j}\}$.

On the basis of this hypothesis, it can be shown from stratified sampling arguments that the vector $\underset{\sim}{n}_h$ follows the product multiple hypergeometric distribution given by the probability model,

$$\Pr(\underset{\sim}{n}_h | H_o) = \frac{\prod_{i=1}^{s} N_{hi\cdot}! \prod_{j=1}^{r} N_{h\cdot j}!}{N_{h\cdot\cdot}! \prod_{i=1}^{s} \prod_{j=1}^{r} n_{hij}!} . \tag{1.1}$$

Let $P_{hi\cdot} = (N_{hi\cdot}/N_{h\cdot\cdot})$ and $P_{h\cdot j} = (N_{h\cdot j}/N_{h\cdot\cdot})$ denote the fixed marginal proportions for the ith row (subpopulation) and the jth column (response category), respectively. For h = 1,2,...,q tables, the set of fixed marginal proportions can be expressed as $\underset{\sim}{P}'_{h*\cdot} = (P_{h1\cdot},\ldots,P_{hs\cdot})$ and $\underset{\sim}{P}_{h\cdot *} = (P_{h\cdot 1},\ldots,P_{h\cdot r})$.

1.2.1 Mean score test

The nature of the distribution for data with response categories $j = 1,2,\ldots,r$ that are ordinally scaled with progressively larger intensities can be summarized for each of the $i = 1,2,\ldots,s$ subpopulations in terms of a mean score. In particular, let $a'_h = (a_{h1},a_{h2},\ldots,a_{hr})$ be the vector of response category scores for the hth stratum (e.g., ranks); and then the mean score for the ith subpopulation in the hth stratum (table) can be expressed as

$$F_{hi} = \frac{1}{N_{hi\cdot}} \sum_{j=1}^{r} a_{hj} n_{hij} , \tag{1.2}$$

with the set of mean scores for the hth stratum being $\underset{\sim}{F}'_h = (F_{h1},F_{h2},\ldots,F_{hs})$. The hypothesis expectation of $\{F_{hi}\}$ corresponds to the mean score on the column marginal proportions

$$\bar{a}_h = \sum_{j=1}^{r} a_{hj} P_{h\cdot j} ; \tag{1.3}$$

The total variance of the response variable in the overall population for the hth stratum with respect to $\underset{\sim}{a}$ can be expressed as

$$S^2_{a,h} = \sum_{j=1}^{r} (a_{hj} - \bar{a}_h)^2 \, P_{h\cdot j} \; . \tag{1.4}$$

Thus, for i = 1,2,...,s subpopulations the weighted sum of the mean scores across the q tables can be expressed as

$$\underset{\sim}{F} = \sum_{h=1}^{q} N_{h\cdot\cdot} \underset{\sim}{D}_{\underset{\sim}{P}_{h*\cdot}} \underset{\sim}{F}_{h\cdot} \; , \tag{1.5}$$

where $\underset{\sim}{D}_{\underset{\sim}{k}}$ is a diagonal matrix with elements of $\underset{\sim}{k}$ on the main diagonal. The corresponding hypothesis expectation and variance are given by:

$$\underset{\sim}{E}\{\underset{\sim}{F}|H_o\} = \sum_{h=1}^{q} N_{h\cdot\cdot} \bar{a}_h \underset{\sim}{P}_{h*\cdot} \tag{1.6}$$

$$\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{F}|H_o\} = \sum_{h=1}^{q} \{N^2_{h\cdot\cdot}/[N_{h\cdot\cdot} - 1]\} S^2_{a,h} \{\underset{\sim}{D}_{\underset{\sim}{P}_{h*\cdot}} - \underset{\sim}{P}_{h*\cdot} \underset{\sim}{P}'_{h*\cdot}\} . \tag{1.7}$$

The generalized mean score Mantel-Haenszel statistic can be expressed in its quadratic form as:

$$Q_{MS} = [\underset{\sim}{F} - \underset{\sim}{E}\{\underset{\sim}{F}|H_o\}]'\underset{\sim}{C}'\{\underset{\sim}{C}[\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{F}|H_o\}\underset{\sim}{C}'\}^{-1} \underset{\sim}{C}[\underset{\sim}{F} - \underset{\sim}{E}\{\underset{\sim}{F}|H_o\}]\} \; , \tag{1.8}$$

where $\underset{\sim}{C} = [\underset{\sim}{I}_{(s-1)}, -\underset{\sim}{1}_{(s-1)}]$ defines the (s − 1) space of contrasts among the weighted mean score sums $\underset{\sim}{F}$, $\underset{\sim}{I}_k$ denotes an identity matrix of rank (k) and $\underset{\sim}{1}_k$ denotes a column vector of (k) 1's. This test statistic is directed at location shift alternatives which correspond to the extent to which the mean scores for certain subpopulations consistently exceed (or are exceeded by) the mean scores for other subpopulations across the respective tables in an 'average partial association' sense. Under H_o, Q_{MS} asymptotically has the chi-square distribution with d.f. = (s − 1). For the case where all scores are either 0 or 1 and s = 2,

$$Q_{MS} = \{ \sum_{h=1}^{q} (n_{h1} - \frac{N_{h1\cdot}N_{h\cdot 1}}{N_{h\cdot\cdot}})\}^2 / \{ \sum_{h=1}^{q} \frac{N_{h1\cdot}N_{h2\cdot}N_{h\cdot 1}N_{h\cdot 2}}{N_{h\cdot\cdot}^2(N_{h\cdot\cdot} - 1)}\} , \quad (1.9)$$

which is identical (except for the lack of a continuity correction) to the statistic recommended in Mantel and Haenszel (1959). For the case where all scores are marginal rank or ridit-type (i.e., ranks divided by sample size) within each stratum (table) with midranks assigned to ties, Q_{MS} in (1.8) can be viewed as equivalent to a partial Kruskal-Wallis ANOVA test on ranks, adjusted for the levels of the stratification set. In particular, if the tables are restricted to one subject per row and none of the responses are tied within a table, then with rank scores for s subjects per block such that N_h = s for each of q blocks, Stanish (1978) showed:

$$Q_{MS} = \{12/qs(s + 1)\} \sum_{i=1}^{s} \{F_i - [q(s + 1)/2]\}^2 , \quad (1.10)$$

which is Friedman's statistic [see Birch (1965) and Stanish (1978) for further details].

1.2.2 Correlation test

The nature of association for data with response categories j = 1,2,...,r and subpopulation categories i = 1,2,...,s which are both ordinally scaled with progressively larger intensities, often can be assessed in terms of a composite mean score formulated as the sum of the (s) (r) cell score products, $a_{hj}c_{hi}n_{hij}$, based on the response scores $\underset{\sim}{a}_h' = (a_{h1},a_{h2},\ldots,a_{hr})$ and the subpopulation scores $\underset{\sim}{c}_h' = (c_{h1},c_{h2},\ldots,c_{hs})$ for h = 1,2,...,q tables. In this situation, the randomization hypothesis can be concerned with the extent of association between the response scores and the subpopulation scores in the respective tables. More specifically, the generalized correlation Mantel-Haenszel statistic can be written as the ratio of the squared weighted sum of the score covariance to the weighted sum of the score variance product,

$$Q_{MA} = \{\sum_{h=1}^{q} N_{h\cdot 1} S_{ac,h}\}^2 / \{\sum_{h=1}^{q} [N_{h\cdot\cdot}^2 / (N_{h\cdot\cdot}^2 - 1)] S_{a,h}^2 S_{c,h}^2\}, \tag{1.11}$$

where

$$S_{ac,h} = \sum_{i=1}^{s} \sum_{j=1}^{r} (a_{hj} - \bar{a}_h)(c_{hi} - \bar{c}_h)(n_{hij}/N_{h\cdot\cdot}), \tag{1.12}$$

$$S_{a,h}^2 = \sum_{j=1}^{r} (a_{hj} - \bar{a}_h)^2 P_{h\cdot j}, \quad \bar{a}_h = \sum_{j=1}^{r} a_{hj} P_{h\cdot j} , \tag{1.13}$$

and

$$S_{c,h}^2 = \sum_{i=1}^{s} (c_{hi} - \bar{c}_h)^2 P_{hi\cdot}, \quad \bar{c}_h = \sum_{i=1}^{s} c_{hi} P_{hi\cdot} . \tag{1.14}$$

Under H_o, Q_{MA} asymptotically has the chi-square distribution with d.f. = 1. This test statistic is directed at assessing the extent to which there is consistent positive (or negative) association between the response scores and the subpopulation scores across the respective tables in an 'average partial association' sense. For the case where all the scores are either 0 or 1, Q_{MA} simplifies to the (2 x 2) case of the Q_{MS} statistic in (1.9). For the case where all the scores are marginal rank or ridit-type scores with midranks assigned for ties, the Q_{MA} statistic is essentially equivalent to a partial Spearman Rank Correlation test adjusted for the levels of the stratification set.

1.3 One-Way Multivariate Analysis of Variance

In this section, the technical aspects of randomization methodology are summarized for j = 1,2,...,d response variables in a framework involving $i = 1,2,\ldots,s$ subpopulations within a single stratum. The discussion is based upon the formulation of randomization methods for the mean score test as reviewed in Koch and Bhapkar (1980) and the computing strategy available with the SAS macro GRMM, documented in Amara and Koch (1980). By expressing the data as a mean vector of responses per subpopulation, the formulation can be viewed as encompassing both categorical and continuous data in the same sense as discussed in section 1.2.

The randomization framework and hypothesis of no association is presented in Amara and Koch (1980). It is expressed in terms of (d x 1) vectors $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n$ for d non-redundant response variables for the subjects in some finite population of size n. Let $i = 1,2,\ldots,s$ index a set of subpopulations for which the relationship to the d response variables is to be investigated. Let $n_1, n_2, \ldots, n_s$ denote the corresponding numbers of subjects so that $\sum_{i=1}^{s} n_i = n$. If the subpopulations are homogeneous in the sense of equivalence to random allocation, then the distribution of the $\{\underset{\sim}{y}_\ell\}$ among them is compatible with the hypothesis of no association:

H_o: There is no relationship between subpopulations and responses in the sense of equally likely realizations for the $(n!/\prod_{i=1}^{s} n_i!)$ possible exhaustive, random partitions of the n data vectors $\{\underset{\sim}{y}_\ell\}$ into successive random samples of sizes $n_1, n_2, \ldots n_s$ for the s subpopulations.

Under the null hypothesis, the expected values $\{\bar{y}_j\}$ of the subpopulation means $\{\bar{y}_{ij}\}$ and the covariance structure $\underset{\sim}{V}$, can be formulated along the lines given for the contingency table case in section 1.2.1. The resulting randomization statistic has the form,

$$Q_{MMS} = \{(n - 1)/n\} \sum_{i=1}^{s} n_i(\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})'\underset{\sim}{V}^{-1}(\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})$$

$$= \{(n - 1)/n\}\{tr(\underset{\sim}{S}_a\underset{\sim}{V}^{-1})\}, \tag{1.15}$$

where $\underset{\sim}{S}_a$ is the among subpopulation sums of products matrix and $\underset{\sim}{V}$ is the assumed non-singular total variance-covariance matrix and tr is the trace matrix operator. Under H_o, Q_{MMS} asymptotically has the chi-square distribution with d.f. = d(s - 1). For the case of normally distributed data, Q_{MMS} corresponds to the Pillai's trace criterion for which exact tables are available for small samples in Pillai (1960). For rank transformed responses, the trace criterion is the multivariate extension of the Kruskal-Wallis analysis of variance statistic discussed in Chatterjee and Sen (1966) and Puri and Sen (1971).

Suppose that the multivariate response profile corresponds to j = 1,2,...,d binary indicators for a (d + 1) = r level univariate categorical response. More specifically, in the raw data format, an observation in the conceptualized (s x r) table for subpopulations versus response levels would have a response profile with zeroes for (d - 1) variables and with a one for the jth variable if the response had the jth value, or all zeroes if the response had the (d + 1) = rth value. The Q_{MMS} statistic is equivalent to the Pearson chi-square statistic except for a multiplicative factor {(n - 1)/n}. Additionally, for this situation, suppose that there is a set of scores, $\underset{\sim}{a}' = \{a_1, a_2, \ldots, a_d\}$, which correspond to the values of the categories for the response; then their product, $\underset{\sim}{a}'\underset{\sim}{y}_\ell = a_{ij\ell}$, is the actual value of the observation's response level. The corresponding univariate analysis of variance randomization statistic is the mean score statistic, Q_{MS}, in section 1.2.1 with d.f. = (s - 1).

Other specific details of the expected value and covariance structure as well as the nature of the multivariate test statistic are given in Chapter II relative to the stratified sample case.

1.4 One-Way Analysis of Covariance

In this section, some aspects of randomization methodology are summarized for one response variable and one or more covariables in a framework involving $i = 1,2,\ldots,s$ subpopulations within a single stratum. The discussion is based upon the formulation of randomization methods for covariance adjustment as presented by Quade (1967) for the case of rank transformed data.

Let $\{y_{i\ell},x_{i\ell}\}$ represent a bivariate set of values regarding a response variable, y, and a corresponding concomitant variable, x for $\ell = 1,2,\ldots,n_i$ subjects in $i = 1,2,\ldots,s$ subpopulations. Furthermore, assume that the distribution of $\{x_{i\ell}\}$ is random with respect to its allocation across subpopulations in the sense of resulting from a set of simple random samples without replacement of sizes $\{n_1,n_2,\ldots,n_s\}$ from the pooled sample $\sum_{i=1}^{s} n_i = n$ values. In this situation, the pooled sample can be used to determine a relationship through which $\{y_{i\ell}\}$ can be predicted from $\{x_{i\ell}\}$. Residual scores can be formulated as the difference between observed and predicted values of $\{y_{i\ell}\}$ such that the scores are negative for observed values smaller than the predicted values and positive for values larger. Under the null hypothesis that the distribution of $\{y_{i\ell}\}$ is random with respect to the subpopulations, the residual scores will be randomly distributed across subpopulations; i.e., they are comparable to a set of simple random samples without replacement of sizes $\{n_1,n_2,\ldots,n_s\}$ from the pooled sample $\sum_{i=1}^{s} n_i = n$ values.

The alternatives to this hypothesis correspond to tendencies for one or more subpopulations to have more negative (or positive) residual scores than other subpopulations. Quade suggests comparing the subpopulations by performing an ordinary one-way analysis of variance on the residual scores.

For the particular case of ranks, Quade requires that each variate be measured on at least an ordinal scale; continuity is not required and even a dichotomy is permitted as an extreme case. Quade defines a rank corrected for the mean as being $\{R_{i\ell} - (n + 1)/2\}$, using average ranks in the case of ties and (for definiteness) ranks from the smallest first. Quade uses the ordinary variance ratio criterion based on the among subpopulations mean squares to within subpopulation mean squares; this criterion asymptotically has an F distribution with $(s - 1, n - s)$ degrees of freedom. With residual scores, correction for the mean is avoided since the scores sum to zero.

Quade suggests that the formulation of residuals can be based on models with more than one covariate. Furthermore, a multivariate response profile of residual scores can be constructed by successive concatenation of residual score vectors associated with univariate regression models. With Q_{MS} in section 1.2 or Q_{MMS} in section 1.3 as an alternative to the variance ratio used by Quade, the analysis of variance on the residual scores of rank-type data can be viewed as a covariate-adjusted Kruskal-Wallis test statistic and a covariate-adjusted multivariate Kruskal-Wallis statistic, respectively.

Other details of the expected value and covariance structure as well as the nature of the covariance-adjusted test statistic are given in Chapter II relative to the stratified sample case.

1.5 Incomplete Data

For multivariate response profiles, an incomplete data problem frequently occurs in the sense that some observational units may be missing one or more of the responses. An alternative to deleting response variables that have missing data or to deleting observational units with missing data involves estimating the missing values. In the situation where the multivariate dimension is a result of having repeated measurements in contrast to multiple variables, estimates can be determined on an observational unit basis. In particular, the estimated value might be either a central value or an interpolated value of neighboring data as in the case of improvement scales relative to successive visits. A more general method useful to both repeated measurements and multiple variables involves an assignment of the expected value, e.g. stratum mean. Beale and Little (1975), proposed a regression on the variables which are observed in each data vector.

A dilemma in estimating values is that subsequent statistical procedures usually are blind to the nature of the individual data value, be it observed or estimated. In particular, both the means and the variance structure can be sensitive to the estimation procedure and the inflated sample size. Addressing this problem, John and Prescott (1975) describe a method of covariance adjustment of responses for the pattern of missing data identified by a set of missing data indicator variables representing the observed status of each value in the "completed" data array. Stanish, Gillings and Koch (1978) show that within the context of functional asymptotic regression methodology, means and variances for ratio estimates based on missing value indicators depend only on the observed data. A similar approach for randomization statistics is presented in Chapter II. For this purpose, the following discussion pertains to the case of the ratio estimator presented by Stanish, Gillings, and Koch (1978).

Let $\underset{\sim}{y}' = (y_1, y_2, \ldots, y_r)$ represent a multivariate response vector obtained for n observations where some data are missing. Let $\underset{\sim}{u}' = (u_1, u_2, \ldots, u_r)$ represent a multivariate vector of binary indicators relative to the missing or observed status of the corresponding response variable. A necessary assumption is that u_j is independent of the actual value of y_j for $j = 1, 2, \ldots, r$.

With u_j assuming the value one if the jth response variable is observed and the value zero otherwise, the product $f_j = y_j u_j$ has the value y_j if the data is observed and zero otherwise. Furthermore, the mean of the jth response variable is simply the ratio estimator

$$R_j = \sum_{\ell=1}^{m} f_{j\ell} / \sum_{\ell=1}^{n} u_{j\ell} = \bar{f}_j / \bar{u}_j \tag{1.16}$$

the mean of the observed data. In accordance with the functional asymptotic regression methodology, if the pertinent random variables are included in one vector $g' = (f_1, f_2, \ldots, f_r, u_1, u_2, \ldots, u_r)$ for each observation, then the multivariate ratio estimator of the multivariate means can be expressed as

$$\underset{\sim}{R} = \underset{\sim}{\exp}\{\underset{\sim}{A}\ \underset{\sim}{\log}_e(\bar{\underset{\sim}{g}})\}, \tag{1.17}$$

where $\bar{\underset{\sim}{g}}$ is the across subject sample mean vector of $\underset{\sim}{g}$, $\underset{\sim}{\log}_e(\underset{\sim}{a})$ transforms each element of $\underset{\sim}{a}$ to its natural logarithm, $\underset{\sim}{A} = [\underset{\sim}{I}_r - \underset{\sim}{I}_r]$ subtracts $\log_e \bar{u}_j$ from $\log_e \bar{f}_j$ and $\underset{\sim}{\exp}(\underset{\sim}{a})$ transforms each element of $\underset{\sim}{a}$ to its antilogarithm. The covariance matrix of $\underset{\sim}{R}$ is a multiplicative computation based on the ratio construction. More specifically

$$\underset{\sim}{V}_R = \underset{\sim}{D}_R \underset{\sim}{A}\, \underset{\sim}{D}_g^{-1} \underset{\sim}{V}_g \underset{\sim}{D}_g^{-1} \underset{\sim}{A}' \underset{\sim}{D}_R, \tag{1.18}$$

where $\underset{\sim}{D}_a$ is a diagonal matrix with elements of $\underset{\sim}{a}$ on the main diagonal. The specific elements of variance, $\underset{\sim}{V}_R$, depend only on the observed data. More specifically,

$$\begin{aligned} \mathrm{Var}(R_j) &= \left(\frac{1}{\bar{u}_j}\right)^2 \sum_{\ell=1}^{n} (f_{j\ell} - R_j u_{j\ell})^2/n^2 \\ &= \sum_{obs(j)} (y_{j\ell} - R_j)^2 n_j{}^2 \end{aligned} \tag{1.19}$$

$$\begin{aligned} \mathrm{Cov}(R_j, R_{j'}) &= \sum_{obs(jj')} (y_{j\ell} - R_j(y_{j'\ell} - R_{j'})/n_j n_{j'} \\ &= \{n_{jj'}^2/n_j n_{j'}\} V_{jj'} \end{aligned} \tag{1.20}$$

where $v_{jj'}$ denotes the conditional covariance between the jth and j'th ratio estimates, based on the $n_{jj'}$ observations which have data present for both responses. With this structure, the results of the subsequent analysis are invariant to whatever constant value might be assigned to missing data.

1.6 Research Effort

The objective of this dissertation is to demonstrate the application of randomization methods as a broad and flexible class of non-parametric procedures to a variety of health sciences applications involving multivariate responses, covariate and stratified samples. Thus this literature review has focused on the randomized allocation of a fixed population and the performance of the test statistics under the null hypothesis of no relationships. There already exists a vast literature pertaining to the properties of nonparametric procedures for given random sampling assumptions from general families of distributions; see such references as Gibbons (1971, 1976), Hollander and Wolfe (1973), and Lehmann (1975). This research addresses the use of comparable forms of statistics pertaining to nonparametric rank methods, categorical methods and parametric methods performed on rank transformed data.

The evolution of the concept of no association has been reviewed in the context of the contingency table setting. The corresponding analysis of variance and correlation statistics have been shown to be a special case of a more general raw-data form of randomization statistics. Thus, the more general formulation provides the extension of categorical techniques to include covariance adjustment. Alternatively, the raw-data form of randomization statistics provides the basis for computing multivariate and covariate partial association extensions of continuous nonparametric statistics. These include Kruskal-Wallis analysis of variance on rank scores, and Spearman rank correlation. Furthermore when scores correspond to modified ridits formulated as a standardized rank (i.e., rank divided by (n + 1)), the partial association randomization statistic in the raw-data setting corresponds to a stratified, combined rank statistic discussed by van Elteren (1960) and Lehmann (1975). Finally, when scores correspond to logranks, the randomization statistics correspond to the nonparametric test procedures proposed by Mantel (1966) for survival data comparisons. As a result of the general formulation, these statistics in turn are extended to the categorical case.

The statistical theory of the general formulation for the mean score test statistics is presented in Chapter II relative to scores, stratification, covariance adjustment, multivariate response profiles and incomplete multivariate response profiles. Because other methods such as weighted least squares and maximum likelihood logistic models are also discussed in the examples, brief descriptions of these techniques are given in the latter sections of Chapter II.

The general application of randomization methodology is discussed in Chapter III relative to two examples. The first example is an unbalanced repeated measurements experiment concerned with intracardiac conduction impedance relative to two diagnostic groups of patients in a single stratum with a small sample of 17 patients. A strategy is presented for summarizing the complex multivariate response profile into an univariate score defined at both an experimental condition level and a more general level. The second example is a clinical trial assessing drug efficacy for chronic joint pain. The percent of patients with little or no pain is contrasted between the two treatment groups, active drug and placebo, for a response variable with adjustment for the concomitant age distribution. Also, strategies concerning stratification and covariance adjustment are presented in coordination with weighted least squares and maximum likelihood procedures.

Chapter IV is concerned with multivariate response profiles relative to a multi-center clinical trial within which four subpopulations are defined for the combined levels of two drugs. Patient response is monitored hourly. Strategies for the multivariate data are presented within a stratification system for clinics and initial condition status. Additional analyses are performed using weighted least squares.

Chapter V pertains to the application of stratification techniques in a cumulative analysis of ten ordered strata. Average partial association statistics used in this analysis include the generalized randomized block chi-square statistic, the average standard normal statistic and the Fisher combined p-value statistic discussed in section 3.1 of Chapter II for the two subpopulation case. The data for this example are from a series of double blind, parallel group randomized clinical trials employing six centers to explore the efficacy of an investigational treatment for depression in comparison with a placebo control treatment and an active control treatment.

Analytical aspects of a nationwide observational study with a historical control are presented in Chapter VI. A strategy is presented for defining the randomization framework wherein an antidote subpopulation is derived from one study and a supportive therapy subpopulation from the other study. Multivariate data are viewed in both their continuous form and their categorical form. The use of different scores is explored in the univariate case.

Special applications of randomization statistics are presented in Chapter VII relative to an incomplete design example and an incomplete data example. The incomplete design involves multiple strata having two to four subpopulations each. The incomplete data example involves clinic data collected over three consecutive clinic visits relative to patient response to two treatments, active drug or placebo. Different clinics, composing separate strata, have different missing value patterns across the three visits. A supportive analysis using a weighted least squares based ratio estimator procedure is also presented.

Chapter VIII concludes this dissertation with a summary of the randomization methodology presented and suggested strategies for future research.

CHAPTER II

MEAN SCORE RANDOMIZATION TEST STATISTICS

The concept of no association regarding the partition of scores among subpopulations has been applied to mean scores from both a contingency table test approach and a nonparametric test approach, see Chapter I. With mean score functions, the Landis, Heyman and Koch (1978) formulations in the contingency table setting can be rewritten in a more general raw-data form. As a result, the more general raw-data formulation of mean score randomization test statistics yields a family of analysis of variance statistics. More specifically, for data with categorical scores relating to a contingency table framework, this general raw-data formulation can be equivalent to tests such as Pearson's test for r x c tables and Mantel-Haenszel's test for sets of s x r tables. With rank-type scores, tests can include Quade's rank analysis of covariance, Friedman's randomized block analysis and Kruskal-Wallis' k-sample analysis of variance. Section 2.1 presents this raw-data formulation for the multivariate response profile case without stratification. For the single contingency table case, the resulting MANOVA randomization statistic is equivalent to Pearson's chi-square test statistic for an (s x r) contingency table with (s - 1)(r - 1) degrees of freedom. For rank-type scores, the resulting MANOVA randomization statistic is equivalent to the multivariate Kruskal-Wallis test statistic for d responses and s subpopulations with d(s - 1) degrees of freedom.

Section 2.2 presents the raw-data formulation for covariance adjustment in the single stratum case. In particular, covariance adjustment is shown to be a partitioning of the multivariate statistic in Section 2.1 into two components: the covariate(s) statistic and the covariate-adjusted response(s) statistic. For rank-type scores, the covariate-adjusted mean score randomization statistic gives equivalent results to Quade's (1967) rank analysis of covariance statistic. For categorical data, the covariate-adjusted mean score randomization statistic provides an extension to categorical techniques to include covariance adjustment in the single table.

Section 2.3 concludes the presentation of the generalized raw-data formulation with the stratified extension of the mean score test statistic and its covariance adjustment. For categorical data, the two-way analysis of variance randomization statistic is the generalized Mantel-Haenszel statistic discussed in Landis, Heyman and Koch (1978). For rank-type scores, the statistic can be viewed as a generalized Friedman statistic. With modified ridits, the statistic yields van Elteran's stratified combined rank statistic. With logrank scores, the statistic is equivalent to the nonparametric test procedure proposed by Mantel (1966) for survival data comparisons. With covariance adjustment and multivariate profiles, the across-strata mean score randomization statistic represents the covariance and multivariate extensions of these test statistics. Furthermore, with the application of rank-type scores to categorical data, the across-strata mean score randomization statistic provides categorical counterparts to nonparametric statistics.

The remaining sections include Section 2.4 which demonstrates the use of binary covariates as a missing data strategy, and Section 2.5 which is concerned with specific types of scores used in this dissertation. The statistical methodology in Section 2.1 through 2.5 is applied in Chapters III, IV, V, VI and VII through the use of a SAS macro GRMM documented in Amara and Koch (1980).

Finally, Sections 2.6 and 2.7 briefly review aspects of weighted least squares and maximum likelihood logistic models which appear in the examples as supportive analyses.

To facilitate the discussion in Sections 2.1 and 2.2, it is assumed that the data are non-redundant such that the overall variance-covariance matrix is non-singular. In Section 2.3, the across-strata covariance matrix is a summation of the overall variance-covariance matrices which can involve stratum matrices that might be singular.

2.1 Multivariate Analysis of Variance

The randomization framework and hypothesis of no association is presented in Amara and Koch (1980). It is expressed in terms of (d x 1) vectors $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n$ for $j = 1,2,\ldots,d$ response variables for the subjects in some finite population of size n. Let $i = 1,2,\ldots,s$ index a set of subpopulations for which the relationship to the d response variables is to be investigated. Let $n_1, n_2, \ldots, n_s$ denote the corresponding numbers of subjects so that $\sum_{i=1}^{s} n_i = n$. If the subpopulations are homogeneous in the sense of equivalence to random allocation, then the distribution of the responses among them is compatible with the hypothesis of no association:

H_o: There is no relationship between subpopulations and responses in the sense of equally likely realizations for the $(n!/\prod_{i=1}^{s} n_i!)$ possible exhaustive, random partitions of the n data response vectors into successive random samples of sizes $n_1, n_2, \ldots, n_s$ for the s subpopulations.

In accordance with finite population randomization, the j-th response of the ℓ-th subject in the i-th subpopulation has the expected value

$$E\{y_{ij\ell}|H_o\} = \{\sum_{i=1}^{s}\sum_{\ell=1}^{n_i} y_{ij\ell}\}/n = \bar{y}_j, \tag{2.1.1}$$

the mean of the j-th response in the pooled sample. If $\bar{\underset{\sim}{y}}' = \{\bar{y}_j\} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_d)$ denotes the pooled sample mean vector, then the covariance structure of the multivariate responses in the pooled sample can be expressed as

$$\begin{aligned} \text{Cov}\{y_{ij\ell}, y_{ij'\ell}|H_o\} &= \{\sum_{i=1}^{s}\sum_{\ell=1}^{n_i}(y_{ij\ell} - \bar{y}_j)(y_{ij'\ell} - \bar{y}_{j'})\}/n \\ &= \underset{\sim}{V}\,. \end{aligned} \tag{2.1.2}$$

Let $\bar{\underset{\sim}{y}}_i = \{\bar{y}_{ij}\} = \{\bar{y}_{i1}, \bar{y}_{i2}, \ldots, \bar{y}_{id}\}$ denote the observed mean vector of the i-th subpopulation for $i = 1,2,\ldots,s$. In accordance with finite population randomization, the expected value vector of $\bar{\underset{\sim}{y}}_i$ is the pooled sample mean vector $\bar{\underset{\sim}{y}}$. Since the $\bar{\underset{\sim}{y}}_i$ are linear functions of the data vectors with the same expected value vector, its covariance structure can be summarized as

$$\underset{\sim}{\text{Cov}}\{\bar{\underset{\sim}{y}}_i, \bar{\underset{\sim}{y}}_{i'}|H_o\} = \{\underset{\sim}{V}/(n-1)\}\{(n\delta_{ii'}/n_i) - 1\}, \tag{2.1.3}$$

where $\delta_{ii'} = 1$ if $i = i'$ and $\delta_{ii'} = 0$ if $i \neq i'$. The vectors $\bar{\underset{\sim}{y}}_i$ have approximate multivariate normal distributions with expected value structure $\bar{\underset{\sim}{y}}$ and covariance structure (2.1.3), see Puri and Sen (1971).

With moderately large samples (e.g., $n_i > 20$), chi-square test statistics for the null hypothesis of randomization can be constructed via quadratic forms in linear functions among the $\bar{\underset{\sim}{y}}_i$.

Let $\underset{\sim}{F}$ denote a composite subpopulation vector,

$$\underset{\sim}{F}' = \{\bar{y}_{11},\ldots,\bar{y}_{1d},\ldots,\bar{y}_{ij},\ldots,\bar{y}_{s1},\ldots,\bar{y}_{sd}\} \quad (2.1.4)$$

From the results for $\bar{\underset{\sim}{y}}_i$, the expected value and covariance structure can be expressed in matrix notation as

$$\underset{\sim}{E}\{\underset{\sim}{F}|H_o\} = \{\bar{y}_1,\ldots,\bar{y}_d\}' \otimes \underset{\sim}{1}_s \quad (2.1.5)$$

$$\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{F}|\underset{\sim}{H}_o\} = \{\underset{\sim}{V}/(n-1)\} \otimes \{\underset{\sim}{D}_{\underset{\sim}{P}}^{-1} - \underset{\sim}{1}_s\underset{\sim}{1}_s'\} \quad (2.1.6)$$

where $\underset{\sim}{1}_s$ denotes an (s x 1) vector of 1's, $\underset{\sim}{D}_{\underset{\sim}{P}}$ denotes a diagonal matrix with elements of the vector

$$\underset{\sim}{P}' = \{P_1,P_2,\ldots,P_s\} = \frac{1}{n}\{n_1,n_2,\ldots,n_s\} \quad (2.1.7)$$

on the main diagonal, and $\otimes$ denotes left Kronecker product multiplication.

The variation among the subpopulations can be expressed as linear functions

$$\underset{\sim}{G} = [\underset{\sim}{I}_d \otimes \underset{\sim}{C}]\underset{\sim}{F} = \begin{bmatrix} \underset{\sim}{F}_1 & - & \underset{\sim}{F}_s \\ \underset{\sim}{F}_2 & - & \underset{\sim}{F}_s \\ & \cdots & \\ \underset{\sim}{F}_{(s-1)} & - & \underset{\sim}{F}_s \end{bmatrix} \quad (2.1.8)$$

where $\underset{\sim}{I}_d$ denotes a (d x d) identity matrix and

$$\underset{\sim}{C} = [\underset{\sim}{I}_{(s-1)} \;,\; \underset{\sim}{1}_{(s-1)}] \quad (2.1.9)$$

denotes the contrast basis. From the results for $\underset{\sim}{F}$, the expected value and covariance structure can be expressed as

$$E\{\underset{\sim}{G}|H_o\} = \underset{\sim}{0}_{(d)(s-1)} \tag{2.1.10}$$

$$\mathrm{Var}\{\underset{\sim}{G}|H_o\} = \{\underset{\sim}{V}/(n - 1)\} \otimes \{\underset{\sim}{C}\ \underset{\sim}{D}_{\underset{\sim}{P}}^{-1}\ \underset{\sim}{C}'\} \tag{2.1.11}$$

Thus, the randomization test statistic can be expressed in its quadratic form

$$Q_{MMS} = \underset{\sim}{G}'[\mathrm{Var}\{\underset{\sim}{G}|H_o\}]^{-1}\ \underset{\sim}{G} \tag{2.1.12}$$

where MMS denotes multivariate mean scores. This statistic, with moderate sample size, asymptotically has a chi-square distribution with degrees of freedom d(s - 1). Aspects of the asymptotic distribution of Q_{MMS} are discussed in Puri and Sen (1969).

