Multiple Testing in Discrete Data Setting

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By

Yan Li, M.S.

Graduate Program in Public Health
With specialty in Biostatistics

The Ohio State University
2010

Dissertation Committee:
Professor Bo Lu, Adviser
Professor Jason C. Hsu
Professor Eloise Kaizar
ABSTRACT

Multiple comparisons are common practice in statistical analysis. How to construct valid and powerful multiple testing procedures that control multiple testing error rates remains an interesting and challenging question.

Resampling methods and modeling are important tools to capturing the data structure, which makes construction of more powerful multiple testing procedures possible. In this work, we study the limitations of permutation tests as well as construct step-down short cut version of the partitioning test based on modeling in the discrete data analysis setting.

For example, in pharmacogenomics, multiple testing for significant association between genetic markers and phenotypes is of interest. Permuting response group labels to generate a reference distribution is often thought of as a convenient thresholding technique that automatically captures dependence in the data. It is shown in this work that, without non-trivial model assumptions, permutation testing may not control unconditional multiple testing error rates. We also show the lack of control for the conditional error rate.

When modeling is possible, analytical derivation of the joint distributions of the test statistics may be feasible. Upon satisfaction of certain sufficient conditions, we show how to construct a more powerful step-down short-cut version of the partitioning test. Discussions and examples are within the context of logistic regression modeling.
for the independent binary outcome variable and GEE modeling for the correlated binary outcome variable.
To my family.
ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of so many people in so many ways. I sincerely appreciate Dr. Jason C. Hsu and Dr. Eloise Kaizar for their guidance and encouragement in this long journey. I also thank Dr. Bo Lu very much for his comments and accepting me as his student at my most difficult time.
VITA

1999 ................................. B.M. Basic Medical Science, Peking University, Health Science Institute

2002 ................................. M.S. Molecular Virology, Case Western Reserve University

2004 ................................. M.S. Biostatistics, The Ohio State University

FIELDS OF STUDY

Major Field: Public Health with specialty in Biostatistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction to multiple testing concepts and existing methods for</td>
<td>1</td>
</tr>
<tr>
<td>various models</td>
<td></td>
</tr>
<tr>
<td>1.1 Linear Regression Models with normal errors</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Multiple Testing in Linear Regressions with normal errors</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Computer Software</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Fundamental Concepts in Multiple Testing</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1 Different Error Rates in Multiple Testing</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 The Partitioning Principle</td>
<td>5</td>
</tr>
<tr>
<td>1.2.3 Shortcut the Partitioning Test</td>
<td>7</td>
</tr>
<tr>
<td>1.3 GLMs (Generalized Linear Models)- Logistic Regression Models</td>
<td>13</td>
</tr>
<tr>
<td>1.3.1 GLM model</td>
<td>13</td>
</tr>
<tr>
<td>1.3.2 Logistic Regression Model</td>
<td>15</td>
</tr>
<tr>
<td>1.3.3 Estimation and Inference of Logistic Regression</td>
<td>15</td>
</tr>
<tr>
<td>1.3.4 Coefficient Interpretation</td>
<td>17</td>
</tr>
<tr>
<td>1.3.5 Multiple Testing in Logistic Regression</td>
<td>18</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>1.3.6 Computer Software</td>
<td>19</td>
</tr>
<tr>
<td>1.4 Notation for Grouped Measurements</td>
<td>19</td>
</tr>
<tr>
<td>1.5 Mixed Effects Models</td>
<td>21</td>
</tr>
<tr>
<td>1.5.1 Linear Mixed Effects Models (LMMs)</td>
<td>22</td>
</tr>
<tr>
<td>1.5.2 Generalized Linear Mixed Effects Models (GLMMs)</td>
<td>24</td>
</tr>
<tr>
<td>1.6 Marginal Models</td>
<td>26</td>
</tr>
<tr>
<td>1.6.1 Model Introduction</td>
<td>26</td>
</tr>
<tr>
<td>1.6.2 Estimation: GEE</td>
<td>27</td>
</tr>
<tr>
<td>1.6.3 Computer Software</td>
<td>29</td>
</tr>
<tr>
<td>1.6.4 Multiple Testing in GEE Setting</td>
<td>30</td>
</tr>
<tr>
<td>2. Permutation tests based multiple testing in discrete data</td>
<td>32</td>
</tr>
<tr>
<td>2.1 Motivating example: Genome-Wide Association Study (GWAS)</td>
<td>34</td>
</tr>
<tr>
<td>2.1.1 Single Nucleotide Polymorphisms (SNPs)</td>
<td>34</td>
</tr>
<tr>
<td>2.1.2 GWAS and pharmacogenomics</td>
<td>35</td>
</tr>
<tr>
<td>2.2 Permutation Tests</td>
<td>41</td>
</tr>
<tr>
<td>2.3 Permutation test in GWAS: 2 SNPs example</td>
<td>44</td>
</tr>
<tr>
<td>2.3.1 ( \text{maxT} ) test</td>
<td>50</td>
</tr>
<tr>
<td>2.3.2 ( \text{minp} ) test</td>
<td>61</td>
</tr>
<tr>
<td>2.3.3 Example of MDJ does not hold</td>
<td>65</td>
</tr>
<tr>
<td>2.4 Discussion</td>
<td>87</td>
</tr>
<tr>
<td>3. Independent Binary Outcomes: Logistic Regression</td>
<td>100</td>
</tr>
<tr>
<td>3.1 Simpler case for binary outcome</td>
<td>100</td>
</tr>
<tr>
<td>3.1.1 ( 2 \times 2 ) contingency table and Exact Conditional tests</td>
<td>100</td>
</tr>
<tr>
<td>3.2 Logistic Regression</td>
<td>103</td>
</tr>
<tr>
<td>3.3 Applying Partition Testing to Logistic Regression Model</td>
<td>104</td>
</tr>
<tr>
<td>3.3.1 Motivating Example</td>
<td>106</td>
</tr>
<tr>
<td>3.3.2 Problem Formulation</td>
<td>106</td>
</tr>
<tr>
<td>3.3.3 The ‘Comprehensive’ Model</td>
<td>107</td>
</tr>
<tr>
<td>3.3.4 The Test Statistics and Critical Values</td>
<td>110</td>
</tr>
<tr>
<td>3.3.5 Apply Partition Testing in Logistic Regression Model</td>
<td>112</td>
</tr>
<tr>
<td>3.3.6 Step-down Short-cut of Partition Test in Logistic Regression</td>
<td>114</td>
</tr>
<tr>
<td>3.3.7 Example analysis</td>
<td>117</td>
</tr>
<tr>
<td>3.4 Likelihood Ratio Tests Cannot Shortcut</td>
<td>124</td>
</tr>
<tr>
<td>4. Multiple Testing in Marginal Models</td>
<td>136</td>
</tr>
<tr>
<td>4.1 Marginal Models and GEE</td>
<td>136</td>
</tr>
<tr>
<td>4.2 Motivating Example</td>
<td>140</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table                                      Page

2.1  Table of dichotomized genotypes and phenotype groups in Abacavir example (Mallal et al., 2008). ............................... 39
2.2  Table of dichotomized genotypes and phenotype groups in Flucloxacillin example (Daly et al., 2009). ............................. 40
2.3  Table of pharmacogenomic examples. ........................................ 41
2.4  Table of biallelic SNP genotypes and phenotypes. ......................... 44
2.5  Table of dichotomized genotypes and phenotype groups. .................. 45
2.6  Marginal underlying probabilities for SNP_i. .............................. 45
2.7  Table of observed samples for k = 2. ...................................... 48
2.8  Table of underlying joint probabilities for k = 2. ......................... 49
2.9  One numerical example where MDJ condition does not hold. .......... 67
2.10 Numerical example from Figure 2.1. ..................................... 72
2.11 Conditional permutation probabilities and true probabilities of T = t. 72
2.12 Joint probabilities configuration of Figure 2.6. ............................ 79
2.13 Joint probabilities configuration of Figure 2.7. ............................ 80
2.14 Joint probabilities configuration of Figure 2.8. ............................ 80
2.15 Joint probabilities configuration of Figure 2.9. 81
2.16 Joint probabilities configuration of Figure 2.10. 83
2.17 Joint probabilities configuration of Figure 2.11. 84

3.1 Test statistics, critical values, P-values and confidence intervals by step-down shortcut of partitioning test for data from (Dasgupta et al., 2000) 122
3.2 Test statistics, critical values, P-values and confidence intervals by single-step Dunnett’s method for data from (Dasgupta et al., 2000) 122
3.3 Test statistics, critical values, P-values and confidence intervals by Holm’s method for data from (Dasgupta et al., 2000) 122

4.1 Summary of Toxicology Data, Paul S.R. Analysis of Proportions of Affected fetuses in Teratological Experiments, Biometrics 1982, 38:2, 361-70. ‘M’-malformed fetuses; ‘T’- total number of fetuses. 141
4.2 Test statistics, critical values, P-values and confidence intervals by step-down shortcut of partition testing for data from (Paul, 1982) 158
4.3 Test statistics, critical values, P-values and confidence intervals by single-step Dunnett’s method for data from (Paul, 1982) 158
4.4 Test statistics, critical values, P-values and confidence intervals by Holm’s method for data from (Paul, 1982) 158
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Maximum difference between the conditional permutation probability and the true null probability of $T = t$ when $n^D \leq n^N, \pi_{++}^D &gt; \pi_{++}^N$</td>
<td>70</td>
</tr>
<tr>
<td>2.2 Maximum difference between the conditional permutation probability and the true null probability of $T = t$ when $n^D \leq n^N, \pi_{++}^D &lt; \pi_{++}^N$</td>
<td>71</td>
</tr>
<tr>
<td>2.3 Maximum difference in p-values between the conditional permutation distribution and the true null distribution when $n^D \leq n^N, \pi_{++}^D &gt; \pi_{++}^N$</td>
<td>76</td>
</tr>
<tr>
<td>2.4 Maximum difference in p-values between the conditional permutation distribution and the true null distribution when $n^D \leq n^N, \pi_{++}^D &lt; \pi_{++}^N$</td>
<td>77</td>
</tr>
<tr>
<td>2.5 Maximum difference in p-values between the conditional permutation distribution and the true null distribution when $n^D \leq n^N, \pi_{++}^D &gt; \pi_{++}^N$</td>
<td>78</td>
</tr>
<tr>
<td>2.6 True null distribution and permutation null distribution of $T = \max_{i=1,2} T_i, T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$ for unequal sample sizes 2-sided test.</td>
<td>92</td>
</tr>
<tr>
<td>2.7 True null distribution and permutation null distribution of $T = \max_{i=1,2} T_i, T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$ for equal sample sizes 1-sided test.</td>
<td>93</td>
</tr>
<tr>
<td>2.8 True null distribution and permutation null distribution of $p = \min_{i=1,2} p_i$ for unequal sample sizes 1-sided test.</td>
<td>94</td>
</tr>
<tr>
<td>2.9 True null distribution and permutation null distribution of $p = \min_{i=1,2} p_i$ for equal sample sizes 1-sided test.</td>
<td>95</td>
</tr>
<tr>
<td>2.10 True null distribution and permutation null distribution of $T = \max_{i=1,2} T_i, T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$ for unequal sample sizes 1-sided test.</td>
<td>96</td>
</tr>
</tbody>
</table>
2.11 True null distribution and permutation null distribution of \( p = \min_{i=1,2} p_i \)
for unequal sample sizes 1-sided test. .................................................. 97

2.12 Difference between smallest p-values of true null distribution and per-
mutation null distribution of \( T = \max_{i=1,2} T_i \) for unequal sample sizes
2-sided test. ........................................................................................................ 98

2.13 Difference between smallest p-values of true null distribution and per-
mutation null distribution of \( T = \max_{i=1,2} T_i \) for equal sample sizes
2-sided test. ........................................................................................................ 99

3.1 Honey bee example result: test statistics and critical values from the
step-down partition test, the single-step extended Dunnett’s method
and Holm’s method. ......................................................................................... 123

3.2 Rejection regions by likelihood ratio statistics: \( H_{0\{1\}} : \theta_1 = 0 \) and
\( \theta_2 = 0 \) (outside of the circle), \( H_{0\{1\}} : \theta_1 = 0 \) or \( H_{0\{2\}} : \theta_2 = 0 \) (outside
of the dashed lines) ..................................................................................... 132

3.3 Rejection regions by maxT statistics: \( H_{0\{1\}} : \theta_1 = 0 \) and \( \theta_2 = 0 \)
(outside of the straight lines), \( H_{0\{1\}} : \theta_1 = 0 \) or \( H_{0\{2\}} : \theta_2 = 0 \) (outside
of the dashed lines) ..................................................................................... 135

4.1 Toxicity example result: test statistics and critical values from the
step-down partition test, the single-step extended Dunnett’s method
and Holm’s method. ......................................................................................... 159
CHAPTER 1

INTRODUCTION TO MULTIPLE TESTING CONCEPTS
AND EXISTING METHODS FOR VARIOUS MODELS

The multiple testing problem, which is about testing more than one null hypothesis simultaneously, is commonly encountered in scientific studies.

In biomedical research and clinical trials, comparisons among various treatment groups, patient sub-populations or clinical endpoints are of great interest. For example, in a phase II clinical trial, different doses of a new medicine could be administered to patients to study the efficacy in reducing level of pain. In a safety study, various adverse effects are to be tested to find if significant differences in the adverse events rates exist between the treated group and the control group. To locate the tumor suppressor genes, binary event LOH (loss of heterozygosity) or ROH (retention of heterozygosity) will be measured on multiple marker loci from a group of patients and compared for LOH probabilities. Modern biotechnologies have developed into the stage that tens of thousands of genetic markers can be measured all at once. Thus tens of thousands of comparisons may be performed simultaneously.
It is well known that without appropriate adjustment, the probability of making erroneous rejection increases with the number of tests. This fact brings up the importance and necessity of developing multiplicity adjustment methods, especially in the discrete data setting.

In this chapter, we make a brief introduction to various modeling techniques in different data settings as well as basic but important multiple testing concepts.

1.1 Linear Regression Models with normal errors

To model independent normally distributed variables, simple linear regression models using the Ordinary Least Square (OLS) estimation method is a classical approach, i.e.

\[ Y = X\beta + \epsilon \quad (1.1) \]

where \( Y \) is the \( n \times 1 \) vector of the outcome, \( X \) is a \( n \times k \) design matrix and parameters of interest are \( \beta \). A normal distribution with equal variance is assumed, i.e. \( \epsilon_i \) independently identically distributed (i.i.d) \( N(0, \sigma^2) \).

1.1.1 Multiple Testing in Linear Regressions with normal errors

Presented in the balanced one-way model form, for independent observations \( Y_{ia} \) from the \( i \)th treatment group, \( Y_{ia} = \mu_i + \epsilon_{ia}, \epsilon_{ia} \sim N(0, \sigma^2); i = 1, \ldots, k; a = 1, \ldots, n_i \) (Hsu, 1999). Denote the multiple comparison parameters of interest by

\[ \mu_i - \mu_j, i = 1, \ldots, k; j = 1, \ldots, k; j \neq i \quad (1.2) \]
Many multiple testing methods are readily available for balanced and unbalanced one-way models as discussed in great details by Hsu, (Hsu, 1999). These include the Bonferroni’s method, Dunnett’s method for multiple comparisons with a control (MCC), multiple comparisons with the best (MCB) based on MCC and Tukey’s method for all pairwise comparisons (MCA).

Let \( \sigma^2 v^i_j \) denote the variance of \( \hat{\mu}_i - \hat{\mu}_j \). If there exists a set of positive constants \( \alpha_1, \ldots, \alpha_k \), such that \( v^i_j = \alpha_i + \alpha_j \) for all \( i \neq j \), then it is called the ‘one-way structure’ model by Hsu (Hsu, 1999). With this structure, multiple comparisons can be executed as were in the one-way model with proper changes in the formula as pointed out in Chapter 7 (Hsu, 1999).

### 1.1.2 Computer Software

Many statistical software packages are available for fitting linear models. For example, the SAS GLM procedure can be used. By specifying ADJUST=BON or =SIDAK in the LSMEANS statement, multiplicity adjustments based on either disjointness or independence will be performed. ADJUST=DUNNETT or TUKEY will implement corresponding Dunnett’s method for MCC or Tukey’s method for MCA. For unbalanced data, the factor analytic method (Hsu, 1992) or the Edwards-Berry simulation based approach (Edwards and Berry, 1987) can be specified.

The next section briefly reviews some basic concepts in the multiple testing problem setting.

### 1.2 Fundamental Concepts in Multiple Testing

In general, the multiple testing problem can be stated as testing null hypotheses about \( k \) parameters simultaneously, \( H_{0i} : \theta_i \in \Theta_i, i = 1, \ldots, k ; \theta_i \) could be a vector
of mean treatment effects, or differences between treatment groups, or proportions, or other parameters of interest.

1.2.1 Different Error Rates in Multiple Testing

Among the $k$ null hypotheses, a number of them will be true and the rest will be false. Let $I$ denote the non-empty set of indices corresponding to true null hypotheses.

Let $V$ denote the number of erroneously rejected true null hypotheses. Given the non-empty collection of true null hypotheses, $\{H_{0i}, i \in I\}$, so that $\theta \in \Theta^I$, the definition of FWER is

$$FWER = \sup_{\theta \in \Theta^I} P_{\theta}(\text{Reject at least one true null hypothesis } H_{0i}, i \in I)$$

$$= \sup_{\theta \in \Theta^I} P_{\theta}(V > 0)$$

(1.3)

where the $sup$ is taken over the entire parameter space $\Theta$, which means, for any possible subset $I \in \{1, \ldots, k\}$, the probability of rejecting at least one true null hypothesis is well controlled at the desired level $\alpha$. Controlling FWER in the ‘strong’ sense means the probability of rejecting any true null hypotheses is kept at a pre-specified low level under any parameter configuration under $\theta \in \Theta^I$.

Generalized familywise error rate (gFWER) is a less stringent error rate in multiple testing,

$$gFWER = \sup_{\theta \in \Theta^I} P_{\theta}(V > m)$$

(1.4)

where $m$ is a small positive number. Therefore, when thousands of tests are performed, the probability of making a small number of false rejections is controlled.

A differently defined error rate in the multiple testing setting is the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). Let $R$ be the number of total rejections larger than 0. Then FDR is defined:
\[ FDR = E\left[ \frac{V}{R} I_{\{R > 0\}} \right] \]

Notice the FDR definition is based on the expectation of the proportion of incorrect rejection of true null hypotheses. Without proper control of the number of false hypotheses, control of the FDR may not guarantee control of the actual number of wrong rejections out of all rejections.

### 1.2.2 The Partitioning Principle

The Partitioning Principle, (Stefansson et al., 1988; Finner and Strassburger, 2002), is a fundamental principle that can be relied on to construct valid multiple testing procedures that control FWER strongly.

Partition testing proceeds as follows:

1: Partition the entire parameter space \( \Theta_1 \times \cdots \times \Theta_k \) into \( 2^k \) disjoint subspaces so that each subspace contains exactly one subset of true null hypotheses, i.e.
\[ H^*_{0(I)} : \theta \in \bigcap_{i \in I} \Theta_i \cap (\bigcap_{j \notin I} \Theta_j^c) \] for all \( I \subseteq \{1, \ldots, k\} \).

2: Test each \( H^*_{0(I)} \) at level-\( \alpha \). There are at most \( 2^k - 1 \) subset tests. Since the \( H^*_{0(I)} \)'s are disjoint, at most one \( H^*_{0(I)} \) can be true. Therefore, even though no multiplicity adjustment is required in testing up to \( 2^k - 1 \) hypotheses, partition testing controls FWER strongly as long as each individual subset is tested at level \( \alpha \).

3: Infer \( \theta \notin \Theta_I, I \subseteq \{1, \ldots, k\} \) if and only if all \( H^*_{0(J)} I \subseteq J \), are rejected. That is, if all null hypotheses subsets \( H^*_{0(J)} \) implying \( H^*_{0(I)} \) are rejected then \( H^*_{0(I)} \) is rejected. This is a logical conclusion since the parameter space of \( H^*_{0(I)} \) is the union of the parameter spaces of all \( H^*_{0(J)} I \subseteq J \).
For example, let \( k = 2, \Theta_i = \{\theta_i \in \mathbb{R} | \theta_i \leq 0\} \). Possible one-sided multiple testing null hypotheses are:

\[
H_{01} : \theta_1 \leq 0 \text{ vs. } H_{a1} : \theta_1 > 0 \\
H_{02} : \theta_2 \leq 0 \text{ vs. } H_{a2} : \theta_2 > 0
\]  
(1.5)

In partition testing, we partition the parameter space into four disjoint subspaces and form the hypotheses:

\[
H^{*}_{0{\{12\}}} : \theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \\
H^{*}_{0\{1\}} : \theta_1 \leq 0 \text{ and } \theta_2 > 0 \\
H^{*}_{0\{2\}} : \theta_2 \leq 0 \text{ and } \theta_1 > 0 \\
H^{*}_{0} : \theta_1 > 0 \text{ and } \theta_2 > 0
\]  
(1.6)

According to the partitioning principle, one can test the first three hypotheses each at level \( \alpha \) to make valid inference for \( \theta_1 \) and \( \theta_2 \). To collate the partition testing results:

- If \( H^{*}_{0\{12\}} \) is not rejected, no inference is given.
- If \( H^{*}_{0\{12\}} \) is rejected but not \( H^{*}_{0\{1\}} \) or \( H^{*}_{0\{2\}} \), then the inference is at least one of \( \theta_1 > 0, \theta_2 > 0 \) is true, but which one cannot be specified.
- If \( H^{*}_{0\{12\}} \) and \( H^{*}_{0\{1\}} \) are both rejected but not \( H^{*}_{0\{2\}} \), then \( \theta_1 > 0 \) is inferred.
- If \( H^{*}_{0\{12\}} \) and \( H^{*}_{0\{2\}} \) are both rejected but not \( H^{*}_{0\{1\}} \), then \( \theta_2 > 0 \) is inferred.
- If \( H^{*}_{0\{12\}} \) and \( H^{*}_{0\{1\}} \) and \( H^{*}_{0\{2\}} \) are all rejected, then \( \theta_1 > 0 \) and \( \theta_2 > 0 \) is inferred.