The statistic Q_{MMS} can be written in terms of $\underset{\sim}{F}$.

$$\begin{aligned} Q_{MMS} &= \underset{\sim}{G}'[\mathrm{Var}\{\underset{\sim}{G}|H_o\}]^{-1}\ \underset{\sim}{G} \\ &= (n - 1)\underset{\sim}{F}'[\underset{\sim}{I}_d \otimes \underset{\sim}{C}]'[\underset{\sim}{V} \otimes \underset{\sim}{C}\ \underset{\sim}{D}_{\underset{\sim}{P}}^{-1}\underset{\sim}{C}']^{-1}[\underset{\sim}{I}_d \otimes \underset{\sim}{C}]\underset{\sim}{F} \\ &= (n - 1)\underset{\sim}{F}'[\underset{\sim}{V}^{-1} \otimes \underset{\sim}{C}'\{\underset{\sim}{C}\ \underset{\sim}{D}_{\underset{\sim}{P}}^{-1}\ \underset{\sim}{C}'\}\underset{\sim}{C}]\underset{\sim}{F}. \end{aligned} \tag{2.1.13}$$

On applying the matrix identity lemma in Koch (1969b), it follows that

$$\begin{aligned} Q_{MMS} &= (n - 1)\underset{\sim}{F}'[\underset{\sim}{V}^{-1} \otimes \{\underset{\sim}{D}_{\underset{\sim}{P}} - \underset{\sim}{P}\ \underset{\sim}{P}'\}]\underset{\sim}{F} \\ &= (n - 1)\{\sum_{i=1}^{s} P_i\underset{\sim}{F}_i'\underset{\sim}{V}^{-1}\ \underset{\sim}{F}_i - [\sum_{i=1}^{s} P_i\underset{\sim}{F}_i]'\underset{\sim}{V}^{-1}[\sum_{i=1}^{s} P_i\underset{\sim}{F}_i]\}. \end{aligned}$$

However,

$$\sum_{i=1}^{s} P_i\underset{\sim}{F}_i = \{\sum_{i=1}^{s} n_i\underset{\sim}{F}_i\}/n = \bar{\underset{\sim}{y}}\ .$$

Thus

$$\begin{aligned} Q_{MMS} &= (n-1)\{\sum_{i=1}^{s} p_i \underset{\sim}{F}'_i \underset{\sim}{V}^{-1} \underset{\sim}{F}_i - \bar{\underset{\sim}{y}}'\underset{\sim}{V}\bar{\underset{\sim}{y}}\} \\ &= (n-1)\{\sum_{i=1}^{s} p_i (\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})'\underset{\sim}{V}^{-1}(\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})\} \\ &= (n-1)\mathrm{tr}\{\sum_{i=1}^{s} p_i (\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})(\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})'\underset{\sim}{V}^{-1}\} \\ &= \{(n-1)\}\mathrm{tr}\{\underset{\sim}{S}_F \underset{\sim}{V}^{-1}\} \end{aligned} \tag{2.1.14}$$

where

$$\underset{\sim}{S}_F = \sum_{i=1}^{s} n_i (\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})(\bar{\underset{\sim}{y}}_i - \bar{\underset{\sim}{y}})' \tag{2.1.15}$$

is the among subpopulations sums of cross-products matrix. In other words, Q_{MMS} can be interpreted as a one-way multivariate analysis of variance statistic based on the multivariate trace criterion applied to $\underset{\sim}{S}_F \underset{\sim}{V}^{-1}$.

2.2 Multivariate Analysis of Covariance

For $i = 1,2,\ldots,s$ subpopulations with $\ell = 1,2,\ldots,n_i$ subjects, let $\bar{\underset{\sim}{y}}_i$ and $\bar{\underset{\sim}{x}}_i$ denote subpopulation mean vectors of $j = 1,2,\ldots,d$ responses and $k = 1,2,\ldots,t$ covariables respectively. With moderate sample size ($n_i > 20$), the mean vectors $\bar{\underset{\sim}{y}}_i$ and $\bar{\underset{\sim}{x}}_i$ have approximately a joint multivariate normal distribution. If $\bar{z}_{i\ell} = \underset{\sim}{a}'\bar{\underset{\sim}{y}}_{i\ell}$ represents a univariate function of the multivariate response profile for the ℓ-th subject in the i-th subpopulation, it follows from Section 2.1 that $\bar{z}_i = \underset{\sim}{a}'\bar{\underset{\sim}{y}}_i$ and $\bar{\underset{\sim}{x}}_i$ have approximately a joint multivariate normal distribution with the expected value and covariance structure

$$\underset{\sim}{E}\{\begin{bmatrix}\bar{z}_i \\ \bar{\underset{\sim}{x}}_i\end{bmatrix}|H_o\} = \underset{\sim}{E}\{\frac{1}{n_i}\sum_{\ell=1}^{n_i}\begin{bmatrix}z_{i\ell} \\ \underset{\sim}{x}_{i\ell}\end{bmatrix}|H_o\} = \frac{1}{n}\sum_{i=1}^{s}\sum_{\ell=1}^{n_i}\begin{bmatrix}z_{i\ell} \\ \underset{\sim}{x}_{i\ell}\end{bmatrix} = \begin{bmatrix}\bar{z} \\ \bar{\underset{\sim}{x}}\end{bmatrix} \tag{2.2.1}$$

$$\mathrm{Cov}\{\begin{bmatrix}\bar{z}_i\\ \bar{\underset{\sim}{x}}_i\end{bmatrix},\begin{bmatrix}\bar{z}_{i'}\\ \bar{\underset{\sim}{x}}_{i'}\end{bmatrix}|H_o\} = \{\sum_{i=1}^{s}\sum_{\ell=1}^{n_i}\begin{bmatrix}(z_{i\ell}-\bar{z})\\ (\underset{\sim}{x}_{i\ell}-\bar{\underset{\sim}{x}})\end{bmatrix}[(z_{i\ell}-\bar{z}),(\underset{\sim}{x}_{i\ell}-\bar{\underset{\sim}{x}})']\}\{\frac{n\delta_{ii'}-n_i}{n_i(n-1)}\}$$

$$= \begin{bmatrix}V_{zz} & \underset{\sim}{V}_{xz}'\\ \underset{\sim}{V}_{xz} & \underset{\sim}{V}_{xx}\end{bmatrix}\{\frac{n\delta_{ii'}-n_i}{n_i(n-1)}\} \tag{2.2.2}$$

where $\delta_{ii'} = 1$ if $i = i'$ and $\delta_{ii'} = 0$ if $i \neq i'$. The multivariate analysis of variance statistic can be expressed as

$$Q(z,\underset{\sim}{x}) = \{(n-1)/n\}\{\sum_{i=1}^{s} n[(\bar{z}_i-\bar{z}),(\bar{\underset{\sim}{x}}_i-\bar{\underset{\sim}{x}})']\begin{bmatrix}V_{zz} & \underset{\sim}{V}_{xz}'\\ \underset{\sim}{V}_{xz} & \underset{\sim}{V}_{xx}\end{bmatrix}^{-1}\begin{bmatrix}(\bar{z}_i-\bar{z})\\ (\bar{\underset{\sim}{x}}_i-\bar{\underset{\sim}{x}})\end{bmatrix}\} \tag{2.2.3}$$

$$= Q_{MMS}$$

which has an approximate chi-square distribution with degrees of freedom $(1 + t)(s - 1)$ under the joint randomization hypothesis. The statistic $Q(z,\underset{\sim}{x})$ can be partitioned (under the hypothesis) into two independent components,

$$Q(z,\underset{\sim}{x}) = Q(\underset{\sim}{x}) + \{Q(z,\underset{\sim}{x}) - Q(\underset{\sim}{x})\}$$

$$= Q(\underset{\sim}{x}) + Q(z|\underset{\sim}{x}) \tag{2.2.4}$$

where $Q(\underset{\sim}{x})$, which has the form,

$$Q(\underset{\sim}{x}) = \{(n-1)/n\}\{\sum_{i=1}^{s} n_i[(\bar{\underset{\sim}{x}}_i-\bar{\underset{\sim}{x}})']\underset{\sim}{V}_{xx}^{-1}[(\bar{\underset{\sim}{x}}_i-\bar{\underset{\sim}{x}})]\} \tag{2.2.5}$$

is the multivariate criterion for testing the randomization hypothesis for the covariables. If the randomization hypothesis for the covariables is true, (which is usually the case when the values of the covariables are available prior to randomization), the univariate statistic $Q(z|\underset{\sim}{x})$

can be interpreted as a covariance-adjusted test statistic. As will be shown next, $Q(z,\underset{\sim}{x})$ represents total variation, $Q(\underset{\sim}{x})$ represents lack of fit and $Q(z|\underset{\sim}{x})$ represents variation explained by the regression model for among subpopulation variation of the $\bar{z}_i$ given no variation among the $\bar{\underset{\sim}{x}}_i$.

Let $\underset{\sim}{F}$ be the composite vector of subpopulation mean vectors $\underset{\sim}{F}_i = \{\bar{z}_i, \bar{\underset{\sim}{x}}_i\}$. The elements of $\underset{\sim}{F}$ can be rearranged into variable vectors of s subpopulations

$$\underset{\sim}{F}^* = \underset{\sim}{A}_0 \underset{\sim}{F}$$

where $\underset{\sim}{A}_0$ is a permutation matrix operator. It follows that

$$\begin{aligned} \underset{\sim}{G} &= [\underset{\sim}{C} \otimes \underset{\sim}{I}_{(1+t)}]\underset{\sim}{A}_0\underset{\sim}{F} \\ &= \begin{bmatrix} \bar{z}_1 & - & \bar{z}_s \\ & \cdots & \\ \bar{z}_{(s-1)} & - & \bar{z}_s \\ \bar{x}_{11} & - & \bar{x}_{s1} \\ & \cdots & \\ \bar{x}_{(s-1)t} & - & \bar{x}_{st} \end{bmatrix} \end{aligned} \tag{2.2.6}$$

denotes the multivariate subpopulation mean vector contrasts presented in Section 2.1 for (1 + t) variates. For this case,

$$\underset{\sim}{E}\{\underset{\sim}{G}|H_o\} = \underset{\sim}{0}_{(1+t)(s-1)} \tag{2.2.7}$$

$$\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{G}|H_o\} = \{\underset{\sim}{C}\ \underset{\sim}{D}_{\underset{\sim}{p}}^{-1}\underset{\sim}{C}'\} \otimes \{\underset{\sim}{V}/(n-1)\}. \tag{2.2.8}$$

A linear model of interest is directed at subpopulation contrasts of the response z while describing the variation of t covariables as being at random in the sense of a simple random sample partition of the pooled sample among the subpopulations. The corresponding model with regard to the subpopulation contrasts can be expressed as:

$$E(\underset{\sim}{G}|H_o) = \underset{\sim}{K}\ \underset{\sim}{B} = [\underset{\sim}{I}_{(s-1)}, \underset{\sim}{0}_{((s-1)xt)}]'\underset{\sim}{B} \tag{2.2.9}$$

$$\begin{bmatrix} \bar{z}_1 & - & \bar{z}_s \\ \bar{z}_2 & - & \bar{z}_s \\ & \cdots & \\ \bar{z}_{(s-1)} & - & \bar{z}_s \\ \bar{x}_{11} & - & \bar{x}_{s1} \\ & \cdots & \\ \bar{x}_{(s-1)t} & - & \bar{x}_{st} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & . & 0 \\ . & & \ddots & . \\ 0 & 0 & . & 1 \\ \hline 0 & 0 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & 0 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_{(s-1)} \end{bmatrix}.$$

The $\underset{\sim}{B}$ is a ((s - 1) x 1) vector of parameters which reflect an adjustment of the observed $\underset{\sim}{G}$ in accordance with the randomization constraint on $\underset{\sim}{x}$. Thus, the randomization hypothesis concerning the covariance adjusted response z can be represented as

$$H_{o,z|\underset{\sim}{x}} : \underset{\sim}{B} = 0. \tag{2.2.10}$$

For this purpose, weighted least squares methods can be used to fit the model and produce estimates, $\underset{\sim}{b}$, for $\underset{\sim}{B}$ of the nature

$$\underset{\sim}{b} = \{\underset{\sim}{K}'[\underset{\sim}{Var}(\underset{\sim}{G}|H_o)]^{-1}\underset{\sim}{K}\}^{-1}\underset{\sim}{K}'[\underset{\sim}{Var}(\underset{\sim}{G}|H_o)]^{-1}\underset{\sim}{G} \tag{2.2.11}$$

for which the covariance matrix is

$$\underset{\sim}{Var}\{\underset{\sim}{b}|H_o\} = \{\underset{\sim}{K}'[\underset{\sim}{Var}(\underset{\sim}{G}|H_o)]^{-1}\underset{\sim}{K}\}^{-1}. \tag{2.2.12}$$

An appropriate test statistic for the weighted least squares parameter estimates is

$$Q(b) = \underset{\sim}{b}'[\underset{\sim}{Var}(\underset{\sim}{b}|H_o)]^{-1}\underset{\sim}{b}. \tag{2.2.13}$$

Moreover, from weighted regression theory it follows that $Q(\underset{\sim}{b}) \equiv Q(z|\underset{\sim}{x})$ where $Q(z|\underset{\sim}{x})$ is the univariate reduction statistic defined in (2.2.4) and correspondingly has an asymptotic chi-square distribution with degrees of freedom $(s - 1)$.

The weighted least squares estimates, $\underset{\sim}{b}$, can also be expressed in terms of the subpopulation means. For this purpose, let

$$\underset{\sim}{G} = [\underset{\sim}{I}_{(1+t)} \otimes \underset{\sim}{C}] \begin{bmatrix} \bar{z}_1 \\ \bar{x}_{11} \\ \cdots \\ \bar{x}_{1t} \\ \cdots \\ \bar{z}_s \\ \bar{x}_{s1} \\ \cdots \\ \bar{x}_{st} \end{bmatrix} = \begin{bmatrix} \bar{z}_1 & - & \bar{z}_s \\ \bar{x}_{11} & - & \bar{x}_{s1} \\ & \cdots & \\ \bar{x}_{1t} & - & \bar{x}_{st} \\ & \cdots & \\ \bar{z}_{(s-1)} & - & \bar{z}_s \\ \bar{x}_{(s-1)1} & - & \bar{x}_{s1} \\ & \cdots & \\ \bar{x}_{(s-1)t} & - & \bar{x}_{st} \end{bmatrix} \qquad (2.2.14)$$

and

$$\mathrm{Var}\{\underset{\sim}{G}|H_o\} = \frac{1}{(n-1)}\{\underset{\sim}{V} \otimes \underset{\sim}{C}\ \underset{\sim}{D}_p^{-1}\underset{\sim}{C}'\} \ . \qquad (2.2.15)$$

Recall,

$$\underset{\sim}{b} = \{\underset{\sim}{K}'[\mathrm{Var}(\underset{\sim}{G}|H_o)]^{-1}\underset{\sim}{K}\}^{-1}\ \underset{\sim}{K}'[\mathrm{Var}(\underset{\sim}{G}|H_o)]^{-1}\underset{\sim}{G}$$

and

$$\mathrm{Var}(\underset{\sim}{b}|H_o) = \{\underset{\sim}{K}'[\mathrm{Var}(\underset{\sim}{G}|H_o)]^{-1}\underset{\sim}{K}\}^{-1}$$

where $\underset{\sim}{K}$ for this permutation of subpopulation means can be expressed as

$$\underset{\sim}{K} = \{\begin{bmatrix} 1 \\ \underset{\sim}{0}_{(t)} \end{bmatrix} \otimes \underset{\sim}{I}_{(s-1)}\} \qquad (2.2.16)$$

By substituting the above expressions regarding $\underset{\sim}{G}$ and $\underset{\sim}{K}$, it follows that

$$\underset{\sim}{b} = (n-1)\{\widehat{\mathrm{Var}}(\underset{\sim}{b}|H_o)\}\{[1,\underset{\sim}{0}_{(1xt)}] \otimes \underset{\sim}{I}_{(s-1)}\}\{\underset{\sim}{V} \otimes \underset{\sim}{C}\ \underset{\sim}{D}_P^{-1}\underset{\sim}{C}'\}^{-1}\{\underset{\sim}{I}_{(1+t)} \otimes \underset{\sim}{C}\}\begin{bmatrix}\bar{z}_1\\ \cdots\\ \bar{x}_{1t}\\ \cdots\\ \bar{z}_s\\ \cdots\\ \bar{x}_{st}\end{bmatrix} \tag{2.2.17}$$

The central matrix product

$$\{[1,\underset{\sim}{0}_{(1xt)}] \otimes \underset{\sim}{I}_{(s-1)}\}\{\underset{\sim}{V} \otimes \underset{\sim}{C}\ \underset{\sim}{D}_P^{-1}\ \underset{\sim}{C}'\}^{-1}\{\underset{\sim}{I}_{(1+t)} \otimes \underset{\sim}{C}\}$$

can be simplified by recognizing the following equivalent re-expressions,

$$\{[1,\underset{\sim}{0}_{(1xt)}] \otimes \underset{\sim}{I}_{(s-1)}\}\{\underset{\sim}{V}^{-1} \otimes (\underset{\sim}{C}\ \underset{\sim}{D}_P^{-1}\underset{\sim}{C}')^{-1}\}\{\underset{\sim}{I}_{(1+t)} \otimes \underset{\sim}{C}\}$$

$$\{[1,\underset{\sim}{0}_{(1xt)}]\underset{\sim}{V}^{-1}\ \underset{\sim}{I}_{(1+t)}\} \otimes \{\underset{\sim}{I}_{(s-1)}(\underset{\sim}{C}\ \underset{\sim}{D}_P^{-1}\underset{\sim}{C}')^{-1}\underset{\sim}{C}\}$$

$$\{[1,\underset{\sim}{0}_{(1xt)}]\underset{\sim}{V}^{-1}\} \otimes \{(\underset{\sim}{C}\ \underset{\sim}{D}_P^{-1}\ \underset{\sim}{C}')^{-1}\ \underset{\sim}{C}\} \tag{2.2.18}$$

Furthermore, the variance of the subpopulation means can be partitioned

$$\underset{\sim}{V} = \begin{bmatrix} V_{zz} & \underset{\sim}{V}'_{xz} \\ \underset{\sim}{V}_{xz} & \underset{\sim}{V}_{xx} \end{bmatrix}.$$

By recognizing that $\underset{\sim}{V}\ \underset{\sim}{V}^{-1} = \underset{\sim}{I}$, the corresponding partitions of $\underset{\sim}{V}^{-1}$ can be derived from four linear equations

$$\begin{bmatrix} V_{zz} & \underset{\sim}{V}'_{xz} \\ \underset{\sim}{V}_{xz} & \underset{\sim}{V}_{xx} \end{bmatrix} \begin{bmatrix} A & \underset{\sim}{B}' \\ \underset{\sim}{B} & \underset{\sim}{C} \end{bmatrix} = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{I}_{(t)} \end{bmatrix}$$

which can be expressed explicitly as

$$\begin{bmatrix} V_{zz} A + \underset{\sim}{V}'_{xz} \underset{\sim}{B} = 1 \\ \underset{\sim}{V}_{xz} A + \underset{\sim\sim}{V}_{xx} \underset{\sim}{B} = \underset{\sim}{0}_{(tx1)} \\ V_{zz} \underset{\sim}{B}' + \underset{\sim}{V}'_{xz} \underset{\sim}{C} = \underset{\sim}{0}_{(1xt)} \\ \underset{\sim}{V}_{xz} \underset{\sim}{B}' + \underset{\sim\sim}{V}_{xx} \underset{\sim}{C} = \underset{\sim}{I}_{(t)} \end{bmatrix} . \qquad (2.2.19)$$

By solving for the partitions A, $\underset{\sim}{B}'$ and $\underset{\sim}{C}$, it follows that

$$\underset{\sim}{V}^{-1} = \begin{bmatrix} A & \underset{\sim}{B}' \\ \underset{\sim}{B} & \underset{\sim}{C} \end{bmatrix} \qquad (2.2.20)$$

where

$$A = (V_{zz} - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx} \underset{\sim}{V}_{xz})^{-1}$$

$$\underset{\sim}{B}'_{(1xt)} = -(V_{zz} - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx} \underset{\sim}{V}_{xz})^{-1} \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}$$

$$\underset{\sim}{C}_{(txt)} = -\underset{\sim\sim}{V}^{-1}_{xx} \underset{\sim}{V}_{xz} (V_{zz} - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx} \underset{\sim}{V}_{xz})^{-1} (\underset{\sim\sim}{V}_{xx} - \underset{\sim}{V}_{xz} V^{-1}_{zz} \underset{\sim}{V}'_{xz})^{-1}$$

and

$$\{[1, \underset{\sim}{0}_{(1xt)}] \underset{\sim}{V}^{-1}\} = \{[V_{zz} - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx} \underset{\sim}{V}_{xz}]^{-1} [1, -\underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}]\} .$$

Thus, the central matrix product (2.2.18) simplifies to

$$\{[V_{zz} - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx} \underset{\sim}{V}_{xz}]^{-1} [1, -\underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}]\} \otimes \{(\underset{\sim}{C} \underset{\sim}{D}^{-1}_{P} \underset{\sim}{C}')^{-1} \underset{\sim}{C}\} \qquad (2.2.21)$$

Recall that

$$\mathrm{Var}(\underset{\sim}{b} | H_o) = \{\underset{\sim}{K}' [\mathrm{Var}(\underset{\sim}{G} | H_o)]^{-1} \underset{\sim}{K}\}^{-1}.$$

By substituting the definition of $\underset{\sim}{K}$ in (2.2.16), the variance can be re-expressed as

$$\frac{1}{(n-1)}[\{[1,\underset{\sim}{0}_{(1xt)}]\ \underset{\sim}{V}^{-1}\begin{bmatrix}1\\ \underset{\sim}{0}_{(tx1)}\end{bmatrix}\}\otimes\{\underset{\sim}{C}\ \underset{\sim}{D}_{P}^{-1}\ \underset{\sim}{C}'\}^{-1}]^{-1}.$$

More specifically in terms of $\underset{\sim}{V}^{-1}$ partitions,

$$\frac{1}{(n-1)}[\{V_{zz} - \underset{\sim}{V}'_{xz}\ \underset{\approx}{V}_{xx}^{-1}\ \underset{\sim}{V}_{xz}\}\otimes\{\underset{\sim}{C}\ \underset{\sim}{D}_{P}^{-1}\ \underset{\sim}{C}'\}]. \tag{2.2.22}$$

By substituting (2.2.22) into (2.2.17), the expression for the estimates $\underset{\sim}{b}$ can now be written as

$$\underset{\sim}{b} = \{[1,\ -\underset{\sim}{V}'_{xz}\ \underset{\approx}{V}_{xx}^{-1}]\otimes\underset{\sim}{C}\}\begin{bmatrix}\bar{z}_1\\ \bar{x}_{11}\\ \cdots\\ \bar{x}_{1t}\\ \cdots\\ \bar{z}_s\\ \bar{x}_{s1}\\ \cdots\\ \bar{x}_{st}\end{bmatrix}$$

$$= \underset{\sim}{C}\begin{bmatrix}\bar{z}_1 - \underset{\sim}{V}'_{xz}\ \underset{\approx}{V}_{xx}^{-1}\ \underset{\sim}{\bar{x}}_1\\ \cdots\\ \bar{z}_s - \underset{\sim}{V}'_{xz}\ \underset{\approx}{V}_{xx}^{-1}\ \underset{\sim}{\bar{x}}_s\end{bmatrix}$$

$$= \underset{\sim}{C}[\underset{\sim}{\bar{z}}_i - \underset{\sim}{\hat{z}}_i] \tag{2.2.23}$$

where $\underset{\sim}{\hat{z}}_i$ is the composite vector of least squares predicted values for the subpopulation means of the response function $\{z_{i\ell}\}$. These predicted values are based on a linear regression model involving the t covariables for the pooled population. In this regard,

$$\underset{\sim}{\hat{B}} = \underset{\approx}{V}_{xx}^{-1}\ \underset{\sim}{V}_{xz} \tag{2.2.24}$$

represents the vector of least squares estimates for the parameter $\underset{\sim}{B}$ of the linear model,

$$\hat{z}_{i\ell} = \bar{z} + \underset{\sim}{B}'(\underset{\sim}{x}_{i\ell} - \bar{\underset{\sim}{x}})$$
$$= \bar{z} + \underset{\sim}{V}_{xz}' \underset{\sim\sim}{V}_{xx}^{-1}(\underset{\sim}{x}_{i\ell} - \bar{\underset{\sim}{x}}) \quad (2.2.25)$$

where $\hat{z}_{i\ell}$ is the predicted value of the response function $z_{i\ell}$ observed for the ℓ-th subject in the i-th subpopulation. It follows then that

$$g_{i\ell} = (z_{i\ell} - \hat{z}_{i\ell}) \quad (2.2.26)$$

is the corresponding residual. By construction,

$$\underset{\sim}{b} = \underset{\sim}{C}\ \underset{\sim}{g}. \quad (2.2.27)$$

From the results in Section 2.1, it follows that

$$\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{b}|H_o\} = \frac{1}{(n-1)}\{V_{gg} \otimes \underset{\sim}{C}\ \underset{\sim}{D}_P^{-1}\ \underset{\sim}{C}'\} \quad (2.2.28)$$

where V_{gg} is the pooled population residual variance under the randomization hypothesis. This result follows directly by noting that

$$\bar{g}_i = \frac{1}{n_i} \sum_{\ell=1}^{n_i} g_{i\ell}$$
$$= (\bar{z}_i - \bar{z}) - \underset{\sim}{V}_{xz}' \underset{\sim\sim}{V}_{xx}^{-1} (\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}}) \quad (2.2.29)$$

is the mean residual for the i-th subpopulation with

$$\bar{g} = \frac{1}{n} \sum_{i=1}^{s} \sum_{\ell}^{n_i} g_{i\ell} = 0$$

as the pooled population mean residual. The variance can then be expressed as

$$V_{gg} = \frac{1}{n} \sum_{i=1}^{s} \sum_{\ell=1}^{n_i} [(g_{i\ell} - \bar{g})(g_{i\ell} - \bar{g})']$$

$$= \frac{1}{n} \sum_{i=1}^{s} \sum_{\ell=1}^{n_i} g_{i\ell} g'_{i\ell}$$

$$= \frac{1}{n} \sum_{i=1}^{s} \sum_{\ell=1}^{n_i} [(z_{i\ell} - \bar{z}) - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}(\underset{\sim}{x}_{i\ell} - \bar{\underset{\sim}{x}})][(z_{i\ell} - \bar{z}) - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}(\underset{\sim}{x}_{i\ell} - \bar{\underset{\sim}{x}})]'$$

$$= \frac{1}{n}\{V_{zz} - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}_{xx} \underset{\sim}{V}_{xz}\}. \tag{2.2.30}$$

Thus, the test statistic $Q(b)$ or equivalent $Q(z|\underset{\sim}{x})$ can be expressed in terms of the residuals as follows,

$$Q(z|\underset{\sim}{x}) = Q(\underset{\sim}{b}) = \underset{\sim}{b}'[\mathrm{Var}(\underset{\sim}{b}|H_o)]^{-1}\underset{\sim}{b}$$

$$= (n-1)\ \underset{\sim}{g}'\underset{\sim}{C}'\{V_{gg} \otimes \underset{\sim}{C}\ \underset{\sim}{D}_{\underset{\sim}{P}}^{-1}\ \underset{\sim}{C}'\}\underset{\sim}{C}\ \underset{\sim}{g} \tag{2.2.31}$$

where $\underset{\sim}{g}' = \{\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_s\}$ is a composite vector of subpopulation mean residuals. The statistic $Q(z|\underset{\sim}{x})$ can further be expressed as a summation across subpopulations,

$$= \frac{(n-1)}{n}\{\sum_{i=1}^{s} n_i \bar{g}'_i\ V^{-1}_{gg} \bar{g}_i\}$$

$$= \frac{(n-1)}{n}\{[\sum_{i=1}^{s} n_i \bar{g}_i \bar{g}'_i] V^{-1}_{gg}\}$$

$$= \frac{(n-1)}{n}\{S_{gg}\ V^{-1}_{gg}\} \tag{2.2.32}$$

where

$$S_{gg} = \sum_{i=1}^{s} n_i [(\bar{z}_i - \bar{z}) - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})][(\bar{z}_i - \bar{z}) - \underset{\sim}{V}'_{xz} \underset{\sim\sim}{V}^{-1}_{xx}(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})]' \tag{2.2.33}$$

denotes the among subpopulations sums of squares. If the residuals are based on a ranks regression model, then $Q(z|\underset{\sim}{x})$ is equivalent to Quade's rank analysis of covariance statistic reviewed in Chapter I. This statistic with moderate sample sizes $\{n_i > 20\}$ asymptotically has a chi-square distribution with degrees of freedom $(s - 1)$.

Since $z_{i\ell} = \underset{\sim}{a}'\underset{\sim}{y}_{i\ell}$ represents a general function of the multivariate response profile for the ℓ-th subject in the i-th subpopulation, then these univariate results for a randomization test statistic can be extended to the multivariate response profile case. In this situation,

$$Q(\underset{\sim}{y}|\underset{\sim}{x}) = \frac{(n-1)}{n}\{\sum_{i=1}^{s} n_i\bar{\underset{\sim}{g}}_i' \underset{\approx}{V}_{gg}^{-1} \bar{\underset{\sim}{g}}_i\}$$

$$= \frac{(n-1)}{n} \operatorname{tr}\{[\sum_{i=1}^{s} n_i\bar{\underset{\sim}{g}}_i\bar{\underset{\sim}{g}}_i']\underset{\approx}{V}_{gg}^{-1}\}$$

$$= \frac{(n-1)}{n} \operatorname{tr}\{\underset{\approx}{S}_{gg}\ \underset{\approx}{V}_{gg}^{-1}\} . \tag{2.2.34}$$

In other words, $Q(\underset{\sim}{y}|\underset{\sim}{x})$ can be interpreted as a one-way multivariate analysis of variance statistic based on the trace of the covariable regression residual matrix product $\{\underset{\approx}{S}_{gg}\ \underset{\approx}{V}_{gg}^{-1}\}$ and, in this sense, represents a one-way multivariate analysis of covariance test statistic. As stated previously, if the sample sizes $\{n_i\}$ are sufficiently large, then $Q(\underset{\sim}{y}|\underset{\sim}{x})$ approximately has a chi-square distribution with $d(s - 1)$ degrees of freedom.

The statistic $Q(\underset{\sim}{y}|\underset{\sim}{x})$ can be viewed as involving a more precise covariance structure than $Q(\underset{\sim}{y})$ given that $\underset{\sim}{x}$ is at random with respect to the subpopulations since for any vector of contrasts $\underset{\sim}{C}$.

$$\underset{\sim}{C}'\underset{\approx}{V}_{yy}\underset{\sim}{C} > \underset{\sim}{C}'\ \underset{\approx}{V}_{yy}\underset{\sim}{C} - \underset{\sim}{C}'\underset{\approx}{V}_{xy}'\ \underset{\approx}{V}_{xx}^{-1}\ \underset{\approx}{V}_{xy}\underset{\sim}{C} = \underset{\sim}{C}'\underset{\approx}{V}_{gg}\underset{\sim}{C}. \tag{2.2.35}$$

In this regard, the preciseness of $Q(\underset{\sim}{y}|\underset{\sim}{x})$ over $Q(\underset{\sim}{y})$ pertains to generally less subject heterogeneity in the sense of providing adjustments which induce equivalence of randomized differences among the subpopulations with respect to the covariables. Furthermore, since the assumed truth of H_o implies $\underset{\sim}{S}_{gg} \doteq \underset{\sim}{S}_{yy}$, it follows that $Q(\underset{\sim}{y}|\underset{\sim}{x})$ is typically larger than $Q(\underset{\sim}{y})$ for those situations when $H_{o,\underset{\sim}{y}|\underset{\sim}{x}}$ is not true. Alternatively, the tendency for the covariance statistic $Q(\underset{\sim}{y}|\underset{\sim}{x})$ to be larger or smaller than its unadjusted counterpart $Q(\underset{\sim}{y})$ in any specific analysis is partly a random event which depends upon whether relatively more or fewer subjects with "less favorable covariable status" are assigned to the more favorable subpopulation. In view of these considerations, the covariance adjusted statistic $Q(\underset{\sim}{y}|\underset{\sim}{x})$ can be regarded as a more effective statistic than the unadjusted statistic $Q(\underset{\sim}{y})$ in detecting true differences. This is clarified later through examples in Chapters III, IV, V and VI.

2.3 Partial Association

Let $h = 1,2,\ldots,q$ index a set of strata for some partition of a finite population. Regardless of whether stratification is undertaken prior to randomization (i.e. by design) or afterwards as an analytical strategy (i.e. post-stratification) the randomization hypothesis of no association is expressed simultaneously for each stratum in the form of a single hypothesis of no partial association. In this regard, partial association corresponds to a two-way analysis of variance.

An average partial association statistic in accordance with the generalized Mantel-Haenszel strategy consists of across-strata summary measures that are components of a quadratic form. For this purpose, within-stratum statistics presented in Sections 2.1 and 2.2 are now expressed with a subscript h indicating the h-th stratum. More specifically, let

$$\underset{\sim}{F}_{*i} = \sum_{h=1}^{q} n_{hi} \bar{\underset{\sim}{y}}_{hi} \tag{2.3.1}$$

denote the vector of across strata sums of $\{y_{hij}\}$, $j = 1,2,\ldots,d$ responses for the i-th subpopulation. The expected value and variance structure for $\underset{\sim}{F} = (\underset{\sim}{F}'_{*1},\ldots,\underset{\sim}{F}'_{*s})$ can be expressed as

$$E\{\underset{\sim}{F}|H_o\} = \sum_{h=1}^{q} \underset{\sim}{D}_{\underset{\sim}{n}_h} \bar{\underset{\sim}{y}}_h \tag{2.3.2}$$

$$\text{Var}\{\underset{\sim}{F}|H_o\} = \sum_{h=1}^{q} \frac{n_h^2}{(n_h-1)} \{\underset{\sim}{V}_h \otimes [\underset{\sim}{D}_{\underset{\sim}{P}_h} - \underset{\sim}{P}_h \underset{\sim}{P}'_h]\} \tag{2.3.3}$$

where $\underset{\sim}{n}_h = (n_{h1}, n_{h2}, \ldots, n_{hs})$, $n_h = \sum_{i=1}^{s} n_{hi}$ and $\bar{\underset{\sim}{y}}_h$ and $\underset{\sim}{V}_h$ are the h-th stratum counterparts of $\bar{\underset{\sim}{y}}$ and $\underset{\sim}{V}$ defined in (2.1.1) and (2.2.2). The average partial association (or generalized randomized block) chi-square test statistic has the form

$$Q_{GRB} = \underset{\sim}{G}' \underset{\sim}{V}_{\underset{\sim}{G}}^{-1} \underset{\sim}{G} \tag{2.3.4}$$

where

$$\underset{\sim}{G} = (\underset{\sim}{I}_d \otimes \underset{\sim}{C})(\underset{\sim}{F} - E(\underset{\sim}{F}|H_o)) \tag{2.3.5}$$

$$\underset{\sim}{V}_{\underset{\sim}{G}} = (\underset{\sim}{I}_d \otimes \underset{\sim}{C})[\text{Var}\{\underset{\sim}{F}|H_o\}] (\underset{\sim}{I}_d \otimes \underset{\sim}{C})' . \tag{2.3.6}$$

Under the randomization hypothesis, Q_{GRB} has an approximate chi-square distribution with d(s - 1) degrees of freedom provided that the sample sizes $\{n_{*i} = \sum_{h=1}^{q} n_{hi}\}$ are sufficiently large for $\underset{\sim}{G}$ to have an approximate multivariate normal distribution, the research design is consistent across strata and the inverse of the across strata variance exists in the sense of being non-singular.

The generalized randomized block statistic Q_{GRB} is directed toward average partial association alternatives in the sense that consistent subpopulation ordering of observed means across the $h = 1,2,\ldots,q$ strata strengthens its sensitivity while contradictory subpopulation ordering of observed means weakens its sensitivity. In view of this consistency tendency, the across strata statistic Q_{GRB} can be applied in a cumulative sense across successive strata which have been arranged according to some criterion of interest such as a decreasing likelihood for subpopulation differences. Such information concerning the extent to which successive strata tend to increase or decrease the p-value for subpopulation comparisons can provide useful insights concerning which strata strongly support the tendency for one subpopulation to exceed others and which strata tend to be contradictory.

The covariance adjustment framework described in Section 2.2 for the separate strata are used to formulate covariance adjusted generalized randomized block statistics. In particular, if the $\underset{\sim}{V}_{h,\underset{\sim}{xx}}$ are nonsingular for all $h = 1,2,\ldots,q$ strata, then the corresponding residuals $\underset{\sim}{g}_h$ can be defined with respect to the h-th stratum set of covariables. These covariables do not have to be the same across strata i.e., the covariables used in stratum h may be different from those used in stratum h'.

For the covariance adjustment framework, let

$$\underset{\sim}{g} = \sum_{h=1}^{q} [n_h \underset{\sim}{I}_d \otimes \underset{\sim}{D}_{\underset{\sim}{p}_h}] \underset{\sim}{g}_h \tag{2.3.7}$$

denote the across-strata sum of the within-stratum residuals $\underset{\sim}{g}_h$, i.e., covariate-adjusted responses, which has an expected value equal to zero and variance,

$$\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{g}|H_o\} = \sum_{h=1}^{q} \frac{n_h^2}{(n_h-1)}\{\underset{\sim}{V}_{h,\underset{\sim\sim}{gg}} \otimes [\underset{\sim}{D}_{\underset{\sim}{P}_h} - \underset{\sim}{P}_h\underset{\sim}{P}_h']\} . \qquad (2.3.8)$$

The (s - 1) subpopulation contrasts regarding these residuals can be expressed as

$$\underset{\sim}{b} = [\underset{\sim}{I}_d \otimes \underset{\sim}{C}]\underset{\sim}{g} . \qquad (2.3.9)$$

It follows that

$$Q_{GRB}(\underset{\sim}{y}|\underset{\sim}{x}) = \underset{\sim}{b}'\{[\underset{\sim}{I}_d \otimes \underset{\sim}{C}][\underset{\sim}{\mathrm{Var}}\{\underset{\sim}{g}|H_o][\underset{\sim}{I}_d \otimes \underset{\sim}{C}]'\}^{-1}\underset{\sim}{b}. \qquad (2.3.10)$$

Under randomization, this statistic asymptotically has a chi-square distribution with d(s - 1) degrees of freedom provided that the overall sample sizes $\{n_{*i} = \sum_{h=1}^{q} n_{hi}\}$ are sufficiently large.

For the case where some or many of the $\underset{\sim}{V}_{h,\underset{\sim\sim}{xx}}$ are singular or the case where there is a large number of strata each with small sample sizes, the covariance adjustment for a fixed set of covariables across strata can have the form

$$Q^*_{GRB}(\underset{\sim}{y}|\underset{\sim}{x}) = Q_{GRB}(\underset{\sim}{y},\underset{\sim}{x}) - Q_{GRB}(\underset{\sim}{x}) . \qquad (2.3.11)$$

$Q^*_{GRB}(\underset{\sim}{y}|\underset{\sim}{x})$ is an average partial association reduction statistic based on $Q_{GRB}(\underset{\sim}{y},\underset{\sim}{x})$ which is the joint multivariate chi-square statistic for responses and covariates with (d + t)(s - 1) degrees of freedom and $Q_{GRB}(\underset{\sim}{x})$ which is the chi-square statistic for the covariates with t(s - 1) degrees of freedom. Thus, $Q^*_{GRB}(\underset{\sim}{y}|\underset{\sim}{x})$ has a chi-square distribution with {(d + t)(s - 1) - t(s - 1)} or d(s - 1) degrees of freedom under the randomization hypothesis with large overall sample sizes, e.g. $\{\sum_{h=1}^{q} n_i > 20\}$.

Because $Q^*_{GRB}(\underset{\sim}{y}|\underset{\sim}{x})$ is not expressed as a quadratic form whose components are across-strata summary measures, $Q^*_{GRB}(\underset{\sim}{y}|\underset{\sim}{x})$ is not an average partial association statistic in the usual sense of pertaining to within stratum contrasts. Alternatively, $Q^*_{GRB}(\underset{\sim}{y}|\underset{\sim}{x})$ represents an across-strata covariate adjustment as a result of the difference between two average partial association statistics. In this regard, the underlying assumption is that the adjustment of the responses for the covariates can be standardized across strata.

2.3.1 The two subpopulation case

For comparisons between a particular pair of subpopulations, the generalized randomized block statistic can be expressed in terms of direct differences between the subpopulation means. For this purpose, let y_{hij} denote the value of a response variable for the $\ell = \{1,2,\ldots,n_{hi}\}$-th subject in the $i = 1,2$-th subpopulation for the $h = 1,2,\ldots,q$-th stratum. Let

$$\begin{aligned} d_h &= \sum_{\ell=1}^{n_{h1}} (y_{h1\ell}/n_{h1}) - \sum_{\ell=1}^{n_{h2}} (y_{h2\ell}/n_{h2}) \\ &= (\bar{y}_{h1} - \bar{y}_{h2}) \end{aligned} \tag{2.3.12}$$

denote the difference between the means of the response variable for the two subpopulations within the h-th stratum. Under the randomization hypothesis, the expected value of that difference is equal to zero and the variance can be expressed as

$$\begin{aligned} \operatorname{Var}\{d_h|H_o\} &= (n_h/n_{h1}n_{h2}) \sum_{i=1}^{s} \sum_{\ell=1}^{n_{hi}} (y_{hi\ell} - \bar{y}_h)^2/(n-1) \\ &= (n_h/n_{h1}n_{h2})V_{h,yy} \end{aligned} \tag{2.3.13}$$

where $n_h = (n_{h1} + n_{h2})$ is the number of subjects in the pooled subpopulations for the h-th stratum, $\bar{y}_h = (\sum_{i=1}^{2} \sum_{\ell=1}^{n_{hi}} y_{hi\ell}/n_h)$ is the finite population mean for all subjects in the h-th stratum and $V_{h,yy}$ is the finite population variance. The generalized randomized block statistic across the q strata has the form

$$Q_{GRB} = \{\sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h} d_h\}^2/\{\sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h} V_{h,yy}\}. \tag{2.3.14}$$

For large overall sample sizes (e.g., $\sum_{h=1}^{q} n_{hi} > 20$ per subpopulation) Q_{GRB} has an approximate chi-square distribution with 1 degree of freedom under the hypothesis of subpopulation equivalence.

Two other methods for combining the differences between two subpopulations across strata are the average standard normal method as described in Snedecor and Cochran (1976) and the combined p-value method as discussed by Fisher (1973) and Kendall and Buckland (1971). Let the standard normal statistic for the h-th stratum be expressed as

$$Z_h = \{\mathrm{sgn}(d_h)\}\, Q_h^{1/2} = d_h/\sqrt{\mathrm{Var}\{d_h\}} \tag{2.3.15}$$

where Z_h has an approximate standard normal distribution with expected value 0 and variance 1 for large sample sizes; e.g. $n_{hi} > 20$ per subpopulation, under the hypothesis of subpopulation equivalence. With the basic principle that the sum of normally distributed variables is also normally distributed, it follows that the average standard normal statistic across strata has the form

$$t = \{\sum_{h=1}^{q} Z_h/\sqrt{q}\} \tag{2.3.16}$$

where t has an approximate standard normal distribution for large overall sample sizes, e.g., $\sum_{h=1}^{q} n_{hi} > 20$ per subpopulation under the randomization hypothesis.

The Fisher combined p-value method is an approximate method for combining the individual stratum probabilities of the respective standard normal statistic, Z_h. More specifically,

$$W = \sum_{h=1}^{q} \{-2\ln(P_h)\} \tag{2.3.17}$$

where $P_h = \Phi(Z_h)$. Similarly, W^* can be defined with $P_h = 1 - \Phi(Z_h)$. For large stratum sample sizes, e.g., $\{n_{hi}\} > 20$ per subpopulation, P_h can be considered continuous with the uniform distribution on the (0,1) interval. In this case, W (W^*) has an approximate chi-square distribution with 2q degrees of freedom under the randomization hypothesis.

The expressions for covariance adjusted statistics in the two subpopulation case have the same form as the direct difference statistics except for the use of the covariate adjusted subpopulation means and variance,

$$\begin{aligned} g_h &= \{d_h - \frac{V_{h,xy}}{V_{h,xx}}(\bar{x}_{h1} - \bar{x}_{h2})\} \\ &= \{(\bar{y}_{h1} - \bar{y}_{h2}) - \frac{V_{h,xy}}{V_{h,xx}}(\bar{x}_{h1} - \bar{x}_{h2})\} \end{aligned} \tag{2.3.18}$$

$$\begin{aligned} \mathrm{Var}\{g_h|H_o\} &= \{\mathrm{Var}\{d_h|H_o\} - \frac{n_h}{n_{h1}n_{h2}} \frac{V^2_{h,xy}}{V_{h,xx}}\} \\ &= \frac{n_h}{n_{h1}n_{h2}}\{V_{h,xy} - \frac{V^2_{h,xy}}{V_{h,xx}}\} \end{aligned} \tag{2.3.19}$$

where $\bar{x}_{hi}$ is the mean of a single covariate for the i-th subpopulation and h-th stratum; and $V_{h,yy}$, $V_{h,xy}$ and $V_{h,xx}$ are the partitions of the finite population covariance matrix

$$V_h = \begin{bmatrix} V_{h,yy} & V'_{h,xy} \\ V_{h,xy} & V_{h,xx} \end{bmatrix}$$

$$= \frac{1}{(n_h-1)} \sum_{i=1}^{2} \sum_{\ell=1}^{n_{hi}} \begin{bmatrix} (y_{hi\ell} - \bar{y}_h)^2 & (y_{hi\ell} - \bar{y}_h)(x_{hi\ell} - \bar{x}_h) \\ \cdot & (x_{hi\ell} - \bar{x}_h)^2 \end{bmatrix} \qquad (2.3.20)$$

where $\bar{y}_h$ and $\bar{x}_h$ are the h-th stratum finite population means for the response and covariate respectively. Thus, the generalized randomized block statistic has the form,

$$Q_{GRB}(y|x) = \{ \sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h} g_h \}^2 / \{ \sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h} [V_{h,yy} - \frac{V_{h,xy}^2}{V_{h,xx}}] \}.$$

$$(2.3.21)$$

This statistic has a chi-square distribution with 1 degree of freedom.

Also, $t(y|x) = \sum_{h=1}^{q} \{g_h/\sqrt{\text{Var}\{g_h|H_o\}}\}/\sqrt{q} = \sum_{h=1}^{q} Z_h(y|x)/\sqrt{q}$ (2.3.22)

has a standard normal distribution; and

$$W(y|x) = \sum_{h=1}^{q} \{-2\ln(\Phi(Z_h(y|x)))\} \qquad (2.3.23)$$

has a chi-square distribution with 2q degrees of freedom.

2.4 Missing Data Adjustment

For multivariate response profiles with missing data, one pertinent hypothesis focuses on the randomized nature of the data which are present with respect to their distribution across subpopulations given the assumption that missing data occur at random for all variables under study (i.e., occur independently of the hypothetical values which such data might have had). The assumption that missing data is a random phenomena implies that all subpopulations have equivalent prevalence for it. For this purpose, let $(\underset{\sim}{y}^*_{hi\ell}, \underset{\sim}{x}_{hi\ell})'$ denote the extended data vector for the $\ell = \{1,2,\ldots,n_{hi}\}$-th subject in the $i = \{1,2,\ldots,s\}$-th subpopulation of the $h = \{1,2,\ldots,q\}$-th stratum, where $\underset{\sim}{y}^*_{hi\ell}$ is the set of $j = \{1,2,\ldots,d\}$ responses with missing data being assigned an arbitrary constant such as zero and $\underset{\sim}{x}_{hi\ell}$ is the set of $j = 1,2,\ldots,d$ binary indicators defined to have the value 1 if the response is observed and the value 0 if the response is missing. More specifically, the data vector for the ℓ-th subject in the i-th subpopulation of the h-th stratum has the form

$$(y^*_{hi1\ell},\ldots,y^*_{hid\ell},x_{hi1\ell},\ldots,x_{hid\ell}) \tag{2.4.1}$$

where

$$y^*_{hij\ell} = \begin{cases} y_{hij\ell} & \text{if j-th response is observed} \\ 0 & \text{if missing} \end{cases}$$

$$x_{hij\ell} = \begin{cases} 1 & \text{if j-th response is observed} \\ 0 & \text{if missing} \end{cases}$$

For this framework, the covariance adjusted statistic $Q(\underset{\sim}{y}^*|\underset{\sim}{x})$ or $Q_{GRB}(\underset{\sim}{y}^*|\underset{\sim}{x})$ in Sections 2.2 and 2.3 is a multivariate randomization test statistic with respect to the observed $\{\underset{\sim}{y}_\ell\}$ given adjustment for the presumed random pattern of missing data. Accordingly, it has an approximate chi-square distribution with $d(s - 1)$ degrees of freedom provided that the sample sizes $\{n_{hi}\}$ are sufficiently large. In practice, $\{\underset{\sim}{x}_h\}$ is the vector of non-redundant binary indicators for the h-th stratum such that the j-th indicator is included in the vector only if missing values occur in that stratum for the j-th response. In Section 2.3 it was shown that the average partial association statistic with respect to stratum residuals does not require the same number of indicators per stratum.

2.4.1 Invariance to an assigned constant

Let $\underset{\sim}{y}^*_\ell(0)$ denote the completed data array in which missing data are assigned the value zero. Let ξ_j denote an arbitrary fixed value such that the linear transformation

$$y^*_{hij\ell}(\xi_j) = y^*_{hij\ell}(0) + \xi_j - \xi_j x_{hij\ell} \tag{2.4.2}$$

has values

$$y^*_{hij_\ell}(\xi_j) = \begin{cases} y_{hij\ell} & \text{if the j-th response variable is observed} \\ \xi_j & \text{if missing} \end{cases}$$

and as before

$$x_{hij\ell} = \begin{cases} 1 & \text{if j-th response variable is observed} \\ 0 & \text{if missing .} \end{cases}$$

It follows then that the vector of subpopulation means $\underset{\sim}{F}(\xi)$ is a linear transformation of $\underset{\sim}{F}(0)$.

$$\begin{aligned} \underset{\sim}{F}(\xi) &= \left\{ \begin{bmatrix} \underset{\sim}{I}_d & -\underset{\sim}{D}_{\underset{\sim}{\xi}} \\ \underset{\sim}{0}_{(dxd)} & \underset{\sim}{I}_d \end{bmatrix} \otimes \underset{\sim}{I}_s \right\} \underset{\sim}{F}(0) + \begin{bmatrix} \underset{\sim}{\xi}_d \\ \underset{\sim}{0} \end{bmatrix} \otimes \underset{\sim}{1}_s \\ &= \{\underset{\sim}{H} \otimes \underset{\sim}{I}_s\}\underset{\sim}{F}(0) + \begin{bmatrix} \underset{\sim}{\xi}_d \\ \underset{\sim}{0}_d \end{bmatrix} \otimes \underset{\sim}{1} \end{aligned} \tag{2.4.3}$$

which has values

$$\underset{\sim}{F}(\xi) = \begin{bmatrix} \bar{\underset{\sim}{y}}^*_1(0) + \underset{\sim}{\xi} - \underset{\sim}{D}_{\underset{\sim}{\xi}}\bar{\underset{\sim}{x}}_1 \\ \bar{\underset{\sim}{x}}_1 \\ \cdots \\ \bar{\underset{\sim}{y}}^*_s(0) + \underset{\sim}{\xi} - \underset{\sim}{D}_{\underset{\sim}{\xi}}\bar{\underset{\sim}{x}}_s \\ \bar{\underset{\sim}{x}}_s \end{bmatrix} . \tag{2.4.4}$$

For (s - 1) subpopulation contrasts of d responses and d indicators

$$
\begin{aligned}
\underset{\sim}{G}(\xi) &= [\underset{\sim}{I}_{(2d)} \otimes \underset{\sim}{C}]\underset{\sim}{F}(\xi) \\
&= [\underset{\sim}{I}_{(2d)} \otimes \underset{\sim}{C}]\,[\underset{\sim}{H} \otimes \underset{\sim}{I}_s]\underset{\sim}{F}(0) \\
&= [\underset{\sim}{H} \otimes \underset{\sim}{C}]\underset{\sim}{F}(0) \\
&= [\underset{\sim}{H} \otimes \underset{\sim}{I}_{(s-1)}]\underset{\sim}{G}(0) \\
&= \underset{\sim}{L}\ \underset{\sim}{G}(0) \qquad (2.4.5)
\end{aligned}
$$

It then follows that the randomization statistic

$$
\begin{aligned}
Q(\underset{\sim}{y}^*|\underset{\sim}{x}:\xi) &= \underset{\sim}{G}(0)'\underset{\sim}{L}'\{\underset{\sim}{L}\ \underset{\sim}{V}_{\underset{\sim}{G}(0)}\underset{\sim}{L}'\}^{-1}\ \underset{\sim}{L}\ \underset{\sim}{G}(0) \\
&= \underset{\sim}{G}'(0)\ \underset{\sim}{V}^{-1}_{\underset{\sim}{G}(0)}\ \underset{\sim}{G}(0) \\
&= Q(\underset{\sim}{y}^*|\underset{\sim}{x}:0) \qquad (2.4.6)
\end{aligned}
$$

does not depend on the value (ξ_j) of the constants assigned to the missing data. Since $Q(\underset{\sim}{x})$ does not involve the y's at all, it does not depend on the $\{\xi_j\}$; and so neither does the difference $Q(\underset{\sim}{y}^*|\underset{\sim}{x})$ between $Q(\underset{\sim}{y}^*,\underset{\sim}{x})$ and $Q(\underset{\sim}{x})$ as in (2.2.4). Moreover, the covariance residual vectors $\underset{\sim}{g}_{hi\ell}$ can be expressed as

$$
\begin{aligned}
\underset{\sim}{g}_{hi\ell}(\xi) &= (\underset{\sim}{y}^*_{hi\ell}(0) + \underset{\sim}{\xi} - \underset{\sim}{D}_{\xi}\underset{\sim}{x}_{hi\ell}) - (\bar{\underset{\sim}{y}}^*_h(0) + \underset{\sim}{\xi} - \underset{\sim}{D}_{\xi}\bar{\underset{\sim}{x}}_h) \\
&\quad - (\underset{\sim}{V}'_{h,\underset{\sim\sim}{xy}*(0)} - \underset{\sim}{D}_{\xi}\underset{\sim}{V}_{h,\underset{\sim\sim}{xx}})\underset{\sim}{V}^{-1}_{h,\underset{\sim\sim}{xx}}\ (\underset{\sim}{x}_{hi\ell} - \bar{\underset{\sim}{x}}_h) \\
&= (\underset{\sim}{y}^*_{hi\ell}(0) - \bar{\underset{\sim}{y}}^*_h(0)) - \underset{\sim}{V}_{h,\underset{\sim\sim}{xy}*(0)}\ \underset{\sim}{V}^{-1}_{h,\underset{\sim\sim}{xx}}\ (\underset{\sim}{x}_{hij} - \bar{\underset{\sim}{x}}_h) \\
&\quad - \underset{\sim}{D}_{\xi}(\underset{\sim}{x}_{hi\ell} - \bar{\underset{\sim}{x}}_h) + \underset{\sim}{D}_{\xi}\underset{\sim}{V}_{h,\underset{\sim\sim}{xx}}\ \underset{\sim}{V}^{-1}_{h,\underset{\sim\sim}{xx}}\ (\underset{\sim}{x}_{hi\ell} - \bar{\underset{\sim}{x}}_h) \\
&= \underset{\sim}{g}_{hi\ell}(0) \qquad (2.4.7)
\end{aligned}
$$

and hence do not depend on ξ.

These results imply that missing values could be assigned the value of the stratum mean of the observed values. Thus, the expected value of the stratum is the mean of the observed data. Furthermore, the elements of the matrix $\underset{\sim}{V}_{h,\underset{\sim}{y}\underset{\sim}{y}}$ have the form

$$v_{h,yyjj} = \frac{1}{n_h} \sum_{i=1}^{s} \sum_{obs(j)} (y_{hij\ell} - \bar{y}_{hj})^2$$

and

$$v_{h,yyjj'} = \frac{1}{n_h} \sum_{i=1}^{s} \sum_{obs(jj')} (y_{hij\ell} - \bar{y}_{hj})(y_{hij'\ell} - \bar{y}_{hj'}) \tag{2.4.8}$$

where obs(j) denotes summation over all subjects for which the j-th response is observed and obs(jj') denotes summation over all subjects for whom both the j-th and j'-th responses are observed. Thus, except for the n_h in the divisor, elements of $\underset{\sim}{v}_{h,\underset{\sim}{y}\underset{\sim}{y}}$ represent the overall population variances and covariances with respect to the response variable data which are actually observed. However, the divisor n_h can be viewed as being consistent with the finite population randomization framework wherein n_h subjects were randomly assigned to s subpopulations. Thus, although the test statistic $Q(\underset{\sim}{y}^*|\underset{\sim}{x})$ in Section 2.2 or 2.3 appear to be based upon the assignment of some arbitrary value to the missing data from a computational point of view, it only specifically depends upon the observed data for the responses together with the complementary set of binary indicator variables for the pattern of missing data.

2.5 Scores

Randomization statistics discussed in the previous sections are formulated as contrasts among subpopulation means. Because these statistics can be applied to any scaling of the data, they represent a general class of mean score randomization statistics of which specific scoring applications yield equivalent results to several categorical or nonparametric tests.

Statistical applications in this dissertation involved actual values, integer, binary, rank, modified ridit, and logrank scores. An example of the actual values would be proportion of hours with little or no pain over some fixed period (see Chapter IV). For categorical data involving an ordinal response, actual values in the usual sense, are not available. Alternatively integer scores are useful with categorical data because these scores correspond to assigning consecutive integers to arranged values or sets of values. In this regard, the scored values or sets of values are sometimes assumed to be equally spaced from one integer to the next, see Koch, Gillings, Stokes (1980) or Koch, et al. (1977); for others, they can be interpreted as a sum of the proportions exceeding successive categories, see Section 2.5.1.2. Binary scores can be viewed as two-level integer scores. They are particularly useful in focusing statistical power on the prevalence of a substantively meaningful set of levels within a profile of possible values. Also, these scores can be applied in principle, to either categorical or continuous data. Rank, modified ridit and logrank scores represent general rank-type scores in the sense that they are associated with ranking the data where ties are assigned mid-rank values for either categorical or continuous data.

2.5.1 Across-strata mean scores

In accordance with Section 2.3.1, the across-strata mean score difference for the univariate response, two subpopulation case can be expressed as

$$S = \sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h}(\bar{y}_{h1} - \bar{y}_{h2}) \tag{2.5.1}$$

and the generalized randomized block statistic can be expressed as

$$Q_{GRB} = S^2/\mathrm{Var}(S) \tag{2.5.2}$$

The S formulation defines the across strata summation of the mean score difference to be weighted by a factor $(n_{h1}n_{h2}/n_h)$. Such weight can be viewed as the number of paired comparisons (in a Mann-Whitney sense) between subjects in the first subpopulation and those in the second subpopulation per subject under study in the h-th stratum. These weights have a maximum value $(n_h/4)$ for the case $n_{h1} = n_{h2} = (n_h/2)$ and a minimum value $(n_h - 1)/n_h$ for the cases $\{n_{h1}, n_{h2}\} = \{(n_h - 1), 1\}$ or $\{1, (n_h - 1)\}$. If all strata have balanced subpopulation sample sizes, then the mean score differences are weighted across strata by their respective stratum sample sizes. More specifically,

$$S = 1/4 \sum_{h=1}^{q} n_h(\bar{y}_{h1} - \bar{y}_{h2})$$

where $n_{h1} = n_{h2} = n_h/2$.

Another general expression for S can be derived as follows:

$$\begin{aligned} S &= \sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h}(\bar{y}_{h1} - \bar{y}_{h2}) \\ &= \sum_{h=1}^{q} \frac{n_{h1}n_{h2}}{n_h}\left(\frac{n_{h2}y_{h1}}{n_{h2}} - \frac{(n_h\bar{y}_h - n_{h1}\bar{y}_{h1})}{n_{h2}}\right) \\ &= \sum_{h=1}^{q} n_{h1}(\bar{y}_{h1} - \bar{y}_h) \end{aligned} \tag{2.5.4}$$

where $\bar{y}_h$ is the overall stratum mean score for $n_h = (n_{h1} + n_{h2})$ subjects. Thus, S can be interpreted as an unweighted across strata sum of the difference between the score total and its expected value for one of the two subpopulations. This unweighted across stratum summation has the same value as its previous formulation in terms of a weighted across-stratum summation of the difference between subpopulation mean scores.

2.5.1.1 Binary scores

For two subpopulations and two response categories for each of $h = 1,2,\ldots,q$ strata, S can be expressed as

$$S = \sum_{h=1}^{q} n_{h1}(\bar{y}_{h1} - \bar{y}_h)$$

$$= \sum_{h=1}^{q} (n_{h11} - m_{h11}) \qquad (2.5.5)$$

for q sets of 2 x 2 tables, where rows correspond to subpopulations, columns correspond to response categories and different tables correspond to specific strata. Here, $m_{h11}(n_{h11})$ is the expected (observed) number of subjects with the first response in the first subpopulation and can be calculated as $(n_{h11}N_{h+1}/N_{h+})$ where N_{h+1} is the number of subjects in the h-th stratum with the first response and N_{h+} is the total number of subjects in the h-th stratum. The generalized randomized block statistic is directed toward an unweighted sum across strata of {observed—expected value} quantities yielding the Mantel and Haenszel (1959) statistic

$$Q_G = \left\{ \sum_{h=1}^{q} (n_{h11} - \frac{N_{h1+}N_{h+1}}{N_{h+}}) \right\}^2 \Bigg/ \left\{ \sum_{h=1}^{q} \frac{N_{h1+}N_{h+1}N_{h2+}N_{h+2}}{N_{h+}^2(N_{h+} - 1)} \right\} . \qquad (2.5.6)$$

2.5.1.2 Integer scores

For categorical data with $j = 1,2,\ldots,r$ categories scored with coinciding integer values, S can be expressed as

$$S = \sum_{h=1}^{q} \sum_{j=1}^{r} j(n_{h1j} - m_{h1j}) \tag{2.5.7}$$

where $m_{h1j}(n_{h1j})$ is the expected (observed) number of subjects with the j-th response in the first subpopulation of the h-th stratum and can be calculated as $n_{h1j}N_{h+j}/N_{h+}$ where N_{h+j} is the number of subjects with the j-th response in the h-th stratum and N_{h+} is the number of subjects in the h-th stratum. It then follows that

$$S = \sum_{h=1}^{q} \sum_{j=1}^{r} \sum_{r=1}^{j} (n_{h1j} - m_{h1j})$$

$$= \sum_{h=1}^{q} \sum_{k=1}^{r} \sum_{j=k}^{r} (n_{h1j} - m_{h1j})$$

$$= \sum_{h=1}^{q} \sum_{k=1}^{r} (N_{h1k} - M_{h1k}) \tag{2.5.8}$$

where $N_{h1k} = \sum_{j=k}^{r} n_{h1j}$ and $M_{h1k} = \sum_{j=k}^{r} m_{h1j}$ are the observed and expected numbers of subjects in the first subpopulation of the h-th stratum who have a response status which is at least as large as the k-th outcome. Thus, for integer scores in the categorical setting, the generalized randomized block statistic is directed toward the unweighted sum across strata and response categories of the cumulative number of subjects with response at least as large as the respective possible outcomes.

2.5.1.3 Rank scores

If the observed responses are ranked across all subjects within a stratum, then the mean score expected value of any subpopulation is the average stratum rank, $(n_h + 1)/2$, such that

$$S = \sum_{h=1}^{q} n_{h1}(\bar{y}_{h1} - \bar{y}_h)$$
$$= \sum_{h=1}^{q} \left(U_{h1} - \frac{n_{h1}(n_h + 1)}{2}\right) \tag{2.5.9}$$

where U_{h1} is the Wilcoxon rank sum statistic with respect to the first subpopulation for the h-th stratum. Thus, for this case, the generalized randomized block statistic is directed toward the unweighted across-strata sum of the Wilcoxon statistics for the respective strata. For categorical data, rank scores yield a contingency table counterpart of the Bernard-van Elteren statistic.

2.5.1.4 Modified ridit scores

Modified ridit scores are within stratum ranks divided by the corresponding $\{n_h + 1\}$. It follows from Section 2.5.1.3 that S has the form

$$S = \sum_{h=1}^{q} \frac{1}{(n_h + 1)} \left\{U_{h1} - \frac{n_{h1}(n_h + 1)}{2}\right\} . \tag{2.5.10}$$

For modified ridit scores, the generalized randomized block statistic is directed toward the across strata sum of the Wilcoxon statistics weighted by the reciprocal of the augmented stratum sample sizes, $1/(n_h + 1)$. For categorical data, modified ridit scores yield a contingency table counterpart of the van Elteren (1960) combined rank statistic emphasized by Lehmann (1975, Chapter III, pp. 132-141).

2.5.1.5 Logrank scores

Logrank scores are a standardized set of values for right skewed data such as survival data from Weibull populations with common shape parameter; see Koch, Sen and Amara (1982). For data which do not involve any censoring and can be ranked without ties, within stratum scores have the form

$$a_{j,N} = 1 - \sum_{k=1}^{j} (N - K + 1)^{-1}$$

$$= \sum_{k=1}^{j} [n_{+k} / \sum_{\ell=k}^{r} n_{+\ell}]$$

$$= 1 - E\{T_j : N\} \qquad (2.5.11)$$

where $j = 1,2,\ldots,N$ indexes the ordering of subjects from smallest to largest and $T_{j:N}$ denotes the j-th order statistic from the unit exponential distribution. For data with censoring, the scores have the form

$$a_{j0,N} = -\sum_{k=1}^{j} (n_{+k1} / \sum_{\ell=k}^{r} n_{+\ell}) \quad \text{where } j = 1,2,\ldots,(r-1)$$

$$a_{j1,N} = (1 + a_{j0,N}) \qquad \text{where } j = 1,2,\ldots,(r-1)$$

$$a_{r1,N},\ a_{r0,N} = a_{(r-1)0,N} \qquad (2.5.12)$$

where $\{a_{j0,N}\}$ are scores for censored observations and $\{a_{j1,N}\}$ are scores for non-censored observations. For these scores, S has a log-rank counterpart to that given for uniform scores in the sense of being an unweighted summation across strata of scores standardized for respective stratum sample sizes; i.e.,

$$S = \sum_{h=1}^{q} n_{h1}(\bar{y}_{h1} - \bar{y}_h)$$

$$= \sum_{h=1}^{q} \sum_{j=1}^{r} \sum_{j'=0}^{1} a_{jj',N}\, n_{1jj'}$$

$$= \sum_{h=1}^{q} \sum_{j=1}^{(r-1)} (n_{1j1} - m_{1j1})$$

$$= \sum_{h=1}^{q} F_{h1}. \qquad (2.5.13)$$

For survival studies, F_{h1} can be interpreted as the sum of the differences between the observed numbers (n_{1j1}) of deaths or failures in successive time intervals for the first subpopulation and their conditional expected values (m_{1j1}) relative to the corresponding numbers of subjects at risk.