For convenience, \( H^{*}_{0\{12\}}, H^{*}_{0\{1\}}, H^{*}_{0\{2\}} \) can be tested by
\[ H_{0(12)} : \theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \]

\[ H_{0(1)} : \theta_1 \leq 0 \]

\[ H_{0(2)} : \theta_2 \leq 0 \]  \hspace{1cm} (1.7)

since a level-\( \alpha \) test for \( H_{0(i)}, i = 1, 2 \) is guaranteed to be a level-\( \alpha \) test for \( H_{0(i)}^*, i = 1, 2 \).

The closed testing principle (Marcus et al., 1976) is similar to the Partitioning Principle but tests less restrictive hypotheses

\[ H_{0(12)} : \theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \]

\[ H_{0(1)} : \theta_1 \leq 0 \]

\[ H_{0(2)} : \theta_2 \leq 0 \]  \hspace{1cm} (1.8)

1: For each \( I \subseteq \{1, \ldots, k\} \), form the closure \( H_0(I) : \theta \in \{\theta_i \in \Theta_i \text{ for all } i \in I\} \).

2: Test each \( H_0(I) \) at level-\( \alpha \).

3: For each \( i \), infer \( \theta_i \not\in \Theta_i \) if and only if all \( H_0(I) \) with \( i \in I \) are rejected.

Closed Testing guarantees strong control of FWER because a level-\( \alpha \) test for \( H_0(I) \) is also a level-\( \alpha \) test for \( H_0(I)^* \).

1.2.3 Shortcut the Partitioning Test

Notice that the number of partitions increases exponentially as the number of parameters of interest \( k \) increases. Therefore, for medium to large \( k \), to test each partitioning hypothesis one by one is computationally infeasible.
One set of sufficient conditions for shortcutting partitioning test has been proposed (Calian et al., 2008; Huang and Hsu, 2007; Xu and Hsu, 2007) and widely used. The sufficient conditions are:

1. The test statistics $T_i$ for $H^*_0I$, $i \in I$ are in the form of rejecting $H^*_0I$ if $\max_{i \in I} T_i \geq c_I$.

2. $\sup_{H^*_0I} P\{\max_{i \in I} T_i \geq c_I\} \leq \alpha$.

3. The values of the test statistics $T_i$, $i = 1, \ldots, k$, are not re-computed for different $H^*_0I$.

4. Critical values $c_I$ have the property that if $J \subset I$ then $c_J \leq c_I$ and if $J = I$ then $c_J = c_I$.

Upon satisfaction of the conditions 1–4, it is possible to construct a step-down multiple testing procedure that tests much fewer null hypotheses. For example, to test $k$ hypotheses, $H_0i : \theta_i \leq 0$, $i = 1, \ldots, k$. Let $T_i$, $i = 1, \ldots, k$ denote test statistics for each single $H_0i$. Let $[1], \ldots, [k]$ be the random indices such that $T_{[1]} \leq \cdots \leq T_{[k]}$. Let $\theta_{[j]}$ be the parameter of test interest corresponding to $T_{[j]}$. Once sufficient conditions 1–4 are satisfied, the step-down version of the partition testing proceeds in following steps:

Step 1: Test $H_{\theta_{[1],\ldots,[k]}} : \theta_{[1]} \leq 0$ and $\theta_{[2]} \leq 0$ and ... and $\theta_{[k]} \leq 0$.

If $T_{[k]} \geq c_{\alpha_{\{[1],[2],\ldots,[k]\}}}$ then infer $\theta_{[k]} > 0$ and go to step 2; else stop.

Step 2: Test $H_{\theta_{[1],\ldots,[k-1]}} : \theta_{[1]} \leq 0$ and $\theta_{[2]} \leq 0$ and ... and $\theta_{[k-1]} \leq 0$.

If $T_{[k-1]} \geq c_{\alpha_{\{[1],[2],\ldots,[k-1]\}}}$, then infer $\theta_{[k-1]} > 0$ and go to step 3; else stop.
Step $k$ : Test $H_{0\{1\}} : \theta_1 \leq 0$.

If $T_{[1]} \geq c_{\alpha,\{1\}}$, then infer $\theta_1 > 0$; else stop.

Therefore, instead of testing up to $2^k - 1$ hypotheses, at most $k$ tests are needed to make valid inference for $H_{0i} : \theta_i \leq 0, i = 1, \ldots, k$ by the step-down short-cutting procedure.

For a specific example, suppose the t-test statistics $T_i = \frac{\hat{\theta}_i - \theta_i}{s_p}, i = 1, 2$ are used for testing $H_{01} : \theta_1 \leq 0, H_{02} : \theta_2 \leq 0$, where $s_p$ is a pooled estimate of the variance.

Suppose $T_1 < T_2$ so that $T_{[1]} = T_1, T_{[2]} = T_2$. The step-down partitioning test is executed as:

Step 1: Test $H_{0\{12\}} : \theta_1 \leq 0$ and $\theta_2 \leq 0$. If $max_{i=1,2} T_i = T_{[2]} = T_2 > t_{\alpha/2,\nu}$, then reject $H_{0\{1,2\}}$ and make inference $\theta_2 > 0$. Else, stop here and make no inference.

(Test $H_{0\{2\}} : \theta_2 \leq 0$. Since $T_{[2]} = T_2 > t_{\alpha/2,\nu}$ and also $t_{\alpha/2,\nu} > t_{\alpha,\nu}$, it is guaranteed that $T_{[2]} = T_2 > t_{\alpha,\nu}$ and $H_{0\{2\}}$ will be rejected. So this step can be skipped.)

Step 2: Test $H_{0\{1\}} : \theta_1 \leq 0$. If $T_{[1]} = T_1 > t_{\alpha,\nu}$ then reject $H_{0\{1\}}$ and make inference $\theta_1 > 0$. Otherwise stop and make no further inference.

Note that only if the test statistics $T_i, i = 1, 2$ do not change with null hypotheses subset $I$, $T_{[2]} > t_{\alpha/2,\nu}$ can guarantee $T_{[2]} > t_{\alpha,\nu}$. Otherwise, the second step may not be skipped.

For example, suppose the pooled standard deviation $s_p$ is used in testing $H_{0\{12\}}$, that is, reject $H_{0\{12\}}$ if $T_{[2]} = max_{i=1,2} \frac{\hat{\theta}_i - \theta_i}{s_p} > t_{\alpha,\nu}$. But suppose the individual
standard deviation for each sample \( i \), i.e. \( s_i \), is used in testing single null hypothesis, i.e. reject \( H_0(i) \) if \( T_i' = \frac{\theta_i - \theta_0}{s_i} > t_{\alpha/2, \nu}, i = 1, 2 \). Then no short-cut can be taken. The reason is that, even \( T_2' > t_{\alpha/2, \nu} \) holds, whether \( T_2'' > t_{\alpha, \nu} \) is not known, since the denominator has changed and so \( T_2'' \neq T_2' \).

Since the test is in the form of rejecting \( H_0I \) if \( \max_{i \in I} T_i \geq c_I \), it is often referred as the ‘\( \text{max}T \)’ test in multiple testing. If \( p_i \) are p-values from testing null hypothesis \( H_0i \) are used and \( H_0I \) is rejected if \( \min_{i \in I} p_i \leq c_I \), this type of test is often refereed as the ‘\( \text{minp} \)’ test. Note that sufficient conditions similar to 1–4 can easily be constructed for the minp tests.

Holm proposed the well-known and widely used step-down testing procedure (Holm, 1979) that is based on the Bonferroni inequality. This method is conservative no matter if the test statistics are positively correlated, or are independent, or even slightly negatively correlated.

The Bonferroni inequality states that:

\[
P\left( \bigcup_{i=1}^{k} E_i^c \right) \leq \sum_{i=1}^{k} P(E_i^c) \quad (1.9)
\]

This provides an upper bound for controlling FWER. If \( k \) null hypotheses are tested simultaneously, by performing each test at level \( \alpha/k \), one could control the overall level of the test at \( \alpha \). Therefore, FWER is controlled at \( \alpha \). Holm’s step-down method is often stated in the p-value version:

Let \( [1], \ldots, [k] \) be the random indices such that \( p_{[1]} \leq \cdots \leq p_{[k]} \), where \( p_{[i]} \) is the p-value resulting from the test of \( H_{0[i]} : \theta_{[i]} \leq 0, i = 1, \ldots, k \).

Step 1: Test \( H_{0[[1],\ldots,[k]]} : \theta_{[1]} \leq 0 \) and \( \theta_{[2]} \leq 0 \) and \( \ldots \) and \( \theta_{[k]} \leq 0 \).

If \( p_{[1]} < \alpha/k \), then reject \( H_{0[1]} \) and infer \( \theta_{[1]} > 0 \) and go to step 2; else stop.
Step 2: Test $H_0\{[2,\ldots,[k]]\} : \theta_{[2]} \leq 0$ and \ldots and $\theta_{[k]} \leq 0$.

If $p_{[2]} < \alpha/(k-1)$, then infer $\theta_{[2]} > 0$ and go to step 3; else stop.

\ldots

Step $k$: Test $H_0\{[k]\} : \theta_{[k]} \leq 0$.

If $p_{[k]} < \alpha$, then infer $\theta_{[k]} > 0$ and stop; else stop.

Notice that Holm’s method is a special version of the partition testing. A level-$\alpha$ test for the null hypothesis $H_{0I}$ tested in each step of the Holm’s method is also a level-$\alpha$ test for the partitioning hypothesis $H_{0I}^*$.

Holm’s method is most useful when the joint distribution of the test statistics is hard to estimate so that to find good rejection regions is not possible. But in the situation where test statistics are positively correlated. One practical example of where we might expect test statistics to be positively correlated is in clinical trials, with the common control group in comparison.

Hochberg’s step up method can be viewed as another special short-cut version of the partitioning test. It is based on the so called Simes-Hochberg test for each partitioning subset $H_{0I}^*$, i.e.

Reject $H_{0I}^*$ if $p_i^I < \frac{\alpha}{|I|-i+1}$ for any $i \in I$ \hfill (1.10)

Using p-values, Hochberg’s step-up test proceeds as:

Let $[1],\ldots,[k]$ be the random indices such that $p_{[1]} \leq \cdots \leq p_{[k]}$, where each $p_{[i]}$ corresponds to testing $H_{0[i]} : \theta_{[i]} \leq 0$, $i = 1,\ldots,k$.

Step 1: If $p_{[k]} \leq \alpha$, then infer $\theta_{[1]},\ldots,\theta_{[k]} > 0$, $i = 1,\ldots,k$ and stop; else go to step 2.
Step 2: If $p_{[k-1]} \leq \alpha/2$, then infer $\theta_{[1]}, \ldots, \theta_{[k-1]} > 0$ and stop; else go to step 3.

\[ \cdots \]

Step $k$: If $p_{[k]} \leq \alpha/k$, then infer $\theta_{[k]} > 0$ and stop; else stop.

Since the testing procedure starts with the least significant looking parameter of interest and goes toward the most significant one, it is called the ‘step-up’ method. This is in contrast with Holm’s method, which is referred to as a ‘step-down’ method since it starts with the most significant looking parameter and moves to the least significant one in steps.

Hochberg’s method has been shown to have a larger rejection region than Holm’s step-down method (Huang and Hsu, 2007). Thus it is slightly more powerful than Holm’s method. However, Hochberg’s method is based on more distributional assumptions for the test statistics such as independence or more generally, the MTP2 property, (Huang and Hsu, 2007). Therefore, it is not correct to claim a step-up type of test is always more powerful than a step-down type of test.

Ideally, if the joint distribution of test statistics can be derived analytically, taking distributional information into account can help construct more powerful step-down shortcut partitioning tests. Varieties of modeling techniques are very important tools to achieve this goal.
1.3 GLMs (Generalized Linear Models)- Logistic Regression Models

1.3.1 GLM model

For a single binary observation $Y_i \sim Bernoulli(\pi)$, $0 \leq \pi \leq 1$, the simple linear regression technique, i.e. $Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$, cannot be simply applied due to some unique features related to discrete data.

First, since $Y_i$ is binary, $\epsilon_i = Y_i - x_i\beta$ only has two values: $1 - x_i\beta$ or $-x_i\beta$. Therefore $\epsilon_i$ is not normally distributed as assumed in simple linear regression models.

Secondly, $Var(Y_i) = E(Y_i)(1 - E(Y_i)) = \pi(x_i)(1 - \pi(x_i))$. That is, the variance of binary data changes with the expectation of the outcome $Y_i$. So $Var(Y_i)$ may change with $x_i$. The constant variance $Var(Y_i) = \sigma^2$ specified by the simple linear regression does not hold.

Thirdly, the expectation of response has a constraint, i.e. $0 \leq E(Y_i) \leq 1$, where $E(Y_i) = x_i\beta$. That is, the constraint applies to all possible values of $x_i\beta$, which is not the case for ordinary linear regressions either.

All of these features make binary, or more generally, discrete data analysis different from ordinary normal outcome variable analysis.

Let

$$Y_{n \times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}. \quad (1.11)$$

Let $X$ be the design matrix,

$$X_{n \times k} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \text{ where } X'_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ik} \end{pmatrix}. \quad (1.12)$$
Let $\beta$ be the vector of parameters,

$$\beta_{k \times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

(1.13)

If the response variable $Y_i, i = 1, \ldots, n$ follow an exponential family distribution, i.e.

$$f(Y_i) = \exp\left(\frac{Y_i \theta_i - c(\theta_i)}{\phi} + g(Y_i, \phi)\right)$$

(1.14)

Generalized Linear Models (GLM) may be used in modeling.

The link function $h(\cdot)$ connects the mean of the response to the ‘linear predictor’ $X_i \beta$, i.e. $h(E[Y_i]) = h(\mu_i) = X_i \beta$, or

$$\mu_i = E(Y_i) = h^{-1}(X_i \beta)$$

(1.15)

Common link functions include:

1. The logit link $h(\mu_i) = \log(\mu_i/(1 - \mu_i))$ and probit link $h(\mu_i) = \Phi^{-1}(\mu_i)$ for Binomial models.

2. The log link: $h(\mu_i) = \log(\mu_i)$ for Poisson models.

3. The identity link $h(\mu_i) = \mu_i$ for normal models.

The maximum likelihood (ML) estimates can be easily obtained assuming independence between observations $Y_i, i = 1, \ldots, n$. 

14
1.3.2 Logistic Regression Model

Logistic regression is the most widely used technique in binary data regression analysis. Effects from both continuous and discrete independent variables on the probability of a binary event can be modeled simultaneously.

Let \( Y' = (Y_1, Y_2, \ldots, Y_n) \) where \( Y_i = 1 \) or \( 0 \). The logistic regression model is:

\[
\log \frac{\pi(x_i)}{1 - \pi(x_i)} = x_i \beta
\]  

(1.16)

Equivalently,

\[
Pr(Y_i = 1) = \pi(x_i) = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)}
\]

(1.17)

\( \log \frac{\pi(x_i)}{1 - \pi(x_i)} \) is the ‘logit’ transformation. For dichotomous data analysis, this transformation has advantages over the linear regression formulation \( Y_i = x_i \beta \). As \( x_i \) ranges from \(-\infty\) to \(+\infty\), \( \log \frac{\pi(x_i)}{1 - \pi(x_i)} \) also ranges from \(-\infty\) to \(+\infty\). Therefore, the constraint for \( x_i \beta \) to be within \([0, 1]\) is accomplished through transformation. In addition, the linear combination of \( x_i \beta \) is linked to the logit transformation of \( E(Y_i) \), which makes explanation of coefficient much easier in the logit scale.

1.3.3 Estimation and Inference of Logistic Regression

For one binary observation \( Y_i \), the likelihood function is

\[
f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}
\]

(1.18)

Due to independence between individuals, the joint likelihood function for \( Y \) is

\[
L(\beta) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}, i = 1, \ldots, n
\]

(1.19)
Taking derivative of the logarithm of the joint likelihood function and solving the score function lead to the maximum likelihood estimates (MLE) of parameters $\beta$. Notice that there is no closed-form solution for the MLE of logistic regression parameters due to the nonlinear logit link function. Numeric optimizations such as the Newton-Raphson algorithm are required to obtain the MLE $\hat{\beta}$. Based on $\hat{\beta}$, the asymptotic variance-covariance matrix of $\hat{\beta}$ which is the inverse of the Fisher’s information matrix $I$, i.e. $\Sigma = I^{-1}$, is further estimated.

In practice, the asymptotic variance-covariance matrix of $\hat{\beta}$ is often approximated by the Hessian matrix $-H_{\hat{\beta}}^{-1}$.

With a large sample size, $I \simeq -H_{\hat{\beta}}$, which contains functions of the second order partial derivatives of the logarithm of the likelihood function evaluated at MLE $\hat{\beta}$, we obtain the following approximate result.

Let

$$H = \begin{pmatrix} s_{11}^2 & \cdots & s_{1p}^2 \\ \vdots & \ddots & \vdots \\ s_{p1}^2 & \cdots & s_{pp}^2 \end{pmatrix} \quad (1.20)$$

be the Hessian matrix for the coefficient estimates with the $ij$th element

$$s_{ij}^2 = \frac{\partial^2 \ln L(\beta)}{\partial \beta_i \partial \beta_j}, i, j = 1, \ldots, p. \quad (1.21)$$

The large-sample variance for the MLE coefficient estimates can be approximated by the Hessian matrix by evaluating

$$V(\hat{\beta}_i) = (-s_{ii}^2)^{-1}. \quad (1.22)$$

at $\beta_i = \hat{\beta}_i$. 

16
With large sample size, under mild conditions, the MLE asymptotic theory guarantees \( \hat{\beta} \) is asymptotically normally distributed:

\[
\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d \text{MVN}(0, I^{-1})
\]  

(1.23)

The likelihood ratio test is the standard test for comparing nested models. Alternatively, the Wald test, based on the asymptotic property of MLEs, can be used.

\[
W = \frac{\sqrt{n}(\hat{\beta}_i - \beta_i)^2}{\hat{V}(\hat{\beta}_i)}
\]  

(1.24)

For large sample size \( n \), \( W \sim \chi^2_1 \).

1.3.4 Coefficient Interpretation

Logistic regression has a convenient and explicit interpretation of the parameters. Since the linear predictor predicts the logit of expectation, the explanation of coefficients is in the ‘logit’ or log odds.

The ‘log odds’ is defined by:

\[
\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = x_i \beta
\]  

(1.25)

Given other independent variables fixed, \( \beta \) is the change in log odds that results from a one unit change in \( x_i \).

\[
\{ \log \frac{P(Y_i = 1)}{P(Y_i = 0)} | X_i = x_i + 1 \} - \{ \ln \frac{P(Y_i = 1)}{P(Y_i = 0)} | X_i = x_i \} = \beta
\]  

(1.26)

Therefore, \( \beta \) is the difference between the log odds of the event given \( X_i = x_i + 1 \) and those given \( X_i = x_i \), which is the ‘log odds ratio’: 17
\[
\log \left( \frac{P(Y_i = 1)/P(Y_i = 0)|X_i = x_i + 1}{P(Y_i = 1)/P(Y_i = 0)|X_i = x_i} \right) = \beta \tag{1.27}
\]

Equivalently,

\[
\text{odds ratio} = \frac{P(Y_i = 1)/P(Y_i = 0)|X_i = x_i + 1}{P(Y_i = 1)/P(Y_i = 0)|X_i = x_i} = \exp(\beta) \tag{1.28}
\]

The odds ratio is a commonly used measurement for association, especially in epidemiology studies where $X$ is binary. It is often used since the ratio of the odds of an event (case), given exposure, to those without exposure is of special interest to epidemiologists and medical researchers.

### 1.3.5 Multiple Testing in Logistic Regression

Simple multiplicity adjustment methods such as Bonferroni’s method and Holm’s method can be utilized in multiple testing for p-values from logistic regressions.

Dasgupta et al. proposed a Multiple Comparison with the Control testing procedure in logistic regression setting based on the likelihood ratio test (Dasgupta et al., 2000). Consider the test $H_{01,\ldots,k} : \beta_1 = \ldots = \beta_k = 0$ versus $H_{a1,\ldots,k} : \text{at least one inequality}$, where $\beta_i, i = 1, \ldots, k$ are parameters from the logistic regression model.

The test compares the test statistic $L_{01,\ldots,k} = -2\ln(\Lambda_{01,\ldots,k})$ to a chi-square distribution with $k$ degrees of freedom, where $\Lambda_{01,\ldots,k}$ is the likelihood ratio statistic of the full logistic regression model versus the reduced model with $k$ fewer parameters.

If $L_{01,\ldots,k} > \chi^2_{\alpha,k}$, $H_{01,\ldots,k}$ is rejected. The test will move to the next series of null hypotheses testing where exactly $k - 1$ parameters are involved. The smaller hypothesis involves fewer parameters including $\beta_i$ will be tested only if the ‘bigger’ intersection null hypothesis where $\beta_i, i \in \{1, \ldots, k\}$ is involved is rejected. Otherwise,
any null hypothesis involving $\beta_i$ will not be tested further. That is, only if all null hypotheses $H_{0i}, i \in I$ are rejected, $H_{0i}$ can be rejected. This is actually a closed testing procedure (Marcus et al., 1976).

Therefore, at most $2^k - 1$ tests are to be performed to draw valid inference for $k$ parameters. Obviously, as $k$ increases, number of hypotheses to be tested will be too large to handle. In Chapter 3, we will use Dasgupta’s example discussing why a short-cut of the multiple testing procedure cannot be taken based on test statistics in the quadratic forms, including the likelihood ratio test statistic.

1.3.6 Computer Software

I am not aware of any statistical package that provides multiple comparison procedures specifically for logistic regression. Simple multiple testing procedures such as the Bonferroni’s adjustment method or Holm’s method can be applied to p-values from logistic regression models without considering the joint distributional information.

1.4 Notation for Grouped Measurements

It is very common for analysts to encounter data that are measured in groups, i.e. clusters. For example, students in the same classroom, or probes on the same gene chip, or agriculture products produced in the same field can be all deemed clustered data. Within each cluster, observations are similar in some sense. Longitudinal studies, in which measurements are made repeatedly on the same subjects over a period of time, is an important source of clustered data. Due to subject-specific effects, the repeated measurements from the same subject are associated with each other.
Data collected from biomedical studies or clinical trials are often measured repeatedly on the same patient or sample. This allows one to study the trend of change in the response variable within and between individuals across time or treatments. For example, to locate TSGs (tumor suppressor genes) that are essential in tumor development and prognosis, multiple marker loci from tumor samples of a group of patients and normal tissue from another group of subjects will be measured for the binary event LOH (loss of heterozygosity)/ROH (retention of heterozygosity). Dozens of markers on the same chromosome will be compared for LOH probabilities. As the gene chip technology develops, the number of genetic markers that can be measured grows. Thousands of markers can now be measured and compared for LOH rates simultaneously now.