2.6 Weighted Least Squares Methodology

This section presents aspects of the Grizzle, Starmer and Koch (1969) method of weighted least squares to the analysis of categorical data. The conceptual sampling framework presumes that the subjects under study are representative of some broad population beyond themselves. More specifically, groups of subjects forming subpopulations correspond to a stratified simple random sampling of some superpopulation partitioned according to the subpopulation structure. For example, subpopulations might be defined as a cross-classification of age intervals and treatments; e.g., subjects within a single subpopulation of subjects that are within the pertinent age interval and that might have received the pertinent treatment. Thus, both age interval for $h = 1,2,\ldots,q$ and treatment for $i = 1,2,\ldots,s$ define $(q)(s)$ independent subpopulations.

For the cross-classified subpopulation framework, linear models can be constructed to describe the variation among mean score vectors representing the response distribution of $\{\ell = 1,2,\ldots,n_{hi}\}$ subjects. The overall variance structure is a subpopulation block diagonal matrix. If $\underset{\sim}{F}$ denotes the composite vector of subpopulation mean score vectors $\{\bar{\underset{\sim}{a}}_{hi}\}$, then the variance $\underset{\sim}{V}_{\underset{\sim}{F}}$ has blocks

$$\underset{\sim}{V}_{a,hi} = \underset{\sim}{A}[\sum_{\ell=1}^{n_{hi}} (\underset{\sim}{y}_{hi\ell} - \bar{\underset{\sim}{y}}_{hi})(\underset{\sim}{y}_{hi\ell} - \bar{\underset{\sim}{y}}_{hi})'] \underset{\sim}{A}'/n_{hi} . \qquad (2.6.1)$$

More generally, $\underset{\sim}{F}$ can be a vector of functions of the mean scores $\{\bar{\underset{\sim}{a}}_{hi}\}$ with respect to exponential, linear, and/or logarithmic operations. A consistent estimate of its covariance structure can be derived by linear Taylor series methods.

Variation among the elements of $\underset{\sim}{F}$ can be described by a linear regression model

$$\underset{\sim}{E}_A\{\underset{\sim}{F}(\underset{\sim}{a})\} = \underset{\sim}{K}\,\underset{\sim}{B} \qquad (2.6.2)$$

where $\underset{\sim}{K}$ is a pre-specified full rank design matrix of known coefficients, $\underset{\sim}{B}$ is a vector of unknown parameters to be estimated and $\underset{\sim}{E}_A$ denotes asymptotic expectation. Let

$$\underset{\sim}{b} = (\underset{\sim}{K}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{K})^{-1}\,\underset{\sim}{K}'\,\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\,\underset{\sim}{F}$$

and

$$\underset{\sim}{V}_{\underset{\sim}{b}} = (\underset{\sim}{K}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{K})^{-1} \qquad (2.6.3)$$

denote the weighted least squares parameter estimates and their covariance matrix. An appropriate statistic for model goodness of fit is the Wald statistic

$$Q_W = (\underset{\sim}{F} - \underset{\sim}{K}\,\underset{\sim}{b})'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\,(\underset{\sim}{F} - \underset{\sim}{K}\,\underset{\sim}{b}) \qquad (2.6.4)$$

which has an approximate chi-square distribution with degrees of freedom equal to the difference between dimensions of $\underset{\sim}{F}$ and $\underset{\sim}{b}$ for moderately large samples (i.e., $n_{hi} > 20$).

If the model adequately characterizes the variation in $\underset{\sim}{F}$, tests of linear hypotheses pertaining to the parameters of the form

$$H_C : \underset{\sim}{C}\,\underset{\sim}{B} = \underset{\sim}{0}$$

can be undertaken using the Wald statistic

$$Q_{W,C} = \underset{\sim}{b}'\underset{\sim}{C}'(\underset{\sim}{C}\ \underset{\sim}{V}_{\underset{\sim}{b}}\ \underset{\sim}{C}')^{-1}\ \underset{\sim}{C}\ \underset{\sim}{b}$$

which has an approximate chi-square distribution with degrees of freedom equal to the rank of the contrast matrix, $\underset{\sim}{C}$.

Predicted values for $\underset{\sim}{F}$ based on the model (2.6.2) can be calculated from

$$\hat{\underset{\sim}{F}} = \underset{\sim}{K}\ \underset{\sim}{b} = \underset{\sim}{K}\ \underset{\sim}{V}_{\underset{\sim}{b}}\ \underset{\sim}{K}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} \tag{2.6.7}$$

with variance

$$\underset{\sim}{V}_{\hat{\underset{\sim}{F}}} = \underset{\sim}{K}\ \underset{\sim}{V}_{\underset{\sim}{b}}\ \underset{\sim}{K}' \ . \tag{2.6.8}$$

These predicted values are useful in showing the composite effect of the parameter set for each subpopulation mean score vector. Because the component parameters for any subpopulation set of predicted values were estimated for the combined data in all subpopulations, the resulting predicted values can be viewed as smoothed estimates of any subpopulation set of mean scores in contrast to the original mean scores, $\underset{\sim}{F}$, which were formulated from data in the single subpopulation only. Furthermore, covariate-adjusted predicted values of response mean scores can be obtained by applying a design matrix $\underset{\sim}{K}$ having a common parameter or set of parameters for covariable mean scores across the pertinent set of subpopulations. This type of analysis is discussed further in Chapter III where a specific application is given.

The weighted least squares analyses presented in this dissertation were undertaken through the use of the program CRISCAT documented as Stanish et al. (1978). The particular applications in Chapters III, IV and VI involve the special capabilities of raw data input, missing value ratio estimates as described in Chapter I, and split-log and split-exponentiation transformations. Alternatively, for less complicated applications, the program GENCAT documented in Landis et al. (1976), provides weighted least squares statistics for contingency table input.

2.7 Logistic Regression

This section summarizes some technical aspects of logistic regression for binary scores observed across (q)(s) distinct subpopulations. This presentation is based upon the discussion in Koch and Edwards (1982).

The data to which binary scores have been applied, are presumed to be conceptually representative of data from some large population in a stratified simple random sampling sense where hi = (11,12,...,qs) indexes the subpopulations. The binary scores for these data can be summarized as frequencies $\{\underset{\sim}{f}_{hi} = (f_{hi1}, f_{hi2})'\}$ which have the product binomial distribution

$$\phi(\underset{\sim}{f}|\theta) = \prod_{h=1}^{q} \prod_{i=1}^{s} n_{hi}!\, \theta_{hi}^{f_{hi1}} (1 - \theta_{hi})^{f_{hi2}} / f_{hi1}!\, f_{hi2}! \qquad (2.7.1)$$

where the $\{\theta_{hi}\}$ denote the respective probabilities of the first category for a randomly selected subject in the hi-th subpopulation. The variation of the $\{\theta_{hi}\}$ across the (q)(s) subpopulations is described by a linear logistic model

$$\theta_{hi} = \{1 + \exp(-B_0 - \underset{\sim}{x}'_{hi}\, \underset{\sim}{B}_1)\}^{-1} \qquad (2.7.2)$$

where B_0 is an intercept parameter and $\underset{\sim}{B}_1$ is a vector of slope parameters relative to a matrix of subpopulation covariates $\underset{\sim}{x}_{hi}$. With the logit transformation of θ_{hi}, the linear model can be expressed as

$$\text{logit}(\theta_{hi}) = \ln(\theta_{hi}/[1 - \theta_{hi}])$$

$$= B_0 + \underset{\sim}{x}'_{hi}\, \underset{\sim}{B}_1 \qquad (2.7.3)$$

If most of the frequencies $\{\underset{\sim}{f}_{hi}\}$ are sufficiently large (i.e., > 20 and none < 5) to have approximately normal distributions, then either the weighted least squares (WLS) methods described in Section 2.6 or maximum likelihood (ML) methods described further in this section could be equivalently used to obtain estimates for the parameters. For smaller frequencies, maximum likelihood methods are preferable. Once the ML or WLS estimates and their corresponding covariance structure have been obtained and goodness of fit has been established for the model (2.7.3), then procedures for the testing hypothesis concerning parameters and formulating predicted values can be undertaken with matrix procedures like those described in Section 2.6. The predicted values calculated from the model (2.7.3) have values in the (0,1) interval for all $\underset{\sim}{x}$.

2.7.1 Maximum likelihood estimation

Maximum likelihood estimates for the linear logistic model (2.7.3) can be expressed as the solution of the nonlinear equations

$$\sum_{h=1}^{q} \sum_{i=1}^{s} (n_{hi1} - n_{hi}\,\hat{\theta}_{hi})\, \underset{\sim}{x}_{hiA} = \underset{\sim}{0} \qquad (2.7.4)$$

where

$$\hat{\theta}_{hi} = \hat{\theta}(\underset{\sim}{x}_{hi}, \hat{\underset{\sim}{B}}_A) = \{1 + \exp(-\hat{B}_0 - \underset{\sim}{x}'\hat{\underset{\sim}{B}}_1)\}^{-1} \qquad (2.7.5)$$

are the model predicted ML estimates of the $\{\theta_{hi}\}$ based on the ML estimates $\hat{\underset{\sim}{B}}_A$ of $\underset{\sim}{B}_A$ and $\underset{\sim}{x}_{hiA} = [1, \underset{\sim}{x}'_{hi}]'$. The computation of $\hat{\underset{\sim}{B}}_A$ generally requires an iterative procedure such as the Newton-Raphson method (or iterative proportional fitting). The procedure is terminated after a convergence criterion is reached for the estimates in the $(\ell + 1)$-th step

$$\hat{\underset{\sim}{B}}_{A,\ell+1} = \hat{\underset{\sim}{B}}_{A\ell} + [\underset{\sim}{V}(\hat{\underset{\sim}{B}}_{A\ell})] [\sum_{h=1}^{q} \sum_{i=1}^{s} \{n_{hi1} - n_{hi}\, \hat{\theta}(\underset{\sim}{x}_{hi}, \hat{\underset{\sim}{B}}_{A\ell})\} \underset{\sim}{x}_{hiA}] \quad (2.7.6)$$

where

$$\underset{\sim}{V}(\hat{\underset{\sim}{B}}_{A\ell}) = \{\sum_{h=1}^{q} \sum_{i=1}^{s} n_{hi}\, \hat{\theta}(\underset{\sim}{x}_{hi}, \hat{\underset{\sim}{B}}_{A\ell}) [1 - \hat{\theta}(\underset{\sim}{x}_{hi}, \hat{\underset{\sim}{B}}_{A\ell})] \underset{\sim}{x}_{hiA} \underset{\sim}{x}'_{hiA}\}^{-1}$$

is the ℓ-th step estimate for the asymptotic covariance matrix for $\hat{\underset{\sim}{B}}_{A}$.

The goodness of fit of the model (2.7.3) can be assessed by using the log-likelihood ratio chi-square statistic

$$Q_L = \sum_{h=1}^{q} \sum_{i=1}^{s} \sum_{j=1}^{2} 2f_{hij} [\ln(f_{hij}/\hat{f}_{hij})],$$

where $\hat{f}_{hi1} = n_{hi}\, \hat{\theta}_{hi}$ and $\hat{f}_{hi2} = n_{hi}(1 - \hat{\theta}_{hi})$ are the ML estimates for the expected values of the $\underset{\sim}{f}_{hi}$ under (2.7.2). An equivalent statistic is the Pearson chi-square statistic

$$Q_P = \sum_{h=1}^{q} \sum_{i=1}^{s} \sum_{j=1}^{2} (f_{hij} - \hat{f}_{hij})^2/\hat{f}_{hij}. \quad (2.7.7)$$

The maximum likelihood analyses presented in Chapter III were made possible through the use of a SAS macro CATMAX which does iterative proportional fitting in the contingency table framework. A SAS procedure LOGIST documented in Harrell (1980) has a raw-data input capability.

CHAPTER III

GENERAL RANDOMIZATION APPLICATIONS

The two examples given in this chapter demonstrate the use of mean score randomization test statistics. The first example involves the Wilcoxon Rank Sum statistic regarding univariate measures derived from a complex repeated measurements study in cardiology. Here, the hypothesis of randomization is applied to a small number of patients in two diagnositc subpopulations for presence or absence of atrial abnormalities. Test p-values are obtained from exact tables for the Wilcoxon Rank Sum statistic as well as from the chi-square approximation for the generalized mean score randomization statistic in Section 1 of Chapter II. A covariate-adjusted result for the mean score randomization statistic is also shown and can be interpreted as a modified covariate-adjusted Wilcoxon Rank Sum statistic.

The second example pertains to a randomized clinical trial with a moderate number of patients being assigned randomly to two treatment subpopulations. This example illustrates aspects of randomization analysis of covariance relative to weighted least squares and maximum likelihood modeling. These methods have different types of information due to differences in their respective underlying sampling processes. The coordination of these methods provide complementary analysis strategies. Randomization statistics provide a rigorous covariance framework for testing treatment group differences with respect to the finite population under study. Weighted least squares and maximum likelihood provide probability models describing the effect of covariates on the response distribution and producing covariate-specific predicted values for generalization to a broad population. In view of these considerations, the analytical procedures are discussed with emphasis on the sampling process, the covariance structure and the extent to which results can be generalized. Other aspects of this example are discussed in Amara, Davis and Koch (1981).

3.1 Intracardiac Conduction Example

The data for this example pertain to a clinical study in arrhythmogenesis. A cause of atrial fibrillation is re-entry of consecutive heartbeats. This re-entry is often initiated by a conduction slowing of a particular heartbeat. Premature beats are more likely to have this conduction slowing than non-premature beats. This study is concerned with identifying patients at risk of developing atrial fibrillation by measuring the extent of conduction slowing in premature beats.

A small number (17) of patients undergoing routine electrophysiological studies formed two diagnostic subpopulations for absence of any abnormality and presence of atrial abnormalities. Their hearts were electrically paced and the intra-atrial conduction of heartbeats was timed. Here, conduction time of non-premature beats was equivalent for the two groups. However, patients with atrial abnormalities required more electrical stimulus in pacing.

The minimum amount of electrical current necessary to induce a beat (threshold current) is a measure of resistance to beat induction.

Threshold current (beat induction resistance) remains at a constant level for beats minimally to moderately premature. For extreme prematurity, resistance increases sharply to a refractory point beyond which no beats can be induced. In this study, changes in conduction time were measured relative to magnitude changes in threshold current across the range of induction prematurity (progressing from late diastole or minimal prematurity to the refractory period or extreme prematurity). Patients with abnormal atria seemed to show a rapid increase in conduction time as a function of induction difficulty (increased threshold current), see Figure 3.1.1. Their rapid increase in conduction time is in contrast to steady conduction times in patients without any abnormalities.

The data for this example were bivariate values of conduction time slowing and order of magnitude change in threshold current in a series of premature beats for each patient. For any given patient, the data predominantly involved a narrow interval of minimal threshold current changes within which conduction time was not expected to vary beyond measurement error. Correspondingly, the remaining data were distributed sparsely across the range of threshold change. With this type of data point distribution together with a complex (i.e., non-linear) relationship of conduction time slowing with threshold current change and samples of size 6 and 11, randomization test statistics represent a practical alternative method to growth curve modeling (with assumptions of normality and variance stability).

3.1.1 Randomization model analysis

The magnitude change in threshold current stimulus over the late diastolic threshold for the series of induced beats was categorized into

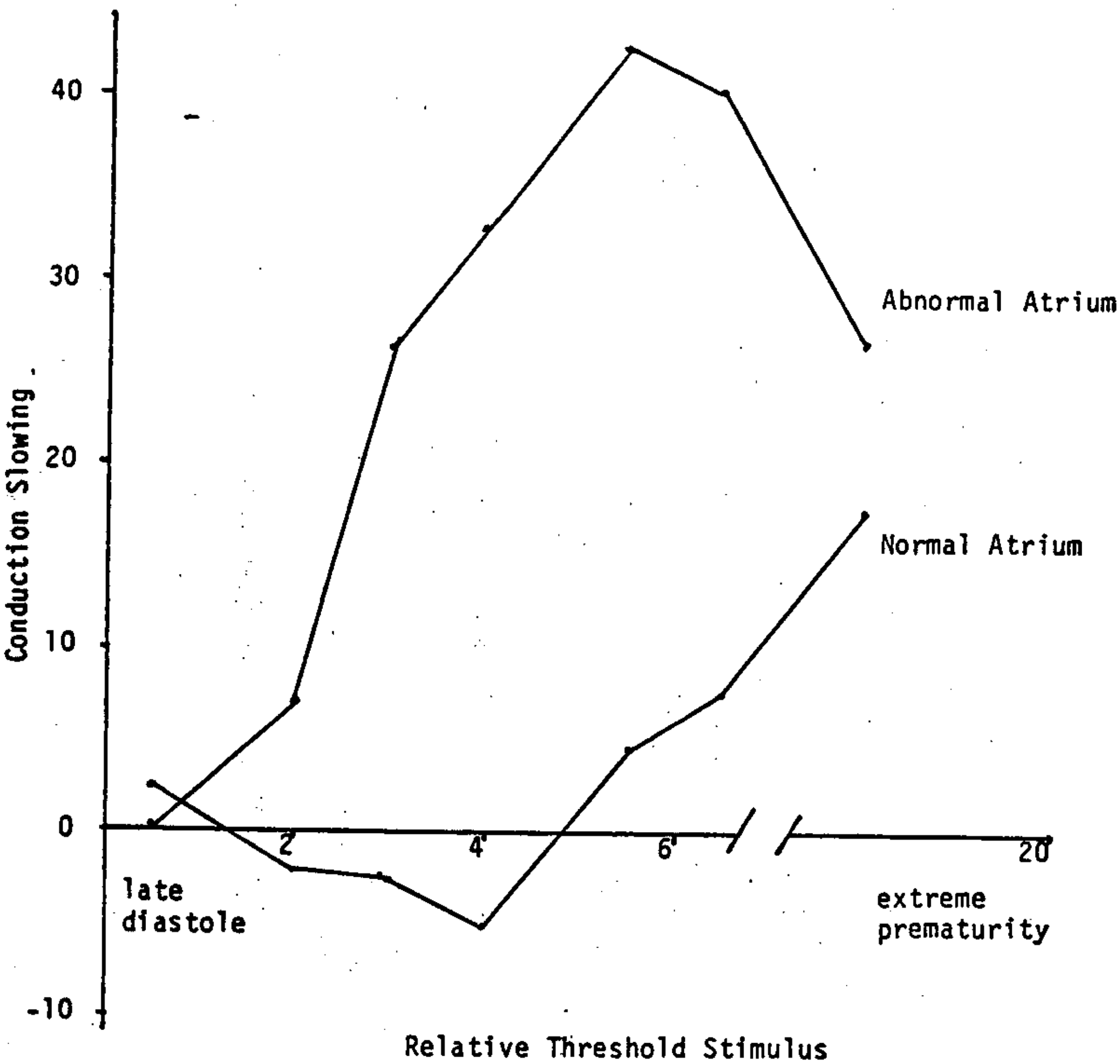

Figure 3.1.1 Mean patient median conduction slowing of premature beats within intervals of threshold stimulus change plotted at the midpoints of the corresponding overlapping threshold stimulus intervals for patients with normal and abnormal atria.

seven overlapping intervals. The first interval consists of all observations at essentially the late diastolic threshold; the negligible increase in difficulty in inducing these beats is expected to be associated with an equally negligible increase in conduction slowing for both diagnostic groups. Group differences in conduction slowing are expected to appear within the moderate range of difficulty in inducing beats associated with the inner threshold stimulus intervals; the extreme threshold stimulus interval might not show group differences because of the uncertainty of the effect of large currents on conduction time relative to the actual site being stimulated.

For each interval, patient median changes in conduction time were computed, summarizing their repeated observations into one measure per patient. As beats were more difficult to induce as in the maximal threshold intervals, very few repeated observations were made; many patients had no observations in at least one of the maximal threshold intervals. The seven overlapping intervals were used such that patient medians represented a continuum of conduction slowing across the full range of increasing threshold stimulus. In this regard, a patient's set of medians can be viewed as a standardized set of observations not as sensitive to threshold interval boundaries as medians derived from non-overlapping boundaries for adjacent threshold intervals.

Central ridits[1] were assigned across patients within each of the seven threshold change intervals as shown in Table 3.1.1. These ridits are distributed uniformly from -0.5 to 0.5 corresponding to least conduction slowing to most conduction slowing with the group-wise median conduction slowing at zero. A patient's central ridit indicates his

[1] central ridit = {(rank (median) - 0.5) / n(patients)} - 0.5.

TABLE 3.1.1

INTRA-THRESHOLD INTERVAL CENTRAL RIDITS FOR MEDIAN INTRA-ATRIAL CONDUCTION TIME CHANGE IN PREMATURE BEATS AND THE NUMBER OF OBSERVATIONS BEING SUMMARIZED FOR PATIENTS WITH NORMAL AND ABNORMAL ATRIA

		Relative Threshold Current Stimulus						
Atrium	Patient	$\Delta T < 1$	$1 \le \Delta T < 3$	$2 \le \Delta T < 4$	$3 \le \Delta T < 5$	$4 \le \Delta T < 7$	$5 \le \Delta T < 8$	$\Delta T \ge 6$
Normal	1	-0.08	-0.47	-0.46	-0.45	-0.42	-0.40	
		(7)	(2)	(1)	(1)	(2)	(2)	
	2	-0.08	-0.06			0.08	0	0
		(3)	(1)			(3)	(3)	(2)
	3		-0.38	-0.35	-0.32			
			(7)	(3)	(2)			
	4		-0.06	-0.19		-0.08	-0.20	-0.20
			(10)	(2)		(1)	(3)	(3)
	5	0.35	-0.06	-0.08	-0.18			
		(1)	(10)	(2)	(1)			
	6	0.35	-0.06	-0.35	-0.32	-0.25		
		(3)	(7)	(3)	(2)	(1)		
Abnormal	7	0.15	-0.06	0.27	0.18	0.25	0.30	0.30
		(2)	(7)	(3)	(3)	(1)	(1)	(1)
	8	0.35	0.35	0.08				
		(5)	(3)	(1)				
	9	-0.08	0.35	0.15	0.09			
		(1)	(6)	(6)	(3)			
	10	-0.46	-0.38					
		(4)	(6)					
	11	-0.08	-0.06					
		(1)	(8)					
	12		0.35	0.46	0.41			
			(12)	(2)	(1)			
	13	-0.35	0.47	0.27	0.41	0.42	0.30	
		(10)	(6)	(3)	(1)	(2)	(1)	
	14	0.35	-0.06	0	0			
		(1)	(6)	(2)	(1)			
	15	-0.35	-0.06	-0.19	-0.09			
		(2)	(20)	(6)	(1)			
	16	-0.08	0.24					-0.40
		(6)	(3)					(1)
	17		-0.06	0.38	0.27			0.30
			(20)	(2)	(1)			(1)

relative conduction slowing position among patients within the threshold interval. A negative central ridit identifies conduction slowing below the median slowing. Alternatively, a positive central ridit identifies conduction slowing above the median slowing. The averaging of central ridits across intervals that a patient has observations results in an estimate of the patient's relative position across the range of his data. As can be seen in Table 3.1.1, patients tend to have similar central ridits across intervals beyond the first. Both an average central ridit across all intervals and an average inner central ridit across the inner threshold intervals ($1 \leq \Delta T < 8$) were computed for each patient.

The randomization model corresponds to viewing the partition of conduction slowing responses between the two diagnostic groups as a random partition of fixed sizes {n(normal), n(abnormal)} from the pooled sample in accordance with the multiple hypergeometric model. Thus, for the average interval measure, the Wilcoxon Rank Sum statistic can be used to test the relative difference in conduction time slowing between patients with normal atria and patients with abnormal atria as shown at the bottom of Table 3.1.2. On average, those patients with abnormal atria had early ($p < .02$) conduction slowing during the inner threshold stimulus intervals. This difference in conduction slowing was not as pronounced ($p < .06$) when applying the average across all stimulus intervals. These results are further strengthened by adjusting for nonsignificant ($p > .25$) differences regarding initial conduction time for non-premature beats as a covariate.

The significant results for the average interval measure can be attributed to the inner interval differences in the central ridits

TABLE 3.1.2

RANDOMIZATION TEST STATISTICS COMPARING INTRA-ATRIAL CONDUCTION TIME CHANGE IN PREMATURE BEATS WITHIN LEVELS OF PREMATURITY INDICATED BY INTERVALS OF INCREASED THRESHOLD CURRENT STIMULUS OVER THE LATE DIASTOLIC THRESHOLD CURRENT FOR PATIENTS WITH NORMAL AND ABNORMAL ATRIA

Relative Threshold Current Interval	Atrium	Numbers of Patients	Relative Conduction Mean Patient Median	Wilcoxon Rank Sum	p-values* Unadjusted Exact	Unadjusted Q_{MHS}	Adjusted Q_{MHS}
$\Delta T < 1$ (Late	Normal	4	2.5	56.0	> .20	.259	.304
Diastolic)	Abnormal	9	0.3				
$1 \leq \Delta T < 3$	Normal	6	-2.0	35.5	< .10	.043	.035
	Abnormal	11	7.0				
$2 \leq \Delta T < 4$	Normal	5	-2.5	16.5	< .01	.006	.003
	Abnormal	8	26.2				
$3 \leq \Delta T < 5$	Normal	4	-5.0	10.0	< .01	.008	.004
	Abnormal	7	32.8				
$4 \leq \Delta T < 7$	Normal	4	4.4	10.0	< .20		
	Abnormal	2	42.5				
$5 \leq \Delta T < 8$	Normal	3	7.5	6.0	< .20		
	Abnormal	2	40.0				
$\Delta T \geq 6$ (Extreme	Normal	2	17.5	5.0	> .20		
Prematurity)	Abnormal	3	26.7				
Inner Interval Average	Normal	6	0.8	30.0	< .02	.016	.011
$(1 \leq \Delta T < 8)$	Abnormal	11	16.9				
Overall Average	Normal	6	1.4	35.0	< .10	.056	.044
	Abnormal	11	14.9				

* Exact p-values were obtained for the Wilcoxon Rank Sum (S) statistic using a table [Conover, 1971] of quantiles of the Mann-Whitney (T) statistic where $T_L = S_L - [n_L(n_L+1)/2]$ and n is the number of patients in S, the subscript L indicates the group of patients with the lower ranks; Q_{MHS} has an asymptotic chi-square distribution with 1 degree of freedom. The adjustment is for conduction time at late diastolic, a baseline conduction value.

between the two groups, see Table 3.1.2. The Wilcoxon Rank Sum statistics show that despite equivalent ($p > .20$) conduction slowing during the initial increase in threshold stimulus ($\Delta T < 1$), those patients with abnormal atria showed more pronounced ($p < .05$) conduction time slowing in each of the next three intervals of increasing threshold current stimulus. Furthermore, this pattern of more pronounced conduction slowing for patients with abnormal atria was somewhat maintained ($p < .20$) for the few patients who had extremely premature beats ($\Delta T > 4$).

3.2 Chronic Joint Pain Example

The frequencies in Table 3.2.1 summarize data collected in a clinical trial in which 84 patients with chronic joint pain were randomly assigned to placebo or active treatment. Here, patient response to treatment (PRS) is defined as being at one of five levels: excellent (1), good(2), moderate(3), fair(4) or poor(5); age and sex are considered to be important covariates. One question of interest is the extent to which active treatment is clearly more effective than placebo in treating chronic joint pain in the sense that a higher proportion of patients on active treatment experience good or excellent response to treatment than patients on placebo. For this purpose, several measures of patient response status can be used; one of these is a binary indicator representing good or excellent (1) vs moderate, fair or poor (0) responses.

3.2.1 Randomization model analysis

The randomization model is directly supported by the design-based randomized assignment of the 84 patients to the two groups within which

TABLE 3.2.1

TABULATION OF PATIENT END OF STUDY GLOBAL EVALUATIONS FOR CHRONIC JOINT PAIN CLINICAL TRIAL, CROSS-CLASSIFIED BY SEX, AGE AND TREATMENT

			Patient Response to Treatment					
Sex	Age	Treatment	Excellent	Good	Moderate	Fair	Poor	Number of patients
Female	≤ 44	Active	0	1	1	1	2	5
Female	≤ 44	Placebo	0	1	1	3	3	8
Female	45-54	Active	0	1	0	0	1	2
Female	45-54	Placebo	0	1	0	1	7	9
Female	55-64	Active	3	8	1	0	1	13
Female	55-64	Placebo	1	1	4	2	1	9
Female	≥ 65	Active	2	1	3	0	1	7
Female	≥ 65	Placebo	1	1	2	1	1	6
Male	≤ 44	Active	1	0	1	0	2	4
Male	≤ 44	Placebo	0	0	0	2	0	2
Male	45-54	Active	0	1	0	0	0	1
Male	45-54	Placebo	0	0	0	0	4	4
Male	55-64	Active	0	2	1	1	3	7
Male	55-64	Placebo	0	1	0	2	2	5
Male	≥ 65	Active	0	1	0	0	1	2
Male	≥ 65	Placebo	0	0	0	0	0	0

43 patients received placebo and 41 patients received active treatment. Given that the 84 patients can be viewed as a finite population from which a fixed size sample of 43 patients are to be randomly selected without replacement, then the possible partitions can be enumerated by the hypergeometric model. In other words, the observed partition under randomization is one of 84!/43! 41! equally likely partitions. Thus, the randomization model is synonymous to the hypothesis of no treatment group difference regarding age, sex or the proportion of patients with good or excellent response in the sense of a single pooled-group population. Correspondingly, tests of the hypothesis involve a covariance structure for patients as a single population in the sense of a pooled total variance. This randomization covariance structure is in contrast to a weighted or unweighted partition average of a within-group covariance structure used in methods such as weighted or unweighted least squares modeling.

The randomization or single population model is supported in the sense that each treatment group has an equivalent allocation of patients relative to their age and sex distributions as indicated by nonsignificant treatment group differences for age and sex in Table 3.2.2. The direct treatment group comparison of the proportions of patients with good or excellent response shows that patients taking the active treatment have significantly ($p = .001$) greater response to treatment than those patients taking placebo. This result is negligibly strengthened by adjusting the comparison for the nonsignificant differences in the age and sex distributions.

Alternatively, stratification by age levels provides a more precise covariance framework in that treatment group comparisons involve

TABLE 3.2.2

RANDOMIZATION (AVERAGE PARTIAL ASSOCIATION) CHI-SQUARE TEST STATISTICS Q, THEIR DEGREES OF FREEDOM, AND APPROXIMATE P-VALUES FOR COMPARISON OF ACTIVE AND PLACEBO TREATMENT GROUPS RELATIVE TO THE PROPORTION OF PATIENTS WITH GOOD OR EXCELLENT RESPONSE TO TREATMENT WITH AND WITHOUT COVARIANCE ADJUSTMENT

Treatment Group comparisons	Unstratified population			Stratified by age (≤ 44, 45-54, 55-64, ≥ 65)		
	Q	D.F.	P	Q	D.F.	P
Good or excellent indicator (GE)	11.39	1	0.001**	8.79	1	0.003**
AGE(A)[1], SEX(S)	1.50	2	0.473	2.29	2	0.318
GE adjusted for A[1], S	11.61	1	0.0001**	8.44	1	0.004**

[1] Age in years

patients who are of similar age. The age stratification analysis shows a significant treatment effect ($p = .003$) which is comparable in strength to that ($p = .001$) which was observed in the unstratified analysis. Moreover, this treatment effect is maintained ($p = .004$) after covariate adjusting for age and sex within each age stratum.

These results in Table 3.2.2 indicate that active treatment is more effective than placebo in the sense that it is associated with a higher proportion of patients with good or excellent responses. Although this result is observed also at an average age stratum, it is not clear that this result occurs at every age. Furthermore, these results pertain to the observed sample only because the randomization sampling frame does not include sampling from some larger population of patients.

3.2.2 Weighted least squares model analysis

The weighted least squares analysis is undertaken in stratified simple random setting relative to an assumed larger population cross classified according to sex and treatment subpopulations. Within each subpopulation, bivariate random sample for PRS and age jointly is assumed.

The proportion of patients with good or excellent response can be modeled in its logit form across the four subpopulations in conjunction with mean age. The corresponding array of functions formulated within CRISCAT is:

$$
\underset{\sim}{F} = \begin{array}{l} \underset{\sim}{F}' \text{ (female, active)} \\ \underset{\sim}{F}' \text{ (female, placebo)} \\ \underset{\sim}{F}' \text{ (male, active)} \\ \underset{\sim}{F}' \text{ (male, placebo)} \end{array}
$$

where $\{F'\} = \{\ln(\bar{y}/(1-\bar{y}), \bar{x}\}$ with $\bar{y}$ being the mean of the good or excellent indicator and $\bar{x}$ being the mean age. In accordance with the assumed sampling structure, the estimated covariance structure, $\tilde{V}_{\tilde{F}}$, is block diagonal where each block is the corresponding within-subpopulation covariance matrix (as opposed to the pooled total variance from the randomization framework). In this regard, the randomization method can be at a disadvantage in detecting true differences among subpopulations that are heterogeneous as opposed to homogeneous. Observed values with standard errors are listed in part a of Table 3.2.3. An initial model of interest which can be expressed as $\{\text{logit}, \text{age}\} = \tilde{K}\tilde{B}$, is shown in part b of Table 3.2.3. The model goodness of fit statistic supports the assumption that mean age is equivalent across sex and treatment groups. Furthermore, the sex by treatment interaction (B_4) does not appear to be significant for the logit good or excellent indicator. The goodness of fit test for a model reduced by the interaction term verifies its nonsignificance. Although the reduced model shows no significant sex effect (B_2), it is useful in providing sex-specific predicted values as shown in part a of Table 3.2.3.

Finally, the hypothesis of no treatment difference based on the reduced model is rejected in the sense that for the sampled population, higher proportions of male and female patients had good or excellent response in the active treatment group when compared to patients in the placebo group. Although this result is adjusted for variation about the mean age, it does not provide information concerning the extent to which treatment effectiveness is age-specific. In this regard, the WLS model-based predicted values may only be indicative of results from

TABLE 3.2.3

WEIGHTED LEAST SQUARES LOGIT ANALYSIS OF PROPORTION OF PATIENTS WITH GOOD OR EXCELLENT RESPONSE FOR ACTIVE AND PLACEBO TREATMENT GROUPS RELATIVE TO SEX WITH COVARIANCE ADJUSTMENT FOR AGE

a. Observed and covariance logit model predicted values with standard errors for percentage of patients with good or excellent response and age for active and placebo groups relative to sex

		Observed values		Covariance logit model WLS predicted values	
Sex	Treatment	% good/ex.	Mean age	% good/ex.	Mean age
Female	Active	59.2 ± 9.5	55.7 ± 2.3	57.4 ± 8.9	53.6 ± 1.3
Female	Placebo	18.8 ± 6.9	51.6 ± 2.3	19.8 ± 6.7	53.6 ± 1.3
Male	Active	35.7 ± 12.8	52.4 ± 4.1	35.5 ± 11.5	53.6 ± 1.3
Male	Placebo	9.1 ± 8.7	53.8 ± 2.4	9.2 ± 5.2	53.6 ± 1.3

b. Covariance linear model specification, parameter estimates and test statistics

Specification matrix (X_1)

$$X_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Parameter interpretation	WLS parameter estimates ± s.e.'s
Reference logit value for females (β_1)	0.29 ± 0.38
Increment for males (β_2)	-0.87 ± 0.68
Increment for placebo (β_3)	-1.68 ± 0.58
Interaction increment of males for placebo (β_4)	-0.07 ± 1.28
Mean age for all groups (β_5)	53.61 ± 1.27

Model goodness of fit statistic Q = 1.71 with D.F. = 3

c. Reduced model parameter estimates and test statistics

Remaining parameters	WLS parameter estimates ± s.e.'s	Statistic	Q	D.F.
β_1	0.30 ± 0.36	Model goodness of fit	1.72	4
β_2	-0.89 ± 0.57	Treatment group difference (β_3=0)	10.63	1
β_3	-1.70 ± 0.52			
β_4	53.62 ± 1.26			

broader populations whose age distribution is similar to that of the assumed conceptually sampled population.

3.2.3 Maximum likelihood model analysis

As a supportive analysis, the maximum likelihood analysis is undertaken in a stratified simple random sampling setting with respect to the same 16 cross-classified age, sex and treatment subpopulations shown in Table 3.2.1. This stratification system is more detailed than the weighted least squares system because sample size restrictions are not as stringent for maximum likelihood estimation when a simplified model is potentially applicable. Within each subpopulation, a random sample for PRS is assumed. Observed proportions with standard errors are shown in Table 3.2.4.

The proportion of patients with good or excellent response can be modeled in its logit form across the 16 subpopulations with age and sex as covariates and treatment as the factor of interest. A general model can be expressed as: $\text{logit} = B_0 + \underset{\sim}{B}_1\ f(\text{age, sex, treatment})$ where B_0 is the intercept parameter for 20 year old female patients on active treatment and $\underset{\sim}{B}_1$ are slope parameters associated with sex, a linear trend for age, treatment, linear age x sex interaction, and sex x treatment interaction. In this regard, age was expressed in the standardized form (age-20)/10 so as to correspond to the number of 10 year intervals above a minimum age of 20. Also, age has been grouped into intervals for < 44, 45-54, 55-64, > 65 with respect to which the "mid-point values" 40, 50, 60, 70 have been assigned. The data for the resulting sex x age group x treatment subpopulations are shown in Table 3.2.4. The proportions of good or excellent response for these subpopulations were analyzed by logistic regression. An initial model that

TABLE 3.2.4

OBSERVED PERCENTAGES OF PATIENTS IN CHRONIC JOINT PAIN CLINICAL TRIAL WITH GOOD OR EXCELLENT END OF STUDY GLOBAL EVALATION AND THEIR REDUCED LOGISTIC MODEL MAXIMUM LIKELIHOOD (ML) PREDICTED ESTIMATES TOGETHER WITH ESTIMATED STANDARD ERRORS FOR CROSS-CLASSIFICATION BY SEX, AGE, AND TREATMENT

			Number or patients with global evaluation		Percentage of patients with or excellent global evaluation			
Sex	Age	Treatment	Good or excellent	Moderate, fair,poor	Observed	s.e.	Model (X_2) ML Predicted	s.e.
Female	≤ 44	Active	1	4	20.0	17.9	40.3	14.4
Female	≤ 44	Placebo	1	7	12.5	11.7	10.8	5.9
Female	45-54	Active	1	1	50.0	35.4	50.8	10.8
Female	45-54	Placebo	1	8	11.1	10.5	15.6	6.1
Female	55-64	Active	11	2	84.6	10.0	61.2	9.2
Female	55-64	Placebo	2	7	22.2	13.9	22.0	7.5
Female	≥ 65	Active	3	4	42.9	18.7	70.7	10.3
Female	≥ 65	Placebo	2	4	33.3	19.2	30.2	11.6
Male	≤ 44	Active	1	3	25.0	21.7	22.5	11.6
Male	≤ 44	Placebo	0	2	00.0	...	5.0	3.6
Male	45-54	Active	1	0	100.0	...	30.8	11.5
Male	45-54	Placebo	0	4	00.0	...	7.4	4.6
Male	55-64	Active	2	5	28.6	17.1	40.5	12.7
Male	55-64	Placebo	1	4	20.0	17.9	10.9	6.9
Male	≥ 65	Active	1	1	50.0	35.4	51.0	15.7
Male	≥ 65	Placebo	0	0	00.0	...	15.7	10.1

adequately describes the variation among the subpopulations, as indicated by nonsignificant goodness of fit test statistics, is shown in Table 3.2.5. The slope parameters pertaining to age by sex interaction (B_5) and age by treatment (B_6) interaction appear nonsignificant, suggesting a parallel line model. Furthermore, there does not appear to be significant sex by treatment interaction (B_7). The goodness of fit tests for a model reduced by the exclusion of the interaction terms confirm their nonsignificance. Although the reduced model shows no significant sex effect (B_2), it is useful in providing sex-specific predicted values shown in Table 3.2.4

Finally, the hypothesis of no treatment differences based on the reduced model is contradicted in the sense that for the sampled population, higher proportions of male and female patients had good or excellent response in the active treatment group when compared to patients in the placebo group. Furthermore, the proportion of patients with good or excellent response increases in a linear logistic sense for males and females alike in both treatment groups with respect to increasing age.

3.2.4 Discussion

The randomization analysis, weighted least squares analysis and maximum likelihood analysis all showed significant treatment effectiveness in the sense that a higher proportion of patients had good or excellent response to treatment. Furthermore, treatment effectiveness was maintained after adjusting for age and/or sex covariates. The randomization results for the age stratification analysis are directed at the average partial association of treatment with the response across the levels of age. The stratified comparison without covariance

TABLE 3.2.5

MAXIMUM LIKELIHOOD LOGISTIC COVARIANCE MODELS FOR PERCENTAGES OF PATIENTS WITH GOOD OR EXCELLENT RESPONSE RELATIVE TO 16 CROSS-CLASSIFIED SEX, AGE AND TREATMENT SUBPOPULATIONS

a. Covariance linear model specification, parameter estimates and test statistics

Model X_1 Specification

$$X_1 = \begin{bmatrix} 1 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 1 & 0 & 2 & 0 \\ 1 & 0 & 3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 3 & 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 & 0 & 0 & 0 \\ 1 & 0 & 4 & 1 & 0 & 4 & 0 \\ 1 & 0 & 5 & 0 & 0 & 0 & 0 \\ 1 & 0 & 5 & 1 & 0 & 5 & 0 \\ 1 & 1 & 2 & 0 & 2 & 0 & 0 \\ 1 & 1 & 2 & 1 & 2 & 2 & 1 \\ 1 & 1 & 3 & 0 & 3 & 0 & 0 \\ 1 & 1 & 3 & 1 & 3 & 3 & 1 \\ 1 & 1 & 4 & 0 & 4 & 0 & 0 \\ 1 & 1 & 4 & 1 & 4 & 4 & 1 \\ 1 & 1 & 5 & 0 & 5 & 0 & 0 \\ 1 & 1 & 5 & 1 & 5 & 5 & 1 \end{bmatrix}$$

Parameter descriptors	ML parameter estimates ± S.E.'s
Reference logit value for females with age < 20 (β_1)	-1.01 ± 1.49
Increment for males (β_2)	-0.60 ± 2.38
Linear increment for (age-20)/10 (β_3)	-0.37 ± 0.38
Increment for placebo treatment (β_4)	-2.58 ± 2.24
Interactin increment of linear age for males (β_5)	-0.08 ± 0.62
Interaction increment of linear age for placebo (β_6)	0.22 ± 0.57
Interaction increment of males for placebo (β_7)	0.16 ± 1.35

Model goodness of fit test statistics: Q_L = 10.66, D.F. = 9

Q_P = 9.84, D.F. = 9

Hypotheses concerning treatment group differences for specific age and sex criteria:

mean age	percent female	parameter contrast	Q_W	D.F.
55	50	$\beta_4 + 3.5\,\beta_6 + 0.5\beta_7$	6.46	1
53.4	29.8	$\beta_4 + 3.34\beta_6 + 0.298\beta_7$	9.01	1

b. Reduced model parameter estimates and test statistics

Remaining parameters	ML parameter estimates ± s.e.'s
β_1	-1.24 ± 1.05
β_2	-0.44 ± 0.26
β_3	-0.42 ± 0.26
β_4	-1.72 ± 0.55

Model goodness of fit test statistics

Q_L = 10.88, D.F. = 12

Q_P = 10.20, D.F. = 12

adjustment was based on the Mantel-Haenszel statistic for four 2 x 2 tables where the rows correspond to treatment groups and the columns correspond to the good/excellent indicator levels with the cells having counts of patients in a treatment group that have or have not good or excellent response to treatment. In this regard, the covariate-adjusted response is a modified covariate-adjusted Mantel-Haenszel test statistic.

The weighted least squares results pertain to the treatment group differences relative to a model for sex and treatment group with age as a covariate. Here, subpopulation sample sizes were not large enough to support levels for age groups within treatment subpopulations. The maximum likelihood results pertain to the treatment group differences relative to a model for sex, age and treatment groups.

The differences in the analysis framework affect the generality of the results. The randomization results refer to an average age comparison which does not necessarily apply equally at each age level. The weighted least squares results specifically refer to the group difference at the central group age and also does not necessarily apply equally across the age range in the sense of parallel group lines. The maximum likelihood results describe group differences along the age range. Furthermore, the randomization results are applicable to the finite population under study only, but do not require any assumptions about a conceptual population they might represent. The weighted least squares results are applicable to some large population from which the data were conceptually sampled. Because age was not a stratification variable, an underlying assumption is that the age distribution is equivalent to that of the large population of interest in the sense of proportional sampling. The maximum likelihood results are applicable to some large population from which the data are conceptually sampled within an age stratification.

The method of covariance adjustment for randomization statistics involves showing that the distribution of age and sex are at random with respect to their partition between the two treatment groups. One mechanism to enhance compatibility with this condition is through post-stratification. In particular, stratification and covariance adjustment can be used in combination such that imbalances in subpopulation allocation of a covariate are controlled with the minimum number of strata necessary to have nonsignificant covariate differences. These remaining covariate differences are then adjusted within each strata. This combination technique is an alternative to having a large number of strata with few observations each. Here, age and sex are not significantly associated with treatment in the single stratum analysis and the age stratified analysis.

Covariance adjustment for the weighted least squares model involved establishing a common mean age which was the weighted average of the group ages. Sex was not a covariate since sex defined levels of treatment subpopulations. Accordingly, the variation between subpopulations for males and for females is described by a model for which sex-specific predicted values are possible.

Covariance adjustment for the maximum likelihood logistic regression involved fitting a parallel line or common slope model with respect to the age gradient across subpopulations. For this construction, age specific predicted values are possible as well as sex specific predicted values. These results suggest a continuous ML model with age assuming a linear relationship with PRS. Also PRS can be analyzed in its full integer score form.

CHAPTER IV

A MULTIVARIATE RANDOMIZATION APPLICATION

The example (see Johnson, Amara, Edwards and Koch [1981]), in this chapter illustrates the multivariate randomization statistics in Sections 1 and 3 of Chapter II. The multivariate profile has 8 repeated measurements in time for patients randomly assigned to one of four treatment groups within each of four clinics. The analysis strategy for comparing treatment groups consists of two stages. One is randomization statistics to provide tests of hypotheses; and the other is weighted least squares models to describe a complex interaction for clinic and to produce corresponding predicted values.

4.1 Obstetrical Pain Example

The following multivariate analysis pertains to data from a repeated measurements design used in a clinical trial to assess effectiveness of four treatments for obstetrically related pain. In particular, four combinations of drugs A and B were identified to provide a placebo treatment, A treatment only, B treatment only and a combination AB treatment. The formulation of these compounds corresponds to the drug A-drug B cross-classification relative to null and active levels; the four treatments are expressed as placebo, A, B, and AB respectively.

The patients participating in the study under consideration were from four clinics and were experiencing pain that started within the first 24 hours subsequent to vaginal delivery. They were randomly assigned one of the four treatments defined previously. Before treatment their pain level was classified as "some pain" or "a lot of pain". After receiving treatment, patients rated their pain as none (0), a little (1), some (2), a lot (3) or terrible (4) for each hour of a four hour period. A second dose was administered at the end of the fourth hour and pain ratings were continued hourly for the next four hours. Average pain intensity ratings with respect to the uniform scores 0, 1, 2, 3, 4 are shown in Table 4.1 for each of the eight hours for the combined patients in all clinics, cross-classified according the the initial pain status and treatment.

The pain rating averages in Table 4.1 suggest that scores are predominantly 0, 1 or 2. Clinically important score distinctions involve scores 0 or 1 versus scores 2, 3 or 4. In view of these considerations, the focus of discussion in the remainder of this chapter will be based upon a binary measure of response indicating patients with little or no pain versus patients with at least some pain. This binary measure is expressed as the proportion of patients with little or no pain.

The proportion of patients with little or no pain is listed hourly for each of the four treatments within the two levels of initial pain status in Table 4.2. It is evident that patients improve with time within each treatment group. More specifically, the first three hours show an increase in the proportion of patients with little or no pain. There is a temporary relapse of pain between the third and fourth hour suggesting a wearing off of the initial dose, and then a noticeable improvement in pain relief between the fourth and fifth hour corresponding to the administration of the second dose. The proportion of patients with little or no pain appears to remain stable through the sixth hour

TABLE 4.1

AVERAGE PAIN INTENSITY CROSS-CLASSIFIED BY INITIAL PAIN STATUS AND TREATMENT

Initial pain status	Treatment	Number of patients	Average Pain Intensity Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Hour 8
Some	Placebo	80	1.54	1.19	1.11	1.30	0.98	0.99	1.06	1.18
Some	B	85	1.38	1.01	0.74	0.93	0.67	0.65	0.73	0.75
Some	A	78	1.44	1.09	0.96	1.10	0.79	0.73	0.81	0.92
Some	AB	85	1.26	0.94	0.73	0.93	0.68	0.55	0.61	0.66
A lot	Placebo	90	2.28	1.77	1.61	1.70	1.41	1.51	1.56	1.68
A lot	B	102	1.99	1.35	1.14	1.12	0.84	0.74	0.81	0.95
A lot	A	98	2.13	1.51	1.23	1.52	1.02	0.95	1.11	1.21
A lot	AB	100	1.88	1.28	1.01	1.13	0.76	0.75	0.74	0.83

TABLE 4.2

PROPORTIONS OF PATIENTS WITH LITTLE OR NO PAIN, CROSS-CLASSIFIED BY INITIAL PAIN STATUS AND TREATMENT

Initial pain status	Treatment	Number of patients	Proportions Of Patients With Little Or No Pain At Hourly Evaluation Times Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Hour 8
Some	Placebo	80	0.425	0.650	0.650	0.589	0.788	0.762	0.675	0.700
Some	B	85	0.518	0.776	0.859	0.776	0.894	0.882	0.847	0.835
Some	A	78	0.449	0.731	0.795	0.667	0.859	0.910	0.885	0.782
Some	AB	85	0.588	0.788	0.894	0.800	0.906	0.941	0.918	0.894
A lot	Placebo	90	0.189	0.367	0.467	0.422	0.567	0.533	0.522	0.456
A lot	B	102	0.275	0.520	0.667	0.706	0.824	0.882	0.833	0.775
A lot	A	98	0.214	0.418	0.592	0.480	0.684	0.745	0.663	0.622
A lot	AB	100	0.380	0.580	0.720	0.670	0.890	0.910	0.900	0.840

with a slight decline in the seventh and eighth hours. This pattern of response appears to be consistent across the four treatments and two levels of initial pain.

Among the treatments, active B treatment and the combination AB treatment appear to be more effective as analgesic compounds than the active A treatment or placebo; i.e., B and AB consistently show higher proportions of patients with little or no pain than A and placebo. Furthermore, B and AB do not show the noticeably reduced effectiveness of A and placebo for patients with a lot of pain initially relative to patients with some pain initially.

Finally, it is apparent that proportions of patients with little or no pain are consistently smaller for those with a lot of pain initially in contrast to those with only some pain initially.

4.2 Randomization Model Analysis

The randomization model is based upon a finite population sampling framework whereby patients are randomly assigned to one of four treatment groups within each clinic. According to this sampling design, each clinic is a distinct stratum. Because initial pain status is known *a priori* relative to treatment, initial pain status can be used as a post-stratification variable, thereby providing a subdivision of each clinic into separate strata for patients with some pain initially and patients with a lot of pain initially. In this regard, the randomization of patients within clinics can be viewed as equivalent to randomization within the eight post-stratification levels of initial pain status within each clinic. The randomization model for equivalent treatment effectiveness corresponds to the observed partition of responses (j) (e.g., defined as a binary indicator for little or no pain for each of the eight hours or 2^8 possible response profiles) among the four treatment groups (i) being one of the equally likely partitions in accordance with the product multiple hypergeometric distribution.

$$\Pr(\{n_{hij}\}) = \prod_{h=1}^{8} [\prod_{i=1}^{4} n_{hi+}! \prod_{j=1}^{256} n_{h+j}!/n_{h++}! \prod_{i=1}^{4} \prod_{j=1}^{256} n_{hij}!] .$$

Also, see the statement of H_o in Section 2.1.

The data are not compatible with this randomization model in the sense that the results in Table 4.3 indicate that at least one treatment hourly and across all hours is associated ($p < 0.05$) with a different proportion of patients with little or no pain relative to the respective finite population strata. The average partial association statistics for pairwise comparisons in Table 4.4 show that the AB treatment is consistently better than placebo and consistently better than the A treatment, but not consistently better than the B treatment. Furthermore, the B treatment is consistently better than placebo, but not consistently better than the A treatment. Alternatively, all pairwise comparisons except AB versus B show significant differences relative to the multivariate eight hour profile with placebo having the least analgesic effect and A having less effect than B or AB.

Hourly pairwise treatment comparisons in Table 4.4 also show that significant differences are more pronounced in the second dosing period (hours 5 through 8) than in the first dosing period (hours 1 through 4). This pattern is also apparent when comparing the stratification statistics having one degree of freedom for the average proportion of hours that patients experience little or no pain over the first four hours and the average proportion over the last four hours, as shown in Table 4.5.

TABLE 4.3

RANDOMIZATION TEST STATISTICS† FOR OVERALL COMPARIONS AMONG 4 TREATMENTS RELATIVE TO PROPORTIONS OF PATIENTS WITH LITTLE OR NO PAIN

Stratum: Center	Stratum: Initial pain status	Number of patients	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Hour 8
1	Some	74	5.22	1.40	5.91	4.32	4.66	8.26*	8.96*	6.45
1	A lot	108	2.22	3.05	2.25	17.11**	11.74**	13.66**	17.05**	19.56**
2	Some	79	2.20	1.32	4.43	4.66	0.40	7.39	5.12	2.65
2	A lot	104	1.46	5.77	12.06**	25.39**	29.92**	24.68**	20.98**	25.23**
3	Some	96	10.09**	4.28	2.68	5.82	5.76	10.44*	9.83*	5.98
3	A lot	96	7.87*	8.43*	2.38	3.51	8.02*	3.24	4.38	5.67
4	Some	79	0.32	3.53	8.24*	2.05	2.54	3.67	3.17	3.93
4	A lot	82	3.73	0.16	2.28	2.78	15.24**	22.77**	20.28**	7.67
Combined		718	16.64**	15.02**	31.55**	40.64**	35.94**	56.88**	57.98**	49.74**

All hours = 94.12**

*Denotes significance at the $\alpha = 0.05$ level; **Denotes significance at the $\alpha = 0.01$ level.

†For the separate hours, these test statistics have approximate chi-square distributions with DF = 3; for all hours, it has an approximate chi-square distribution with DF = 24.

TABLE 4.4

AVERAGE PARTIAL ASSOCIATION RANDOMIZATION TEST STATISTICS FOR PAIRWISE TREATMENT COMPARISONS OF PROPORTIONS OF PATIENTS WITH LITTLE OR NO PAIN

Treatment Comparison	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Hour 8	All hours†
B vs Placebo	3.15	7.54**	16.88**	24.76**	18.02**	29.77**	27.36**	24.23**	39.47**
A vs Placebo	0.14	1.36	6.56*	1.69	3.85*	15.07**	12.56**	6.54*	20.45**
AB vs Placebo	12.48**	11.75**	25.98**	25.74**	28.68**	43.62**	49.04**	41.78**	60.62**
A vs B	2.00	2.53	2.66	14.31**	6.49*	3.24	3.79	6.46*	16.56*
AB vs B	3.90*	0.60	1.11	0.03	1.52	1.87	3.99*	2.79	8.79
AB vs A	10.51**	5.38*	6.89**	13.99**	12.69**	9.31*	14.54**	16.47**	28.76**

*Denotes significance at the $\alpha = 0.05$ level; **Denotes significance at the $\alpha = 0.01$ level.

†For the separate hours, these test statistics have approximate chi-square distributions with DF = 1; for all hours, they have approximate chi-square distributrions with DF = 8.

Although the chi-square values are larger for the second dosing period in all comparisons except A versus B, they are not consistently larger than the chi-square values for the average proportion over the eight hours. Furthermore, the increase in the average proportion for the last four hours over the first four hours is minimal in most treatment pairs with the exception of {A versus placebo} and {AB versus placebo}. For these two comparisons, this dosing period difference can be interpreted as a significant ($p < .05$) increase in the last four hour comparison. This dosing period difference is attributable to {A versus placebo} showing a weak ($.05 < p < 10$) response difference in the first four hours but a strong ($p < .01$) difference in the second four hours, and {AB versus placebo} showing strong ($p < .01$) differences in the first four hours but substantially stronger differences in the second four hours. Despite these dosing period differences, treatment group differences detected when averaging the proportions of hours that patients experience little or no pain across the eight hours are similar to results of averaging the first four hours and the second four hours separately. The eight hour average is also a comparable measure of treatment differences relative to the bivariate profile of the four hour averages.

The appropriateness of the eight hour average is also apparent in Table 4.6 relative to the overall comparisons among the four treatments. Essentially, treatment group differences can be relatively constant over time in the sense that patients have roughly parallel but not equivalent response to the different treatments over the eight hour period. This interpretation is supported by Table 4.7 where hourly orthogonal contrasts within each dosing period show only a few isolated significant

TABLE 4.5

RANDOMIZATION TEST STATISTICS FOR OVERALL COMPARISONS AMONG 4 TREATMENTS

Stratum: Center	Stratum: Initial pain	Degrees of freedom	Average proportion hours 1-4	Average proportion hours 5-8	Average proportion hours 1-8	Difference for hours 5-8 vs 1-4	Degrees of freedom	Average proportions for hours 1-4 and hours 5-8
1	Some	3	5.21	8.65*	9.45*	1.78	6	11.07
1	A lot	3	9.21*	18.25**	17.89**	8.01*	6	18.93**
2	Some	3	4.48	3.73	5.36	1.67	6	7.08
2	A lot	3	15.92**	31.64**	28.37**	9.98*	6	35.33**
3	Some	3	7.70	6.96	8.05*	3.99	6	11.92
3	A lot	3	6.90	5.80	6.66	4.13	6	10.99
4	Some	3	4.37	4.60	5.51	0.24	6	5.72
4	A lot	3	2.62	20.37**	11.93**	11.98**	6	22.77**
Combined		3	45.21**	65.42**	70.67**	6.83	6	74.16**

*denotes significance at the $\alpha = 0.05$ level; **denotes significance at the $\alpha = 0.01$ level.

TABLE 4.6

AVERAGE PARTIAL ASOCIATION RANDOMIZATION TEST STATISTICS FOR PAIRWISE TREATMENT COMPARISONS

Treatment Comparison	Univariate Comparisons: Degrees of freedom	Univariate Comparisons: Average proportion hours 1-4	Univariate Comparisons: Average proportion hours 5-8	Univariate Comparisons: Average proportion hours 1-8	Univariate Comparisons: Difference for hours 5-8 vs 1-4	Multivariate Comparisons: Degrees of freedom	Multivariate Comparisons: Average proportions for hours 5-8 and hours 1-4
B vs Placebo	1	22.04**	31.56**	34.33**	3.35	2	34.90**
A vs Placebo	1	3.66	11.81**	9.85**	3.99*	2	11.83**
AB vs Placebo	1	33.21**	51.86**	54.83**	4.91*	2	56.54**
A vs B	1	9.21**	6.93**	10.22**	0.05	2	10.46**
AB vs B	1	1.68	3.46	3.22	0.26	2	3.59
AB vs A	1	17.64**	18.69**	23.81**	0.03	2	23.83**

*denotes significance at the $\alpha = 0.05$ level; **denotes significance at the $\alpha = 0.01$ level.

TABLE 4.7

RANDOMIZATION TEST STATISTICS FOR OVERALL COMPARISONS AMONG FOUR TREATMENTS FOR ACROSS HOURS CONTRASTS WITHIN DOSAGE PERIODS

Stratum		Univariate Tests For Contrasts						Multivariate Test for All Contrasts
Center	Initial Pain	Hour 1 vs Average Hours 2-4	Hour 2 vs Average Hours 3-4	Hour 3 vs Hour 4	Hour 5 vs Hour 6	Hours 5-6 Average vs Hour 7	Hours 5-7 Average vs Hour 8	
1	Some	3.75	1.32	1.76	2.20	6.29	1.35	16.60
1	A lot	6.02	3.78	9.04*	6.43	2.87	3.07	31.21*
2	Some	2.16	0.64	3.41	8.44*	3.29	0.20	19.63
2	A lot	13.05**	3.11	4.50	0.84	0.76	1.15	22.88
3	Some	5.06	3.15	2.82	5.93	5.45	13.82**	38.81**
3	A lot	2.67	3.72	5.89	7.60	0.10	3.29	25.73
4	Some	2.33	1.84	1.62	3.75	0.54	1.60	16.38
4	A lot	3.21	2.28	0.37	5.75	3.28	1.42	15.69
Combined		6.54	4.93	6.15	6.98	3.88	2.25	27.30

*Denotes significance at the α = 0.05 level; **Denotes significance at the α = 0.01 level.
†For the separate contrasts, these test statistics have approximate chi-square distributions with DF = 3; for all contrasts, they have approximate chi-square distributions with DF = 18.

differences among the treatment groups relative to these components of response pattern over time. In view of these findings the multivariate time-response curve for each patient can be summarized as an overall average for comparison purposes.

Finally, it can be seen in Table 4.6 relative to the overall comparison among the four treatments, that strata involving patients with a lot of pain initially have predominantly stronger treatment group differences than strata involving patients with some pain initially. Also, significant ($p < .05$) treatment group differences in the increase in the average proportion for the last four hours over the first four hours are found only in strata with patients having a lot of pain initially. However, as noted previously when combined with the other strata, these treatment group differences in the increase in the average proportion for the last four hours over the first four hours do not have a consistent pattern since the combined average partial association statistic is not significant ($p > 0.05$). These results relative to initial pain status support the post stratification strategy in the sense of providing a more precise covariance structure for treatment group comparisons.

In summary, the response to treatment is not at random relative to the different treatment groups observed at the four clinics. More specifically, treatments AB and B are associated with the highest proportions of hours with little or no pain status; treatment A is associated with higher proportions of hours with little or no pain status than placebo. These differences tend to be roughly uniform across hours such that the multivariate eight hour profile can be adequately summarized by the average response over the eight hours. These randomization results are applicable to the finite population under study only, but do not require any assumptions about a conceptual population they might represent.

The detailed randomization analysis provides a basis for defining an initial model of a WLS analysis. The important sources of variation indicated in this randomization analysis are:

1. clinic
2. initial pain status
3. treatment (incremental effects)
4. treatment by initial pain status interaction.

These sources of variation require a 32 subpopulation framework. With 32 subpopulations, the CPU (central processing unit) space allocation in GENCAT restricts the response profile to two variates. These variates could be the first four hour and second four hour average proportion of patient hours with little or no pain. The orthogonal hourly contrasts in Table 4.7 show little significant information remaining in the eight hour profile beyond the bivariate four hour averages. These results justify summarizing the eight hour profile into the bivariate four hour averages. However, Tables 4.5 and 4.6 show that the bivariate four hour average can be summarized as the eight hour average. This result is based upon re-expressing the bivariate response as the eight hour average and the dose period difference. This dose period difference was found to be inconsistent with regard to treatment comparisons across strata. Thus, the univariate eight hour average proportion of patient hours with little or no pain represents an appropriate response measure for the eight hour multivariate response profile.

4.3 Weighted Least Squares Model Analysis

The weighted least squares analysis for this study is undertaken by viewing the patients as being representative of a general population of patients suffering obstetrical related pain within the first 24 hour period for a vaginal delivery without complications. In this regard, the variation among treatments can be modeled as distinct subpopulations within the eight levels of clinic and initial pain status.

For a resulting $i = 1,2,\ldots,32$ subpopulations, a model of interest (see Section 4.2) describing the variation in the average proportion of hours $\{\bar{y}_i\}$ over the eight hour period that patients experienced little or no pain, involves main effects for clinic, initial pain status and treatment as well as interaction effects for treatment with initial pain status. The model equation for a reference cell parameterization ($\underset{\sim}{X}_1$) shown in Table 4.8 can be written:

$$E\{\bar{y}_i\} = \sum_{k=1}^{11} B_k \, x_{ik}$$

where the $i = 1,2,\ldots,32$ rows of $\underset{\sim}{X}$ and short descriptors of $\underset{\sim}{B}$ are given in Table 4.8. More specifically, $\underset{\sim}{B}$ is the vector of $k = 1,2,\ldots,11$ parameters where

b_{11} is the average proportion with little or no pain associated with the first clinic, some initial pain and placebo treatment;

b_{12} is the increment associated with a lot of pain initially relative to some pain initially;

b_{13} is the increment associated with the second clinic relative to the first clinic;

b_{14} is the increment associated with the third clinic relative to the first clinic;

b_{15} is the increment associated with the fourth clinic relative to the first clinic;

b_{16} is the increment associated with treatment B relative to placebo;

b_{17} is the increment associated with treatment A relative to treatment B;

b_{18} is the increment associated with treatment AB relative to treatment A;

b_{19} is the increment associated with b_{16} and a lot of pain initially relative to b_{16} and some pain initially;

$b_{1,10}$ is the increment associated with b_{17} and a lot of pain initially relative to b_{17} and some pain initially;

$b_{1,11}$ is the increment associated with b_{18} and a lot of pain initially relative to b_{18} and some pain initially.

Because the Wald statistic for goodness of fit is nonsignificant, the model provides an appropriate environment for testing hypotheses concerning clinic effects, initial pain status effect and the interrelationship among the four treatments. Wald statistics for the relevant hypotheses are shown in Table 4.8. These statistics suggest that there are significant clinic variation, significant initial pain status effect and significant interaction between treatment and initial pain status. The parameter estimates involving treatment effects indicate a moderate increase in the proportion of patient hours with little or no pain for treatment B relative to placebo, a slight decrease for treatment A relative to treatment B and a modest increase for treatment AB relative to

TABLE 4.8

ESTIMATED PARAMETERS, STANDARD ERRORS, AND TEST STATISTICS FOR PRELIMINARY MODEL FOR AVERAGE PROPORTION OF PATIENT HOURS WITH LITTLE OR NO PAIN RESPONSE

Specification matrix ($\underset{\sim}{X}_1$)

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}
$$

Parameter Interpretation	WLS parameters estimates ± s.e.'s
Predicted reference value for Placebo for initially some pain patients in center 1 (b_{11})	0.610 ± 0.038
Increment for initially a lot of pain status (b_{12})	−0.209 ± 0.050
Increment for center 2 (b_{13})	0.067 ± 0.023
Increment for center 3 (b_{14})	0.007 ± 0.027
Increment for center 4 (b_{15})	0.120 ± 0.024
Increment for B vs Placebo (b_{16})	0.151 ± 0.042
Increment for A vs B (b_{17})	−0.027 ± 0.033
Increment for AB vs A (b_{18})	0.066 ± 0.032
(B vs Placebo) × Initial status (b_{19})	0.118 ± 0.059
(A vs B) × Initial status ($b_{1,12}$)	−0.132 ± 0.050
(AB vs A) × Initial status ($b_{1\cdot 11}$)	0.141 ± 0.048

Wald test statistics for hypotheses	
No initial status variation ($b_{12} = 0$)	Q(DF = 1) = 17.49
No center variation ($b_{13} = b_{14} = b_{15} = 0$)	Q(DF = 3) = 29.02
No treatment × initial status interaction ($b_{19} = b_{1,10} = b_{1,11} = 0$)	Q(DF = 3) = 12.63
†Model reduction ($b_{17} = b_{19} = + b_{1,10} = b_{19} - b_{1,11} = 0$)	Q(DF = 3) = 0.93
Model goodness of fit	Q(DF = 21) = 26.90

†The model reduction hypothesis corresponds to no A vs B difference no (A vs Placebo) × Initial Status Interaction, and no (AB vs B) × Initial Status Interaction.

treatment A. For patients with a lot of pain initially there exists a second moderate increase for treatment B relative to placebo, a second moderate decrease for treatment A relative to treatment B and a second moderate increase of treatment AB relative to treatment A. A more succinct model describing these treatment effects could be expressed as an increment for treatments A, B, or AB relative to placebo, an increment for treatment B or AB relative to treatment A or placebo for patients with a lot of pain initially, and an increment for treatment AB relative to treatment B. The corresponding model reduction contrast hypothesis shown in Table 4.8 is nonsignificant.