With multiple measurements made on the same individual, different notation is required. Let the vector of response variable from $n$ individuals be:

$$Y = \begin{pmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_n' \end{pmatrix}$$ \hfill (1.29)

with expectations

$$\mu = \begin{pmatrix} \mu_1' \\ \mu_2' \\ \vdots \\ \mu_n' \end{pmatrix}$$ \hfill (1.30)

There are $m_s$ observations from the $s$th individual, $s = 1, \ldots, n$:

$$Y_s = \begin{pmatrix} Y_{s1} \\ Y_{s2} \\ \vdots \\ Y_{sm_s} \end{pmatrix}$$ \hfill (1.31)
The $m_s \times m_s$ variance-covariance matrix for $Y_s$ is:

$$V_s = \begin{pmatrix}
v_{11} & v_{12} & \cdots & v_{1m_s} \\
v_{21} & v_{22} & \cdots & v_{2m_s} \\
\vdots & \vdots & \ddots & \vdots \\
v_{m_s1} & v_{m_s2} & \cdots & v_{m_sm_s}
\end{pmatrix}$$

(1.32)

Since observations from different individuals are independent, the variance-covariance matrix of the entire vector of observations (1.4) is block diagonal:

$$V = \text{Var}(Y) = \begin{pmatrix}
V_1 & 0 & \cdots & 0 \\
0 & V_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & V_n
\end{pmatrix}$$

(1.33)

The design matrix for individual $s$ is an $m_s \times k$ matrix with $k$ independent variables:

$$X_s = \begin{pmatrix}
x_{s1}' \\
x_{s2}' \\
\vdots \\
x_{sm_s}'
\end{pmatrix} = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{m_s1} & x_{m_s2} & \cdots & x_{m_sk}
\end{pmatrix}$$

(1.34)

The design matrix for vector (1.4) is:

$$X = \begin{pmatrix}
X_1' \\
X_2' \\
\vdots \\
X_n'
\end{pmatrix}$$

(1.35)

1.5 Mixed Effects Models

To model the random effects introduced by heterogeneity between subjects, ‘linear mixed effects models’ (LMM) for normally distributed data and ‘generalized linear mixed effects models’ (GLMM) for other exponential family members including discrete data are widely used.
They are named ‘mixed effects’ models since two sources of effects, i.e. the ‘fixed’
effects and the ‘random’ effects, are modeled in addition to the measurement errors.
The former models the effects from factors that can be completely controlled in a
designed experiment, such as dosage, treatment groups, etc. On the other hand,
since subjects in a study cannot cover the entire population, they must be some
representatives drawn randomly from it. The heterogeneity across individuals are
modeled by the ‘random’ effects, which are assumed to follow certain probability
distributions.

1.5.1 Linear Mixed Effects Models (LMMs)
Model Introduction

Linear Mixed-Effects Models (LMM) assume the expectation of responses to be
a linear function of both fixed effects and random effects varying from subject to
subject (Laird and Ware, 1982). One important assumption is that the common
term(s)—subject-specific random effects shared among observations from the same
subject explain the source of within-subject correlation.

Let \( Z_s \) be a \( m_s \times q \) matrix for random effects of subject \( s \):

\[
\begin{pmatrix}
Z_{s1} \\
Z_{s2} \\
\vdots \\
Z_{sm_s}
\end{pmatrix}
\]

(1.36)

Typically, for the individual \( s, s = 1, \ldots, n, \)

\[
Y_s = X_s \beta + Z_s b_s + \epsilon_s
\]

(1.37)

with ‘fixed effects’ \( \beta \) and ‘random effects’ or ‘subject-specific effects’ \( b_s \) i.i.d. \( \sim \)
\( N(0, \Sigma_b) \). \( \epsilon_s \sim N(0, V_s) \), where \( V_s \) is defined in (1.4). \( b_s, \epsilon_s \) are independent. \( \Sigma_b \)
is the ‘between-subject variance’ and typically assumed unstructured. \( V_s \) is the co-
vvariance matrix that reflects analyst’s assumption on the within-subject correlation. Commonly used structures include ‘independent’, ‘autoregressive’, ‘compound sym-
metry and ‘unstructured’ structures.

Therefore \( Y_s \) is conditionally normally distributed

\[
Y_s | b_s \sim N(X_s \beta + Z_s b_s, V_s)
\]

\[
Var(Y_s) = Z_s \Sigma b Z_s' + V_s
\]

Individual level changes in the response variables are explained by fixed and random
effects together.

**Estimation and Inference**

Parameter estimation in LMM can be achieved by Maximum Likelihood method
based on marginal distribution of \( Y_s \): i.e.

\[
f_s(Y_s) = \int f_s(Y_s | b_s) f_s(b_s) db_s
\]

Numerical optimization is often required. REML estimator of the variance compo-
nents is often favoured over MLEs. Null hypotheses of the parameters can be tested
using the Wald test based on the multivariate normal distribution \( \hat{\beta} \sim N(\beta, \Sigma \beta) \). If
two models to be tested are nested, the likelihood ratio (LR) test is also applicable.

**Multiple Testing in LMMs**

The multiple comparison procedures for null hypotheses on the fixed effects have
been developed in the LMM setting. For example, the Bonferroni’s adjustment
method, or the adjustment based on Sidak’s inequality which assumes independence
are available for p-values of the fixed effects. In the case of multiple comparison
with the control, the factor analytic method by Hsu (Hsu, 1992) can be employed for correlated fixed effects.

**Computer Software**

The PROC MIXED procedure in SAS is a popular program for fitting LMMs with different correlation structures. Under the LSMEANS statement, the different options for the ADJUST statement can implement multiple testing methods on the fixed effects mentioned above. The MCB macro in SAS sample library based on Hsu’s MCB method is readily applicable in the PROC MIXED settings as well.

### 1.5.2 Generalized Linear Mixed Effects Models (GLMMs)

#### Model Introduction

To handle correlated discrete data, Generalized Linear Mixed Models (GLMM) can be used. The word ‘generalized’ indicates generalization from linear case to non-linear case. Compared to GLMs, GLMMs add the random effects \( b_s \) to deal with the correlation within individuals. The general specifications of GLMMs are:

1) \( Y_{st}|b_s \sim \text{exponential family with mean } \mu_{st}. \)

2) The linear predictor \( \eta_s = X_s \beta + Z_s b_s \), with \( b_s \) i.i.d. \( \sim N(0, \Sigma_b) \).

3) \( h(E[Y_{st}|b_s]) = h(\mu_{st}) = \eta_{st} \), where \( h \) is the link function.

Therefore,

\[
h(E[Y_s|b_s]) = X_s \beta + Z_s b_s \tag{1.40}
\]

Expectations in GLMM are linked with the linear predictor through the link function just as in GLMs. Logit, probit, and log link functions are all commonly
used. If the identity link function is used for continuous data, GLMM is equivalent to LMM.

**Estimation and Inference**

For most of the non-Gaussian data, there are no corresponding definitions of the multivariate distributions, thus no closed-form integrations are available. Numerical approximations are required. Inference of parameters is readily available according to asymptotic normality property of the parameter estimates.

**Multiple Testing in GLMM Setting**

The methods of multiplicity adjustment in the GLMM setting are very similar to the ones introduced in the LMM section.

**Computer Software**

The PROC GLIMMIX macro from SAS with DIST= option in the MODEL statement is useful in fitting GLMMs. It can apply various correlation structures for the random effects using RANDOM statement and for the repeated measurements with the RSIDE option under the RANDOM statement. Similar to PROC GLM from SAS, under the LSMEANS statement, ADJUST= with different options implement different multiple comparisons methods such as the Bonferroni inequality, Dunnett's intervals by Hsu's factor analytic method and Tukey's method.

The PROC NLMIXED from SAS is another choice of fitting nonlinear mixed effects models. The difference between it and the GLMMIX procedure is that PROC NLMIXED implements a different estimation technique – it directly maximizes an approximate integrated likelihood.
1.6 Marginal Models

1.6.1 Model Introduction

Liang and Zeger (Zeger and Liang, 1986; Liang and Zeger, 1986) proposed the marginal models using Generalized Estimating Equations (GEE) estimation method for both continuous and discrete longitudinal/clustered data analysis. Although this method only requires first-order moments estimation, robust and consistent parameter estimates are guaranteed.

In longitudinal data literature, modeling the change of response variable due to group effects is often referred as modeling the response variable in the ‘population’ level or modeling the ‘marginal’ mean. The population consists of subjects who share the same group information $X$.

For example, to study the impact of pregnancy smoking on low birth weight rate, the focus is the relative risk of giving birth to low weight babies between the smoking and non-smoking women population. In other words, the major interest is not to predict subject-specific outcome, but to explore the impact of ‘smoking’ on a group of women as a rather homogeneous population.

Let $Y_{st}$ denote the $t$th observation from the $s$th subject, $s = 1, \ldots, n, t = 1, \ldots, m_s$. Link functions are utilized in marginal models to link the ‘marginal mean’ $E[Y_s]$ to the ‘linear predictor’ $\eta_s = X_s \beta$, i.e.

$$h(E[Y_s]) = X_s \beta \quad \text{(1.41)}$$

As in GLMs, various link functions may be used for various models, such as logit link or probit link function. Marginal models make assumptions that for individual $s$ from certain population, the expected outcome $E[Y_s]$ is predicted by the group
effects $\beta$. Therefore, heterogeneity $b_s$ between individuals are not explicitly involved in the expression of $E[Y_s]$.

### 1.6.2 Estimation: GEE

The GEE (Generalized Estimating Equations) method was proposed for estimating marginal model parameters (Zeger and Liang, 1986; Liang and Zeger, 1986). It is an extension of the quasi-likelihood (QL) theory (McCullagh, 1983) to repeated measurements, so called since they are not derived from the true likelihood.

Similar to the QL method, to estimate parameters in marginal models by GEE, one only needs to specify the mean and variance of $Y_s$. The key assumption of the GEE method is that this mean structure is correctly specified, i.e.

$$h(E[Y_s]) = h(\mu_s) = X_s\beta$$  \hfill (1.42)

Instead of obtaining the joint density function and taking derivatives, a series of equations called ‘quasi score equations’ are solved. That is, the GEE estimates $\hat{\beta}$ for parameters $\beta$ are the solution of the ‘Generalized Estimating Equations’:

$$U(\beta) = \sum_{s=1}^{n} D_s'V_s^{-1}(Y_s - \mu_s) = 0$$  \hfill (1.43)

where

$$D_s = \frac{\partial \mu_s}{\partial \beta} = \frac{\partial h^{-1}(X_s\beta)}{\partial \beta}$$  \hfill (1.44)

$$V_s = A_s^{1/2}R_s(\alpha)A_s^{1/2}$$  \hfill (1.45)

27
V_s is often called the ‘working’ correlation matrix, where \( A_s^{1/2} \) is a diagonal matrix with standard deviations of \( Y_s \) on the diagonal. \( R_s(\alpha) \) is the assumed correlation structure of \( Y_s \) defined by parameters \( \alpha \).

\[
A_s = \begin{pmatrix}
V(Y_{s1}) & 0 & \cdots & 0 \\
0 & V(Y_{s2}) & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & V(Y_{sm})
\end{pmatrix}
\]  \hspace{1cm} (1.46)

\[
R_s(\alpha) = \begin{pmatrix}
1 & \alpha \\
1 & \ddots \\
\alpha & 1
\end{pmatrix}
\]  \hspace{1cm} (1.47)

By specification of the working correlation, the dependence between repeated measurements is taken into account. Many choices for the correlation structure are available for \( R_s(\alpha) \), such as the ‘compound symmetry’ structure, the ‘autoregressive’ structure and the ‘unstructured’ structure.

By the method of moments, Liang and Zeger showed that when (1.41) is correctly specified, under very mild conditions the GEE estimates of \( \beta \) are consistent. Their variance-covariance matrix \( V_G \) are also consistently estimated from the so called ‘sandwich’ estimator \( V \) i.e.

\[
V = \lim_{n \to \infty} nI_0^{-1} I_1 I_0^{-1}
\]  \hspace{1cm} (1.48)

\[
I_0 = \sum_{s=1}^{n} D_s' V^{-1}_s D_s
\]

\[
= \sum_{s=1}^{n} \left( \frac{\partial h^{-1}(X_s\beta)}{\partial \beta} \right)' \left[ A_s^{1/2} R(\alpha) A_s^{1/2} \right]^{-1} \left( \frac{\partial h^{-1}(X_s\beta)}{\partial \beta} \right)
\]  \hspace{1cm} (1.49)
\[ I_1 = \sum_{s=1}^{n} D_s' V_s^{-1} \text{Cov}(Y_s) V_s^{-1} D_s \]
\[ = \sum_{s=1}^{n} \left( \frac{\partial h^{-1}(X_s\beta)}{\partial \beta} \right)' A_s^{1/2} R(\alpha) A_s^{1/2}^{-1} \text{Cov}(Y_s) [A_s^{1/2} R(\alpha) A_s^{1/2}]^{-1} \left( \frac{\partial h^{-1}(X_s\beta)}{\partial \beta} \right) \]  

(1.50)

With a reasonably good approximation of the correlation structure, the GEE estimates are nearly as efficient as MLEs. But misspecification of the correlation structure might cost efficiency of estimation. Therefore, careful exploration of the data structure or obtaining preliminary information before modeling is still important.

Finding estimates by GEE is an iterative procedure. An initial estimate of \( \beta \) can be obtained by assuming independence as in an ordinary GLM. Based on residuals \( r_{st} = Y_{st} - E[Y_{st}] \), \( \hat{R}_s(\hat{\alpha}) \) can be computed and \( \hat{\beta} \) is updated. Since \( \hat{\beta} \) and \( \hat{\alpha} \) cannot be estimated separately, updates between them will take place iteratively. This is called ‘iterative re-weighted least squares’ estimation (IRLS). Once \( \hat{\beta} \) and other variance \( V_s \) related parameters are estimated, \( V_G \) could be further estimated by the ‘sandwich estimator’.

The inference for GEE models are based on the asymptotic normal distribution of the estimates. Wald-type tests and score-type tests are used.

1.6.3 Computer Software

In SAS, the procedure GENMOD can fit GEE models with different link functions and various working correlation structures and output parameter estimates along with the ‘sandwich’ variance estimates. But no multiple testing algorithms has been included in PROC GENMOD.
1.6.4 Multiple Testing in GEE Setting

In the one-way model setting, where observations are independent, many multiplicity adjustment methods are available for multiple testing problem. However, in the more complicated general linear model setting, different approximations are needed to obtain critical values for constructing simultaneous confidence intervals (Hsu, 1999). How to further generalize multiple comparisons methods to GEE model setting, where the response variables are discrete and correlated are still under development.

Orelien et al. generalized Dunnett’s MCC method in the GEE model setting for correlated binary data (Orelien et al., 2002). Hsu’s factor analytic method was used to reduce the integration dimension. They showed that the MCC null hypotheses $H_0: \mu_i = \mu_0, i = 1, \ldots, k$ can be tested by linear combinations of parameter estimates $\hat{\beta}$ from GEE models, i.e. $L\hat{\beta} = 0$. By utilizing the asymptotic multivariate normal joint distribution of $\hat{\beta}$, the single-step Dunnett’s method is generalized to GEE setting. Orelien et al. simulated on generated correlated binary data that followed a beta binomial distribution and showed validity of generalization of Dunnett’s method in the GEE model setting with false rejection rate close to the nominal $\alpha$ level.

In this thesis work, we discuss multiple testing in the discrete data setting. In Chapter 2, we study the validity of using permutation test for tests on discrete data with discrete multivariate explanatory variables, for example, when we are testing associations between some binary genetic markers versus phenotype groups. If the response variable is binary and the sample size is relatively large compared to the number of explanatory variables, then chapter 3 discusses multiple testing procedures based on logistic regression models. If the binary response variable is further
correlated, then chapter 4 discusses multiple testing procedures based on GEE models.
CHAPTER 2

PERMUTATION TESTS BASED MULTIPLE TESTING IN DISCRETE DATA

To construct a multiple testing procedure that properly controls FWER, computing critical values according to the reference distribution under the null hypothesis is the key to finding acceptance and rejection regions.

Simple multiplicity adjustment methods, such as Bonferroni’s method and Holm’s method, which is a step-down partitioning test based on the Bonferroni’s inequality, do not make use of distributional information of the test statistics. Therefore they tend to be over conservative when positive associations exist among test statistics.

If the joint distribution of the test statistics can be derived analytically, taking the distributional information into account will lead to more powerful test procedures. Three techniques for constructing a reference distribution in order to compute the critical values are:

1. Model the data
2. Re-sample data within treatment groups
3. Re-sample data pooled across treatment groups
To model the data using well established modeling techniques with reasonable
assumptions is one way familiar to statisticians. When modeling is possible, the
joint distribution of the test statistics can sometimes be derived analytically, allowing
the critical values $c^I$ to be computed by numerical analytic methods. When the
test statistics have a multivariate normal or a multivariate $t$ distribution, the factor
analytic technique by Hsu (Hsu, 1992) is applicable if the correlation structure has
exactly or approximately a 1-factor structure. Otherwise, the variance-reduced Monte
Carlo technique of Genz and Bretz is applicable (Genz and Bretz, 1999).

Quite often, modeling is not feasible. For example, in genomewide association
studies (GWAS), SNP chips can genotype 500,000 SNPs at once. Modeling all SNPs
together and obtaining the correlation structure could be difficult using current tech-
niques.

In the continuous data setting, if the joint distribution of the test statistics cannot
be derived analytically, then a popular technique is to build a reference distribution
by re-sampling the residuals, re-computing the test statistics each time the data is
re-sampled (Westfall and Young, 1993).

Resampling the original data has been considered one way of maintaining the
association structure of the data, so that the correlations between test statistics are
kept. Resampling could be performed either between or within groups of comparison
interests. Pollard and van der Lann (Pollard and van der Lann, 2004) proposed to
resample data within each treatment group and then properly center the re-scaled
test statistics. This method can only asymptotically control the type I error rate in
resampling based multiple testing (not FWER).
Permutation tests are another popular choice based on resampling observations across treatment groups without replacement. We study the limitations of permutation tests based multiple testing procedures in this chapter.

2.1 Motivating example: Genome-Wide Association Study (GWAS)

2.1.1 Single Nucleotide Polymorphisms (SNPs)

The human genome is highly polymorphic, which means, there are a large number of alternative DNA sequences. If a single base at a position (a ‘locus’) in a single strand of DNA shows variations within a species, this is called a ‘Single Nucleotide Polymorphism’ (SNP). There are approximately 10 to 30 million SNPs in the human genome.

Human chromosomes have two copies. Each copy is a double-stranded DNA helix, consisting of bonded base pairs of nucleotides A (adenine), C (cytosine), T (thymine) and G (guanine). Base A only bonds with base T. Base C only bonds with base G.

SNPs are typically in the biallelic form which consists of two out of the four bases. For example, if the alleles are composed of A and G, then three SNP genotypes may exist: AA, AG and GG. In general, we could simplify the notation and use ‘A’ representing one allele (e.g. A) and ‘B’ for another (e.g. G). Then the three biallelic SNP genotypes are distinguished by AA, AB, and BB.

It is well known that the most essential role of DNA is to serve as a ‘coding’ system for proteins. The genetic code is composed by three-letter ‘codons’, which are combinations of three nucleotides. That is, three letter nucleotides have to satisfy certain combinations and sequences to code for meaningful amino acids, which compose proteins.
A variety of problems could result if the DNA sequence is changed due to SNP. For example, the DNA transcription may stop too early. A different amino acid could be translated. The polypeptide synthesis could stop at the wrong location. The properties of a protein, such as binding feature, binding strength, or structure, all could be severely affected. It is not difficult for one to imagine that many diseases are related to SNPs. Thus the main interest of SNP study is to find the SNP(s) associated with a phenotype such as a disease so that eventually a treatment could be found.

2.1.2 GWAS and pharmacogenomics

As modern technologies develop, screening for phenotype-predictive SNPs in the scale of the whole genome is not hard technically. This type of study is called the Genome-Wide Association Study (GWAS). A typical GWAS involves a large number of genetic features (e.g. tens of thousands of SNPs). The comparison of interest is usually differences in probabilities of mutation between two groups: the control group and the case group.

Numerous GWAS are carried out every year. Some of them have led to findings of candidate genes that can predict disease prognosis or efficacy of treatment.

For example, a GWAS of symptom response in an open-label study of the methylphenidate transdermal system (MTS) was conducted by Mick et al. on 187 ADHD subjects (Mick et al., 2008). 29 potentially gene candidates out of 319722 SNPs were found to be possibly associated with positive response for the MTS treatment.

Arnett et al. (Arnett et al., 2009) conducted a GWAS on the left ventricular (LV) mass in the Family Blood Pressure Program data. LV mass is a sensitive predictor
of cardiovascular mortality and morbidity in all genders, races, and ages. 101 cases and 101 controls were involved. 11 out of 12 potential SNP candidates detected by GWAS were successfully validated in a follow-up study by fitting GEE models on a new dataset.

A GWAS by Song et al. identified an ovarian cancer susceptibility locus on 9p22.2 with 1817 ovarian cancer cases and 2353 controls (Song et al., 2009). 12 SNPs were initially identified to be associated with higher disease risk. The most significant SNP was further confirmed for its association by fitting logistic regression models in new samples.

Finding biomarkers by GWAS is useful in many areas. For example, in pharmacogenomics, biomarkers are used as classifiers identifying subgroups of patients suitable for different therapeutic interventions or prognoses. For example, the mutation status of genes could be an indicator for better drug efficacy or greater risk of severe adverse effects.

**Epidermal Growth Factor Receptor (EGFR) and cancer treatment**

One important mechanism of carcinogenesis is losing control of the cell cycle. EGFR (Epidermal Growth Factor Receptor) has an important role in controlling cell cycles.

Specifically, EGFR acts as a tyrosine kinase. Once activated, it can transfer one phosphate group from one ATP molecule to a tyrosine residue of its downstream target so that the target is further activated. The \textit{KRAS} gene encodes a small protein which is a downstream target of EGFR. Together, KRAS and EGFR regulate cell proliferation, adhesion and migration. Over-activation of EGFR due to genetic
mutations have been related to the development of many human tumors. Therefore, EGFR inhibitors could be used for treating cancer patients.