Parameter estimates and tests of hypotheses for the reduced model ($\underset{\sim}{X}_2$) are shown in Table 4.9. Treatment A and B are more effective than placebo in the sense of having significantly higher average proportions of patient hours with little or no pain. For patients experiencing a lot of pain initially, treatment B is significantly more effective than treatment A. Treatment AB can be considered the most effective treatment in that it is significantly more effective than treatment B. Observed and predicted values are listed in Table 4.10 together with their estimated standard errors.

Another aspect of this example which was of interest was the increase in the last four hour average over the first four hour average relative to the proportion of patient hours with little or no pain. Modeling of this quantity showed interactions existed at a fourth order level involving time, clinic, initial pain status and treatment. An initial model with a rank of 20 involved a block diagonal main effects matrix where each block represented a clinic. Within each clinic, the parameters corresponded to a reference value for placebo and initially

TABLE 4.9

ESTIMATED PARAMETERS, STANDARD ERRORS, AND TEST STATISTICS FOR REDUCED MODEL FOR AVERAGE PROPORTION OF PATIENT HOURS WITH LITTLE OR NO PAIN RESPONSE

Specification matrix ($\underset{\sim}{X}_2$)

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 1 & 1
\end{bmatrix}
$$

Parameter Interpretation	WLS parameters estimates ± s.e.'s
Predicted reference value for Placebo for initially some pain patients in center 1 (b_{21})	0.622 ± 0.032
Increment for initially a lot of pain status (b_{22})	-0.229 ± 0.028
Increment for center 2 (b_{23})	0.064 ± 0.022
Increment for center 3 (b_{24})	0.004 ± 0.026
Increment for center 4 (b_{25})	0.117 ± 0.024
Increment for A, B, or AB relative to placebo (b_{26})	0.127 ± 0.030
Increment for B or AB relative to A or placebo for initially a lot of pain patients (b_{27})	0.151 ± 0.031
Increment for AB relative to B (b_{28})	0.051 ± 0.020

Wald test statistics for hypotheses

Hypothesis	Statistic
No initial status variation ($b_{22} \hat{=} 0$)	Q(DF = 1) = 66.76
No center variation ($b_{23} \hat{=} b_{24} \hat{=} b_{25} \hat{=} 0$)	Q(DF = 3) = 28.56
A vs Placebo ($b_{26} \hat{=} 0$)	Q(DF = 1) = 18.26
AB vs B ($b_{28} \hat{=} 0$)	Q(DF = 1) = 6.65
B vs A for initially a lot of pain patients ($b_{27} \hat{=} 0$)	Q(DF = 1) = 23.29
Model goodness of fit	Q(DF = 24) = 27.83

some pain, an increment for initially a lot of pain, an increment for treatment B relative to placebo, an increment for treatment A relative to treatment B and finally, an increment for treatment AB relative to treatment A. Using the model predicted values together with the parameter estimates and standard errors, the model was reduced to having the following (32 x 2) design matrix:

$$\underset{\sim}{X}_3' = \begin{bmatrix} 1212 & 2323 & 1111 & 1222 & 1111 & 1111 & 1111 & 1111 \\ 0000 & 0000 & 0000 & 0000 & 0000 & 0000 & 1111 & 1000 \end{bmatrix}$$

where the first parameter represents a complex clinic x initial pain status x treatment interaction and the second parameter represents a decrement for the fourth clinic for patients with some pain initially or a lot of pain and placebo. The goodness of fit statistic was not significant ($Q = 30$, d.f. $= 30$); the observed and predicted values are listed in Table 4.10 along with their estimated standard errors.

In view of the complex interactions found concerning the increase in average proportion of patient hours with little or no pain for the last four hours beyond that of the first four hours, the multivariate model involving both the eight hour average proportion and the increase in the last four hour average was considered as a more complete framework for evaluating treatment group differences. Also, this provides predicted values for not only the eight hour average proportions of patient hours with little or no pain, but also the first four hour average proportions and last four hour average proportions. For this purpose, patient proportions of the eight hour period with little or no pain and the difference between last four hour proportion and first four hour proportion were submitted to MISCAT. The resulting response vector

TABLE 4.10

OBSERVED ESTIMATES AND PREDICTED VALUES FOR MODELS FOR AVERAGE PROPORTIONS OF PATIENT HOURS AND DIFFERENCES WITH LITTLE OR NO PAIN RESPONSE FOR 8 HOUR EVALUATION PERIODS IN POST-PARTUM PAIN CLINICAL TRIAL

Subpopulations			Number	Average proportions little or no pain				Average proportions little or no pain difference for hours 1-4 vs. hours 5-8			
Center	Initial pain status	Treatment	of subjects	Observed estimates	s.e.'s	Model X_2 estimates	predicted s.e.'s	Observed estimates	s.e.'s	Model X_1 estimates	predicted s.e.'s
1	Some	Placebo	18	0.639	0.067	0.622	0.032	0.194	0.112	0.158	0.007
1	Some	B	19	0.776	0.033	0.749	0.020	0.342	0.056	0.317	0.014
1	Some	A	18	0.611	0.072	0.749	0.020	0.250	0.073	0.158	0.007
1	Some	AB	19	0.822	0.035	0.800	0.021	0.250	0.059	0.317	0.014
1	A lot	Placebo	26	0.466	0.063	0.394	0.031	0.279	0.061	0.317	0.014
1	A lot	B	28	0.580	0.051	0.671	0.023	0.393	0.053	0.475	0.021
1	A lot	A	28	0.429	0.057	0.520	0.029	0.286	0.090	0.317	0.014
1	A lot	AB	26	0.740	0.028	0.722	0.020	0.500	0.073	0.475	0.021
2	Some	Placebo	20	0.644	0.072	0.686	0.032	0.188	0.070	0.158	0.007
2	Some	B	20	0.775	0.059	0.813	0.020	0.100	0.069	0.158	0.007
2	Some	A	19	0.809	0.041	0.813	0.020	0.250	0.074	0.158	0.007
2	Some	AB	10	0.800	0.042	0.864	0.023	0.150	0.094	0.158	0.007
2	A lot	Placebo	26	0.404	0.067	0.458	0.032	0.154	0.071	0.158	0.007
2	A lot	B	26	0.774	0.030	0.735	0.021	0.394	0.082	0.317	0.014
2	A lot	A	26	0.644	0.057	0.584	0.030	0.365	0.051	0.317	0.014
2	A lot	AB	26	0.798	0.036	0.786	0.021	0.288	0.044	0.317	0.014
3	Some	Placebo	23	0.609	0.076	0.627	0.034	0.152	0.065	0.158	0.007
3	Some	B	26	0.769	0.065	0.754	0.024	0.058	0.046	0.158	0.007
3	Some	A	22	0.722	0.053	0.754	0.024	0.216	0.040	0.158	0.007
3	Some	AB	25	0.855	0.037	0.804	0.024	0.150	0.057	0.158	0.007
3	A lot	Placebo	22	0.432	0.068	0.398	0.033	0.250	0.074	0.158	0.007
3	A lot	B	26	0.611	0.075	0.675	0.029	0.125	0.046	0.158	0.007
3	A lot	A	22	0.477	0.074	0.525	0.032	0.114	0.066	0.158	0.007
3	A lot	AB	26	0.668	0.066	0.726	0.027	0.240	0.048	0.158	0.007
4	Some	Placebo	19	0.737	0.071	0.740	0.033	0.079	0.053	0.050	0.024
4	Some	B	20	0.881	0.040	0.867	0.020	0.062	0.046	0.050	0.024
4	Some	A	19	0.895	0.041	0.867	0.020	0.079	0.056	0.050	0.024
4	Some	AB	21	0.881	0.051	0.917	0.025	0.048	0.040	0.050	0.024
4	A lot	Placebo	16	0.469	0.087	0.511	0.033	-0.156	0.103	0.050	0.024
4	A lot	B	22	0.801	0.040	0.788	0.024	0.216	0.078	0.158	0.007
4	A lot	A	22	0.676	0.060	0.638	0.031	0.216	0.049	0.158	0.007
4	A lot	AB	22	0.739	0.064	0.839	0.025	0.136	0.060	0.158	0.007

was permuted such that the first 32 functions pertained to the eight hour average proportion for the 32 distinct subpopulations and the second 32 functions pertained to the corresponding first four hours versus second four hours differences. The multivariate model is simply a block diagonal matrix, shown in Table 4.11, with a final model ($\underset{\sim}{X}_2$) for the eight hour average proportions as the first block and the final model ($\underset{\sim}{X}_3$) for the differences as the second block. The goodness of fit test statistic indicates that the model adequately describes the variation among subpopulations. Furthermore, the standard errors for predicted values of the eight hour averages and the four hour average differences shown in Table 4.12, are smaller than those from the corresponding separate univariate models shown in Table 4.10.

The multivariate model indicates that treatments A and B are more effective than placebo in the sense of having significantly greater average proportions of patient hours with little or no pain; for patients experiencing a lot of pain initially, treatment B is significantly more effective than treatment A. Treatment AB shows a slight increase in effectiveness relative to treatment B, which is reasonably suggestive ($p = 0.08$) even though not strictly significant. Therefore, it can be concluded that treatments B and AB are the most effective for patients suffering from pain within the 24-hour period following vaginal delivery.

4.4 Discussion

Both the randomization analysis and weighted least squares analysis showed that treatments AB and B are most effective for patients suffering obstetrical pain with treatment AB tending to be slightly more effective than B. Treatment A, although less effective than B or AB was found to be more effective than placebo. The randomization results are directed

TABLE 4.11

ESTIMATED PARAMETERS, STANDARD ERRORS, AND TEST STATISTICS FOR MULTIVARIATE MODEL FOR AVERAGE PROPORTION OF PATIENTS WITH LITTLE OR NO PAIN RESPONSE RELATIVE TO THE 8 HOUR AVERAGE PROPORTION OF PATIENT HOURS AND THE INCREASE IN THE LAST 4 HOUR AVERAGE OVER THE FIRST 4 HOURS

Specification matrix ($\underset{\sim}{X}_4$)	Parameter interpretation	WLS parameters estimates s.e.'s
$\begin{bmatrix} \underset{\sim}{X}_2 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{X}_3 \end{bmatrix}$ 64×10	Predicted reference value for Placebo for initially some pain patients in center 1	0.638 ± 0.030
	Increment for initially a lot of pain status	-0.233 ± 0.026
	Increment for center 2	0.068 ± 0.017
	Increment for center 3	0.007 ± 0.022
	Increment for center 4	0.112 ± 0.021
	Increment for A relative to Placebo	0.122 ± 0.028
	Increment for B relative to A for initially a lot of pain patients	0.176 ± 0.028
	Increment for AB relative to B	0.030 ± 0.017
	Initial pain status × center × treatment increase in last 4 hours over first 4 hours	0.169 ± 0.061
	Increment in increase in last 4 hours over first 4 hours for center 4 with active treatment for some pain patients and center 4 with placebo	-0.115 ± 0.024

Wald test statistics for hypotheses

A vs placebo	Q(DF = 1) = 18.82
AB vs B	Q(DF = 1) = 3.06
B vs A for initially a lot of pain patients	Q(DF = 1) = 38.62
Model goodness of fit	Q(DF = 54) = 66.53

TABLE 4.12

MULTIVARIATE MODEL PREDICTED VALUES FOR AVERAGE PROPORTION OF PATIENT HOURS WITH LITTLE OR NO PAIN RESPONSE FOR 8 HOUR EVALUATION PERIODS IN POST-PARTUM PAIN CLINICAL TRIAL

Center	Initial Pain Status	Treatment	Number of Patients	8 hour average		Increase in last 4 hour average over first 4 hour average		First 4 hour average		Last 4 hour average	
				Estimates	s.e.'s	Estimates	s.e.'s	Estimates	s.e.'s	Estimates	s.e.'s
1	Some	Placebo	18	0.638	0.030	0.169	0.006	0.554	0.060	0.723	0.060
1	Some	B	19	0.761	0.018	0.338	0.012	0.592	0.039	0.929	0.037
1	Some	A	18	0.761	0.018	0.169	0.006	0.676	0.037	0.845	0.036
1	Some	AB	19	0.790	0.016	0.338	0.012	0.621	0.038	0.959	0.031
1	A lot	Placebo	26	0.405	0.027	0.338	0.012	0.237	0.055	0.574	0.057
1	A lot	B	28	0.703	0.018	0.506	0.018	0.450	0.045	0.956	0.035
1	A lot	A	28	0.527	0.026	0.338	0.012	0.359	0.053	0.696	0.055
1	A lot	AB	26	0.733	0.012	0.506	0.018	0.480	0.040	0.986	0.017
2	Some	Placebo	20	0.707	0.030	0.169	0.006	0.622	0.060	0.791	0.060
2	Some	B	20	0.829	0.016	0.169	0.006	0.744	0.033	0.913	0.033
2	Some	A	19	0.829	0.016	0.169	0.006	0.744	0.033	0.913	0.033
2	Some	AB	20	0.858	0.021	0.169	0.006	0.774	0.044	0.943	0.042
2	A lot	Placebo	26	0.474	0.029	0.169	0.006	0.389	0.058	0.558	0.060
2	A lot	B	26	0.772	0.016	0.338	0.012	0.603	0.037	0.940	0.032
2	A lot	A	26	0.596	0.027	0.338	0.012	0.427	0.054	0.764	0.057
2	A lot	AB	26	0.801	0.018	0.338	0.012	0.632	0.044	0.970	0.035
3	Some	Placebo	23	0.646	0.032	0.169	0.006	0.561	0.064	0.730	0.065
3	Some	B	26	0.768	0.022	0.169	0.006	0.683	0.044	0.852	0.044
3	Some	A	22	0.768	0.022	0.169	0.006	0.683	0.044	0.852	0.044
3	Some	AB	25	0.797	0.021	0.169	0.006	0.713	0.044	0.882	0.042
3	A lot	Placebo	22	0.412	0.031	0.169	0.006	0.328	0.062	0.497	0.064
3	A lot	B	26	0.710	0.026	0.169	0.006	0.626	0.053	0.795	0.051
3	A lot	A	22	0.534	0.030	0.169	0.006	0.450	0.060	0.619	0.062
3	A lot	AB	26	0.740	0.023	0.169	0.006	0.656	0.049	0.824	0.046
4	Some	Placebo	19	0.750	0.031	0.054	0.024	0.723	0.069	0.777	0.065
4	Some	B	20	0.872	0.018	0.054	0.024	0.845	0.047	0.899	0.041
4	Some	A	19	0.872	0.018	0.054	0.024	0.845	0.047	0.899	0.041
4	Some	AB	21	0.902	0.023	0.054	0.024	0.875	0.054	0.928	0.049
4	A lot	Placebo	16	0.517	0.031	0.054	0.024	0.490	0.068	0.544	0.066
4	A lot	B	22	0.815	0.020	0.169	0.006	0.730	0.042	0.899	0.041
4	A lot	A	22	0.639	0.029	0.169	0.006	0.554	0.057	0.723	0.059
4	A lot	AB	22	0.844	0.022	0.169	0.006	0.760	0.047	0.929	0.044

at the average partial association of treatments with response across the finite population cross-classification of clinic and initial pain status. The weighted least squares results are directed at the treatment group differences relative to a model of clinic, initial pain status and treatment group variation. Also, this model is viewed as being representative of the variation in some larger population from which these data are conceptually sampled.

The tests of hypotheses undertaken with the randomization model provided a rationale useful in constructing the initial weighted least squares model. Separate weighted least squares models were developed for the eight hour average and the difference between the two dosing periods. The multivariate model for the eight hour average and the dosing period difference was simply a combination of the two separate models with one model as the first block and the second model as another block forming a two block diagonal design matrix. In actuality, the multivariate profile of the eight hour average and the dosing period difference is equivalent to the multivariate profile of the two four hour dosing periods. However, with the complex interactions between the two periods, an equally parsimonious model for the two dosing periods would have been more difficult to specify. The weighted least squares multivariate model provided predicted values and standard errors for the $i = 1,2,\ldots,32$ subpopulations.

The randomization analysis strategy consisted of viewing the hourly responses univariately for the four treatment comparison involving tests with three degrees of freedom and the six sets of paired-treatment comparisons involving tests with one degree of freedom. The eight hourly response profile was also examined simultaneously involving a test with 24 degrees of freedom for the four treatment comparison and tests with eight degrees of freedom for the set of six paired-treatment comparisons.

A more parsimonious multivariate analysis consisted of the average response across the first four hours (the first dosing period) and an average response across the second four hours (the second dosing period). This analysis involved tests with six degrees of freedom for the four treatment group comparison and tests with two degrees of freedom for the set of six paired-treatment comparisons.

The two dosing period averages were also expressed as the overall average and the difference between the two periods. If the difference between the two periods was at random across treatment groups, then the eight hour average would be sufficient for comparing the treatment groups. For this purpose, the eight hour average and the dosing period difference were examined univariately. Although there were two paired-treatment comparisons that were significant ($p < .05$) for the dosing period difference, dosing period differences were not found to have a consistent pattern of association with any of the four treatments. An additional analysis of hourly differences (orthogonal hourly contrasts) within the dosing periods indicate that treatment group differences are roughly comparable from hour to hour, particularly within dosing periods. The results of the eight hour average comparisons of the four treatment and the set of six paired-treatment comparisons were similar to the univariate and multivariate results for the two dosing periods.

The randomization tests of hypothesis for comparing the four treatments involved the overall comparison as well as a set of six paired-treatment comparisons. The variance structure for each comparison varied in that the overall comparison included all the data and each paired-treatment comparison included data for the pair under consideration only. The weighted least squares model provided a unified covariance framework within which paired comparisons were modeled. The weighted least squares tests of hypothesis concerning treatment group comparisons involved testing linear combinations of the model parameters.

The randomization analysis provided tests of significance within a stratification system which provided a precise covariance structure in the sense of matched comparisons without having to describe the interrelationship of the strata. However, the within-stratum results identified a pattern for this interrelationship of the strata. The WLS analysis clarified the interrelationship of strata in terms of main effects and interactions among clinics, initial pain status, treatment and time. Within this completely specified model, the WLS analysis provided tests of significance and predicted values with estimated standard errors.

CHAPTER V

AN APPLICATION OF STRATIFICATION TECHNIQUES

The example in this chapter illustrates the average partial association statistics presented in Section 2.3.1 of Chapter II for the two subpopulation case. Moreover, various aspects of cumulative stratification analyses are discussed with regard to an ordering of ten strata. These strata correspond to a cross-classification of clinic and baseline level within which patients are viewed as randomly assigned to one of three treatments, two of which are control treatments. The control treatments are used with the cumulative stratification strategy to provide an aggressive evaluation of an investigational treatment as an antidepressant drug. For this purpose, the analysis includes the generalized randomized block chi-square statistic, the average standard normal statistic and the Fisher combined p-value statistic.

5.1 Psychopharmacology Example

This example is concerned with data from a series of double blind, parallel group randomized clinical trials employing six centers to explore the efficacy of an investigational treatment for depression in comparison to a placebo control treatment and an active control treatment. In all, 393 patients were involved, and these were randomly assigned to three treatments on a within clinic basis. The questions of interest for this example pertain to whether the investigational treatment is more effective than placebo and equally as effective as the active control. In this regard, the extent to which the active control is more effective then placebo in a particular clinic can be interpreted as an indicator of the extent to which the investigational treatment might be expected to exceed the placebo level of effectiveness. More specifically, clinics with the active control showing more pronounced increases in effectiveness over placebo are more stringent in testing the efficacy of the investigational treatment. Contrarily, clinics with the active control showing less effectiveness than placebo are possibly less relevant in testing the efficacy of the investigational treatment in the sense that the placebo response is potentially at an unexpectedly more favorable level or that the corresponding group of patients are possibly atypical.

The response variable which was used to assess antidepressant efficacy was the proportion of patients with at least 50% improvement in the final Hamilton[1] total score over baseline. As will be indicated later, baseline severity was not uniform across the treatments and thus needed to be taken into account. In particular, a significantly ($p < .05$) higher proportion of placebo patients had less severe depression at baseline than those with the investigational or active control treatments even though there was randomized assignment.

5.2 Randomization Model Analysis

Since both clinic and baseline Hamilton total were significant sources of variation in the analysis of the final Hamilton total score,

---

[1]M. Hamilton. "Rating Scales in Depression." Depressive Illness: Diagnosis, Assessment, Treatment. Edited by P. Kielholz. Baltimore, Md.: Williams and Wilkins, 1972.

the randomization framework consists of ten strata. These strata are defined by the cross-classification of four clinics with less severe versus severe levels of baseline Hamilton total together with a classification of the remaining two clinics which have patients predominantly at the severe level of the baseline Hamilton total. Within each stratum, patients can be considered randomly assigned to treatment both in design with regard to the within-clinic randomization and in post-stratification with regard to baseline severity since the baseline was known before treatment assignment. Thus, for the ten strata, it is expected that patients are homogeneous relative to their chances of improved outcome under the null hypothesis of no treatment differences. In other words, the partition of responses between paired treatment groups in each clinic-baseline status stratum is equivalent to a stratified random sample with fixed sample sizes $\{n_{h_1}, n_{h_2}\}$ from the pooled samples $\{n_{h_1} + n_{h_2}\}$ in accordance with the product multiple hypergeometric model.

Two issues that determine the structure of the randomization analysis here are as follows:

1. The tests pertaining to the comparison of the active treatments with placebo are one-sided so as to provide greater power against alternatives in the direction of the more favorable outcomes expected for them.
2. Strata can be ordered according to the respective strengths of the comparison of the active control with placebo.

In view of these issues, randomization statistics for the paired-treatment comparisons include the generalized randomized block chi-square test statistic, the average standard normal statistic and the Fisher combined p-value statistic presented in Section 2.3.1 of Chapter II. The average standard normal statistic and the Fisher combined p-value statistic are directed specifically toward assessing the efficacy probability. In particular, the across-strata average randomization form of these statistics is used to show the extent to which the investigational treatment versus placebo comparisons are strengthened or weakened as successive strata with weaker active control efficacy are included.

The distribution of 50% improvement on the Hamilton total score is shown in Table 5.1 for patients with respect to their treatment assignment within each clinic-baseline status stratum. Within-stratum active control efficacy p-values for the proportion of patients with at least 50% Hamilton improvement determine the relative ordering of strata for the investigational treatment versus placebo comparisons. This relative ordering of strata, also shown in Table 5.1, is determined separately for direct comparisons and baseline-adjusted comparisons. Two strata show patients taking placebo as having a higher proportion with 50% improvement than patients taking the active control for both the direct comparison of active control to placebo and the baseline-adjusted comparison. These two strata are ordered last in both the direct and baseline-adjusted sets of analyses according to the ordering system.

The across-strata statistics for the direct and baseline-adjusted comparisons of active control versus placebo, investigational treatment versus placebo and investigational treatment versus active control are shown in Table 5.2 through 5.7, utilizing the corresponding stratum-ordering identified in Table 5.1. More specifically, p-values for direct comparisons involving average partial association chi-square statistics, average standard normal statistics and combined Fisher p-values are

TABLE 5.1

DISTRIBUTION OF THE 50% HAMILTON IMPROVEMENT INDICATOR FOR PATIENTS CROSS-CLASSIFIED ACCORDING TO TREATMENT AND THE STRATA FRAMEWORK, AND THE STRATIFICATION ORDERING RELATIVE TO DIRECT AND COVARIANCE ADJUSTED WITHIN STRATUM COMPARISONS FOR ACTIVE CONTROL VS. PLACEBO

Clinic	Baseline Hamilton Total	Treatment*	Final Hamilton Improvement Number ≥50%	Number <50%	Proportion ≥50%	Direct Analyses A vs P p-value (order)	Adjusted Analyses A vs P p-value (order)
A	<27	Invest Active Placebo	14 16 13	9 14 16	.61 .53 .45	.259 (3)	.241 (2)
A	≥27	Invest Active Placebo	5 2 4	5 1 3	.50 .67 .57	.395 (6)	.384 (7)
B	All	Invest Active Placebo	2 2 1	5 6 7	.29 .25 .13	.268 (4)	.341 (5)
C	<27	Invest Active Placebo	13 13 13	8 5 9	.62 .72 .59	.196 (2)	.242 (3)
C	≥27	Invest Active Placebo	11 10 12	5 4 2	.69 .71 .86	.817 (10)	.849 (10)
D	<27	Invest Active Placebo	1 2 2	2 4 5	.33 .33 .29	.429 (7)	.349 (6)
D	≥27	Invest Active Placebo	1 1 1	7 4 5	.13 .20 .17	.446 (8)	.433 (8)
E	<27	Invest Active Placebo	12 11 15	7 6 12	.63 .65 .56	.276 (5)	.288 (4)
E	≥27	Invest Active Placebo	12 6 5	8 7 4	.60 .46 .56	.664 (9)	.572 (9)
F	All	Invest Active Placebo	7 6 1	0 2 6	1.00 .75 .14	.012 (1)	.008 (1)

* Invest, Investigational Treatment

Active, Active Control

TABLE 5.2

TWO-SIDED P-VALUES FROM THE 10 STRATA FRAMEWORK FOR SUCCESSIVE ACROSS STRATA AVERAGE PARTIAL ASSOCIATION RANDOMIZATION CHI-SQUARE STATISTICS FOR PAIRWISE TREATMENT COMPARISONS CONCERNING 50% HAMILTON IMPROVEMENT - NO COVARIANCE ADJUSTMNET

Stratum			50% Hamilton Total Improvement		
Order†	Investigator	Hamilton Baseline	A vs P	I vs P	I vs A
1	F	All	.023 (A)	.002 (I)	.170 (I)
2	C	<27	.050 (A)	.080 (I)	.967 (A)
3	A	<27	.070 (A)	.040 (I)	.707 (I)
4	B	≥27	.055 (A)	.029 (I)	.686 (I)
5	E	<27	.050 (A)	.033 (I)	.760 (I)
6	A	≥27	.049 (A)	.053 (I)	.865 (I)
7	D	<27	.052 (A)	.054 (I)	.868 (I)
8	D	≥27	.053 (A)	.065 (I)	.927 (I)
9	E	≥27	.090 (A)	.069 (I)	.694 (I)
10	C	≥27	.173 (A)	.160 (I)	.752 (I)
9*	E	≥27		.045 (I)	
10*	C	≥27		.065 (I)	

A Favors Active Control; I Favors Investigational Treatment;
P Favors Placebo

† The strata are ordered according to within stratum p-values for A vs P for 50% Hamilton Improvement with the first eight favoring A and the last two favoring P. For each order level, the p-values correspond to average partial association statistics across all strata which occur at or before that order position.

* These results pertain to the comparison of active control vs placebo for strata 1-8 and investigational treatment for strata 9-10.

TABLE 5.3

ONE-SIDED P-VALUES FROM THE 10 STRATA FRAMEWORK FOR SUCCESSIVE ACROSS STRATA AVERAGE WITHIN STRATUM RANDOMIZATION STANDARD NORMAL STATISTICS FOR PAIRWISE TREATMENT COMPARISONS CONCERNING 50% HAMILTON IMPROVEMENT - NO COVARIANCE ADJUSTMENT

Stratum			50% Hamilton Total Improvement		
Order†	Investigator	Hamilton Baseline	A vs P	I vs P	I vs A
1	F	All	.012 (A)	.001 (I)	.085 (I)
2	C	<27	.014 (A)	.010 (I)	.310 (I)
3	A	<27	.015 (A)	.005 (I)	.236 (I)
4	B	$\geq$27	.014 (A)	.005 (I)	.234 (I)
5	E	<27	.013 (A)	.005 (I)	.281 (I)
6	A	$\geq$27	.016 (A)	.013 (I)	.370 (I)
7	D	<27	.020 (A)	.018 (I)	.380 (I)
8	D	$\geq$27	.024 (A)	.029 (I)	.435 (I)
9	E	$\geq$27	.043 (A)	.032 (I)	.341 (I)
10	C	$\geq$27	.090 (A)	.077 (I)	.367 (I)
9*	E	$\geq$27		.021 (I)	
10*	C	$\geq$27		.030 (I)	

A Favors Active Control; I Favors Investigational Treatment;
P Favors Placebo

† The strata are ordered according to within stratum p-values for A vs P for 50% Hamilton Improvement with the first eight favoring A and the last two favoring P. For each order level, the p-values correspond to average partial association statistics across all strata which occur at or before that order position.

* These results pertain to the comparison of active control vs placebo for strata 1-8 and investigational treatment for strata 9-10.

TABLE 5.4

ONE-SIDED P-VALUES FROM THE 10 STRATA FRAMEWORK FOR SUCCESSIVE ACROSS STRATA FISHER COMBINED P-VALUES FOR PAIRWISE TREATMENT COMPARISONS CONCERNING 50% HAMILTON IMPROVEMENT - NO COVARIANCE ADJUSTMENT

Stratum			50% Hamilton Total Improvement		
Order†	Investigator	Hamilton Baseline	A vs P	I vs P	I vs A††
1	F	All	.012 (A)	.001 (I)	.085 (I)
2	C	<27	.016 (A)	.003 (I)	.239 (I)
3	A	<27	.021 (A)	.003 (I)	.241 (I)
4	B	≥27	.025 (A)	.004 (I)	.294 (I)
5	E	<27	.028 (A)	.005 (I)	.370 (I)
6	A	≥27	.038 (A)	.010 (I)	.479 (I)
7	D	<27	.051 (A)	.015 (I)	.528 (0)
8	D	≥27	.065 (A)	.025 (I)	.607 (0)
9	E	≥27	.098 (A)	.031 (I)	.530 (0)
10	C	≥27	.150 (A)	.056 (I)	.584 (0)
9*	E	≥27		.023 (I)	
10*	C	≥27		.033 (I)	

A Favors Active Control; I Favors Investigational Treatment;
P Favors Placebo; 0†† No favoring;

† The strata are ordered according to within stratum p-values for A vs P for 50% Hamilton Improvement with the first eight favoring A and the last two favoring P. For each order level, the p-values correspond to average partial association statistics across all strata which occur at or before that order position.

†† When these p-values exceed 0.500, investigational treatment is no longer favored; but this does not necessarily imply that active control is favored because of the manner in which such one-sided test statistics are constructed.

* These results pertain to the comparison of active control vs placebo for strata 1-8 and investigational treatment for strata 9-10.

TABLE 5.5

TWO-SIDED P-VALUES FROM THE 10 STRATA FRAMEWORK FOR SUCCESSIVE ACROSS STRATA AVERAGE PARTIAL ASSOCIATION RANDOMIZATION CHI-SQUARE STATISTICS FOR PAIRWISE TREATMENT COMPARISONS CONCERNING 50% HAMILTON IMPROVEMENT COVARIANCE ADJUSTMENT FOR BASELINE HAMILTON TOTAL

Stratum			50% Hamilton Total Improvement		
Order†	Investigator	Hamilton Baseline	A vs P	I vs P	I vs A
1	F	All	.016 (A)	.002 (I)	.276 (I)
2	A	<27	.083 (A)	.012 (I)	.522 (I)
3	C	<27	.069 (A)	.045 (I)	.943 (I)
4	E	<27	.065 (A)	.036 (I)	.938 (I)
5	B	All	.059 (A)	.025 (I)	.954 (I)
6	D	<27	.054 (A)	.027 (I)	.962 (A)
7	A	$\geq$27	.052 (A)	.043 (I)	.867 (A)
8	D	$\geq$27	.053 (A)	.043 (I)	.825 (A)
9	E	$\geq$27	.076 (A)	.054 (I)	.930 (I)
10	C	$\geq$27	.163 (A)	.130 (I)	.973 (I)
9*	E	$\geq$27		.032 (I)	
10*	C	$\geq$27		.048 (I)	

A Favors Active Control; I Favors Investigational Treatment;
P Favors Placebo

† The strata are ordered according to within stratum p-values for A vs P for 50% Hamilton Improvement with the first eight favoring A and the last two favoring P. For each order level, the p-values correspond to average partial association statistics across all strata which occur at or before that order position.

* These results pertain to the comparison of active control vs placebo for strata 1-8 and investigational treatment for strata 9-10.

are given in Tables 5.2, 5.3 and 5.4 respectively; the p-values for the corresponding baseline-adjusted set of comparisons are given in Tables 5.5, 5.6 and 5.7. In each table, the p-value for a stratum is the across-strata result based on that stratum and all previously listed strata. Statistical significance is defined as two-tail p-values less than .05 and one-tail p-values less than .025.

Because patients in the last two strata in both the direct and baseline-adjusted analyses have an unusually more favorable response to placebo than the active control, the investigational treatment is not necessarily likely to show a more favorable response to placebo there. An alternative analysis of efficacy involves the replacement of placebo with active control for these strata. Thus, the across-strata statistics for the last two strata are based on results from the investigational treatment versus placebo comparison of the first eight strata together with results from the investigational treatment versus active control comparison of the ninth strata and the ninth and tenth strata respectively, as shown at the bottom of the investigational treatment versus placebo comparison in each table.

Relative to all across-strata comparisons, the average standard normal statistic shows more ability in detecting differences than the average partial association chi-square statistic in the sense that its one-tail p-value is predominantly less than half the respective two-tail p-value. Furthermore, the average standard normal statistic shows more ability in detecting differences than the combined Fisher p-value. Although both the standard normal and the combined Fisher p-value pertain to the efficacy tail, their orientation is different. More specifically, the average standard normal p-value is still the complement of the opposing tail in that the summation of both tails adds to unity. In contrast, the summation of the combined Fisher p-value for the one-tail with its opposing tail combined Fisher p-value may not add to unity and can exceed unity. Furthermore, the combined Fisher p-value shows monotonic increases in the p-value for the active control versus placebo comparison, and quickly loses significance ($p > .025$) as each successive stratum is included in the across-stratum formulation. The corresponding average standard normal and average partial association chi-square statistics fluctuate with the average partial association statistic quickly losing significance ($p > .05$) in contrast to the average standard normal maintaining significance ($p < .025$) over the first eight strata.

The orientation of the combined Fisher p-value to show the extent to which each successive stratum contributes larger or smaller p-values to the combination of across-strata p-values is useful in justifying the stratum ordering relative to the investigational treatment versus placebo comparison. More specifically, the strata were ordered according to expected strengths of efficacy (over placebo). Because the combined Fisher p-values show a monotonic increase for the investigational treatment versus placebo comparison with each successive strata being included in the across-strata result, the _a priori_ assumption that the strength of the active control efficacy (over placebo) reflects the strength of the investigational treatment efficacy (over placebo) has been substantiated.

For the pairwise comparison of the investigational treatment and placebo, the baseline-adjusted statistics in Tables 5.5, 5.6 and 5.7, are more sensitive to detecting differences than the unadjusted statistics in Tables 5.2, 5.3 and 5.4. The average partial association of the baseline-adjusted comparison is significant ($p < .05$) through the first eight strata and approach significance ($p = .054$) when the ninth is also included. The average standard normal and combined Fisher p-values show somewhat stronger evidence in favor of the investigational treatment over placebo with significance ($p < .025$) being maintained through the first nine strata. However, when the tenth stratum with its surprisingly large placebo response is included, the tendency to favor the investigational treatment is no longer significant. With the modified comparison, having exchanged active control for the placebo as the control drug in the last two strata, the investigational treatment is shown to be efficacious by all three statistics across the ten strata.

The p-values for the active control versus placebo comparisons are mostly larger than the investigational treatment versus placebo for both the unadjusted and the baseline-adjusted comparisons. The baseline adjustment increases the sensitivity slightly in detecting differences. The results of both the unadjusted and baseline-adjusted comparisons are significant or nearly significant through the first eight strata.

Finally, none of the p-values for either unadjusted or baseline-adjusted comparisons of the investigational treatment versus the active control approached significance. The baseline adjustment increased the p-values showing a weaker difference between the investigational treatment and active control. These results supported the general equivalence of these treatments with respect to Hamilton improvement.

5.3 Discussion

After adjusting for differences in the baseline severity, the investigational treatment was found to have a significant antidepressant effect over the response to placebo across the first nine strata. With the active control comparison in the last two strata, these results were significant across all ten strata. The results for the investigational treatment versus placebo comparison appeared stronger in the sense of smaller p-values than the results for the active control versus placebo comparison. However, the investigational treatment was not found to be significantly different from the active control. The unadjusted results showed weaker association between depression improvement and treatment.

The generalized randomized block (average partial association) chi-square statistic, the average standard normal statistics and the Fisher combined p-value statistic were examined cumulatively across an ordered set of strata. This ordering of strata determined by the strength of their respective active control versus placebo comparison was found to coincide with the strength ordering of the investigational treatment versus placebo comparison. The cumulative average partial association statistics for this ordering was useful in specifying an alternative analysis to the investigational treatment versus placebo comparison which involved the last two strata with unusually strong placebo response to treatment. This alternative analysis defines the comparison as being relative to a potentially more appropriate control treatment framework. Because the active control is considered efficacious as an antidepressant, its substitution for placebo in the last two strata for the comparison with the investigational treatment is viewed appropriate in the sense that these two strata showed unexpectedly greater placebo effectiveness over the active control.

For this example, the average standard normal statistics showed more ability in detecting differences than the average partial association chi-square statistic in the sense that its one-tail p-value was predominantly less than half the respective two-tail p-value. Furthermore, the average standard normal statistic showed more ability in detecting differences than the Fisher combined p-value statistic.

Thus, for this type of example for which there was some variation in the extent of treatment differences across the respective strata, the average standard normal statistic may be preferable to the other methods discussed here.

CHAPTER VI

RANDOMIZATION METHODS FOR AN OBSERVATIONAL STUDY

Post stratification methods are used to construct homogeneous strata within which the two study subpopulations are compared. These populations correspond to two distinct groups of patients who were treated in different locations and time periods with an approved regimen. Three physiological measures of liver dysfunction are examined under the hypothesis of randomization. Different scores are used to examine one of these measures which had a right skewed distribution. The different scores showed somewhat different results across the four strata, partially because of their particular across-strata weighting of within-stratum statistics.

6.1 Acetaminophen Overdose Example

The data for this example pertain to the comparison of health outcomes across two different studies for which patients with acetaminophen overdose were treated with appropriate clinical procedures at the corresponding time periods. One study involved patients from throughout the United States. It began in 1976 and was coordinated by the Rocky Mountain Poison Center in Denver, Colorado. In this study, eligible patients were treated with NAC (N-acetylcysteine) and supportive therapy after standard emergency procedures were given. The other study involved a single center in the United Kingdom during an earlier time period. In this latter study, patients were treated with standard emergency procedures followed with supportive therapy. Comparability of the two studies was based on the availability of reasonably equivalent methods for judging patient risk.

In both studies, detailed liver function measures were obtained regularly in addition to assays for the amount of acetaminophen (APAP) in the blood stream. Over the initial 72 hour critical period, the liver is expected to show progressively reduced function. More specifically, bilirubin absorption by the liver is retarded, allowing an accumulation of bilirubin in the blood stream; liver cells are damaged such that metabolic enzymes, e.g. SGOT, are released into the blood stream; and clotting factor synthesis in the liver is retarded, prolonging the prothrombin clotting time. As the body recovers from the APAP toxicity, these functions return showing reduced blood levels of APAP, bilirubin and SGOT, and normal prothrombin time ratios. In this regard, NAC efficacy can be evaluated in terms of the extent to which NAC treated patients at high risk to hepatotoxicity show less extreme values of serum bilirubin, SGOT and prothrombin time ratio than supportive therapy only patients. For this purpose, high risk patients were identified as those patients whose APAP blood concentrations remained at a toxic level over the first twelve hours. Furthermore, bonafide NAC patients had to have received the first dose before the sixteenth hour since the overdose and been administered a minimal regimen of seventeen doses.

6.2 Randomization Model Analysis

Because patients were not randomly assigned to treatments, the randomization model corresponds to an induced randomization framework by hypothesis relative to homogeneous subpopulations for risk of hepatotoxicity; in particular, the hypothesis is that patients are expected to have similar medical experiences if the treatments are equally effective. Risk of hepatotoxicity is taken into account by matching high risk patients according to moderately toxic or highly toxic APAP blood levels at twelve hours and early or late care. Thus, for the four strata, the hypothesis of no treatment differences is equivalent to viewing the partition of responses between NAC treatment and supportive therapy under the hypothesis of their equivalence as resulting from a stratified random sample with fixed sample sizes $\{n_{h_1}, n_{h_2}\}$ from the samples $\{n_h\}$ in accordance with the product multiple hypergeometric model.

Descriptive statistics concerning the distributions of the maximum values for serum bilirubin, SGOT Ratio (to upper limit of normal range) and prothrombin time ratio are shown in Tables 6.1, 6.2 and 6.3 relative to treatment groups cross-classified with respect to moderately toxic and highly toxic APAP blood levels at 12 hours and early or late care strata. From this information, it can be seen that:

a) patients with moderate toxicity at 12 hours and early care experience a lesser degree of liver dysfunction than patients in the other risk strata.

b) patients with moderate toxicity at 12 hours regardless of when care was initiated show less pronounced treatment group differences than patients with high toxicity at 12 hours.

c) patients taking NAC experience a lesser degree of liver dysfunction than patients with supportive therapy only in each of the four strata.

d) serum bilirubin and SGOT show highly skewed distributions.

Serum bilirubin, SGOT ratio and prothrombin time ratio can also be described in terms of clinically-defined elevation levels corresponding

TABLE 6.1

DESCRIPTIVE STATISTICS PERTAINING TO THE DISTRIBUTIONS OF SERUM BILIRUBIN FOR PATIENTS WITH NAC TREATMENT OR SUPPORT THERAPY AT TWO LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum				Extrema and quartiles						
Blood toxicity at 12 hours	Hours to medical intervention	Treatment	Number of patients	Minimum	First quartile	Median	Third quartile	Maximum	Mean	Standard deviation
< 100	≤ 6	NAC	59	0.40	0.70	0.90	1.30	10.00	1.19	1.27
		Support	22	0.50	0.60	0.95	1.40	5.50	1.34	1.18
< 100	> 6	NAC	11	0.50	0.85	1.20	2.28	3.70	1.57	1.03
		Support	7	0.60	0.95	1.30	3.05	5.30	2.11	1.68
≥ 100	≤ 6	NAC	18	0.70	1.00	1.55	2.50	7.30	1.88	1.51
		Support	8	1.30	1.85	3.55	24.75	33.50	11.89	13.72
≥ 100	> 6	NAC	9	1.10	1.10	1.30	2.58	6.00	2.14	1.60
		Support	9	1.50	2.08	3.70	7.25	9.20	4.67	2.87

TABLE 6.2

DESCRIPTIVE STATISTICS PERTAINING TO THE DISTRIBUTIONS OF SGOT RATIOS FOR PATIENTS WITH NAC TREATMENT OR SUPPORT THERAPY AT TWO LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum				Extrema and quartiles						
Blood toxicity at 12 hours	Hours to medical intervention	Treatment	Number of patients	Minimum	First quartile	Median	Third quartile	Maximum	Mean	Standard deviation
< 100	≤ 6	NAC	59	0.4	0.8	1.1	2.6	393.8	20.9	73.8
		Support	22	0.5	0.8	1.7	15.6	117.0	15.2	30.3
< 100	> 6	NAC	11	0.4	1.0	18.0	57.7	112.3	29.4	38.0
		Support	7	0.6	4.5	25.0	108.4	292.5	71.0	108.1
≥ 100	≤ 6	NAC	18	1.2	1.7	6.5	28.5	88.0	23.1	31.5
		Support	8	3.8	36.5	69.8	168.0	250.0	100.3	94.6
≥ 100	> 6	NAC	9	0.6	2.6	7.8	52.8	136.8	33.6	45.0
		Support	9	25.0	66.0	93.0	176.9	240.0	121.8	75.6

TABLE 6.3

DESCRIPTIVE STATISTICS PERTAINING TO THE DISTRIBUTIONS OF PROTHROMBIN TIME RATIO FOR PATIENTS WITH NAC TREATMENT OR SUPPORT THERAPY AT TWO LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum				Extrema and quartiles						
Blood toxicity at 12 hours	Hours to medical intervention	Treatment	Number of patients	Minimum	First quartile	Median	Third quartile	Maximum	Mean	Standard deviation
< 100	≤ 6	NAC	59	0.90	1.00	1.10	1.20	2.30	1.16	0.26
		Support	22	1.20	1.30	1.40	1.40	2.40	1.48	0.32
< 100	> 6	NAC	11	1.00	1.00	1.20	1.28	1.60	1.19	0.21
		Support	7	1.20	1.28	1.50	1.75	3.10	1.70	0.65
≥ 100	≤ 6	NAC	18	0.90	1.10	1.20	1.30	2.40	1.26	0.31
		Support	8	1.40	1.65	2.25	3.05	5.00	2.54	1.18
≥ 100	> 6	NAC	9	1.10	1.18	1.40	1.50	1.60	1.34	0.18
		Support	9	1.80	1.88	2.30	2.52	3.20	2.28	0.50

to normal range, mildly elevated, moderately elevated, severely elevated and very severely elevated. These categories not only standardize the response scale with uniform scores {0, 1, 2, 3, 4} across the three functional measures, but also clarify the interpretation of treatment group differences in the sense of associating elevation levels with treatment groups. With these categories, statistical power can be directed toward detecting clinically significant differences with an adjustment for skewness in the sense that the extreme values have been assigned the maximum score. This also enhances the applicability of chi-square approximations via Central Limit Theory. The corresponding tabulations of patients in elevation levels are shown in Tables 6.4, 6.5 and 6.6. All three functional measures show more tendency of very severe elevation relative to normal for supportive therapy patients than for NAC patients when progressing from moderate toxicity at 12 hours and early care to high toxicity at 12 hours and late care. Although early to late care has some association with a shift in the direction of very severe elevation, the dominant shift in the direction of very severe elevation appears to occur with the shift from moderate to high toxicity.

Randomization statistics pertaining to the four strata framework are shown in Table 6.7 for both the continuous values and the five point elevation scales having uniform scores {0, 1, 2, 3, 4} for the three functional measures. Here, the continuous values of serum bilirubin SGOT ratio have been transformed to the log scale to enhance the applicability of the chi-square approximations via central limit theory. The univariate direct tests across strata show that on average, NAC patients experience significantly ($p < .05$) less liver function damage regarding SGOT, serum bilirubin and prothrombin time than supportive therapy

TABLE 6.4

NUMBERS OF PATIENTS WITH ELEVATED LEVEL OF SERUM BILIRUBIN RELATIVE TO TREATMENT RECEIVED AND LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum			Elevation levels					
Blood toxicity at 12 hours	Hours to medical intervention	Treatment	Normal[a]	Mild[b]	Moderate[c]	Severe[d]	Very Severe[e]	Total
< 100	≤ 6	NAC	42	9	5	2	1	59
		Support	16	2	0	1	3	22
< 100	> 6	NAC	6	2	0	2	1	11
		Support	3	1	0	1	2	7
≥ 100	≤ 6	NAC	7	4	3	3	1	18
		Support	0	1	2	0	5	8
≥ 100	> 6	NAC	3	2	1	1	2	9
		Support	0	1	1	1	6	9

[a] normal level has values not exceeding 1.2.

[b] mild elevation has values greater than 1.2 but not exceeding 1.6.

[c] moderate elevation has values greater than 1.6 but not exceeding 2.

[d] severe elevation has values greater than 2 but not exceeding 3.

[e] very severe elevation has values exceeding 3.

TABLE 6.5

NUMBERS OF PATIENTS WITH ELEVATED LEVEL OF SGOT RATIO RELATIVE TO TREATMENT RECEIVED AND LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum			Elevation levels					
Blood toxicity at 12 hours	Hours to medical intervention	Treatment	Normal[a]	Mild[b]	Moderate[c]	Severe[d]	Very Severe[e]	Total
< 100	≤ 6	NAC	29	9	10	5	6	59
		Support	8	3	4	3	4	22
< 100	> 6	NAC	3	1	1	3	3	11
		Support	1	1	0	1	4	7
≥ 100	≤ 6	NAC	0	5	2	5	6	18
		Support	0	0	1	1	6	8
≥ 100	> 6	NAC	1	1	1	2	4	9
		Support	0	0	0	0	9	9

[a] normal level has values not exceeding the upper normal limit

[b] mild elevation has values greater than the upper normal limit but less than 2 times the upper normal limit

[c] moderate elevation has values between 2 and less than 5 times the upper normal limit

[d] severe elevation has values between 5 and less than 25 times the upper normal limit

[e] very severe elevation has values equal or greater than 25 times the upper normal limit

TABLE 6.6

NUMBERS OF PATIENTS WITH ELEVATED LEVEL OF PROTHROMBIN TIME RATIO RELATIVE TO TREATMENT RECEIVED AND LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum			Elevation levels					
Blood toxicity at 12 hours	Hours to medical intervention	Treatment	Normal[a]	Mild[b]	Moderate[c]	Severe[d]	Very Severe[e]	Total
< 100	≤ 6	NAC	39	15	1	3	1	59
		Support	0	10	9	1	2	22
< 100	> 6	NAC	5	4	2	0	0	11
		Support	0	2	3	1	1	7
≥ 100	≤ 6	NAC	6	10	1	0	1	18
		Support	0	0	2	1	5	8
≥ 100	> 6	NAC	2	2	5	0	0	9
		Support	0	0	0	4	5	9

[a] normal level has values not exceeding 1.1.

[b] mild elevation has values greater than 1.1 but not exceeding 1.3.

[c] moderate elevation has values greater than 1.3 but not exceeding 1.6.

[d] severe elevation has values greater than 1.6 but not exceeding 2.

[e] very severe elevation has values exceeding 2.

TABLE 6.7

RANDOMIZATION STATISTICS COMPARING SERUM BILIRUBIN, SGOT RATIO AND PROTHROMBIN TIME RATIO BETWEEN PATIENTS RECEIVING NAC TREATMENT AND PATIENTS RECEIVING SUPPORT THERAPY ONLY WITHIN LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Functional Form	Stratum: Blood toxicity at 12 hours	Stratum: Hours to medical intervention	Unadjusted test statistics[a]: Serum bilirubin	SGOT ratio	Prothrombin time ratio	Multivariate	Hours to medical intervention	Time adjusted test statistics[†a]: Serum bilirubin	SGOT ratio	Prothrombin time ratio	Multivariate
Continuous values[1]	< 100	≤ 6	0.4	0.9	11.3**	12.6**	2.8	0.1	0.5	10.4**	12.3**
	< 100	> 6	1.0	0.6	4.6*	4.8	1.5	1.4	0.4	3.2	3.5
	≥ 100	≤ 6	8.3**	6.7**	11.1**	11.9**	0.6	8.9**	6.8**	11.7**	12.4**
	≥ 100	> 6	4.7*	7.3**	10.8**	11.8**	3.1	5.0*	5.2*	8.8**	9.1*
	Combined		12.8**	10.3**	36.1**	36.1**	5.4*	12.9**	8.1**	33.0**	33.1**
Elevation levels (5 point scale)	< 100	≤ 6	1.1	1.8	22.7**	24.3**	2.8	0.5	1.4	21.0**	23.0**
	< 100	> 6	1.1	0.7	6.7**	8.6*	1.5	1.6	0.2	5.3*	7.2
	≥ 100	≤ 6	8.0**	3.6	15.3**	16.0**	0.6	7.9**	3.8	15.6**	16.2**
	≥ 100	> 6	4.8*	4.7*	12.4**	13.2**	3.1	3.8	3.3	9.9**	10.7*
	Combined		12.0**	7.8**	57.1**	59.7**	5.4*	10.8**	5.7*	51.7**	54.5**

[1]values for serum bilirubin and SGOT ratio were transformed to log(1 + value)

[a]univariate statistics have 1 degree of freedom; multivariate statistics have 3 degrees freedom

[†]statistics are adjusted for the exact number of hours between the overdose incident and medical intervention

* .01 ≤ p ≤ .05   ** p < .01

patients. The within stratum statistics show that SGOT and serum bilirubin differences between treatments were significant only in patients with highly toxic APAP levels at 12 hours, while prothrombin time differences were significant in all strata. The multivariate tests for direct differences in all three functional measures show that NAC patients have significantly more favorable status than supportive therapy patients for three of the four levels of risk to hepatotoxicity for the continuous values; each of the four levels for the 5 point elevation scales; and across all levels in an average partial association sense.

Randomization tests for the actual hours to initial medical intervention show that NAC and supportive therapy patients had comparable arrival times to a medical facility within the levels of early and late care. However, on average, supportive therapy patients arrived at significantly ($p < .05$) later hours than NAC patients. Although, in this regard, supportive therapy patients can be considered at greater risk to hepatotoxicity than NAC patients, time adjusted univariate and multivariate statistics show that differences between NAC and supportive therapy patients persist for three of the four strata and on average across the four strata.

For both direct and time adjusted analyses, the five point elevation levels are more sensitive to treatment group differences than the continuous values. This greater sensitivity of the five point elevation scales suggests that treatment differences observed in the continuous values are also clinically significant.

Randomization statistics for a two strata framework corresponding to moderately toxic and highly toxic APAP levels at 12 hours are shown in Table 6.8 for both the continuous values and the five-point elevation

TABLE 6.8

RANDOMIZATION STATISTICS COMPARING SERUM BILIRUBIN, SGOT RATIO AND PROTHROMBIN TIME RATIO BETWEEN PATIENTS RECEIVING NAC TREATMENT AND PATIENTS RECEIVING SUPPORT THERAPY ONLY WITHIN LEVELS OF BLOOD TOXICITY

Functional Form	Stratum: Blood troxicity at 12 hours	Unadjusted test statistics[a]: Serum bilirubin	Unadjusted: SGOT ratio	Unadjusted: Prothrombin time ratio	Unadjusted: Multivariate	Hours to medical intervention	Time adjusted test statistics[†a]: Serum bilirubin	Time adjusted: SGOT ratio	Time adjusted: Prothrombin time ratio	Time adjusted: Multivariate
Continuous values[1]	< 100	1.5	2.2	16.0**	16.4**	4.1*	0.7	0.7	13.9**	15.1**
	≥ 100	12.8**	15.2**	20.6**	23.1**	2.9	12.3**	12.7**	18.8**	20.5**
	Combined	13.4**	12.3**	36.5**	36.8**	6.8**	11.5**	8.0*	32.6**	32.9**
Elevation levels (5 point scale)	< 100	2.7	3.3	30.5**	32.5**	4.1*	1.6	1.5	26.9**	29.6**
	≥ 100	13.8**	9.2**	28.6**	29.6**	2.9	11.8**	7.4**	25.7**	26.7**
	Combined	14.0**	9.3**	59.1**	61.2**	6.8**	10.6**	5.7*	52.6**	55.3**

[1]values for serum bilirubin and SGOT ratio were transformed to log(1 + value)

[a]univariate statistics have 1 degree of freedom; multivariate statistics have 3 degrees of freedom

[†]statistics are adjusted for the exact numbers of hours between the overdose incident and medical intervention

* $.01 \le p \le .05$ ** $p < .01$

levels. The results reinforce the four strata findings relative to conclusions for patients with moderate toxicity and patients with high toxicity. More specifically, both strata show that NAC patients have significantly more favorable status than supportive therapy patients relative to the multivariate profile of the three functional measures. However, the univariate tests indicate that NAC patients with moderate toxicity are comparable to supportive therapy patients in SGOT and serum bilirubin while having significantly ($p < .05$) more favorable prothrombin time ratios; NAC patients with high toxicity have significantly ($p < .05$) less severe values of all three functional measures than supportive therapy patients.

The results for patients with moderate toxicity should be viewed cautiously because without the additional stratification for early and late care the actual hours to initial medical intervention are significantly ($p < .05$) longer for supportive therapy patients than for NAC patients. Time adjusted randomization statistics for patients with highly toxic APAP levels at 12 hours show that treatment group differences persist. As noted in the four strata analysis, the five point elevation levels are more sensitive to treatment group differences than the continuous values for both direct and time-adjusted comparisons.

6.3 Randomization Analysis of Scored Data

An alternative analysis for serum bilirubin involves randomization test statistics applied to standardized values or scores. These scores include the logs of (1 + value) or uniform scores, ranks, modified ridits and logranks for the serum bilirubin values and the corresponding five point elevation scales. For this purpose, the observed values were ranked within each stratum with ties being assigned a midrank value; the ranks were standardized to modified ridits by division of each rank by $\{1 + n_h\}$ where n_h is the sample size of the h-th stratum; the observed values were assigned within-stratum logranks by use of a SAS macro LOG-RANK with ties being assigned the logrank score of an average rank.

The randomization test statistics for the four strata framework corresponding to the cross-classification of moderate or high toxicity at twelve hours and early or late care are shown in Table 6.9 for serum bilirubin scores based on the continuum values and in Table 6.10 for serum bilirubin scores based on the five point elevation scales. For the continuous values, the logs of (1 + value) provide the strongest results with those for logrank scores having similar magnitude. In contrast, for the five point elevation scales, the logrank scores provide the strongest results with those for the uniform scores having similar magnitude.

Both the logs of (1 + value) or uniform scores and the logrank scores are dramatically more sensitive to differences in treatment groups than the rank scores. This occurs because both of these quantities are sensitive to one treatment having more of the larger serum bilirubin values than the other relative to the presence of treatment similarity in the distribution of the smaller serum bilirubin values. Ranks impose equal interval spacings for small values and large values alike. In this regard, ranks mask a difference that exists predominantly in the upper extreme range. The modified ridits have the same within-stratum results as the ranks, but are standardized relative to sample sizes such that the range of values within each stratum is restricted to the [0,1] interval. Since, most patients in this study were in the moderate toxicity strata which did not tend to show treatment group differences,

TABLE 6.9

SCORES FOR RANDOMIZATION STATISTICS COMPARING SERUM BIBIRUBIN VALUES BETWEEN PATIENTS RECEIVING NAC TREATMENT AND PATIENTS RECEIVING SUPPORT THERAPY ONLY WITHIN LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum		Sample sizes, NAC support	Test function	Scores			
Blood toxicity at 12 hours	Hours to medical intervention			Log[1]	Rank	Modified ridit	Logrank
< 100	≤ 6	59,22	unadjusted SB	0.4	0.1	0.1	0.6
			time covariate[a]	2.8	2.8	2.8	0.8
			adjusted SB	0.1	0.1	0.1	0.4
< 100	> 6	11, 7	unadjusted SB	1.0	0.4	0.4	0.8
			time covariate	1.5	0.9	0.9	2.2
			adjusted SB	1.4	0.6	0.6	1.4
≥ 100	≤ 6	18, 8	unadjusted SB	8.5**	7.4**	7.4**	8.6**
			time covariate	0.6	0.8	0.8	1.3
			adjusted SB	8.9**	7.3**	7.3**	8.6**
≥ 100	> 6	9, 9	unadjusted SB	4.7*	5.5*	5.5*	4.4*
			time covariate	3.1	3.0	3.0	3.7
			adjusted SB	5.0*	4.7*	4.7*	4.9*
Combined		97,46	unadjusted SB	12.8**	1.4	6.4*	8.9**
			time covariate	5.4*	3.2	3.5	2.2
			adjusted SB	17.9**	1.2	6.0*	9.3**

[1]values were transformed to log(1 + value)

[a]actual hours to the first medical intervention relative to the overdose incident

* $.01 \leq p < .05$ ** $p < .01$

TABLE 6.10

SCORES FOR RANDOMIZATION STATISTICS COMPARING 5-POINT ELEVATION LEVELS OF SERUM BIBIRUBIN VALUES BETWEEN PATIENTS RECEIVING NAC TREATMENT AND PATIENTS RECEIVING SUPPORT THERAPY ONLY WITHIN LEVELS OF BLOOD TOXICITY AND EARLY OR LATE CARE

Stratum: Blood toxicity at 12 hours	Stratum: Hours to medical intervention	Sample sizes, NAC support	Test function	Scores: Uniform	Scores: Rank	Scores: Modified ridit	Scores: Logrank
< 100	≤ 6	59,22	unadjusted SB	1.1	< 0.1	< 0.1	1.7
			time covariate[a]	2.8	2.8	2.8	0.8
			adjusted SB	0.5	< 0.1	< 0.1	1.3
< 100	> 6	11, 7	unadjusted SB	1.1	0.7	0.7	1.1
			time covariate	1.5	0.9	0.9	2.2
			adjusted SB	1.6	1.0	1.0	1.9
≥ 100	≤ 6	18, 8	unadjusted SB	8.0**	7.9**	7.9**	9.5**
			time covariate	0.6	0.8	0.8	1.3
			adjusted SB	7.9**	7.8**	7.8**	9.5**
≥ 100	> 6	9, 9	unadjusted SB	4.8*	4.8*	4.8*	4.3*
			time covariate	3.1	3.0	3.0	3.7
			adjusted SB	3.8	3.3	3.3	2.9
Combined		97,46	unadjusted SB	12.0**	1.4	7.2*	13.2**
			time covariate	5.4*	3.2	3.5	2.2
			adjusted SB	10.8**	0.9	6.0*	12.5**

[a] actual hours to the first medical intervention relative to the overdose incident

* $.01 \le p < .05$ ** $p < .01$

the ranks are insensitive to treatment group differences while the modified ridits are sensitive at the average partial association level.

The randomization test statistics for the two strata framework corresponding to moderate toxicity and high toxicity at twelve hours are shown in Tables 6.1 and 6.12 for the continuous values and five point elevation scales respectively. The results show that for the moderately toxic stratum which contains most of the patients, the treatment groups have equivalent distributions of serum bilirubin and significantly ($p < .05$) different distributions of hours to first medical intervention. Results, for both the continuous values and the five point elevation levels are similar to those found with the four strata framework. However, these results should be viewed with caution because of the strength of treatment group differences in the time to the first medical intervention.

6.4 <u>Discussion</u>

The randomization model has been used to compare results from two observational studies in which patients fortuitously participated in the particular treatment program. For this purpose, homogeneous subpopulations for risk of hepatotoxicity were defined relative to APAP serum blood levels at 12 hours and the time to initial medical care. Within this framework, patients with similar risk to hepatotoxicity were expected to have similar levels of liver dysfunction if the treatments were equally effective. In other words, levels of liver dysfunction would not be at random with regard to treatment. Here, randomization was imposed by the hypothesis in contrast to design randomization. Randomization by hypothesis does not require any assumptions about a

TABLE 6.11

SCORES FOR RANDOMIZATION STATISTICS COMPARING SERUM BIBIRUBIN VALUES BETWEEN PATIENTS RECEIVING NAC TREATMENT AND PATIENTS RECEIVING SUPPORT THERAPY ONLY WITHIN LEVELS OF BLOOD TOXICITY

Stratum: Blood toxicity at 12 hours	Sample sizes, NAC support	Test function	Scores: Log[1]	Rank	Modified ridit	Logrank
< 100	70,29	unadjusted SB	1.5	0.6	0.6	1.5
		time covariate[a]	4.1*	3.8*	3.8*	3.8*
		adjusted SB	0.7	0.2	0.2	0.7
≥ 100	27,17	unadjusted SB	12.8**	14.0**	14.0**	13.4**
		time covariate	2.9	1.4	1.4	4.0*
		adjusted SB	12.3**	13.0**	13.0**	12.7**
Combined	97,46	unadjusted SB	13.4**	3.4	7.7**	10.0**
		time covariate	6.8**	4.9*	5.2*	7.6**
		adjusted SB	11.5**	2.3	6.0*	8.0**

[1]values were transformed to log (1 + value)

[a]actual hours to the first medical intervention relative to the overdose incident

* $.01 \leq p < .05$ ** $p < .01$

TABLE 6.12

SCORES FOR RANDOMIZATION STATISTICS COMPARING 5-POINT ELEVATION LEVELS OF SERUM BIBIRUBIN VALUES BETWEEN PATIENTS RECEIVING NAC TREATMENT AND PATIENTS RECEIVING SUPPORT THERAPY ONLY WITHIN LEVELS OF BLOOD TOXICITY

Stratum: Blood toxicity at 12 hours	Sample sizes, NAC support	Test function	Scores: Uniform	Rank	Modified ridit	Logrank
< 100	70,29	unadjusted SB	2.7	0.6	0.6	3.7
		time covariate[a]	4.1*	3.8*	3.8*	3.8*
		adjusted SB	1.6	0.1	0.1	2.4
≥ 100	27,17	unadjusted SB	13.8**	14.4**	14.4**	14.9**
		time covariate	2.9	1.4	1.4	4.0*
		adjusted SB	11.8**	13.2**	13.2**	12.2**
Combined	97,46	unadjusted SB	14.0**	4.1*	9.0**	15.8**
		time covariate	6.8**	4.9*	5.2*	7.6**
		adjusted SB	10.6**	2.6	6.7**	11.8**

[a]actual hours to the first medical intervention relative to the overdose incident

* $.01 \leq p < .05$ ** $p < .01$

conceptual population these patients might represent. Correspondingly, the randomization results of testing hypotheses refer only to the finite population under study.

The three aspects of liver dysfunction were found to be significantly more severe for patients on supportive therapy only than patients receiving NAC. This finding was particularly strong for patients who had highly toxic APAP levels at 12 hours; also, these results were maintained after adjusting for hours to care.

Within strata of early or late care, time to initial care was not found to be significant. However, patients across strata with supportive therapy only consistently tended to have care initiated at later time intervals from the overdosing such that the average partial association of care hours with treatment was significant. In this regard, covariate adjusted results should be regarded with some caution for average partial association. However, within-stratum results are unaffected.

Logrank scores and logs of (1 + value) or uniform scores showed stronger results than rank or ridit scores for the NAC versus supportive therapy only comparison of serum bilirubin. Their strength may be due in part to the right skewness of the data.

Although ranks and modified ridits give essentially the same within-stratum results, they have different results for the across-strata average partial association statistic. In particular, as discussed in Section 5 of Chapter II, the average partial association statistic for ranks is based on an unweighted sum of the Wilcoxon Rank Sum statistics for the respective strata; for modified ridits, it is based on a weighted sum of the Wilcoxon Rank Sum statistics with the weights being the quantities $\{1/(n_h + 1)\}$, where the n_h are the stratum sample sizes.

Thus, if stratum sample sizes vary from one stratum to the next, ranks and modified ridits would typically show different across strata results.