For example, Iressa is an anti-cancer drug that works as an EGFR inhibitor. It binds to the ATP binding site of EGFR so that EGFR loses its tyrosine kinase activity. It has been shown that patients with certain EGFR mutation(s) are more sensitive to Iressa compared to those who have normal EGFR (Pao et al., 2004).

Some mutations in the \textit{KRAS} gene, however, produces \textit{KRAS} protein that is always in the ‘active’ status. Thus, even when anti-EGFR treatment shuts down EGFR activity, \textit{KRAS} is still active and this causes poor-response to EGFR-inhititors. Panitumumab (by Amgen) and Cetuximab (by ImClone Systems, Inc.) are both monoclonal antibodies of the human EGFR. After the drugs were approved for the general colon cancer population, studies showed lack of response to these two drugs in the \textit{KRAS} mutant positive patients subgroup (Amado et al., 2008). Subsequently, their labels have been updated (July 2009) to reflect this new information.

\textbf{Human Leukemia Antigens HLAs and drug sensitivities}

The HLA (the Human Leukocyte Antigen), which is the major histocompatibility complex (MHC) in humans, consists of a large number (more than two hundred) of genes located on Chromosome 6 that are related to the immune system function. HLA region is mainly divided into three regions, class I, II, and III. Each class involves many genes that have different functions. For example, the HLA-A, HLA-B, HLA-C are genes belong to the HLA class I and are responsible for presenting antigens from inside the cell.

The HLA genes are the most polymorphic ones in the human genome. Most of the these genes have large number of alleles. For example, there are at least one hundred
of alleles for each of the HLA-B, -DQ and -DR loci. Therefore, mutations in HLA genes have been related to immune system diseases.

The allele HLA-B*5701 belongs to the serotype HLA-B57. Mutations in HLA-B*5701 have been related to some rare but life-threatening severe adverse effects (SAE), such as hypersensitivity to a treatment.

Abacavir (brand name Ziagen), is a potent antiviral treatment for HIV-1. About 8% of the patients treated with Abacavir develop a potentially fatal SAE of hypersensitivity.

Mallal et al. reported on a retrospective study of Abacavir-treated patients, with 200 subjects including 18 cases of hypersensitive reaction to look for biomarkers that can screen out patients prone to this SAE (Mallal et al., 2002).

Using conventional techniques (primarily PCR) and Chi-square tests with Bonferroni correction, they discovered three markers (HLA-B*5701, HLA-DR7, HLA-DQ3) to be highly associated with the occurrence of the hypersensitivity SAE (corrected p-values < 0.0001).

Martin et al. (Martin et al., 2004) reported on a follow-up study in which five markers (HLA-B*5701, C4A6, HLA-DRB1*0701, HLA-DQ3 and Hsp70-HomM493T) were found to be highly associated with the occurrence of the hypersensitivity SAE (corrected p-values < 0.0001).

Subsequently, in accordance with the pharmacegenomic concept of the FDA (2005a, b), a double-blind, prospective, randomized study involving 1956 patients from 19 countries was then conducted to validate the HLA-B*5701 biomarker (Hughes et al., 2008; Mallal et al., 2008).
By immunological analysis, the confirmed numbers of hypersensitivity cases in the HLA-B*5701 positive and negative patient groups are shown in Table 2.1. The resulting sensitivity is 100% with 95% CI = (85.2%, 100%) and the specificity equals 96.9% with 95% CI = (95.5%, 98.0%). The estimated odds ratio is infinity with 95% CI = (165.7, ∞).

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>HLA-B*5701</th>
<th></th>
<th></th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Hypersensitivity reaction)</td>
<td>23</td>
<td>0</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>0 (No Hypersensitivity reaction)</td>
<td>25</td>
<td>794</td>
<td></td>
<td>819</td>
</tr>
<tr>
<td>Column Total</td>
<td>48</td>
<td>794</td>
<td></td>
<td>842</td>
</tr>
</tbody>
</table>

Table 2.1: Table of dichotomized genotypes and phenotype groups in Abacavir example (Mallal et al., 2008).

With sufficient PPV (47.9%) and NPV (100%) demonstrated, a warning was added to the Abacavir label in July 2008:

Patients who carry the HLA-B*5701 allele are at high risk for experiencing a hypersensitivity reaction to abacavir. Prior to initiating therapy with abacavir, screening for the HLA-B*5701 allele is recommended; this approach has been found to decrease the risk of hypersensitivity reaction.

Interestingly, GWAS has also helped identify the same allele HLA-B*5701 as a major predictor of drug-induced liver injury (DILI) by Flucloxacillin (Daly et al., 2009).

In this study, 866,399 SNPs were genotyped for two groups of patients. One group of 51 patients had DILI (cases). The other group involves 282 controls matched for sex and ancestry. HLA-B*5701 was identified as the most promising allele for predicting
DILI with a p-value in the magnitude of $10^{-33}$ by Cochran-Armitage trend test and Fisher’s exact test.

The final results of genotypes in DILI and normal groups are shown in Table 2.1.2:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>HLA-B*5701</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Normal)</td>
<td>4 60</td>
<td>64</td>
</tr>
<tr>
<td>0 (DILI)</td>
<td>43 8</td>
<td>51</td>
</tr>
<tr>
<td>Column Total</td>
<td>47 68</td>
<td>115</td>
</tr>
</tbody>
</table>

Table 2.2: Table of dichotomized genotypes and phenotype groups in Flucloxacillin example (Daly et al., 2009).

The resulting sensitivity is 84.3% with $95\% CI = (71.4\%, 93.0\%)$ and the specificity equals 93.8% with $95\% CI = (84.8\%, 98.3\%)$. The estimated odds ratio is 74.9 with $95\% CI = (20.4, 366.1)$ The estimated PPV is 91.5% and NPV is 88.2%.

This GWAS could serve as our motivating example for exploring the limitation of permutation tests in testing multiple null hypotheses on discrete data. Since the genotype of mutation/non-mutation is dichotomous, the statistical problem is to test equality in probabilities of mutation between the case group and the control group for 866,399 SNPs.

Table 2.1.2 summarizes some pharmacogenomic examples.
Table 2.3: Table of pharmacogenomic examples.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Subgroup</th>
<th>Prevalence</th>
<th>Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erbitux</td>
<td>Non-responders</td>
<td>0.3−0.5</td>
<td>K-RAS</td>
</tr>
<tr>
<td>Vectibix</td>
<td>Non-responders</td>
<td>0.3−0.5</td>
<td>K-RAS</td>
</tr>
<tr>
<td>Iressa</td>
<td>Diarrhea</td>
<td>0.28</td>
<td>ABCG2Q141K</td>
</tr>
<tr>
<td>Flucloxacillin</td>
<td>Hepatotoxicity</td>
<td>&lt; 0.001</td>
<td>HLA-B*57:01</td>
</tr>
<tr>
<td>Abacavir</td>
<td>Skin Hypersensitivity Reaction</td>
<td>0.08</td>
<td>HLA-B*57:01</td>
</tr>
<tr>
<td>Carbamazepine</td>
<td>Steven Johnson</td>
<td>&lt; 0.001</td>
<td>HLA-B*15:02</td>
</tr>
</tbody>
</table>

2.2 Permutation Tests

When modeling is not feasible, permutation tests are one alternative resampling method. Permutation testing has had a long history in applications. For example, Fisher’s exact test for $2 \times 2$ tables can be thought of as a permutation test with complete enumeration (Fisher, 1934).

Suppose a case-control study is carried out. Let $n^D$ and $n^N$ be numbers of subjects in the disease and normal group, respectively. Let $Y^N = (Y_{1}^{N}, Y_{2}^{N}, \ldots, Y_{n^N}^{N})$, $Y^D = (Y_{1}^{D}, Y_{2}^{D}, \ldots, Y_{n^D}^{D})$ be vectors of observations from each group, respectively. Assume each subject has $k$ observations.

The basic idea behind permutation tests is that if there is no difference between the joint distributions of $Y^N$ and $Y^D$, one can randomly relabel the subjects with either group label. That is, the observations are exchangeable across groups.

To test for associations between multiple genetic features and the phenotype, the common procedure of permutation tests is to calculate the observed test statistics for each genetic feature first. Then one randomly resamples observations from the pooled samples across groups. Multiple observations from the same subject will be sampled together as a vector. For each realized permutation sample, test statistics
are re-computed. After a large number of permutations are performed, ideally, with complete enumeration, the observed test statistic is compared to its permutation distribution to make inference for the parameters of interest.

Permutation tests are popular because of the rather simple idea and convenience in application. By ‘shuffling’ each subject’s SNP data as a vector, it is perceived the correlation structure within subject is maintained. Also, whenever complete enumeration is possible, permutation tests can generate exact significance levels. Therefore, in analysis of complex data where the association pattern is hard to estimate, such as microarray data analysis or GWAS, permutation tests are prevalent (Balding, 2006; Ziegler et al., 2008; Cardon and Bell, 2001).

However, not enough attention has been paid to the fact that permutation tests are testing for the null hypothesis of identity in joint distributions (‘exchangeability’) between groups. Suppose the null hypotheses of $k$ genetic features are of interest. Let $Y_I^D \sim F_I^D, Y_I^N \sim F_I^N, I \subseteq \{1, \ldots, k\}$. The null hypothesis of joint distributional identity is:

$$H_{0I}^J : F_I^D = F_I^N.$$  \hspace{1cm} (2.1)

However, the multiple testing problem usually considers marginal identity of distributions, i.e.

$$H_{0i} : F_i^D = F_i^N \text{ and } F_j^D \neq F_j^N, i \in I, j \notin I, I \subseteq \{1, \ldots, k\}$$  \hspace{1cm} (2.2)

If permutation tests are used for testing the marginal null hypotheses, the false rejection rate may not be controlled.
The MDJ (marginals-determine-the-joint) condition (Xu and Hsu, 2007) is a sufficient condition for permutation tests to control FWER strongly. Let $D, N$ denote two groups in comparison. The MDJ condition states: “Let $I_j, j = 1, \ldots, n$, be any collection of disjoint subsets of $1, \ldots, k$, $I_j \subseteq \{1, \ldots, k\}$. If the marginal distributions of the observations are identical between the two groups, $F_{I_j}^D = F_{I_j}^N$ for all $j = 1, \ldots, n$, then the joint distributions are identical as well, $F_{I_U}^D = F_{I_U}^N$ where $I_U = \bigcup_{j=1,\ldots,n} I_j$.”

The MDJ condition refers to a property of the observed data that identity in marginal distributions between groups implies equality in their joint distributions. If the joint distributions are not identical, the permutation test may detect unintended differences other than the marginal differences of interest and make false positive findings (Huang et al., 2006; Calian et al., 2008; Xu and Hsu, 2007; Kerr, 2009). That is, FWER may not be controlled if MDJ is not satisfied.

Numeric examples and simulations within the multivariate continuous data setting have been employed to show that permutation tests could be liberal under various parameter configurations where MDJ does not hold (Huang et al., 2006; Kerr, 2009). Theorem 2.2 by Huang et al. (Huang et al., 2006) gives insight into the limitations of permutation tests using test statistics in the form of differences between group means, i.e. $T = \bar{X} - \bar{Y}$, for any arbitrary data distribution $F_X, F_Y$. Specifically, the theorem shows that for equal sample sizes, the even order cumulants of the permutation distribution of $T$ will be equal to those of the true distribution. Unless $F_X = F_Y$, regardless of whether the sample sizes are equal or not, the odd order cumulants will not be identical. Therefore, in general, the permutation reference distribution differs from the true distribution unless $F_X, F_Y$ have identical cumulants.
Due to the continuous data setting, Theorem 2.2 (Huang et al., 2006) was derived in the form of unconditional distributions. That is, the result does not depend on any observed data set specifically. Since it is derived for arbitrary data distributions $F_X$ and $F_Y$, we believe such differences between the permutation reference distribution and the true null distribution of the test statistics still exist when $X$ and $Y$ are discrete.

In the next section, through a simplified GWAS example, we study the limitation of permutation tests for testing multiple null hypotheses in the discrete data setting if MDJ does not hold.

### 2.3 Permutation test in GWAS: 2 SNPs example

To test whether any of the three possible genotypes of a SNP is significantly associated with a binary phenotype (e.g. disease versus normal), a $2 \times 3$ contingency table can be used. For example, if $N$ subjects are tested for SNP genotypes denoted by AA, AB, BB, the observed table is shown in Table 2.4:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>SNP_i</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Disease)</td>
<td>AA $n_{11}$</td>
<td>$n_1^R$</td>
</tr>
<tr>
<td></td>
<td>AB $n_{12}$</td>
<td>$n_2^R$</td>
</tr>
<tr>
<td></td>
<td>BB $n_{13}$</td>
<td>$n_3^R$</td>
</tr>
<tr>
<td>0 (Normal)</td>
<td>$n_{21}$</td>
<td>$n_1^N$</td>
</tr>
<tr>
<td></td>
<td>$n_{22}$</td>
<td>$n_2^N$</td>
</tr>
<tr>
<td></td>
<td>$n_{23}$</td>
<td>$n_3^N$</td>
</tr>
<tr>
<td>Column Total</td>
<td>$n_{.1}$</td>
<td>$N$</td>
</tr>
<tr>
<td></td>
<td>$n_{.2}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n_{.3}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4: Table of biallelic SNP genotypes and phenotypes.

Let ‘B’ denote a mutation in Table 2.4. Very often, by counting frequencies of presence or absence of a mutation, a $2 \times 2$ contingency table is formed, for the purpose of testing whether carrying a mutation significantly affects the binary phenotype.
In a typical case-control study, sample sizes \( n^D, n^N \) are fixed. Let “+” indicate a mutation, “−” indicate no mutation. Let \( n_i^D, n_i^N, i = 1, \ldots, k \) denote the number of mutations of SNP \( i \) in the disease or normal group and \( n_i^+ \) be total number of SNP \( i \) mutations. Once the data are observed, \( n_i^+ \), the number of mutations for SNP \( i \), is also fixed. The observed table for SNP \( i \) is shown in Table 2.5:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>SNP ( i )</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Disease)</td>
<td>( n_i^D )</td>
<td>( n^D - n_i^D )</td>
</tr>
<tr>
<td>0 (Normal)</td>
<td>( n_i^N )</td>
<td>( n^N - n_i^N )</td>
</tr>
<tr>
<td>Column Total</td>
<td>( n_i^+ )</td>
<td>( N - n_i^+ )</td>
</tr>
</tbody>
</table>

Table 2.5: Table of dichotomized genotypes and phenotype groups.

Marginally, the numbers of mutations of SNP \( i \) in each phenotype group follows a binomial distribution, i.e. \( B(n^D, \pi_i^D), B(n^N, \pi_i^N), i = 1, \ldots, k \). \( \pi_i^N = \pi_i^D \) indicates equality of the two binomial distributions as shown in Table 2.3.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>SNP ( i )</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Disease)</td>
<td>( \pi_i^D )</td>
<td>( 1 - \pi_i^D )</td>
</tr>
<tr>
<td>0 (Normal)</td>
<td>( \pi_i^N )</td>
<td>( 1 - \pi_i^N )</td>
</tr>
</tbody>
</table>

Table 2.6: Marginal underlying probabilities for SNP \( i \).

Jointly, the observations from each group follow a multinomial distribution and the joint probability vectors \( \pi^D \) and \( \pi^N \) have dimensions \( 2^k \), i.e.

\[
Y^D \sim MN(n^D, \pi^D) \tag{2.3}
\]
and

\[ Y^N \sim MN(n^N, \pi^N) \]  \hspace{2cm} (2.4)

The two multinomial random variables are independent in a case-control study setting.

Multiple testing of the associations between \( k \) SNPs and the phenotype is to test equality between the marginal probabilities \( \pi^D_i = \pi^N_i \), \( i = 1, \ldots, k \). Depending on the goal of the study, one or two-sided null hypotheses can be used, i.e.

\[ H_{0i} : \pi^D_i - \pi^N_i \leq 0, \quad \text{vs.} \quad H_{ai} : \pi^D_i - \pi^N_i > 0, \quad i = 1, \ldots, k. \]

or

\[ H_{0i} : \pi^D_i - \pi^N_i = 0, \quad \text{vs.} \quad H_{ai} : \pi^D_i - \pi^N_i \neq 0, \quad i = 1, \ldots, k. \]  \hspace{2cm} (2.5)

Such a problem is challenging. First, since the number of genetic features is usually much larger than the number of subjects, the underlying contingency table is highly sparse. This also requires a very large number of tests to be conducted. Therefore, multiplicity adjustment is essential. In addition, for this many genetic features, one cannot simply rule out the possibility that some of the genes are correlated with others. Thus taking the correlation structure into account will help one construct more powerful testing procedure and find more significant results. However, obtaining the joint distribution of the test statistics through modeling is difficult in the GWAS setting, where \( k \) is very large compared to \( N \).

Permutation tests are prevalent in GWAS, due to its rather simple procedure and the perception that by permuting the observations from each subject as a vector,
the correlation structure is preserved (Balding, 2006; Ziegler et al., 2008; Cardon and Bell, 2001).

For example, Wang and You conducted a GWAS looking for SNPs that are associated with spontaneously occurring lung tumorigenesis (SLT) (Wang and You, 2005). They obtained data on approximately 135,900 SNPs from 10 resistant strains and 3 susceptible strains of mice.

For each SNP, a two-sided Fisher’s exact test was applied to the $2 \times 2$ contingency table of counts of the four possible combinations between each SNP status and SLT susceptibility.

They conducted permutation testing by completely enumerating all 286 possible permutations by rearranging the phenotype labels, susceptible or resistant, among all the mouse strains but keeping their SNP data unchanged.

Using the smallest possible p-value from their data set as the critical value, they reported five regions ($SLT_1$ to $SLT_5$) as having significant associations, quoting a (conditional) FWER of $\alpha = 3/286 \approx 0.01$.

Manenti et al. conducted a subsequent study, aiming to confirm previously reported associations between SNPs and lung tumor susceptibility (Manenti et al., 2009). They obtained data from panels of 12,959 and 138,793 SNPs, from 27 strains of mice which included the 13 strains in the Wang and You study (Wang and You, 2005). Treating spontaneous incidence multiplicity as a continuous random variable, they used t-tests to assess association between each SNP and lung tumor phenotype.

However, none of the five regions ($SLT_1$ to $SLT_5$) reported by Wang and You showed an association in this subsequent study at FWER of $\alpha = 0.10$, with multiplicity adjustment using either the Bonferroni inequality or permutation.
To study the validity of permutation tests used in multiple testing for discrete data, we simplify the motivating GWAS example and let $k = 2$. Marginally, numbers of mutations of SNP1 and SNP2 in each group follow binomial distributions $B(n^D, \pi^D_1), B(n^N, \pi^N_1), B(n^D, \pi^D_2), B(n^N, \pi^N_2)$, respectively. The observed tables and underlying probabilities are shown in Table 2.7 and Table 2.8.

The joint distributions of SNP1, SNP2 mutation status follow multinomial distributions with four cell joint probabilities, i.e.
\[ F^D = MN(n^D, \pi^D_+\pi^D_-, \pi^D_+\pi^D_-) \] and \[ F^N = MN(n^N, \pi^N_+\pi^N_-, \pi^N_+\pi^N_-), \]
where the first subscript of $\pi$ indicates mutation status of SNP1 and the second subscript indicates mutation status of SNP2. Let $\pi^D = (\pi^D_+\pi^D_-, \pi^D_+\pi^D_-)$ and \[ \pi^N = (\pi^N_+\pi^N_-, \pi^N_+\pi^N_-). \]

Identical joint distributions indicate $H^I_0 : F^D = F^N$.

\[
\begin{array}{c|cc|c}
\text{Disease(Phenotype = 1)} & \text{SNP 1} & \text{SNP 2} \\
\hline
+ & n^D_+ & n^D_- & n^D \\
- & n^D_+ & n^D_- & n^D \\
\hline
\text{Normal(Phenotype = 0)} & \text{SNP 1} & \text{SNP 2} \\
\hline
+ & n^N_+ & n^N_- & n^N \\
- & n^N_+ & n^N_- & n^N \\
\hline
\end{array}
\]

Table 2.7: Table of observed samples for $k = 2$.

Let $\theta_i = \pi^D_i - \pi^N_i, i = 1, 2$. Assume the one-sided null hypotheses are of interest,
Disease (Phenotype = 1) | SNP 2 | Normal (Phenotype = 0) | SNP 2
|---|---|---|---
| + | \(\pi^D_1\) | + | \(\pi^N_1\)
| - | \(\pi^D_2\) | - | \(\pi^N_2\)

Table 2.8: Table of underlying joint probabilities for \(k = 2\).

\[
H_{01} : \theta_1 = \pi^D_1 - \pi^N_1 \leq 0 \text{ vs. } H_{a1} : \theta_1 = \pi^D_1 - \pi^N_1 > 0
\]

\[
H_{02} : \theta_2 = \pi^D_2 - \pi^N_2 \leq 0 \text{ vs. } H_{a2} : \theta_2 = \pi^D_2 - \pi^N_2 > 0
\]

This is a typical multiple testing problem, where inferences for \(k = 2\) parameters are to be made. Various multiple testing procedures can be used. We choose to use the fundamental principle for constructing valid multiple tests that strongly controls the FWER – the Partitioning Principle (Stefansson et al., 1988; Finner and Strassburger, 2002) to demonstrate the problem.

To test \(H_{01}, H_{02}\) by the partitioning test, we partition the parameter space into four disjoint subspaces and form the partitioning hypotheses:

\[
H^*_0(12) : \theta_1 \leq 0 \text{ and } \theta_2 \leq 0
\]

\[
H^*_0(1) : \theta_1 \leq 0 \text{ and } \theta_2 > 0
\]

\[
H^*_0(2) : \theta_2 \leq 0 \text{ and } \theta_1 > 0
\]

\[
H^*_0 : \theta_1 > 0 \text{ and } \theta_2 > 0
\]

According to the partitioning principle, each partitioning hypothesis should be tested at level-\(\alpha\) to control FWER at level-\(\alpha\).
One can test the first three hypotheses to make valid inference for $\theta_1$ and $\theta_2$. For convenience, $H^*_{0(12)}$, $H^*_{0(1)}$ and $H^*_{0(2)}$ can be tested by level-$\alpha$ tests of:

\begin{align*}
H_{0(12)} & : \theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \\
H_{0(1)} & : \theta_1 \leq 0 \\
H_{0(2)} & : \theta_2 \leq 0.
\end{align*}

One can test the above hypotheses instead since a level-$\alpha$ test for each of these is guaranteed to be a level-$\alpha$ test for each partitioning hypotheses $H^*_{0(12)}$, $H^*_{0(1)}$ and $H^*_{0(2)}$.

To make inference for $\theta_1$ and $\theta_2$, $H_{0(12)}$, $H_{0(1)}$ and $H_{0(2)}$ are individually tested by permutation tests using test statistic in the form of the difference between group means. Since the actual number of SNPs $k$ is large, we consider test statistics that may lead to a step-down shortcut of the partitioning test. Thus the maxT or minp type of test will be employed, i.e. the null hypothesis $H_{0I}$ will be rejected if $\max_{i \in I} T_i > c_I$ or $\min_{i \in I} p_i < c_I$.

2.3.1 maxT test

If the maxT type of test is employed, the null hypothesis $H_{0I}$ will be rejected if $\max_{i \in I} T_i > c_I$.

Test $H_{0(12)}$

If the maxT test based on the permutation test using test statistic in the form of $\bar{X} - \bar{Y}$ is employed to test $H_{0(12)} : \theta_1 \leq 0 \text{ and } \theta_2 \leq 0$, the test procedure could be:

1. Calculate test statistics $T_i = \hat{\alpha}_i^D - \hat{\alpha}_i^N = \frac{n_i^D}{n_i^D} - \frac{n_i^N}{n_i^N}$, $i = 1, 2$ for SNP1, SNP2.

Let $T = \max_{i=1,2} T_i$. 

50
2. To test the one-sided partitioning hypothesis $H_{0\{12\}}$, permute subjects between the case and normal groups by complete enumeration. For each permuted sample, re-compute $T^b_1, T^b_2$ and retain the maximum value between them, i.e. $T^b = \max_{i=1,2} T^b_i, b = 1, \ldots, B$.

3. Compare the observed $T$ with the $\alpha$ upper percentile of the $B$ $T^b$'s. That is, use the $\alpha$ upper percentile of $T^b$'s, i.e. $c = \inf \{c : Pr(T^b \geq c) \leq \alpha\}$, as the critical value. If $T \geq c$, then reject $H_{0\{12\}}$.

Notice that the valid control of the type I error rate in testing $H_{0\{12\}}$ depends on the rejection regions based on the permutation distribution having the correct size. Suppose the rejection region based on the permutation null distribution is larger than that of the true null distribution. Then if one chooses the critical value $c$ according to the permutation distribution, i.e. $c = \inf \{c : P_{\text{perm}}(T \geq c) \leq \alpha\}$, and rejects $H_{0\{12\}}$ if $T \geq c$, because $P_{\text{true}}(T \geq c) > P_{\text{perm}}(T \geq c)$, the level of the test could be larger than the desired level $\alpha$. This indicates one could reject the partitioning null hypothesis $H_{0\{12\}}(H^*_{0\{12\}})$ too often.

**Test $H_{0\{1\}}, H_{0\{2\}}$**

In practice, the partitioning hypotheses

$$H^*_{0\{1\}} : \theta_1 \leq 0 \text{ and } \theta_2 > 0$$

$$H^*_{0\{2\}} : \theta_2 \leq 0 \text{ and } \theta_1 > 0$$

(2.9)

can be tested through level-$\alpha$ tests of:
\[ H_{0(1)} : \theta_1 \leq 0 \]
\[ H_{0(2)} : \theta_2 \leq 0. \]  

(2.10)

Noticing that a 2×2 contingency table is used between each SNP and the phenotype, a popular choice is to use Fisher’s exact test for each table if the expected cell counts are not large enough to apply large sample approximation based methods. If the p-value from the Fisher’s exact test is small, then the corresponding \( H_{0(i)} \) is rejected. We show that the Fisher’s exact test is indeed a permutation test using the test statistic in the \( \bar{X} - \bar{Y} \) form.

Fisher’s exact test is an exact test for independence between two binary variables in a 2×2 contingency table setting with fixed margins. It is based on the hypergeometric distribution under the null hypothesis of independence, which is equivalent to testing whether the odds ratio \( \phi \) is less than or equal to 1.

Given row and column margins \( n^N, n^D \) and \( n^+_i \) fixed, under the null hypothesis, the conditional distribution of \( n^D_i \) is

\[
P(n^D_i | n^N, n^D, n^+_i) = \binom{n^+_i}{n^D_i} \frac{\binom{n^D + n^N - n^+_i}{n^D - n^D_i}}{\binom{n^D + n^N}{n^D}} \]  

(2.11)

By permuting the observations between groups with complete enumeration under the marginal constraints, Fisher’s exact test computes the exact p-value by summing over the probabilities of all tables that appear to be at least as favorable to the alternative hypothesis compared to current observation, i.e.
\[ p\text{-value} = \sum_{n=n_i^D}^{\min(n^D,n_i^+)} P\{n|n^N,n^D,n_i^+\} \]
\[ = \sum_{n=n_i^D}^{\min(n^D,n_i^+)} \left( \frac{n_i^+}{n} \right) \left( \frac{n^D + n^N - n_i^+}{n^D - n} \right) / \left( \begin{array}{c} n^D + n^N \end{array} \right) \]

Notice that the p-value is determined by \( n_i^D \) only if \( n_i^N, n^D \) and \( n_i^+ \) are all fixed.

Also, since \( n_i^D \sim B(n^D, \pi_i^D) \), \( n_i^N \sim B(n^N, \pi_i^N) \) are two independent binomials, the null hypothesis of independence is:

\[ H_0 : \pi_i^D = \pi_i^N \] (2.13)

Let \( X = (X_1, \ldots, X_{n^D}) \) and \( Y = (Y_1, \ldots, Y_{n^N}) \) be two independent random samples,

\[ X_j \sim Bernoulli(1, \pi_i^D), j = 1, \ldots, n^D \]
\[ Y_k \sim Bernoulli(1, \pi_i^N), k = 1, \ldots, n^N \] (2.14)

Let

\[ \bar{X} = \frac{1}{n^D} \sum_{j=1}^{n^D} X_j \]
\[ \bar{Y} = \frac{1}{n^N} \sum_{k=1}^{n^N} Y_k \]
\[ T_i = \bar{X} - \bar{Y} = \frac{n_i^D}{n^D} - \frac{n_i^N}{n^N} = n_i^D \left( \frac{1}{n^D} + \frac{1}{n^N} \right) - \frac{n_i^+}{n^N} \] (2.15)

Given that \( n_i^N, n^D \) and \( n_i^+ \) are fixed, \( T_i \) is a function of \( n_i^D \) only. Also, since \( \frac{1}{n^N} + \frac{1}{n^D} > 0 \), \( T_i \) monotonically increases with \( n_i^D \). Therefore, the p-value of Fisher’s exact test and the permutation test based on \( T_i \) are identical.
Fisher’s exact test for hypotheses $H_{0\{1\}}$, $H_{0\{2\}}$ is hereby a permutation test with complete enumeration using test statistics $T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$, $i = 1, 2$, which has the form of group means difference $\bar{X} - \bar{Y}$. Rejection of $H_{0\{i\}}, i = 1, 2$ by Fisher’s exact test is equivalent to rejection by $T_i \geq c_i, i = 1, 2$, where $c_i$ is the upper $\alpha$ percentile of the permutation distribution of $T_i$.

Notice that, if Fisher’s exact test is used to test the association between one SNP, i.e. $k = 1$, and the binary phenotype, then the type I error rate is properly controlled. From (2.12), one can tell under the null hypothesis of marginal prevalence identity, the distribution of the observed table is determined by $n_i^D$ and marginal constraints of the study only. It does not involve any marginal or joint probabilities. Therefore, the MDJ condition is satisfied in $k = 1$ case, which means type I error rate is controlled.

**FWER may not be controlled**

Notice that, in the permutation test for $H_{0\{12\}} : \theta_1 \leq 0$ and $\theta_2 \leq 0$ with complete enumeration, $T^b = \max_{i=1,2}(T^b_i), b = 1, \ldots, B$. Obviously, $T^b \geq T^b_i, i = 1, 2$. Therefore, for any constant $t$,

$$Pr(T^b \geq t) \geq Pr(T^b_i \geq t), i = 1, 2.$$

Let $c$ be the upper $\alpha$ percentile of $T^b$, i.e. $c = \inf \{c : Pr(T^b \geq c) \leq \alpha\}$. Let $c_i$ be the upper $\alpha$ percentile of $T^b_i$, i.e. $c_i = \inf \{c_i : Pr(T^b_i \geq c_i) \leq \alpha\}, i = 1, 2$. It is easy to see $c \geq c_i, i = 1, 2$.

Let $T_{[1]} \leq T_{[2]} = \max_{i=1,2}T_i = T$. Therefore, it is guaranteed if $T \geq c$, then at least $T = T_{[2]} \geq c_{[2]}$, where $c_{[2]}$ is one of $c_i, i = 1, 2$ that corresponds to $T_{[2]}$. That is, at least one of the two events $\{T_i \geq c_i, i = 1, 2\}$ occurs if $\{T \geq c\}$ occurs.
Recall that Fisher’s exact test is a permutation test with complete enumeration also in the form of $X - Y$. That is, rejection of $H_{0(i)}: \theta_i \leq 0, i = 1, 2$ by Fisher’s exact test is equivalent to rejection by $T_i \geq c_i$ by permutation test, where $T_i = \hat{\pi}^D_i - \hat{\pi}^N_i, i = 1, 2$.

The above results show that whenever $H_{0\{12\}}$ is rejected, i.e. $\{T \geq c\}$, it is guaranteed that at least one of $H_{0(i)}, i = 1, 2$ is also rejected, i.e. at least one of $\{T_i \geq c_i, i = 1, 2\}$ occurs. That is, whenever $H^*_{0\{12\}}$ is rejected, it is guaranteed that at least one of $H^*_{0(i)}, i = 1, 2$ is also rejected.

Therefore, according to the partitioning principle, collating the results leads to:

$\{\text{Reject } H^*_{0\{12\}}\} = \{\text{Reject at least one of } H^*_{0(i)}, i = 1, 2\}$

$= \{\text{Reject at least one of } H_{0i}, i = 1, 2\}$ \hspace{1cm} (2.17)

Recall the definition of FWER is

$$sup_{\theta \in \Theta} P_{\theta}(\text{Reject at least one of true null hypothesis } H_{0i})$$ \hspace{1cm} (2.18)

where the $sup$ is taken over the entire parameter space $\Theta$. Given $I$, the non-empty collection of true null hypotheses, so that all $H_{0i}, i \in I$ are true and $\theta \in \Theta^I$.

Suppose $I = \{1, 2\}$,

$$FWER = sup_{\theta \in \Theta^I} P_{\theta}(\text{Reject at least one of } H_{0i}, i \in I)$$

$= sup_{\theta_1 \leq 0, \theta_2 \leq 0} P_{\theta}(\text{Reject at least one of } H_{01}, H_{02})$

$\geq P_{\theta_1 \leq 0 \cap \theta_2 \leq 0}(\text{Reject at least one of } H_{01}, H_{02})$

$= P_{\theta_1 \leq 0 \cap \theta_2 \leq 0}(\text{Reject } H^*_{0\{12\}})$

$= P^{true}(T \geq c)$ \hspace{1cm} (2.19)
where the \( sup \) is computed when the parameters of interest, i.e. \( \theta_1 \) and \( \theta_2 \), are under the constraint of the null hypotheses. No constraints are put on the other parameters such as the joint probabilities.

Suppose \( P^{perm}(T \geq c) = \alpha \). As discussed in the section on testing \( H_{0(12)} \), if the rejection region based on the permutation distribution is larger than the true rejection region, i.e. \( P^{true}(T \geq c) > P^{perm}(T \geq c) = \alpha \), then FWER may not be controlled for permutation tests based partitioning test procedure.

**Conditional and unconditional distribution calculation**

To explore whether permutation tests based multiple testing procedure controls FWER, one needs to compare the rejection region based on the permutation reference distribution with that of the true null distribution.

The work by Huang et al. is within the multivariate continuous data frame (Huang et al., 2006). Therefore, their cumulants results in Theorem 2.2 were derived in terms of the unconditional distribution of the test statistics, which is more sensible in the continuous data setting.

Compared to the continuous data setting, conditional distributions are perhaps more meaningful for discussion within the discrete data setting. In reality, permutation test based multiple testing is performed on an observed table margin, which is conditional. Although computation of the unconditional distributions leads to examples where the permutation tests are liberal given sample sizes and joint probabilities, it is also sensible for one to compare the conditional permutation distribution and the conditional true distribution for specific margins and show that the difference in these distributions may lead to anti-conservative results conditionally for some sample tables.
Given relatively small sample sizes $n^D, n^N$, one can compute the conditional and unconditional true null distribution and the permutation distribution of the test statistic $T = \max_{i=1,2} T_i$ exactly based on the underlying joint probabilities.

Here by ‘conditional’ we refer to conditioning on the observed frequencies of all possible combinations of $k = 2$ SNPs mutation status after combining the two phenotype groups together, i.e. margin $m = (n_{++}, n_{+-}, n_{-+}, n_{--})$, where $n_{++} = n^{D++} + n^{N++}$, $n_{+-} = n^{D+} + n^{N-}$, $n_{-+} = n^{D-} + n^{N+}$ and $n_{--} = n^{D--} + n^{N--}$.

Assume a case-control study design. Let $\pi = (\pi^D, \pi^N) = (\pi^{D++}, \pi^{D+}, \pi^{D-}, \pi^{D--}, \pi^{N++}, \pi^{N+}, \pi^{N-}, \pi^{N--})$ be the vector of joint probabilities. The probability for a margin $m$ equals the sum of probabilities of all joint tables that are subject to $m$, where the probability of each table is calculated based on the multinomial distribution each phenotype group follows, i.e. $MN(n^D, \pi^{D++}, \pi^{D+}, \pi^{D-}, \pi^{D--})$ and $MN(n^N, \pi^{N++}, \pi^{N+}, \pi^{N-}, \pi^{N--})$. This probability is shown in (2.20).

\[
P(m|n^N, n^D, \pi) = \sum_{n^D, n^N|m \in \Xi} \frac{n^D!}{n^{D++}! n^{D+}! n^{D-}! n^{D--}!} \times \frac{n^N!}{n^{N++}! n^{N+}! n^{N-}! n^{N--}!} \times \pi^{D++} n^{D+} n^{D-} n^{D--} \pi^{N++} n^{N+} n^{N-} n^{N--}
\]

where for simplicity of notation, $n^D$ denotes $n^{D++}, n^{D+}, n^{D-}, n^{D--}$, $n^N$ denotes $n^{N++}, n^{N+}, n^{N-}, n^{N--}$. Together, $n^D, n^N$ denote a sample table as in Table 2.7. $\Xi$ indicates all possible margins $m$ that sum to $n^D + n^N = N$. Thus all tables considered for the calculations satisfy $\sum n^D = n^D, \sum n^N = n^N$ and margin $m$. 

57
Under the null hypothesis that there is no difference in distributions between the two groups, the conditional permutation probability of \( T = t \) is derived by summing over permuted sample tables which have the test statistic value \( t \) out of all possible permutations:

\[
P_{\text{perm}}(T = t|m) = \sum_{n_D,n^N|D,n^N,m} I_{(T|n_D,n^N,m)} \times P(n_D^+,n^N|n_D^+,n^N,m)
\]

\[
= \sum_{n_D,n^N|D,n^N,m} I_{(T|n_D,n^N,m)} \times \left\{ \left( \frac{n_{++}}{n_D^{++}} \right) \left( \frac{N - n_{++}}{n_D^{++} - n_{++}} \right) \left( \frac{N - n_{++}}{n_D^{++} - n_{++}} \right) \right\} \times \left\{ \left( \frac{n_{+ -}}{n_D^{+-}} \right) \left( \frac{N - n_{+ -} - n_{+-}}{n_D^{+-} - n_{+-}} \right) \left( \frac{N - n_{+ -} - n_{+-}}{n_D^{+-} - n_{+-}} \right) \right\}
\]

\[
= \sum_{n_D,n^N|D,n^N,m} I_{(T|n_D,n^N,m)} \left( \frac{n_{++}}{n_D^{++}} \right) \left( \frac{n_{+ -}}{n_D^{+-}} \right) \left( \frac{n_{+-}}{n_D^{-+}} \right) \left( \frac{n_{- -}}{n_D^{--}} \right) \left( \frac{N}{n_D^{--}} \right)^{-1}
\]

where \( I_{(T|n_D,n^N,m)} \) is the indicator function for \( T = t \) given a permuted sample table with margin \( m \).

The conditional true null distribution of \( T \) is derived from the probability of observing test statistic \( T = t \) given margin \( m \):

\[
P_{\text{true}}(T = t|m) = \frac{\sum_{n_D,n^N|m} I_{(T|n_D,n^N)} P(n_D^+,n^N|n_D^+,n^N,\pi)}{P(m|n^N,n^D,\pi)}
\]

\[
= \left\{ \sum_{n_D,n^N|m} I_{(T|n_D,n^N)} \right\} \left( \frac{n_D^!}{n_{++}!n_{++}!n_{--}!} \right) \left( \frac{n_D^!}{n_{+ -}!n_{+ -}!n_{- -}!} \right) \left( \frac{n_D^!}{n_{- +}!n_{- +}!n_{--}!} \right) \left( \frac{n_D^!}{n_{--}!n_{--}!n_{--}!} \right) \left( \frac{N}{n_D^{--}} \right)^{-1}
\]

\[
= \left\{ \frac{n_D^!}{n_{++}!n_{++}!n_{--}!} \right\} \left( \frac{n_D^!}{n_{+ -}!n_{+ -}!n_{- -}!} \right) \left( \frac{n_D^!}{n_{- +}!n_{- +}!n_{--}!} \right) \left( \frac{n_D^!}{n_{--}!n_{--}!n_{--}!} \right) \left( \frac{N}{n_D^{--}} \right)^{-1} \times \left\{ \frac{n_N^!}{n_{++}!n_{++}!n_{--}!} \right\} \left( \frac{n_N^!}{n_{+ -}!n_{+ -}!n_{- -}!} \right) \left( \frac{n_N^!}{n_{- +}!n_{- +}!n_{--}!} \right) \left( \frac{n_N^!}{n_{--}!n_{--}!n_{--}!} \right) \left( \frac{N}{n_D^{--}} \right)^{-1}
\]
where \( I_{(T=t|m,n^D,n^N)} \) is an indicator function for \( T = t \) of an observable sample table with constraints \( n^D, n^N \) and \( \pi \). \( P(m|n^N,n^D,\pi) \) is from formula (2.20).

Under the marginal null hypothesis that \( \pi_i^D = \pi_i^N = \pi_i, i = 1, 2 \), by some simple algebra, equation (2.22) can be further simplified as:

\[
P_{\text{true}}(T = t|m) = \frac{\sum_{n^D,n^N|m} I_{(T=t|n^D,n^N)} A \cdot \left( \frac{\pi_+^D}{\pi_{++}^+} \right)^{n_+^D} \left( \frac{\pi_-^D}{\pi_{-+}^+} \right)^{n_-^D} \left( \frac{\pi_+^N}{\pi_{++}^-} \right)^{n_+^N} \left( \frac{\pi_-^N}{\pi_{-+}^-} \right)^{n_-^N} } {\sum_{s^D,s^N|m \in \Xi} A \cdot \left( \frac{\pi_+^D}{\pi_{++}^+} \right)^{s_+^D} \left( \frac{\pi_-^D}{\pi_{-+}^+} \right)^{s_-^D} \left( \frac{\pi_+^N}{\pi_{++}^-} \right)^{s_+^N} \left( \frac{\pi_-^N}{\pi_{-+}^-} \right)^{s_-^N} } \]  

(2.23)

where \( A = \binom{n_+^D}{n_+^+} \binom{n_-^D}{n_-^-} \binom{n_+^N}{n_+^+} \binom{n_-^N}{n_-^-} (n_D)^{-1} \).

Notice that the conditional permutation distribution based on (2.21) does not depend on the underlying joint probability vector \( \pi \). It is only determined by the probabilities of all sample tables with \( T = t \) by complete enumeration. While the conditional true null distribution of the test statistic is determined by the underlying joint probabilities \( \pi \). This may lead to the difference between the conditional permutation distribution and the true distribution of the test statistics.

If the joint probabilities from the two groups are equal, i.e. \( \pi^D = \pi^N \), then formula (2.23) is further reduced to be equivalent to formula (2.21), that is, the permutation distribution matches the true distribution. However, if the marginal null hypothesis holds, where \( \pi_i^D = \pi_i^N, i = 1, 2 \), but \( \pi^D \neq \pi^N \), then formula (2.22) can be only reduced to formula (2.23).

Computation of the unconditional permutation distribution of \( T \) with complete enumeration is based on all possible margins that satisfy the sample size constraint, i.e. \( m \in \Xi, \Xi = \{ m \in \mathbb{N}(2^k) : \sum m = N \} \).
\[ P_{\text{perm}}(T = t|n^N, n^D, \pi) \]
\[ = \sum_{m \in \Xi} \sum_{n^D, n^N|n^D, n^N, m} I_{(T = t|n^D, n^N, m)} \times P(n^D, n^N|m) \times P(m|n^N, n^D, \pi) \]  

(2.24)

where \( I_{(T = t|n^D, n^N, m)} \) is an indicator function for \( T = t \) given a permuted sample table under margin \( m \). \( P(m|n^N, n^D, \pi) \) is computed as in (2.20).

The computation of the unconditional true null distribution of \( T \) is based on all joint tables which consist of two independent multinomial distributions.

\[ P_{\text{true}}(T = t|n^D, n^N, \pi) = \sum_{n^D, n^N|m \in \Xi} I_{(T = t|n^D, n^N)} \times P(n^D, n^N|m) \times P(m|n^N, n^D, \pi) \]

\[ = \sum_{n^D, n^N|m \in \Xi} I_{(T = t|n^D, n^N)} \times \frac{n^D!}{n^D_{++}!n^D_{++}!n^D_{--}!n^D_{--}!} \frac{n^D_{++}^{n^D_{++}}n^D_{+-}^{n^D_{+-}}n^D_{-+}^{n^D_{-+}}n^D_{--}^{n^D_{--}}}{\pi^D_{++}\pi^D_{+-}\pi^D_{-+}\pi^D_{--}} \times \frac{n^N!}{n^N_{++}!n^N_{++}!n^N_{--}!n^N_{--}!} \frac{n^N_{++}^{n^N_{++}}n^N_{+-}^{n^N_{+-}}n^N_{-+}^{n^N_{-+}}n^N_{--}^{n^N_{--}}}{\pi^N_{++}\pi^N_{+-}\pi^N_{-+}\pi^N_{--}} \]  

(2.25)

where \( I_{(T = t|n^D, n^N)} \) is indicator function for \( T = t \) given any table that satisfies the marginal total \( n^D, n^N \).