CHAPTER VII

INCOMPLETE DESIGNS AND INCOMPLETE DATA

This chapter consists of two examples which demonstrate the application of randomization statistics to data with various aspects of incompleteness. The first example has an incomplete blocks design for which the blocks (strata) have subsets of the subpopulations to be compared. The second example has incomplete data in a multivariate response profile. These latter data were previously analyzed by Stanish, Gillings and Koch (1978) to demonstrate a weighted least squares strategy for missing data through ratio estimators as reviewed in Section 5 of Chapter I. A randomization analysis is presented for these data using the missing value pattern adjustment presented in Section 4 of Chapter II.

7.1 Incomplete Design Example

The data for this example are from a multi-center clinical trial concerned with assessing the efficacy of an analgesic drug for chronic joint pain. For this purpose, patients who suffer from one of four classifications of chronic joint disease, are randomly assigned to placebo or active treatment within each clinic. Both patient and physician global pain evaluations are made just prior to treatment and again after four weeks of treatment. The global evaluations have ordinally scaled values 1, 2, 3, 4, 5 for the range from poor to excellent pain status.

Four clinics, identified as investigators 1, 2, 3, 4 have incomplete randomization within a classification of joint disease since patients are assigned to one treatment only as shown in Table 7.1.1. The patient pool from these four clinics however would provide a reasonable basis for a within-affected joint analysis of treatment group differences. The alternative within-investigator analysis has difficulties not only with small sample comparisons but also with excluding comparisons for which there is data for only one treatment.

Although patients are randomized within clinics, their pooling does not directly represent a randomly assigned group of patients in the sense of a single large clinic. More specifically, patients from one clinic may have more severe pain than patients from another clinic; patients from one clinic may respond more favorably to placebo relative to active treatment than patients from another clinic.

An analysis supporting the pooling of clinics involves the comparison of investigators within the eight levels of the joint disease treatment cross-classification. In this regard, the randomization model corresponds to viewing the partition of responses among these investigators as random samples of sizes $\{n_{h_1}, n_{h_2}, n_{h_3}, n_{h_4}\}$ from the pooled samples $\{n_h\}$, where h indexes the eight levels of joint disease and treatment cross-classifications. The across-level average partial association statistic with three degrees of freedom, combines results of five strata comparing all four investigators, two strata comparing three investigators and one strata comparing two investigators.

The four investigators were found to be comparable relative to the baseline physician global evaluation ($Q_{MMS}(3) = 1.4$). Furthermore, they were found to be comparable at the end of study on both direct comparisons of patient ($Q_{MMS}(3) = 4.6$) and physician ($Q_{MMS}(3) = 5.6$) global evaluations and baseline-adjusted comparisons of patient ($Q(y|x:3) = 3.2$)

Table 7.1.1

NUMBERS OF PATIENTS FOR EACH INVESTIGATOR WITHIN THE CLASSIFICATION OF DISEASE AND TREATMENT ADMINISTERED

Principally effected joint	Treatment	Investigator 1	2	3	4
Lumbar spine	Placebo	2	1	4	1
	Active	3	1	6	0
Hip	Placebo	2	1	1	2
	Active	0	0	1	3
Knee	Placebo	5	3	1	1
	Active	8	4	0	2
Cervical Spine	Placebo	5	4	6	1
	Active	4	3	2	2

global evaluations. In view of these findings, the pooled group of patients from the four clinics can be conceptualized as having been randomly assigned to treatment within one large clinic for the purposes of testing group differences.

7.2 Incomplete Data Example

The data for this example pertain to a multi-center clinical trial concerned with the effectiveness of a drug in alleviating a certain skin condition. These data (which are presented fully in Stanish, Gillings and Koch [1978]) represent the experience of patients randomly assigned to drug or placebo treatments within each of six clinics. Just prior to treatment, the severity of the skin condition was evaluated as fair, poor or exacerbated. At the three follow-up visits, improvement was evaluated as rapidly improving, slowly improving, stable, slowly worsening, or rapidly worsening.

Not all of the 172 patients completed the follow-up visit schedule. More specifically, 169 patients had data for the first visit, 156 patients had data for the second visit and 142 patients had data for the third visit. Missing data were not evenly distributed across the six clinics. In particular, three clinics had missing data for all three visits, two clinics had missing data for just the second and third visits, and one clinic had no missing data. The one clinic with no missing data included only four patients in contrast to the other five clinics which included approximately 30 patients each.

7.2.1 Weighted least squares model analysis

A weighted least squares analysis using ratio estimators was presented in Stanish, Gillings and Koch (1978). Ratio estimators as reviewed in Section 5 of Chapter I are constructed such that means and covariances are based on the observed data. This method assumes that the missing data pattern is independent of the hypothetical values of missing responses for any one visit. This assumption implies that missing data are not associated with patients experiencing rapid improvement, for instance.

The weighted least squares analysis is undertaken in a stratified simple random sampling setting relative to an assumed larger population for the pooled clinics. The stratification corresponds to the cross-classification of treatment and initial severity status, where initial severity has been reduced to two levels, moderate and severe. The four subpopulations of the pooled clinic data are viewed as a multivariate random sample regarding the three visits from each of the four strata in the large population. The final model reported by Stanish, Gillings and Koch involved four parameters:

1. first follow-up visit for drug
2. increment for second follow-up visit for drug over first visit
3. increment for both third follow-up visit for drug over second visit and second follow-up visit for placebo over first visit
4. first follow-up visit for placebo.

The results showed that over three follow-up visits, patients taking the test drug had more favorable outcome than patients taking placebo ($Q_W(3) = 148$). Furthermore, there was a significant ($p < 0.05$) increase in improvement between the first and second follow-up visits for patients taking the test drug and a somewhat less but significant increase in improvement between the second and third follow-up visit for patients taking the test drug. This latter increase in improvement was of the same magnitude as the improvement between the first and second follow-up visit for patients taking placebo. There was no significant improvement noted between the second and third follow-up visits for patients taking placebo. These results were consistent across the two levels of initial severity.

7.2.2 Randomization model analysis

The randomization model is based upon the random assignment of the finite population to two treatment groups within each clinic. Under the null hypothesis of treatment group equivalence, the observed partition of the multivariate improvement scores ($j = 5^3$) into ($i = 2$) treatment groups can be regarded as equivalent to a successive set of random samples $\{n_{h1+}, n_{h2+}\}$ patients from the pooled samples $\{n_{h++}\}$ (where h indexes clinics) in accordance with the product multiple hypergeometric model:

$$\Pr(\{n_{hij}\}) = \prod_{h=1}^{6} [\prod_{i=1}^{2} n_{hi+}! \prod_{j=1}^{5^3} n_{h+j}!/n_{h++}! \prod_{i=1}^{2} \prod_{j=1}^{5^3} n_{hij}!] .$$

Also, see the statement of H_0 in Section 2.1. In the presence of missing values, the joint distribution of improvement scores and corresponding missing value indicators is considered to be at random relative to the partition of response and prevalence of missing data into treatment groups in accordance with the product multiple hypergeometric model. As in the weighted least squares analysis of ratio estimators, the pattern of missing data is assumed to be independent of the improvement scores. Three methods used in managing the missing values are

1. assigning a within-stratum constant to the missing data and constructing a missing-value-pattern adjusted test statistic
2. resolving the missing data by filling in an interpolated value

based on adjacent values of a patient's data

3. resolving the missing data by filling in the within-stratum mean response per visit.

The first method results in a statistic based on observed values with an apparent augmented sample size but adjusted for the pattern of missing data so as to depend on only the available information. The second method provides an estimate of a complete data matrix. The statistics are based on an estimate of variance that is not as sensitive to the increase in sample size in the sense that values being added have variance in contrast to within-stratum constants. The third method results in a statistic based on expected values constructed from the observed data as a result of assigning missing data the within-stratum expected value. However, when subpopulation means differ from the stratum expected value (i.e. stratum means) the mean difference between subpopulations with modified missing data is no longer equivalent to the mean difference between subpopulations with observed data only. Furthermore, the variance based on the augmented sample is conservative in the sense that the added data have a constant value (i.e. the within-stratum mean, which is also the expected value of the two subpopulations).

The first method involved assigning the missing data with a within-stratum constant (the value zero was used throughout visits and strata) and constructing a set of missing value indicators $\underline{u} = (u_1, u_2, u_3)$ where u_j had the value one if the j-th visit was missing and zero otherwise for each patient. Randomization statistics, shown in Table 7.2.1, are based on the within-stratum ranks with the missing value indicators as a covariate structure. More specifically, the direct treatment group comparison is equivalent to the treatment group comparison adjusted for the

Table 7.2.1

MULTIVARIATE RANDOMIZATION CHI-SQUARE STATISTICS WITH 3 DEGREES OF FREEDOM FOR UNADJUSTED AND INITIAL-SEVERITY ADJUSTED TREATMENT GROUP COMPARISONS RELATIVE TO 3 DIFFERENT STRATEGIES FOR MANAGING MISSING VALUES ACROSS THE THREE FOLLOW-UP VISITS

	Missing value pattern[1]			Patient estimated values[2]			Assigned means[3]		
Clinic	Missing pattern adjusted	Joint missing pattern and covariate	Joint adjusted	Direct	Covariate	Covariate adjusted	Direct	Covariate	Covariate adjusted
5	13.4**	7.8(4)[a]	13.0**	13.8**	0.5	13.7**	15.8*	0.5	15.3**
6	6.2	5.0(3)	6.2	10.3*	0.3	10.0*	10.9*	0.3	10.5*
8	16.2**	1.9(4)	16.0**	16.4**	0.3	16.5**	15.0**	0.3	15.1**
9[b]		1.0(1)			1.0			1.0	
10	13.7**	4.6(3)	12.5**	16.3**	1.1	15.2**	14.2**	1.1	13.2**
11	17.0**	3.5(4)	17.2**	15.1**	0.4	14.7**	15.3**	0.4	14.9**
Combined	60.3**	C	58.9**	66.4**	0.3	64.8**	64.5**	0.3	62.9**

1 the direct comparison involves an ajustment of the within-clinic ranks for the corresponding missing value pattern identified by two or three indicators for all clinics but clinic 9; the covariate adjusted comparison involves an adjustment of the within-clinic ranks for the missing value pattern and the covariate jointly.

2 missing data are resolved within each patient according to an interpolation of adjacent visit response prior to ranking.

3 missing data are assigned a within-clinic mean rank.

()[a] degrees of freedom. Because this varies across strata, there is no average partial association statistic.

b Within-stratum statistics are not available because the overall stratum variance is singular. However, within-stratum results are included in the across-strata statistics.

C The across strata statistics for varying numbers of stratum covariates is undefined.

* $0.01 \leq p \leq 0.05$ ** $p \leq 0.01$

pattern of missing data. The baseline-severity adjusted treatment group comparison involves a joint adjustment for baseline severity and missing value patterns. Because clinics varied according to the number of visits with missing data, the number of missing value indicators included in the within-clinic covariate adjustment varied. In this situation, the residual formulation of the average partial association statistic is most useful. The within-stratum residual construction reduces the statistical dimension from the response profile together with the covariate to just the response profile which has a uniform dimension across strata. The missing value pattern together with the baseline severity covariate comprised four degrees of freedom for three strata, three degrees of freedom for two strata, and one degree of freedom for one strata; the within-stratum multivariate residual statistic has three degrees of freedom corresponding to adjusted responses for the three visits for each stratum.

Patient estimated values in the second method were derived by assigning the maximum (or least favorable) response of the first and third visit to a missing second visit and then assigning the second visit response to either a missing first or third visit. Randomization statistics shown in Table 7.2.1 are based on within-stratum ranks for the three visit improvement and the baseline severity. Mean ranks in the third method were obtained by assigning the within-clinic median response for each visit to missing values for the corresponding visit. Ranking the data resulted in tied mean ranks for missing values. The corresponding randomization statistics are also shown in Table 7.2.1.

The three methods show comparable multivariate randomization statistics in Table 7.2.1. The statistics clearly show that the data do not support the randomization model in that the test drug has significantly ($p < 0.05$) more favorable outcome than placebo. The slightly diminished results for the assigned means method relative to the patient estimated values method might be due partially to the use of the same constant for both placebo and test drug when the means for the two drugs are clearly different. The missing value pattern adjusted statistics appear slightly more conservative than the statistics based on modified data, but this might be anticipated since it does not involve the use of information beyond that in the available data.

Although managing missing data is a multivariate data issue, the corresponding univariate application illustrates the statistical impact of these methods. More specifically, the randomization statistics in Table 7.2.2 show that results based on the observed data alone for any one visit are comparable but not equivalent to the results based on the observed data with an augmented sample size adjusted for the pattern of missing data. Furthermore, in this example, the method of assigned means appears to be a comparable strategy to the missing value pattern adjustment relative to the proximity of its results to the observed data only results. The main distinction between the two methods is in the representation of the data. The missing value pattern adjustment accurately reflects the structure of the data; the assigned value methods assume that such data were equivalent to what might have been observed.

7.3 Discussion

Randomization methods discussed in Sections 1, 2 and 3 of Chapter II are applicable to incomplete blocks designs for which blocks (strata) have subsets of the subpopulations to be compared. In this regard, the formulation of statistics such as Friedman or Mantel-Haenszel statistics

Table 7.2.2

UNIVARIATE RANDOMIZATION CHI-SQUARE STATISTICS WITH 1 DEGREE OF FREEDOM FOR TREATMENT GROUPS COMPARISONS REGARDING (1) RANKS OF THE OBSERVED DATA ONLY WITH THE OBSERVED SAMPLE SIZE INDICATED PARENTHETICALLY; (2) RANKS OF MODIFIED DATA ADJUSTED FOR THE PATTERN OF MISSING VALUES WITH THE TOTAL SAMPLE SIZE INDICATED PARENTHETICALLY; AND (3) RANKS OF MODIFIED DATA WITHOUT PATTERN ADJUSTMENT AT EACH CLINIC VISIT

	Visit 1 Comparisons[a]	Visit 2 Comparisons			Visit 3 Comparisons		
Clinic	Adjusted[1]	Observed	Adjusted[1]	Assigned means[2]	Observed	Adjusted[1]	Assigned means[2]
5	11.3**	14.5**	14.6**	14.8**	12.6**	12.0**	12.4*
	(19,18)[3]	(16,16)	(19,18)	(19,18)	(15, 9)	(19,18)	(19,18)
6	9.6**	8.0**	8.0**	8.0**	5.0*	4.8*	4.5*
	(17,16)	(15,13)	(17,16)	(17,16)	(16,10)	(17,16)	(17,16)
8	6.4*	7.5**	7.4**	7.5**	14.1**	14.1**	14.4**
	(17,13)	(16,11)	(17,13)	(17,13)	(16,12)	(17,13)	(17,13)
9	2.0	3.0	3.0	3.0	3.0	3.0	3.0
	(1, 3)	(1, 3)	(1, 3)	(1, 3)	(1, 3)	(1, 3)	(1, 3)
10	7.4**	10.5**	10.5**	10.3**	13.7**	13.5**	14.2**
	(18,17)	(18,16)	(18,17)	(18,17)	(18,14)	(18,17)	(18,17)
11	14.3**	16.1**	16.0**	14.2**	12.2**	12.2**	12.6**
	(16,17)	(14,17)	(16,17)	(16,17)	(13,15)	(16,17)	(16,17)
Combined	48.0**	56.2**	56.0**	54.6**	56.3**	55.4**	56.0**
	(88,84)	(80,76)	(88,84)	(88,84)	(76,63)	(88,84)	(88,84)

1 adjusted for missing value patterns which were not significantly ($p>.10$) different between treatment groups.

2 missing values were assigned the within-clinic mean rank.

3 sample sizes: (test drug, placebo).

a observed and assigned means statistics show little variation because only 3 patients had missing data across clinics

* $.01 \leq p \leq .05$ ** $p \leq .01$

can be viewed as a modification for the case of incomplete randomized blocks or tables respectively.

The incomplete block strategy was useful in assessing average investigator comparability across principally affected joints and treatment blocks. The average partial association statistic was useful in combining information across blocks which had very small sample sizes. Moreover, the randomization methods can be viewed as a useful strategy for justifying the pooling of samples especially for reasons related to limited sample sizes. The strategy involves defining the comparison groups (treatments) as a stratification factor and defining the separate samples to be pooled as the subpopulations.

For incomplete data in multivariate response profiles, the randomization methods can include a missing value pattern adjustment which accurately describes the structure of the modified data in contrast to assuming that the modified missing values were observed. Furthermore, as shown in Section 4 of Chapter II, randomization statistics based on missing values which are assigned a within-stratum constant are invariant to the value being assigned. The overall covariance formulation however still involves the augmented (total) sample size in contrast to the observed sample size as was the case for the weighted least squares ratio estimator approach (see Section 5 of Chapter I). This difference in using the observed sample size or total sample size is consistent with the different sampling frameworks for weighted least squares and randomization models respectively. More specifically, weighted least squares methods assume that the data are a result of a stratified random sampling of some large population for which the observed sample size is the appropriate basis of variance estimation. Alternatively, randomization models pertain to the partition of responses from a fixed size finite population among the fixed-sized subpopulations. In this regard, the occurrence of a missing value can be viewed as an occurrence of an unknown but real value which is assumed to be independent of the hypothetical value and occurring at random across subpopulations. For these aspects of the randomization model, the total sample size (which has been termed the augmented sample size) can be viewed as an appropriate basis of variance estimation.

CHAPTER VIII

SUMMARY AND FUTURE RESEARCH

8.1 Summary

The randomization concept has been shown to be central to a broad class of nonparametric and categorical procedures. The commonality of these procedures is the hypothesis of no association or no partial association to be tested within a finite population sampling framework. The finite population assumption establishes a distribution-free environment for which the observed partition of responses among subpopulations is viewed as one of a number of equally likely possible partitions of the pooled population set of responses. For categorical data, the possible partitions are enumerated according to the hypergeometric model. For continuous data, the enumeration corresponds to partitions of rank-type scores for the pooled population set of responses. Correspondingly, the expected values and variance-covariance for each subpopulation are based upon the overall population means and variance-covariance for the responses. In Chapter I, the literature review covered categorical data statistics such as Pearson's test for r x c contingency tables and Mantel-Haenszel's mean score and correlation tests for q sets of s x r contingency tables, and nonparametric procedures such as the Kruskal-Wallis, Friedman and Spearman rank procedures. Given the commonality of randomization across the categorical and nonparametric procedures, a set of related strategies were reviewed. These strategies included a multivariate trace criterion applied to ranks, Quade's rank analysis of covariance and Mantel-Haenszel's stratification average partial association. Also, a weighted least-squares ratio estimate strategy for incomplete data was reviewed as a method for computing statistics based on observed data and the pattern of missing data in a manner invariant to the value assigned to missing data.

The technical structure of a general raw-data formulation for mean score randomization statistics was presented in Chapter II. This structure included the multivariate trace criterion, analysis of covariance and stratification techniques. These techniques added complementary extensions to the existing categorical and nonparametric procedures reviewed in Chapter I. In addition to the generalized randomized block (average partial association) statistic, methods in stratification included Fisher's combined P-value statistic and an average standard normal statistic for one-tailed alternatives. Other special topics addressed in Chapter II included the definition of scores and their association with existing statistics and the missing values strategy employing binary indicators as covariates. For this missing value strategy, the randomization statistic was shown to be invariant to the constant assigned missing values. Also, the maintaining of the sample size as the total number of observations in contrast to the number of observed values was viewed as consistent with the finite population theory. More specifically, the total number of observations were observed and a fixed number had missing values at random.

The applications of the mean score randomization statistics in Chapters III, IV, V, VI and VII involved two concepts of randomization.

CHAPTER VIII

SUMMARY AND FUTURE RESEARCH

8.1 Summary

The randomization concept has been shown to be central to a broad class of nonparametric and categorical procedures. The commonality of these procedures is the hypothesis of no association or no partial association to be tested within a finite population sampling framework. The finite population assumption establishes a distribution-free environment for which the observed partition of responses among subpopulations is viewed as one of a number of equally likely possible partitions of the pooled population set of responses. For categorical data, the possible partitions are enumerated according to the hypergeometric model. For continuous data, the enumeration corresponds to partitions of rank-type scores for the pooled population set of responses. Correspondingly, the expected values and variance-covariance for each subpopulation are based upon the overall population means and variance-covariance for the responses. In Chapter I, the literature review covered categorical data statistics such as Pearson's test for r x c contingency tables and Mantel-Haenszel's mean score and correlation tests for q sets of s x r contingency tables, and nonparametric procedures such as the Kruskal-Wallis, Friedman and Spearman rank procedures. Given the commonality of randomization across the categorical and nonparametric procedures, a set of related strategies were reviewed. These strategies included a multivariate trace criterion applied to ranks, Quade's rank analysis of covariance and Mántel-Haenszel's stratification average partial association. Also, a weighted least-squares ratio estimate strategy for incomplete data was reviewed as a method for computing statistics based on observed data and the pattern of missing data in a manner invariant to the value assigned to missing data.

The technical structure of a general raw-data formulation for mean score randomization statistics was presented in Chapter II. This structure included the multivariate trace criterion, analysis of covariance and stratification techniques. These techniques added complementary extensions to the existing categorical and nonparametric procedures reviewed in Chapter I. In addition to the generalized randomized block (average partial association) statistic, methods in stratification included Fisher's combined P-value statistic and an average standard normal statistic for one-tailed alternatives. Other special topics addressed in Chapter II included the definition of scores and their association with existing statistics and the missing values strategy employing binary indicators as covariates. For this missing value strategy, the randomization statistic was shown to be invariant to the constant assigned missing values. Also, the maintaining of the sample size as the total number of observations in contrast to the number of observed values was viewed as consistent with the finite population theory. More specifically, the total number of observations were observed and a fixed number had missing values at random.

The applications of the mean score randomization statistics in Chapters III, IV, V, VI and VII involved two concepts of randomization.

These concepts are design based randomization with random assignment to subpopulations and randomization as a hypothesis for observational studies. The role of post stratification is explored in both design cases. In Chapter VI, stratification was used to define homogeneous levels of risk to hepatotoxicity so that within each level the hypothesis of randomization could be tested contrasting the outcome from one observational study with the outcome from an historical control study. While in the second example of Chapter III, post stratification on age levels was used to enhance the randomized design by providing a more precise covariance structure for testing the randomization hypothesis. Here, covariance adjustment for age in years was also provided within each age level stratum to demonstrate how stratification and covariance adjustment might be used jointly for a single variate. This joint use of the same variate can minimize the number of strata necessary to provide an effective covariance structure. In this regard, this strategy can represent an alternative to matched pair comparisons in the sense of expanding comparisons to include more than one observation per group.

Another issue of comparability was addressed in the first example of Chapter III. In this example, repeated measures were grouped according to levels of a third variable which has a non-linear relationship over the range of response. The grouping of data points within levels of the third variable resembles a stratification strategy in the sense that strata defined a more homogeneous framework within which to make comparisons. However, here the same patients could enter more than one stratum. An average stratum result (called an average central ridit) was computed for each patient to adjust for this lack of across-strata independence. A univariate Wilcoxon-Rank Sum statistic was applied to the across-patient set of average central ridits.

The statistical procedures demonstrated in this dissertation included covariance adjustment (second example in Chapter III), multivariate response profiles (Chapter IV), stratification strategies (Chapter V), scoring (Chapter VI), and incomplete design and incomplete data (Chapter VII). These applications illustrated the extensions of existing statistics such as an incomplete covariate-adjusted multivariate extended-Friedman statistic (Chapter VII) or a multivariate covariate-adjusted Mantel-Haenszel statistic (Chapter VI). Also illustrated, is the diversity of the randomization-model role in analyses. In observational studies such as in Chapter VI, the randomization model provided a viable framework for comparing subpopulations. In the first example of Chapter VII, the randomization model provided the justification for combining several small populations into one population for further analysis. In Chapter IV, the randomization model analysis was the basis of an extensive descriptive analysis of a multivariate response profile. The results were used to formulate a WLS\ model as well as to compare subpopulations in a complex stratification system without having to describe the inter-relationship of the strata nor assume multivariate-normality of the binary response measures. Here, results from orthogonal contrasts across the multivariate profile indicated justification for a reduction of the eight measures to two subset averages. Results of expressing the two subset averages as an overall average and the difference between the two subset averages supported a reduction to a single overall average. This overall average was the univariate summary measure in the WLS model for the eight dimensional response profile.

Applications in the second example of Chapter III and in Chapter IV emphasized the complementary role of randomization methods in contrast to its being an alternative for modeling procedures such as WLS or MLE. The issues here were limitations of the randomization model. More specifically, the generality of results pertained only to the population observed; covariate adjustment required covariates to be at random with respect to the subpopulations; the inter-relationship of strata could not be evaluated in the modeling sense; subpopulation comparisons involved an average stratum result in contrast to an across stratum model-smoothed result; and across strata model-smoothed predicted values and estimated standard errors were not relevant to the finite population sampling frame. Alternatively, randomization model analysis provided an appropriate framework for significance tests in a straightforward manner across a range of research design complexity. Modeling procedures can be used to describe nonrandom variation detected through the rejection of the randomization model. Other roles of randomization statistics have been mentioned above.

8.2 Direction for Future Research

Both mean score and correlation randomization statistics were reviewed in Chapter I. However, the general raw-data formulation and its applications in this dissertation pertained only to the mean score statistic. Thus, a similar presentation of correlation randomization statistics would be useful. Here, the methodology for a small number of subpopulations ($s < 6$) with moderate sample sizes ($n_{hi} \geq 20$) could be extended to a large number of subpopulations with small sample sizes provided that these subpopulations can be ordinarily scaled and the composite sample size is at least moderate. Correlation randomization statistics would include the Spearman Rank correlation for continuous data with rank-type scores and the measures of association available through PARCAT for either one or several strata of s x r tables; see Landis, Heyman and Koch (1978). In the contingency table setting, the measure of association based on rank scores corresponds to the Spearman Rank correlation extended to handle ties with mid-rank values. Extensions to raw-data files, multivariate responses and covariates would follow lines involving a merger of the generalized formulation for mean score statistics in this dissertation with the current PARCAT capability for randomization correlation analysis in the Spearman sense.

Another area for future research is the extension of the application of randomization methods in variable selection for categorical data. Higgins and Koch (1977) describe such a variable selection procedure for the case of binary response variables. Here, the variable selection procedure has a series of steps each with two stages. The first stage identifies an ordering of eligible variables with regard to the total explained variation in a multi-way cross-classification of the response variable. The second stage uses partial association randomization methods to calculate the significance of the eligible variable after adjustment for other variables already included. This adjustment is made without an assumed model. This strategy could incorporate the more general mean score and correlation randomization statistics. As a result of the more general formulation, extensions would include ordinal and multivariate response profiles with or without covariance adjustment. Furthermore, significance tests could involve either correlation or mean score statistics.

A third area for future research is the definition of additional types of scores. For repeated measurement experiments, a group by response interaction test, called a profile test in the discussion of split-plot analyses by Koch (1969), can be constructed as a multivariate Kruskal-Wallis statistic applied to a set of scores representing the (d - 1) orthogonal contrasts of d response measures. The profile test is directed toward assessing the homogeneity of treatment group differences across a repeated measurement dimension of response. For example, in Chapter IV, treatment differences tended not to be homogeneous across time from one dosing period to the next although they were generally similar especially within dosing periods. A multivariate profile test was constructed as orthogonal contrasts within each dosing period. The results which did not show significant variation across time in treatment group differences indicated the sufficiency of dose period averages as summary measures. Also in Chapter VII, the missing data pattern across visits in the second example represents an additional consideration. The profile test described by Koch (1969) could be undertaken relative to multiple strata and potentially with adjustment for covariates. Finally, both the correlation and variable selection strategies previously described are as applicable to scores representing profile contrasts as they are to response variables in their own right.

LIST OF REFERENCES

Amara, I. A., Davis, G. W. and Koch, G. G. (1981). Some statistical methods for covariance analysis of categorical data. Proceedings of the Statistical Computing Section of the American Statistical Association, 96-101.

Amara, I. A. and Koch, G. G. (1980). A macro for multivariate randomization analyses of stratified sample data. Proceedings of the Fifth Annual SAS Users Group International Conference, 134-144.

Barr, A., Goodnight, J. H., Sall, J. P., Blair, W. H. and Chilko, D. M. (1979). SAS User's Guide, 1979 Edition, Jane T. Helwig and Kathryn A. Council, eds., Raleigh, N. C.: SAS Institute, Inc.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society B 37, 129-146.

Birch, M. W. (1964). The detection of partial association, I: the 2 x 2 case. Journal of the Royal Statistical Society B 26, 313-324.

Birch, M. W. (1965). The detection of partial association, II: the general case. Journal of the Royal Statistical Society B 27, 111-124.

Chatterjee, S. K. and Sen, P. K. (1966). Non-parametric tests for the multivariate multi-sample location problem. S. N. Roy Memorial Volume. Edited by R. C. Bose et al.

Cochran, W. G. (1954). Some methods for strengthening the common tests. Biometrics 10, 417-451.

Friedman, M. (1937). The use of ranks to avoid the assumptions of normality implicit in the analysis of variance. Journal of the American Statistical Association 32, 675-701.

Gerig, T. (1969). A multivariate extension of Friedman's test. Journal of the American Statistical Association 64, 1595-1608.

Gibbons, J. D. (1971). Nonparametric Statistical Inference. New York: McGraw-Hill.

Gibbons, J. D. (1976). Nonparametric Methods for Quantitative Analysis. New York: Holt, Reinehart and Winston.

Hajek, J. (1969). Nonparametric Statistics. San Francisco: Holden-Day.

Harrell, F. (1980). The LOGIST procedure. SAS Supplemental Library User's Guide, 1980 Edition, Patti S. Reinhardt, ed., Raleigh, N. C.: SAS Institute, Inc.

Higgins, J. E. and Koch, G. G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. International Statistical Review, 45, 51-62.

Hopkins, C. E. and Gross, A. J. (1971). A generalization of Cochran's procedure for the combining of r x c contingency tables. Statistica Neerlandica 25, 57-62.

John, J. A. and Prescott, P. (1975). Estimating missing values in experiments. Applied Statistics 24, 190-192.

Johnson, C., Amara, I. A., Edwards, S. and Koch, G. G. (1981). Some computational strategies for the analysis of longitudinal categorical data. Proceedings of the Statistical Computing Section of the American Statistical Association, 293-298.

Kempthorne O. (1955). The randomization theory of experimental inference. Journal of the American Statistical Association 50, 946-967.

Koch, G. G. (1969). Some aspects of the statistical analysis of "split plot" experiments in completely randomized layouts. Journal of the American Statistical Association 64, 485-505.

Koch, G. G. (1970). The use of non-parametric methods in the statistical analysis of complex split plot experiment. Biometrics 26, 105-128.

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. and Bhapkar, V. P. (1982). Chi-square tests. Encyclopedia of Statistical Sciences. Norman L. Johnson and Samuel Kotz, eds., New York: John Wiley & Sons, Inc., 1, 442-457.

Koch, G. G. and Edwards, S. (1983). Logistic regression. Entry to appear in Encyclopedia of Statistical Sciences, Norman L. Johnson and Samuel Kotz, eds., to be published by John Wiley & Sons, Inc.

Koch, G. G. and Gillings, D. B. (1983). Inference, design based vs. model based. Entry to appear in Encyclopedia of Statistical Sciences, Normal L. Johnson and Samuel Kotz, eds., to be published by John Wiley & Sons, Inc.

Koch, G. G., Gillings, D. B., and Stokes, M. E. (1980). Biostatistical implications of design, sampling, and measurement to health science data analysis. Annual Review of Public Health 1, 163-225.

Koch, G. G. and Reinfurt, D. W. (1974). An analysis of the relationship between driver injury and vehicle age for automobiles involved in North Carolina accidents during 1966-1970. Accident Analysis and Prevention 6, 1-18.

Koch, G. G., Sen, P. K. and Amara, I. (1983). Logrank scores, statistics, and tests. Entry to appear in Encyclopedia of Statistical Sciences, Norman L. Johnson and Samuel Kotz, eds., to be published by John Wiley & Sons, Inc.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis. Journal of the American Statistical Association 47, 583-621. Addendum: Ibid. (1953), 907-911.

Landis, J. R., Cooper, M. M., Kennedy, T., and Koch, G. G. (1977). A computer program for testing average partial association in three-way contingency tables (PARCAT). Computer Programs in Biomedicine 9, 223-246.

Landis, J. R., Heyman, E. R., and Koch, G. G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. International Statistical Review 46, 237-254.

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 squares (GENCAT). Computer Programs in Biomedicine 6, 196-231.

Lehmann, E. L. (1975). Nonparametrics: Statistical Methods Based on Ranks. San Francisco: Holden-Day.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 18, 50-60.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association 58, 690-700.

Mantel, N. (1966). Models for complex contingency tables and polychotomous dosage response curves. Biometrics 22, 83-95.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22, 719-748.

Puri, M. L. and Sen, P. K. (1969). Analysis of covariance based on general rank scores. The Annals of Mathematical Statistics 40, 610-618.

Puri, M. L. and Sen, P. K. (1971). Non-parametric Methods in Multivariate Analysis. New York: John Wiley.

Quade, D. (1967). Rank analysis of covariance. Journal of the American Statistical Association 62, 1187-1200.

Quade, D. (1972b). Analyzing randomized blocks by weighted rankings. Afdeling Mathematische Statistiek, Amsterdam, July 1972.

Stanish, W. M. (1978). Adjustment for covariables in categorical variable selection and in multivariate partial association tests. University of North Carolina Ph.D. Dissertation.

Stanish, W. M., Gillings, D. B., and Koch, G. G. (1978). An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data. Biometrics 34, 305-317.

van Elteren, P. H. (1960). On the combination of independent two-sample tests of Wilcoxon. Bulletin of the International Statistical Institute 37, 351-361.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics 1, 80-83.