Formulas (2.24) and (2.25) compute the unconditional probabilities of \( T = t \), since they are computed over tables with all possible margins \( m \in \Xi \) given observed sample sizes \( n^D + n^N = N \).

In reality, permutation test based multiple testing is performed on an observed table, which is conditional on margin \( m \). However, the results of anti-conservativeness in unconditional distributions are still quite useful in general. Given sample sizes and underlying joint probabilities, liberal unconditional results indicate that for at least one table among all possible joint tables that are subject
to the marginal constraints, the FWER is not controlled. That is, the conditional 
FWER is not controlled for at least one joint table. Therefore, to study the unconditional distributions gives one insight on when a permutation test may lead to invalid reference.

On the other hand, if for every margin, the conditional permutation test is con-
servative, then over all possible margins the unconditional permutation test will be conservative.

2.3.2 \textit{minp test}

If the minp type of test is employed, the null hypothesis $H_{0I}$ will be rejected if $\min_{i \in I} p_i < c'$. 

The test statistic $T = \max_{i=1,2} T_i, T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$ is not scaled. Under the null hypothesis that $\pi_i^D = \pi_i^N = \pi_i$,

$$Var(T_i) = Var(\hat{\pi}_i^D) + Var(\hat{\pi}_i^N)$$

$$= \frac{1}{n^D} \hat{\pi}_i^D(1 - \hat{\pi}_i^D) + \frac{1}{n^N} \hat{\pi}_i^N(1 - \hat{\pi}_i^N)$$

$$= \left( \frac{1}{n^D} + \frac{1}{n^N} \right) \hat{\pi}_i(1 - \hat{\pi}_i)$$

The scale of $T_i$ may be substantially different due to different prevalence $\pi_i, i = 1, 2$. Obviously, such test statistics are only appropriate if $\pi_1$ and $\pi_2$ are close to each other. Otherwise, the scale of $T$ could be mainly determined by the single SNP with much larger prevalence. That is, if one rejects $H_{0\{12\}}$ by $T \geq c$, where $c$ is the upper $\alpha$ percentile of $T$ under the null hypothesis, the information of the test may come from one SNP mainly.
P-values can be deemed as scaled test statistics in some sense. It is asymptotically distributed as $Unif(0,1)$ under the null hypothesis. We will explore whether partitioning test based on p-values from permutation test in the form of $\bar{X} - \bar{Y}$ controls FWER.

It has been shown that Fisher’s exact test is a permutation test in the $\bar{X} - \bar{Y}$ form with complete enumeration. Thus, p-values from the proposed type of permutation test are exactly those from Fisher’s exact tests.

**Test $H_{0\{12\}}$: minp permutation test**

To test $H_{0\{12\}}$, the permutation test procedure could be:

1. Calculate test statistics $p_i, i = 1, 2$ by Fisher’s exact test. Let $p = min_{i=1,2} p_i$.

2. To test the one-sided null hypothesis $H_{0\{12\}} : \theta_1 \leq 0$ and $\theta_2 \leq 0$, permute subjects between the disease and normal groups by complete enumeration. For each permuted sample, re-compute $p^b_1, p^b_2$ and retain the minimal value between them, i.e. $p^b = min_{i=1,2} p^b_i, b = 1, \ldots, B$.

3. Compare the observed $p$ with the $\alpha$ lower percentile of the $B$ $p^b$’s. That is, if $p \leq d$, then reject $H_{0\{12\}}$ where $d = max\{d : Pr(p^b \leq d) \leq \alpha\}$, as the critical value.

**Test $H_{0\{1\}}, H_{0\{2\}}$: Permutation test by minp**

As shown in an earlier section, p-values from Fisher’s exact test for SNP1 and SNP2 are equivalent to p-values from a permutation test in the $\bar{X} - \bar{Y}$ form.

Let $p_i, i = 1, 2$ be p-values obtained from Fisher’s exact test for SNP1, SNP2, respectively. $H_{0\{i\}}, i = 1, 2$ is rejected if corresponding $p_i \leq \alpha$. 
**FWER is not controlled**

Similarly as in the maxT permutation test, notice that in the minp permutation test for $H_{0(12)}: \theta_1 \leq 0$ and $\theta_2 \leq 0$ by complete enumeration, $p^b = \min_{i=1,2}(p_i^b), b = 1, \ldots, B$, so that $p^b \leq p_i^b, P(p^b \leq p) \geq P(p_i^b \leq p), i = 1, 2$.

Let $d$ be the lower $\alpha$ percentile of $p^b$, i.e. $d = \max\{d : Pr(p \leq d) \leq \alpha\}$. Let $d_i$ be the lower $\alpha$ percentile of $p_i^b$, i.e. $d_i = \max\{d_i : Pr(p_i^b \leq d_i) \leq \alpha\}, i = 1, 2$. Therefore $d \leq d_i, i = 1, 2$.

Let $p = \min_{i=1,2}p_i = p_{[1]} \leq p_{[2]}$. Since $d \leq d_i, i = 1, 2$, it is guaranteed if $p \leq d$, then at least $p = p_{[1]} \leq d_{[1]}$, where $d_{[1]}$ is one of $c_i, i = 1, 2$ that corresponds to $p_{[1]}$. That is, at least one of the two events \{p_i \leq d_i, i = 1, 2\} occurs if \{p \leq d\} occurs.

The above result shows that whenever $H_{0(12)}$ is rejected, it is guaranteed that at least one of $H_{0(i)}, i = 1, 2$ is also rejected. That is, whenever $H_{0(12)}^*$ is rejected, it is also guaranteed that at least one of $H_{0(i)}^*, i = 1, 2$ is also rejected. Therefore, FWER is not controlled for minp test in this case, due to a similar reason as we have shown in the maxT example. Suppose $I = \{1, 2\}$,

\[
\text{FWER} = \sup_{\theta_1 \leq 0, \theta_2 \leq 0} P_{\theta}(\text{Reject at least one of } H_{01}, H_{02}) \\
\geq P_{\theta_1 \leq 0, \theta_2 \leq 0}(\text{Reject at least one of } H_{01}, H_{02}) \\
= P_{\theta_1 \leq 0, \theta_2 \leq 0}(\text{Reject } H_{0(12)}^*) \\
= P^{true}(P \leq d)
\]

(2.27)

if the permutation distribution is used and $P^{true}(p \leq d) > P^{perm}(p \leq d) = \alpha$, then FWER is not controlled.
Conditional and unconditional distribution calculation

Given relatively small $n_D, n_N$, one can calculate the unconditional and conditional true null distribution and the exact permutation distribution of the test statistic $p = \min_{i=1,2} p_i$ exactly based on given underlying joint probabilities.

Very similar to the calculations in $maxT$ permutation test section, given margins $m$ subject to $n_D, n_N$, the conditional permutation distribution of $P = p$ is computed based on

$$P_{\text{perm}}(P = p|m) = \sum_{n_D, n_N|m} I_{(P = p|n_D, n_N)} \times P(n_D, n_N|n_D, n_N, m)$$ (2.28)

where $I_{(P = p|n_D, n_N)}$ is the indicator function for $P = p$ given a permuted sample joint table.

The conditional true null distribution of $p$ is computed based on

$$P_{\text{true}}(P = p|m) = \frac{\sum_{n_D, n_N|m} I_{(P = p|n_D, n_N)} P(n_D, n_N|n_D, n_N, \pi)}{P(m|n_N, n_D, \pi)}$$ (2.29)

where $I_{(P = p|n_D, n_N, \pi)}$ is an indicator function for $P = p$ for an observed sample table defined by $n_D$ and $n_N$ and satisfying sample sizes $n_D$ and $n_N$, and $\pi$.

The unconditional permutation distribution with complete enumeration is based on the computation of the probabilities of joint tables with certain margins:

$$P_{\text{perm}}(P = p|n_N, n_D, \pi)$$

$$= \sum_{m \in \Xi} \sum_{n_D, n_N|m} I_{(P = p|n_D, n_N, m)} \times P(n_D, n_N|n_D, n_N, m) \times P(m|n_N, n_D, \pi)$$ (2.30)
where \( I_{(P=p|n^D, n^N, m)} \) is indicator function for \( P = p \) given a permuted sample and \( P(m|n^N, n^D, \pi) \) is computed by (2.20).

The computation of the true null distribution of \( p \) is based on the two independent multinomial distributions that observations of the normal and disease group follow respectively:

\[
P^{true}(P = p|n^D, n^N, \pi) = \sum_{n^D, n^N|m\in\Xi} I_{(P=p|n^D, n^N, \pi)} \times P(n^D, n^N|n^D, n^N, \pi)
\]

\[
= \sum_{n^D, n^N|m\in\Xi} I_{(P=p|n^D, n^N, \pi)} \times \frac{n^D!}{n^D_+!n^D_-!n^N_+!n^N_-!} \pi^D_+^{n^D_+} \pi^D_-^{n^D_-} n^D_+^D \pi^N_+^{n^N_+} \pi^N_-^{n^N_-} n^N_+^N
\]

\[
\times \frac{n^N!}{n^N_+!n^N_-!n^N_+!n^N_-!} \pi^N_+^{n^N_+} \pi^N_-^{n^N_-} n^N_+^N
\]

where \( I_{(P=p|n^D, n^N, \pi)} \) is an indicator function for \( P = p \) given any table that satisfies the marginal total \( n^D, n^N \).

### 2.3.3 Example of MDJ does not hold

The next question is where to find examples that permutation tests based multiple testing procedures may not control FWER. Recall the MDJ condition (Xu and Hsu, 2007) is a sufficient condition for permutation tests to be valid for multiple hypotheses testing of marginal distributional identity. That is, violation of the MDJ condition may lead to an uncontrolled false rejection rate.

We take the very simple numerical example of 2 SNPs to show that if the MDJ condition is not satisfied, even if the null hypotheses of identity in marginal probabilities hold, the partition testing procedure based on permutation tests in the form of
$X - Y$ may not be a level-$\alpha$ test for $H_{0(12)}$. Thus the overall FWER is not guaranteed to be controlled.

Recall that the joint distribution of SNP1, SNP2 mutation status for each group follows a multinomial distribution, i.e. $F^D = MN(n^D, \pi^D_{++}, \pi^D_{+-}, \pi^D_{-+}, \pi^D_{--})$ and $F^N = MN(n^N, \pi^N_{++}, \pi^N_{+-}, \pi^N_{-+}, \pi^N_{--})$. The joint distributional identity indicates $H_0^J: F^D = F^N$.

However, equality in the marginal proportions $\pi^D_1 = \pi^N_1, \pi^D_2 = \pi^N_2$ does not necessarily imply $F^D = F^N$. Under the marginal identity, the difference in the joint probabilities is determined by $\pi^D_{++}$ and $\pi^N_{++}$, i.e. the correlation between SNP1 and SNP2 in each phenotype group:

$$corr(SNP_i, SNP_j | Disease) = \pi^D_{++} - \pi^D_i \times \pi^D_j, \ i, j \in \{1, \ldots, k\}, \ i \neq j$$

$$corr(SNP_i, SNP_j | Normal) = \pi^N_{++} - \pi^N_i \times \pi^N_j, \ i, j \in \{1, \ldots, k\}, \ i \neq j \quad (2.32)$$

Therefore, if $\pi^D_{++} = \pi^N_{++}$, identity in marginal prevalence implies identity in the joint distributions, i.e. $F^D = F^N$. The bigger the difference between $\pi^D_{++}$ and $\pi^N_{++}$, the bigger the difference in the joint distributions between the two phenotype groups.

This is an example of a violation of the MDJ condition. Thus within the 2 SNPs data setting, it is possible to find examples that do not satisfy the MDJ condition under the null hypothesis of marginal probability identity. For example, we can construct examples reflecting SNP-SNP interactions.

Let SNP1 and SNP2 have interactive effects on the probability of being in the disease group, for example, mutations in both SNPs increase the risk of disease synergistically. This could happen when mutations in both SNPs affect normal functions of two genes which are in the same or related pathways of a disease. Then it may be
reasonable to expect more observations with both SNPs mutated in the disease group than in the normal group, i.e. $\pi_{++}^D > \pi_{++}^N$. In the meanwhile, identities in marginal prevalence may still hold.

For example, at least eight BBS genes (BBS 1-BBS 8) have been identified to interactively modify the onset and/or clinical severity of the Bardet-Biedl syndrome (BBS). One more locus, MGC1203, was also found to have an epistatic effect on the phenotype of BBS (Badano et al., 2006). It was confirmed in vivo, that zebrafish embryos with mutations in both the MGC1203 locus and one of the BBS gene loci had much higher proportions of BBS with worse severity compared to embryos with a single mutation in either locus, suggesting interactive effects from multiple genes on one phenotype. The other example is that it was found that double mutations in Peripherin/RDS and ROM1 loci lead to retinitis pigmentosa while single mutation in either locus leads to normal phenotypes (Kajiwara et al., 1994).

One simple numerical example of violation of the MDJ condition is shown in table 2.9.

<table>
<thead>
<tr>
<th>Disease (Phenotype = 1)</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP 1</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>0.45</td>
</tr>
<tr>
<td>−</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal (Phenotype = 0)</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP 1</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>0.25</td>
</tr>
<tr>
<td>−</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 2.9: One numerical example where MDJ condition does not hold.
In this example, the marginal probabilities of SNP1 or SNP2 mutation in the disease group and normal group are equal respectively, i.e. $\pi_1^D = \pi_1^N = 0.5$, $\pi_2^D = \pi_2^N = 0.47$. However, the joint probabilities in the two phenotype groups are not equal, i.e. $\pi^D \neq \pi^N$.

**Conditional results**

Following this idea, one can explore the differences between the conditional and unconditional permutation reference distribution and the true reference distribution of the test statistics $\max T$ or $\min p$ under various sample sizes as well as joint probabilities configurations.

In practice, once a GWA study has been conducted, the observed table is fixed. The permutation test will be performed conditioning on the marginal constraint $m$. It is certainly meaningful to study anti-conservative examples of permutation tests conditionally in the discrete data analysis setting.

We first explored the magnitude of the difference between the conditional permutation probability and the true null probability of the test statistics $T = t$. Under the marginal null hypothesis that $\pi_i^D = \pi_i^N = \pi_i, i = 1, 2$, we searched over a grid of marginal probabilities $\pi_1$ and $\pi_2$ in the parameter space of $\pi_1 \in [0, 1], \pi_2 \in [0, 1]$. For each combination over the grid, sample size from the disease group $n^D$ was fixed at 2, while $n^N$ was allowed to change from 2 to 20. The maximum difference of $P^{true}(T = t) - P^{perm}(T = t)$ over all tables that are compatible with the sample sizes and all possible test statistic values $t$ were plotted in Figures 2.1 and 2.2.

Recall that the correlations between the two SNPs are determined by $\pi_{++}^D$ and $\pi_{++}^N$ under the null hypothesis of marginal identity. The bigger the difference between $\pi_{++}^D$ and $\pi_{++}^N$, the bigger the difference in the joint distributions between the two
phenotype groups. Therefore, we did the numerical exploration under the worst-case scenario, where the difference in the joint distributions is maximized, i.e. \((\pi^{D}_{++} - \pi^{N}_{++})\) is maximized. That is, \(\pi^{D}_{++}\) and \(\pi^{N}_{++}\) are fixed at their boundary values \(\min(\pi_1, \pi_2)\) or \(\max(0, \pi_1 + \pi_2 - 1)\). In Figure 2.1, \(\pi^{D}_{++} < \pi^{N}_{++}\). In Figure 2.2, \(\pi^{D}_{++} > \pi^{N}_{++}\).

It can be seen that the magnitude of the differences between the conditional permutation probability and the true probability can be quite large (close to 1). This warns us that for some observed samples, permutation test based multiple testing procedures could be severely misleading.
Figure 2.1: Maximum difference between the conditional permutation probability and the true null probability of $T = t$ when $n^D \leq n^N$, $\pi_{++}^D > \pi_{++}^N$. 

$1\text{-sided max}T,n^D=2,\pi_{++}^D > \pi_{++}^N$
Figure 2.2: Maximum difference between the conditional permutation probability and the true null probability of $T = t$ when $n^D \leq n^N, \pi^D_{++} < \pi^N_{++}$
One numerical example from Figure 2.1 is based on sample size \( n^D = 2, n^N = 4 \) and the observed margin \( m = (n_{++} = 2, n_{+-} = 3, n_{-+} = 0, n_{--} = 1) \). The underlying joint probabilities are \( \pi^D = (0.3, 0.3, 0, 0.4), \pi^N = (0, 0.6, 0.3, 0.1) \), as shown in Table 2.3.3.

\[
\begin{array}{c|cc}
\text{Disease(Phenotype = 1)} & \text{SNP 2} & \text{SNP 1} \\
+ & + & 0.3 \\
- & + & 0.3 \\
+ & - & 0.6 \\
- & - & 1 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{Normal(Phenotype = 0)} & \text{SNP 2} & \text{SNP 1} \\
+ & + & 0.6 \\
- & + & 0.1 \\
+ & - & 0.4 \\
- & - & 1 \\
\end{array}
\]

Table 2.10: Numerical example from Figure 2.1.

Through conditional probability calculations, the test statistic \( T \) has conditional probabilities as shown in Table 2.3.3.

\[
\begin{array}{c|c|c|c}
\text{Test Statistics} & P_{\text{perm}}(T = t|m) & P_{\text{true}}(T = t|m) & \text{Difference} \\
T = -0.5 & 0.2 & 0 & 0.2 \\
T = 0.25 & 0.733333 & 0 & 0.733333 \\
T = 1 & 0.666667 & 1 & -0.333333 \\
\end{array}
\]

Table 2.11: Conditional permutation probabilities and true probabilities of \( T = t \).

Notice that under the true underlying joint probabilities, \( T = 1 \) is the only possible test statistic since a sample table can be only observed if it follows the underlying joint probabilities. Given \( \pi_{-+}^D = \pi_{++}^N = 0 \), it is determined that \( n_{-+}^D = n_{++}^N = 0 \) in
any consistent sample table. Given \( n_{++} = 2, n_{++} = n^D_{++} + n^N_{++} \), thus \( n^D_{++} = 2 \). Also, since \( n^D = 2 \) and \( n^D_{++} = 2, n^D_{++} = n^D_{+} = n^D_{-} = 0 \). Further, since \( n_{+} = 3, n_{-} = 0, n_{-} = 1 \), we obtain \( n^N_{+} = 3, n^N_{-} = 0, n^N_{-} = 1 \). That is, for the given margin, the observable true joint table must be:

<table>
<thead>
<tr>
<th>Disease(Phenotype = 1)</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Normal(Phenotype = 0)</th>
<th>SNP 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

In this case, the test statistics based on the true sample can only be 1:

\[
T_1 = \hat{\pi}^D_1 - \hat{\pi}^N_1 = (2 + 0)/2 - (0 + 3)/4 = 0.25
\]

\[
T_2 = \hat{\pi}^D_2 - \hat{\pi}^N_2 = (2 + 0)/2 - (0 + 0)/4 = 1
\]

\[
T = \max_{i=1,2} T_i = 1
\]  

(2.33)

On the other hand, the permutation samples allow three different test statistics \( T = -0.25, T = 0.50 \) and \( T = 1 \). Recall that as shown in (2.21), the conditional permutation distribution of the test statistics does not depend on the underlying joint probabilities \( \pi \). Once the margin is observed, i.e. \( m = (n_{++} = 2, n_{+} = 3, n_{-} = 0, n_{-} = 1) \), the permutation test will generate samples that are only subject to constraints \( m \) and \( n^D = 2, n^N = 4 \), but is not limited by \( \pi = (\pi^D, \pi^N) \). Therefore, there are more permuted samples that are possible. That is, there could be more
$T$ values based on the permutation samples than which are truly observable. This results in the difference between the conditional permutation distribution and the true reference distribution.

Specifically, for the above numerical example, the reported p-values based on the conditional permutation distribution and the conditional true null distribution are

$$P_{\text{perm}}(T \geq 1|m = (2, 3, 0, 1)) = 0.067$$

$$P_{\text{true}}(T \geq 1|m = (2, 3, 0, 1)) = 1$$

That is, if $T = 1$ is used as the test statistic for the observed joint table with margins $m = (2, 3, 0, 1)$ and underlying joint probabilities as shown in Table 2.3.3, the p-value obtained by the permutation test is 0.067, while the true significance level is 1. Therefore, if one uses $T = 1$ as the critical value to test $H_{0(12)}$ in permutation tests based partitioning test, the conditional version of FWER is not controlled.

Figure 2.3 and Figure 2.4 further show the difference in reported p-values based on the permutation test and the true distribution of the test statistics. The same parameter configurations as in Figure 2.1 and 2.2 were used.

As we used the worst case scenario condition, the conditional test result shows that in practice, the problem of using permutation tests for null hypotheses on marginal differences could be quite severe.

As shown in all above graphs, for certain sample sizes $n^D$ and $n^N$ and underlying joint probabilities, we can find reported p-values based on the conditional permutation test that are smaller than that based on the true null distribution for at least some margin $m$. That is, $P_{\text{true}}(T \geq c|m) > P_{\text{perm}}(T \geq c|m)$ or $P_{\text{true}}(p \leq d|m) > P_{\text{perm}}(p \leq d|m)$ for constant $c$ or $d$. 74
The maximum difference in p-values changes with sample sizes. Compared to other sample size configurations, when \( n^D = n^N \), the magnitude of the difference is the smallest given fixed underlying joint probabilities.

Further exploration shows that the maximum differences also can increase with sample sizes. The reason is that the biggest possible difference in p-values between the conditional permutation distribution and the true null distribution depends on the total number of permutations, i.e. \( \binom{N}{n^D} \). Notice that the most extreme case always occurs when the biggest possible test statistic, \( maxT \) is the only observable test statistic for the true table. Therefore, \( P^{true}(T \geq maxT) = 1 \). Meanwhile, in the permuted samples, only one table, i.e. one out of all \( \binom{N}{n^D} \) equally likely permuted samples has \( T = maxT \) (as shown in the numerical example in Table 2.11). Therefore, \( P^{perm}(T \geq maxT) = \left( \frac{N}{n^D} \right)^{-1} \). Thus the difference in p-values between the two distributions is no larger than \( 1 - \left( \frac{N}{n^D} \right)^{-1} \). As sample size increases, this difference can also increase. Figure 2.5 shows when \( n^D = 8, n^N \geq n^D \), the maximum difference in p-values are even bigger compared to Figures 2.3 and 2.4. The checks in Figure 2.5 correspond to the maximum differences in conditional p-values calculated by \( 1 - \left( \frac{N}{n^D} \right)^{-1} \). They match exactly with the results obtained from the numerical search over the parameter space.
Figure 2.3: Maximum difference in p-values between the conditional permutation distribution and the true null distribution when $n^D \leq n^N$, $\pi^D_{++} > \pi^N_{++}$
$1\text{-sided max}T, n^D, \pi^D_{++} < \pi^N_{++}$

Figure 2.4: Maximum difference in p-values between the conditional permutation distribution and the true null distribution when $n^D \leq n^N, \pi^D_{++} < \pi^N_{++}$
Figure 2.5: Maximum difference in p-values between the conditional permutation distribution and the true null distribution when $n_D \leq n_N, \pi_{++}^D > \pi_{++}^N$
Unconditional results

We explore the unconditional permutation distribution and the true reference
distribution under different parameter configurations. It is meaningful since anti-
conservative results in comparison of unconditional permutation reference distribution
and the true null distribution imply that there exists at least one observable margin
where the conditional permutation distribution is also liberal.

Four numerical examples are considered with sample sizes and joint probabili-
ties shown in tables 2.12, 2.13, 2.14 and 2.15. For each example, the exact un-
conditional true null distribution and permutation distribution of the test statistic
\( T = \max_{i=1,2} T_i, T_i = \hat{\pi}_i^D - \hat{\pi}_i^N \) are calculated. Figures 2.6, 2.7, 2.8 and 2.9, respectively, display these distributions.

[1 ] 2-sided \( \max T \) test for \( H_0(12) \) with unequal sample size \( n^N = 2, n^D = 4 \) and
joint probabilities as shown in earlier tables:

\[
\begin{array}{c|cc}
\text{Disease(Phenotype = 1)} & \text{SNP 2} & \\
\text{SNP 1} & + & - \\
+ & 0.2 & 0.05 & 0.25 \\
- & 0.55 & 0.2 & 0.75 \\
& 0.75 & 0.25 & 1 \\
\text{Normal(Phenotype = 0)} & \text{SNP 2} & \\
\text{SNP 1} & + & - \\
+ & 0.05 & 0.2 & 0.25 \\
- & 0.7 & 0.05 & 1 \\
& 0.75 & 0.25 & 1 \\
\end{array}
\]

Table 2.12: Joint probabilities configuration of Figure 2.6.

[2 ] 1-sided \( \max T \) test for \( H_0(12) \) with equal sample size \( n^N = n^D = 3 \):
### Table 2.13: Joint probabilities configuration of Figure 2.7.

3] 1-sided minp test for $H_{0(12)}$ with unequal sample size $n^N = 2, n^D = 4$:

<table>
<thead>
<tr>
<th></th>
<th>SNP 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Disease(Phenotype = 1)</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Normal(Phenotype = 0)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

### Table 2.14: Joint probabilities configuration of Figure 2.8.

4] 1-sided minp test for $H_{0(12)}$ with equal sample size $n^N = n^D = 3$:

<table>
<thead>
<tr>
<th></th>
<th>SNP 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Disease(Phenotype = 1)</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Normal(Phenotype = 0)</td>
<td>0.5</td>
</tr>
<tr>
<td>SNP 1</td>
<td>SNP 2</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 2.15: Joint probabilities configuration of Figure 2.9.

As shown in Figures 2.6, 2.7, 2.8 and 2.9, the rejection region based on the permutation reference distribution may be larger than that of the true null distribution for at least the tail area of the test statistic $T$ or $p$ that correspond to small p-values. That is, $P^{true}(T \geq c) > P^{perm}(T \geq c)$ or $P^{true}(p \leq d) > P^{perm}(p \leq d)$ for at least some $c$ or $d$. As discussed earlier, these numerical examples show permutation tests based partitioning test procedures do not always control FWER.

Also note that compared to the conditional results, the magnitude of the difference between the unconditional permutation distribution and the true null distribution is much smaller. The reason is that the unconditional distributions are obtained by averaging the conditional differences over all possible margins subject to the sample size constraints. For some margins, the permutation test based procedure is conservative, while for others the permutation test is anti-conservative. Averaging over all possible margins, the unconditional difference between the permutation distribution and the true distribution is much smaller.
These examples show that the permutation distribution may not be an appropriate reference distribution for tests of maxT or minp tests of $H^*_{0(1,2)}$, regardless of sample size.

Further exploration of the parameter space shows that the difference between the unconditional distributions in the tail area depends on a few parameters. In the simplified 2 SNPs example, in addition to the joint probabilities $\pi$, the marginal prevalence of the genetic features, i.e. $\pi^D_i = \pi^N_i = \pi_i, i = 1, 2$, as well as the sample sizes $n^D, n^N$, all impact the unconditional differences in distributions.

Notice that for both one-sided and two-sided maxT or minp type of tests, we are able to find anti-conservative numerical examples. Under some parameter configurations, exchanging sample sizes $n^D$ and $n^N$ between groups results in the permutation reference distribution changing from anti-conservative to predominantly conservative. In this case, using the permutation distribution to define cut-off values may conservatively control the error rate at a true level smaller than the nominal level.

The difference between the two distributions tends to shrink with the difference in sample size between the two phenotype groups. However, we emphasize that the two distributions remain different even when the sample sizes for the two phenotypes are equal. Recall that it was shown that distributions for the one-sided test statistics may differ in the odd order cumulants even when the sample size for the two phenotype groups are equal (Huang et al., 2006). For the binary case, the odd order cumulants are also determined by the joint probabilities $\pi$. Thus, as shown in Figure 2.7 and Figure 2.9, the p-values calculated under the true and permutation distributions differ even when the sample sizes are equal.
Of course the difference between the joint probabilities of the two phenotype groups is crucial for the difference between permutation distribution and true distribution of the test statistic. Under the marginal identity $\pi^D_i = \pi^N_i, i = 1, 2$, the difference in the joint probabilities is determined by $\pi^{D+}_+ + \pi^{N+}_+$. If $\pi^{D+}_+ = \pi^{N+}_+$, identity in marginal prevalence implies $F^D = F^N$. If $\pi^{D+}_+ \neq \pi^{N+}_+$, permutation test may or may not result in a valid level-$\alpha$ test for the marginal null hypotheses.

However, violation of the MDJ condition may or may not lead to anti-conservative results. Conservative examples of both maxT test and minp test based on permutation tests can be constructed too. For example, consider the following two numerical examples:

[1 ] 1-sided maxT test for $H_0(12)$ with unequal sample size $n^N = 4, n^D = 2$ and underlying probabilities as shown in Table 2.16.

```
<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>0.15</td>
<td>0.05</td>
</tr>
<tr>
<td>-</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.85</td>
<td>0.15</td>
</tr>
</tbody>
</table>
```

Disease(Phenotype = 1) | SNP 2  
SNP 1 | + | - |
--- | --- | --- |
+ | 0.15 | 0.05 | 0.2 |
- | 0.7 | 0.1 | 0.8 |
| 0.85 | 0.15 | 1 |

Normal(Phenotype = 0) | SNP 2  
SNP 1 | + | - |
--- | --- | --- |
+ | 0.2 | 0 | 0.2 |
- | 0.65 | 0.15 | 0.8 |
| 0.85 | 0.15 | 1 |

Table 2.16: Joint probabilities configuration of Figure 2.10.

[2 ] 1-sided minp test for $H_0(12)$ with equal sample size $n^N = 2, n^D = 3$ and underlying probabilities as shown in Table 2.17.

83
As shown in Figures 2.10 and 2.11, the rejection region based on the permutation reference distribution in these two examples is not larger than that of the true null distribution, for at least the tail areas of the test statistics \( \text{max} T \) or \( \text{min} p \) corresponding to small \( p \)-values. That is, \( P^{\text{true}}(T \geq c) \leq P^{\text{perm}}(T \geq c) \) or \( P^{\text{true}}(p \leq d) \leq P^{\text{perm}}(p \leq d) \) for at least some \( c \) or \( d \). These numerical examples indicate that permutation test based partitioning test procedures may control FWER.

We further explore the magnitude of difference in the unconditional permutation distribution and true distribution of the test statistics in the tail area corresponding to small \( p \)-value under the marginal null hypothesis \( \pi^D_i = \pi^N_i = \pi_i, i = 1, 2 \). Within the parameter space of \( \pi_1 \in [0, 1], \pi_2 \in [0, 1] \), the same strategy was used: we search over a grid within the parameter space under the worst-case scenario, where the difference in \( \pi^D_{++} \) and \( \pi^N_{++} \) is maximized, i.e. \( \pi^D_{++} \) and \( \pi^N_{++} \) are fixed at their boundary values \( \min(\pi_1, \pi_2) \) or \( \max(0, \pi_1 + \pi_2 - 1) \).

We explored the parameter space for different sample size combinations of both 1-sided and 2-sided \( \text{max} T \) and \( \text{min} p \) tests. For example, Figure 2.12 shows that for the two-sided \( \text{max} T \) test with sample sizes \( n^D = 2 \) and \( n^N = 5 \), \( \pi^D_{++} = \max(0, \pi_1 + \pi_2 - 1) \).
and $\pi_{++}^N = \min(\pi_1, \pi_2)$, it is possible to find positive differences between p-values based on the true null distribution and those based on the permutation distribution, i.e. $P^{\text{true}}(T \geq t) > P^{\text{perm}}(T \geq t)$, for both the $\pi_1 + \pi_2 > 1$ and the $\pi_1 + \pi_2 < 1$ cases. Therefore, the permutation test based multiple testing procedures could be liberal if $\pi_1 + \pi_2 \neq 1$ since the p-value based on a permutation test could be smaller than that of the true distribution for the same test statistic value $t$. That is, If one rejects the null hypothesis based on the critical value $c$ from the permutation distribution, i.e. reject if $T > c$ where $P^{\text{perm}}(T \geq c) = \alpha$, then the true significance level of the test could be $P^{\text{true}}(T \geq c) > \alpha$.

However, for some parameter configurations we may not be able to find examples where permutation tests are liberal. One example is that when two-sided maxT or minp tests are used, in the case where the sample sizes are equal, i.e. $n^D = n^N$. Figure 2.13 shows one such example where $n^D = n^N = 3$ and a two-sided maxT test is used. Also, notice that these plots only show the differences in the smallest p-values between the permutation distribution and the true distribution. Not being able to find anti-conservative examples in the smallest p-values does not necessarily indicate not being able to find such examples for other cases, for example, between the second smallest p-values.

Numerical exploration cannot exhaust every possible value within the parameter space. Confirmatory conclusions can be derived only based on the theoretical distribution of the test statistics.
To find whether anti-conservative numerical examples exist for permutation test under different parameter configurations, we explored within the parameter space $\pi_1 \in [0, 1], \pi_2 \in [0, 1]$ in the worst case scenario, where the difference between $\pi_{++}^D, \pi_{++}^N$ is maximized. The following tables show the results, where ‘Yes’ indicates we were able to find anti-conservative examples and ‘No’ indicates we were not able to find anti-conservative examples of permutation tests. Three different sample size configurations were considered.

1. Unequal sample size: $n^D < n^N$

<table>
<thead>
<tr>
<th>Test</th>
<th>$\pi_1 + \pi_2$</th>
<th>$\pi_{++}^D &lt; \pi_{++}^N$</th>
<th>$\pi_{++}^D &gt; \pi_{++}^N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-sided</td>
<td>$&lt; 1$</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>$= 1$</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>$&gt; 1$</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2-sided</td>
<td>$&lt; 1$</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>$= 1$</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>$&gt; 1$</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

2. Equal sample size: $n^D = n^N$

<table>
<thead>
<tr>
<th>Test</th>
<th>$\pi_1 + \pi_2$</th>
<th>$\pi_{++}^D &lt; \pi_{++}^N$</th>
<th>$\pi_{++}^D &gt; \pi_{++}^N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-sided</td>
<td>$&lt; 1$</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>$= 1$</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>$&gt; 1$</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2-sided</td>
<td>$&lt; 1$</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>$= 1$</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>$&gt; 1$</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

3. Unequal sample size: $n^D > n^N$
2.4 Discussion

As shown in Theorem 2.2 and Corollary 2.3 by Huang et al. (Huang et al., 2006), the cumulants of the permutation distribution of $T = \bar{X} - \bar{Y}$ and those of the true null distribution will be different in general, regardless of the sample size and the dimensions of $X$ and $Y$, unless $F_X$ and $F_Y$ have the same cumulants.

With the simplified GWAS example of 2 SNPs, we showed that even when the null hypotheses are all true, the difference between the permutation distribution and the true null distribution could lead to anti-conservative multiple testing procedures for both maxT and minp types of tests. In the analysis of data with higher dimension, the difference between the two distributions still exists. That is, in the cases where the data has more than two dimensions, multiple testing based on the particular permutation test could still be anti-conservative due to similar reasons.

Many multiple testing procedures are developed based on the partitioning principle. For example, both Holm’s step-down method and Hochberg’s step up method are special short-cut versions of partitioning tests. Our result indicates that caution needs to be executed when using simple permutation tests in partitioning principle-based testing procedures. Closed testing procedures can be considered as one special case of partitioning tests. Thus, permutation tests in conjunction with procedures based on the closed testing principle are also not recommended.
Generalized FWER (gFWER) is a generalization of the FWER concept to the probability of making at least $m$ false rejections under any parameter configurations, where $m > 0$ is a pre-specified small positive integer. Since the procedure of controlling for gFWER (Xu and Hsu, 2007) is constructed based on the partitioning principle, permutation tests may lead to liberal gFWER controlling procedures as well. FDR equals FWER when all null hypotheses are true. Our results indicate using permutation tests to control FDR may have similar limitations as well.

The MDJ condition is the sufficient condition for permutation tests to be level-$\alpha$. It is a condition based on the observations. Therefore, one needs to check whether the MDJ condition holds at the data level. For example, biological verification is necessary. In fact, it is quite common that a disease or phenotype is determined by epistatic effects of multiple genes or SNPs, as in the BBS example mentioned earlier.

In GWAS, since the number of genetic features is usually huge, SNPs are closely located with relatively high density. Under such circumstances, it is probably not reasonable to assume the SNPs having exactly the same association structure between phenotype groups. Thus our concern of using permutation tests is justified.

Although our motivating example is within the GWAS setting, we caution against using simple permutation tests in testing multiple null hypotheses in general. For example, in a drug safety study, where the prevalence of dozens of adverse effects are of comparison interest between the treatment group and the control group, since relevant adverse effects could be associated with each other, the MDJ condition possibly does not hold. Violation of the MDJ condition may lead to lack of control for FWER if permutation tests are used in this case.
If observations are in the continuous setting, such as gene expression levels in a microarray study, modeling each gene separately is possible. If diagnostic analysis of the linear models reveals that the error terms can be reasonably assumed to be independent and identically distributed, then it was shown that using permutation tests in testing multiple hypotheses can actually control FWER (Calian et al., 2008). But the test statistics must be in the form of the maximum of least square estimates of parameters.

However, there is no corresponding concept of ‘error terms’ in the discrete data setting. The resampling within group method (Pollard and van der Lann, 2004) may be one choice. But, this method does not have strong control of FWER, i.e. the asymptotic control of false rejection rate depends on the sample size.

Our results also apply to adjusting for multiplicity for binary outcome variables while controlling for other variables of interest. In this case, logistic regression is a useful tool. Step-down permutation within strata in the logistic regression setting has been proposed (Troendle, 2005). However, the strong control of the FWER relies on the assumption of exchangeability of observations across treatment groups. Therefore, only if exchangeability is implied under appropriate null hypotheses, permutation test is valid.

Troendle also proposed a step-down multiple hypotheses testing procedure based on permutation test (Troendle, 1995). In section 3 on page 371, the author put assumptions on the data such that for $X_1, \ldots, X_N$ be independent $k-$dimensional observations from the control group. And they follow a common joint distribution $X_i \sim F(\mu_1; x), i = 1, \ldots, N$. Similarly, the observations from the treatment group follows another common joint distribution $X_j \sim F(\mu_2; y), j = N + 1, \ldots, 2N$. However,
although observations within the same treatment group follow a common distribution, no further assumption was made on the relationship between these two joint distributions.

Based on our result, the multiple testing procedure for testing $H_{0i} : \mu_{1i} = \mu_{2i}$ for $i = 1, \ldots, k$ constructed based on such assumptions may or may not control FWER, since there is no implication from the assumption in the paper that once $\mu_1 = \mu_2$, identity between $x$ and $y$ is established. However, all of the numerical simulations used to show control of FWER by the proposed method assume $x$ and $y$, in which case, permutation test based multiple testing procedure should be valid.

Good cited the procedure proposed by Troendle (Good, 2006). However, he makes explicit emphasis on that the permutation test is constructed based on the ‘exchangeability’ assumption (page 70), which is not necessarily the same condition as specified in Troendle’s paper.

In the case where the MDJ condition is not satisfied or hard to verify, the fast, simple and guaranteed to be conservative way of strong control of FWER is Holm’s method. Appropriate statistical tests for each SNP will be applied separately. Notice that, for a single SNP, the null hypothesis of marginal identity, i.e. $\pi_{iD} = \pi_{iN}$ implies distributional identity. Thus, using permutation tests on single SNP, such as Fisher’s exact test for $2 \times 2$ table of a single SNP is valid. Then Holm’s method based on the Bonferroni’s inequality can be applied to the resulting test statistics. Thus FWER is controlled strongly no matter what the actual data structure is. However, this method is also over conservative if some of the genetic features are positively correlated.
Developing new techniques that can either model with large numbers of parameters or check for violation of the MDJ condition could be useful research in the future.
Figure 2.6: True null distribution and permutation null distribution of $T = \max_{i=1,2} T_i; T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$ for unequal sample sizes 2-sided test.
One-sided maxT test

$Pr(\text{max} T \geq t)$

$n^D = 3, n^N = 3$

$-1 -0.667 -0.333 0 0.333 0.667 1$

0.013 0.142 0.542 0.901 0.993 1.000

True null distribution

Permutation distribution

Figure 2.7: True null distribution and permutation null distribution of $T = \max_{i=1,2} T_i; T_i = \hat{\pi}^D_i - \hat{\pi}^N_i$ for equal sample sizes 1-sided test.
Figure 2.8: True null distribution and permutation null distribution of $p = \min_{i=1,2} P_i$ for unequal sample sizes 1-sided test.
Figure 2.9: True null distribution and permutation null distribution of $p = \min_{i=1,2} p_i$ for equal sample sizes 1-sided test.
Figure 2.10: True null distribution and permutation null distribution of $T = \max_{i=1,2} T_i; T_i = \hat{\pi}_i^D - \hat{\pi}_i^N$ for unequal sample sizes 1-sided test.
Figure 2.11: True null distribution and permutation null distribution of $p = \min_{i=1,2} p_i$ for unequal sample sizes 1-sided test.
Figure 2.12: Difference between smallest p-values of true null distribution and permutation null distribution of $T = \max_{i=1,2} T_i$ for unequal sample sizes 2-sided test.
Figure 2.13: Difference between smallest p-values of true null distribution and permutation null distribution of $T = \max_{i=1,2} T_i$ for equal sample sizes 2-sided test.
CHAPTER 3

INDEPENDENT BINARY OUTCOMES: LOGISTIC REGRESSION

Binary outcome variables only have two possible values, such as ‘Yes’ versus ‘No’, or ‘female’ versus ‘male’, etc. Traditionally, ‘1’ is used to indicate ‘success’ - the event of interest has happened, while ‘0’ indicates it has not happened. For example, in a case-control study, ‘1’ could represent a case while ‘0’ represents a control. In a randomized clinical trial, while ‘1’ stands for ‘efficacious effect’, ‘0’ could stand for ‘no effect’. It also applies to the case where a continuous response variable is dichotomized.

For a single binary observation $Y_i$,

$$Y_i \sim Bernoulli(\pi), \ 0 \leq \pi \leq 1 \quad (3.1)$$

For $Y_i, i = 1, \ldots, n$, the total number of ‘success’ follows a Binomial distribution $B(n, \pi)$.

3.1 Simpler case for binary outcome

3.1.1 $2 \times 2$ contingency table and Exact Conditional tests

One common problem for a binary outcome $Y$ is to test if another binary variable $X$ is significantly associated with it, so that the probability $\pi = P(Y_i = 1)$ is different
according to values of $X$. A $2 \times 2$ contingency table can be used to summarize the frequencies of the combinations of $X$ and $Y$ values as in Table 3.1.1.

<table>
<thead>
<tr>
<th>X/Y</th>
<th>Y=1</th>
<th>Y=0</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>X=0</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Column Total</td>
<td>$n_{1}$</td>
<td>$n_{2}$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Let the underlying probabilities for the $2 \times 2$ contingency table be as shown in Table 3.1.1:

<table>
<thead>
<tr>
<th>X/Y</th>
<th>Y=1</th>
<th>Y=0</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=1</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$\pi_1$</td>
</tr>
<tr>
<td>X=0</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$\pi_2$</td>
</tr>
<tr>
<td>Column Total</td>
<td>$\pi_{1}$</td>
<td>$\pi_{2}$</td>
<td>1</td>
</tr>
</tbody>
</table>

To test the null hypothesis that the value of $X$ has no impact on the value of $Y$ is equivalent to testing independence between $X$ and $Y$. Under the null hypothesis of independence:

$$H_0 : \pi_{ij} = \pi_i \times \pi_j; i = 1, 2, j = 1, 2.$$ (3.2)

Conditioning on the total number of observations $N$, row totals $n_{1.}, n_{2.}$, and columns totals $n_{1.}, n_{2.}$, the expected frequencies $f_{ij}; i = 1, 2; j = 1, 2$ under the null hypothesis satisfy

$$f_{ij} = n_{i.} \times n_{.j}/N = N\pi_i\pi_j$$

$$\sum_{j=1}^2 f_{ij} = n_{i.}, i = 1, 2; \sum_{i=1}^2 f_{ij} = n_{.j}, j = 1, 2$$ (3.3)
Notice that if ‘success’ is indicated by $Y = 1$, then the number of successes in the $X = 1$ and $X = 0$ group can be deemed as two independent binomials, i.e. $n_{11} \sim B(n_1, P(Y = 1|X = 1)), n_{21} \sim B(n_2, P(Y = 1|X = 0))$. Independence between $X$ and $Y$ is equivalent to the statement that the probabilities of success in the $X = 1$ and $X = 0$ group are equivalent:

$$H_0 : P(Y = 1|X = 1) = P(Y = 1|X = 0)$$  \hspace{1cm} (3.4)

That is,

$$H_0 : \frac{\pi_{11}}{\pi_1} = \frac{\pi_{21}}{\pi_2}. \hspace{1cm} (3.5)$$

(3.4) also implies that the ‘odds ratio’ of success in the two $X$ categories is 1, where odds in the two categories are defined as $P(Y = 1|X = 1)/P(Y = 0|X = 1)$ and $P(Y = 1|X = 0)/P(Y = 0|X = 0)$, respectively. That is,

$$H_0 : \phi = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1 \hspace{1cm} (3.6)$$

To test the null hypothesis that two dichotomous variables $X$ and $Y$ are independent, conditional exact tests based on different statistics can be employed. The exact p-value of an observed table is computed according to the order of all possible tables that satisfy the marginal constraints based on the statistics. Therefore different methods could order the same table differently.

For example, Freeman’s method calculates the p-values based on hypergeometric probabilities of all tables subject to the marginal constraints by ordering them according to their hypergeometric probabilities (Freeman and Halton, 1951).
Other conditional exact methods include the exact Pearson’s chi-square tests and the exact likelihood ratio tests.

The advantage of conditional exact approaches is that since they are exact, they are not subject to losing information due to large sample approximation or combining small-frequency cells. When the sample size is large, exact calculations might be too time consuming to perform. When the expected cell frequencies $f_{ij}$ are not too small, asymptotic method such as the chi-square method can be readily applied.

### 3.2 Logistic Regression

Because the normal distribution assumption is not appropriate, the simple linear regression technique, i.e. $Y_i = X_i \beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$, cannot be applied to binary outcomes.

Logistic regression is a widely used tool in binary data regression analysis. Let

$$Y_{n \times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$  \hspace{1cm} (3.7)

For the $i$th subject, $Y_i \sim Bernoulli(\pi(x_i)), i = 1, \ldots, n$, with $\pi(x_i) = Pr(Y_i = 1|x_i) = E[Y_i|x_i]$, where $X$ is the design matrix,

$$X_{n \times k} = \begin{pmatrix} X'_1 \\ \vdots \\ X'_{n_i} \end{pmatrix}, X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ik} \end{pmatrix}$$  \hspace{1cm} (3.8)

And $\beta$ is the vector of parameters

$$\beta'_{k \times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$  \hspace{1cm} (3.9)
Logistic regression models the conditional log odds as a linear function of the independent variables:

\[
\log \frac{\pi_i(x_i)}{1 - \pi_i(x_i)} = x_i \beta
\] (3.10)

Since the linear predictor predicts the logit of expectations, the interpretation of the parameters is in the ‘logit’ scale. That is, \( \beta_i \) estimates the change in the ‘log odds’ of experiencing the event of interest that results from every one unit change in \( X_i \).

\[
\log \frac{P(Y_i = 1)/P(Y_i = 0)|X_i = x_i + 1}{P(Y_i = 1)/P(Y_i = 0)|X_i = x_i} = \beta_i
\] (3.11)

The joint likelihood function of \( Y \) is

\[
L(\beta) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, i = 1, \ldots, n.
\] (3.12)

Numerical optimization methods can be used to obtain the MLEs for the coefficients. With large sample size, under mild conditions, the MLE asymptotic theory guarantees MLEs of \( \beta \) are asymptotically normally distributed and have asymptotic variance-covariance \( \Sigma = I^{-1} \), where \( I \) is the information matrix. That is,

\[
\sqrt{n}(\hat{\beta} - \beta) \rightarrow_D MVN(0, \Sigma)
\] (3.13)

where \( N(\cdot) \) denotes cumulative distribution function of the normal distribution.

The Wald test or likelihood ratio test can be utilized to test null hypotheses.

### 3.3 Applying Partition Testing to Logistic Regression Model

As far as I am aware, few multiple testing procedures have been developed in major statistical software packages specifically for the logistic regression model setting.
Simple multiplicity adjustment methods such as Bonferroni’s method and Holm’s method can be utilized in multiple testing for p-values from logistic regressions.

Dasgupta et al. proposed a Multiple Comparison with the Control testing procedure in logistic regression setting based on the likelihood ratio test (Dasgupta et al., 2000). The paper considers testing \( H_{01,\ldots,k} : \beta_1 = \ldots = \beta_k = 0 \) versus \( H_{a1,\ldots,k} : \) at least one inequality, where \( \beta_i, i = 1, \ldots, k \) are parameters from the logistic regression model. The test compares the likelihood ratio statistic \( L_{01,\ldots,k} = -2\ln(\Lambda_{01,\ldots,k}) \) to a chi-square distribution with \( k \) degrees of freedom.

The test is constructed based on the Closure Testing Principle (Marcus et al., 1976). Only if all null hypotheses \( H_{0i}, i \in I \) are rejected, \( H_{0i} \) can be rejected.

Therefore, at most \( 2^k - 1 \) tests are to be performed to draw valid inference for \( k \) parameters. Obviously, as \( k \) increases, number of hypotheses to be tested will be too large.

If the independent variable is discrete, observations can be grouped by strata. Resampling within each strata without replacement to adjust for multiplicity in testing treatment effects in logistic regression models (Troendle, 2005) requires the very strong assumption of equal joint distributions across groups (for example, equal-correlations were used in all simulations in section 5 on page 3588).

Bootstrapping the whole data vector consists of both the outcome variable and the independent variables may control the multiple testing error rates asymptotically.

We show that in Chapter 2, when the rather strong ‘exchangeable’ assumption is not guaranteed, the permutation based resampling method cannot control multiple error rates. Also, whether the assumption is satisfied may not be obvious.
In this section, we propose applying the partitioning principle for multiple testing in the logistic regression model setting and further show how to construct a step-down short-cut version of the partitioning test procedure once certain sufficient conditions are satisfied. This can provide asymptotic control of FWER when exact resampling method such as permutation test is not valid. This will be illustrated through a motivating example on multiple comparisons with the control problem.

3.3.1 Motivating Example

The motivating example is an experiment designed to study if the treatment mint patty, mint oil or oil patty are effective in controlling tracheal mites infection in honey bees. The data is in Table II (Dasgupta et al., 2000). Specifically, four colonies of bees were collected and counted for infection with tracheal mites at time 0. Treatments mint patty, mint oil or oil patty were randomly assigned to three colonies while the last colony was kept untreated. At time points 1.5, 3.5, 8.5 and 10 months, thirty bees from each colony were examined for infection status.

The research question is whether the three treated colonies had similar infection rates compared to the untreated standard.

3.3.2 Problem Formulation

Let \( y_{ij} = 1 \) or \( 0, i = 0, \ldots, k, j = 1, \ldots, n \), indicating the binary outcome observed from the \( j \)th bee from the \( i \)th treatment group, each containing \( n \) independent subjects. For the \( j \)th subject of the \( i \)th dose group, define:

\[
Y_{ij} = \begin{cases} 
1 & \text{infection} \\
0 & \text{no infection}
\end{cases}
\]  

(3.14)

In general subject numbers could differ among groups.
Let $i = 0$ indicate the untreated control group to be compared with. For the $i$th group, the outcome vector is a $n \times 1$ column vector:

\[
Y_i = \begin{pmatrix}
Y_{i1} \\
Y_{i2} \\
\vdots \\
Y_{in}
\end{pmatrix}
\]  

(3.15)

Let $\pi_i$ be the probability of infection in group $i$, i.e. $\Pr(Y_{ij} = 1) = \pi_i$, for $j = 1, \ldots, n$. Assume group 0 is the ‘control’ and the remaining $k$ groups are the ‘treatments’. A typical two-sided multiple comparisons with the control problem is to compare $\pi_i, i = 1, \ldots, k$ of each treatment group to $\pi_0$, i.e.

\[
H_{0i} : \pi_i = \pi_0 \text{ vs. } H_{ai} : \pi_i \neq \pi_0; i = 1, \ldots, k
\]  

(3.16)

In the following sections, we will discuss how to test them with a valid and efficient multiple testing procedure based on the partitioning principle.

3.3.3 The ‘Comprehensive’ Model

A logistic regression model models the conditional logit transformation of the probability of a dichotomous variable, i.e. the probability of experiencing a binary event, through a linear function of independent variables as shown in (3.10).

For the multiple comparisons with a control problem, the design matrix $X$ could be formed in such a way so that for $i = 0$, each $x_{ij}$ is a $(k + 1) \times 1$ vector with the first element being equal to 1 and all the rest elements being 0, i.e.
For other treatment groups where \( i = 1, \ldots, k \), along with the first element, the \((i + 1)\)th element of \( x_{ij} \) is also equal to 1, for \( j = 1, \ldots, n \). For example,

\[
X_2 = \begin{pmatrix}
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 0 & 1 & 0 \\
\end{pmatrix}
\] (3.18)

Therefore, by (3.10), we have

\[
\begin{align*}
\pi_i(X_0) &= P(Y_{ij} = 1|X_0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}; \quad i = 0 \\
\pi_i(X_i) &= P(Y_{ij} = 1|X_i) = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)}; \quad i = 1, \ldots, k
\end{align*}
\] (3.19)

We refer to this model as the ‘comprehensive’ logistic regression model in the rest of the chapter since it estimates the effects of all treatments in a single model.

In general, \( \beta_i \) estimates the effects caused by unit change in \( x_{ij} \) on \( \pi_i \) in the logit scale. Thus \( \beta_i \) estimates the additional ‘log odds’ of infection in group \( i, i = 1, \ldots, k \) compared to those in the control group \((i = 0)\), i.e.

\[
\beta_i = \log \frac{\pi_i/(1 - \pi_i)}{\pi_0/(1 - \pi_0)} = \log \frac{\pi_i}{1 - \pi_i} - \log \frac{\pi_0}{1 - \pi_0}
\] (3.20)

As shown in (3.20), by fitting the ‘comprehensive’ logistic regression model, it follows that \( \pi_i = \pi_0 \), if and only if \( \beta_i = 0 \). Also, \( \pi_i - \pi_0 \) monotonically increase with \( \beta_i \) since the first derivative of \( \pi_i - \pi_0 \) with respect to \( \beta_i \) is positive. This is simply because
\[ \pi_i - \pi_0 = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \] (3.21)

Thus,

\[ \frac{\partial(\pi_i - \pi_0)}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \left\{ \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right\} = \frac{\exp(\beta_0 + \beta_i)}{(1 + \exp(\beta_0 + \beta_i))^2} > 0 \] (3.22)

strictly for finite \( \beta \).

Therefore testing \( \pi_i - \pi_0 = 0 \) vs. \( \pi_i - \pi_0 \neq 0, \; i = 1, \ldots, k \) is equivalent to testing

\[ H_{0i} : \beta_i = 0 \; \text{vs.} \; H_{ai} : \beta_i \neq 0, \; i = 1, \ldots, k \] (3.23)

where \(-\infty < \beta_i < \infty\).

That is, by testing the null hypotheses of the ‘comprehensive’ logistic regression model parameters \( \beta_i, \; i = 1, \ldots, k \), one could make inference on the differences in treatment effects between \( k \) treatment groups and the control group.

It is well known that the probability of making false positive findings by chance increases with the number of tests. Relatively simpler multiplicity adjustment methods can be used, such as the Bonferroni adjustment, which ignores the fact that in comparing each dose group with the same control, the test statistics are positively correlated. Hence Bonferroni method is too conservative in this case.

If the partitioning principle (Stefansson et al., 1988; Finner and Strassburger, 2002) or the closed testing principle (Marcus et al., 1976) is used to test for \( k \) null hypotheses simultaneously, the total number of tests needed is at most \( 2^k - 1 \), which increases fast with \( k \). If no short-cut version can be applied, it will be too time consuming to test all \( 2^k - 1 \) hypotheses even for medium sized \( k \).
To construct a multiple testing procedure that properly controls FWER, one needs to define appropriate test statistics and find critical values according to their reference distribution under the null hypothesis. One important and familiar technique for constructing a reference distribution is to model the data. Through modeling, the joint distribution of the test statistics can sometimes be derived analytically, so that computation of the critical value $c_f$ could be also readily available. The ‘comprehensive’ logistic regression model is such an example.

### 3.3.4 The Test Statistics and Critical Values

In the ‘comprehensive’ logistic regression model, testing $\pi_i - \pi_0 = 0, i = 1, \ldots, k$ is equivalent to testing $\beta_i = 0$ for $i = 1, \ldots, k$. One way to form test statistics is to utilize the asymptotic distributional property of the MLE $\hat{\beta}_i$ and its asymptotic standard deviation $\hat{\sigma}_i$.

Asymptotically the MLE $\hat{\beta}$ from the logistic regression model follows a multivariate normal distribution.

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_D \text{MVN}(0, \Sigma). \quad (3.24)$$

By Slutsky’s theorem,

$$\text{diag}(\hat{\Sigma}^{-\frac{1}{2}})\sqrt{n}(\hat{\beta} - \beta) \rightarrow_D \text{MVN}(0, R). \quad (3.25)$$

where $\Sigma$ is the variance-covariance matrix of $\hat{\beta}$ and $R$ is the correlation matrix of $\hat{\beta}$. This asymptotic joint distribution can be used for testing null hypotheses on $\beta$.

Based on the above asymptotic distributional results, we define the test statistic for testing each two-sided single null hypothesis $\beta_i = 0$ vs. $\beta_i \neq 0$ as:
The test statistics can be readily obtained from the ‘comprehensive’ logistic regression model output, where $\hat{\sigma}_i$ is the square root of the $i$th element of the diagonal of the estimated $(k+1) \times (k+1)$ variance-covariance matrix $\hat{\Sigma}$.

Under the null hypothesis $\beta_i = 0, i = 1, \ldots, k$, the vector $S = (S_1, \ldots, S_k)$ follows a multivariate normal distribution with mean vector $0$ and covariance matrix which equals the correlation matrix of $\hat{\beta}$, i.e $R$.

Following the partitioning principle, and to avoid the possibly large number of subset null hypotheses testings, we choose to test each null hypothesis subset $H_{0I} : \beta_i = 0, i \in I, I \subseteq \{1, \ldots, k\}$ by using test statistics in the ‘maxT’ form. This way we may be able to construct a ‘short-cut’ version of the partition testing procedure. In addition to the ‘maxT’ type of test statistic, the ‘minp’ test statistic has been shown to be able to short cut as well. Here we illustrate the proposed procedure in terms of ‘maxT’ type of test statistic:

$$\text{Reject } H_{0I} : \beta_i = 0, i \in I \text{ if } \max_{i \in I} |S_i| > c_{\alpha,I}$$  \hspace{1cm} (3.27)

where $c_{\alpha,I}$ is the proper critical value according to the error rate $\alpha$ and subset $I$.

The p-value of the test for the two-sided null hypothesis $H_{0\{1,\ldots,k\}} : \beta_1 = 0$ and $\beta_2 = 0$ and $\ldots$ and $\beta_k = 0$ is obtained by:

$$\text{p-value} = Pr(\max|S_i| \geq s_i)$$

$$= 1 - Pr(|S_1| < s_1, |S_2| < s_2, \ldots, |S_k| < s_k)$$

$$= 1 - \int_{-s_k}^{s_k} \ldots \int_{-s_1}^{s_1} f(S_1, \ldots, S_k) dS_1 \ldots dS_k$$  \hspace{1cm} (3.28)
The critical value $c_{\alpha,I}$ for testing null hypothesis subset $I$ can be obtained from multivariate normal distribution with dimension $|I|$, where $|I|$ indicates the number of elements in $I$. For example, the critical value $c_{\alpha,\{1,\ldots,k\}}$ satisfies the equation:

$$1 - \int_{-c_{\alpha}}^{c_{\alpha}} \cdots \int_{-c_{\alpha}}^{c_{\alpha}} f(S_1, \ldots, S_k) dS_1 \cdots dS_k = \alpha$$

(3.29)

where $f(S_1, \ldots, S_k)$ is the multivariate normal joint distribution $MVN(0, R)$ under the null hypothesis. $R$ needs to be further estimated by $\hat{R}$ in practice.

### 3.3.5 Apply Partition Testing in Logistic Regression Model

Based on the famous Partitioning Principle (?), one could construct a multiple testing procedure that controls FWER strongly.

To apply the partition principle in testing $H_0^i: \beta_i = 0$, $i = 1, \ldots, k$ in the logistic regression ‘comprehensive’ model setting, proceed as follows:

P1: Partition null parameter space $\mathbb{B}_1 \times \cdots \times \mathbb{B}_k$ into $2^k$ disjoint subspaces and form subset hypotheses for each subspace:

$$H^*_{0(1,\ldots,k)}: \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \ldots \text{ and } \beta_k = 0$$

$$H^*_{0(1,\ldots,k-2,k)}: \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \ldots \text{ and } \beta_{k-2} = 0 \text{ and } \beta_{k-1} \neq 0 \text{ and } \beta_k = 0$$

$$\vdots$$

$$H^*_{0(1,k)}: \beta_1 = 0 \text{ and } \beta_k = 0 \text{ and } \beta_3 \neq 0 \ldots \text{ and } \beta_{k-1} \neq 0$$

$$\vdots$$
\[ H_{0(1,2)}^* : \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \beta_3 \neq 0 \ldots \text{ and } \beta_k \neq 0 \]

\[ H_{0(k)}^* : \beta_1 \neq 0 \ldots \text{ and } \beta_{k-1} \neq 0 \text{ and } \beta_k = 0 \]

\[ \vdots \]

\[ H_{0(1)}^* : \beta_1 = 0 \text{ and } \beta_2 \neq 0 \ldots \text{ and } \beta_k \neq 0 \]

P2: Test each of above null hypothesis at level-\( \alpha \). There are at most \( 2^k - 1 \) hypotheses to be tested.

P3: For each \( i, i = 1, \ldots, k \), infer \( \beta_i \neq 0 \) if and only if all \( H_{0I}^* \) with \( i \in I, I \subseteq \{1, \ldots, k\} \) are rejected.

There are various ways of testing each subset null hypothesis at level-\( \alpha \). Specifically, we choose to compute the statistics \( |S_i| = \frac{|\hat{\beta}_i - 0|}{\hat{\sigma}_i}, i \in \{1, \ldots, k\} \) and use the \( \max T \) type of test statistics. We first order \( |S_i| \) so that \( |S_{i[1]}| \leq \ldots \leq |S_{i[k]}| \). The \( \max T \) type of test is in the form of rejecting each \( H_{0I}^* \) if \( \max_{i \in I} |S_i| > c_{\alpha, I} \).

If all intersection null hypotheses \( H_{0I, i \in I}^* \) which imply \( H_{0i}^* \) are rejected, then the null hypothesis \( H_{0i}^* : \beta_i = 0 \) is rejected. At most \( 2^k - 1 \) subset hypotheses are required to be tested for making inference of \( k \) parameters.

If the critical value obtained from the joint distribution of all \( k \) test statistics, i.e. \( c_{\alpha, \{1, \ldots, k\}} \) as in (3.29) is used, it is the largest critical value corresponding to testing all \( k \) parameters. One can test and reject each \( H_{0i}^* : \beta_i = 0 \) if the corresponding test statistic \( |S_i| > c_{\alpha, \{1, \ldots, k\}} \). This is a single-step multiple comparison with the control procedure that extends Dunnett’s method in the logistic regression model setting.
3.3.6 Step-down Short-cut of Partition Test in Logistic Regression

Notice that to test the original null hypothesis \( H_{0i}, i = 1, \ldots, k \) for \( k \) parameters, up to \( 2^k - 1 \) subset hypotheses may be required by the partitioning principle. Obviously this is not feasible if the number \( k \) is big. Now we check if multiple testing on the parameter estimates from the ‘comprehensive’ logistic regression model satisfies the sufficient conditions 1–4 introduced in Chapter 1 for a shortcut version of the partition testing. Upon satisfaction of the conditions, we further show that a valid step-down multiple comparison procedure can be constructed.

1: The test for \( H_{0i}^* \) is of the ‘maxT’ form.

Check: To test \( H_{0i} : \beta_i = 0, i \in I \), the statistics \( |S_i| = \frac{|\hat{\beta}_i - 0|}{\hat{\sigma}_i}, i \in I \) are ordered so that \( |S_{[1]}| \leq \ldots \leq |S_{[l]}| \). The test is in the form of rejecting \( H_{0i} \) if \( \max_{i \in I} |S_i| > c_{\alpha,I} \). Therefore, the test statistic is in the ‘maxT’ form.

2: \( \sup_{H_{0i}^*} P\{\max_{i \in I} S_i > c_I\} \leq \alpha \).

Check: The critical value for rejecting \( H_{0i} : \beta_i = 0, i \in I \) if \( \max_{i \in I} |S_i| > c_{\alpha,I} \) satisfies the equation

\[
1 - \int_{-c_{\alpha,I}}^{c_{\alpha,I}} \ldots \int_{-c_{\alpha,I}}^{c_{\alpha,I}} f(S_1, \ldots, S_p) dS_1 \ldots dS_p = \alpha, I = \{1, \ldots, p\} \subseteq \{1, \ldots, k\}
\]

According to the MLE asymptotic theory and Slutsky’s theorem, with large sample size, parameter estimates by maximum likelihood in logistic regression models are approximately normally distributed.
\[ \sqrt{n} \text{diag}(\Sigma^{-\frac{1}{2}})(\hat{\beta} - \beta) \rightarrow_D \text{MVN}(0, R) \quad (3.30) \]

Therefore, \( \sup_{\beta_i = 0, i \in I} P\{\max_{i \in I}|S_i| > c, i\} \leq \alpha \) as long as the asymptotic result holds. The false rejection rate is controlled at level \( \alpha \) for each subset hypothesis test. By the partitioning principle, the overall FWER is controlled.

3: The values of the test statistics \( S_i, i = 1, \ldots, k \), are not re-computed for different \( H_{0I} \).

Check: Since only one model, the so called ‘comprehensive’ logistic regression model containing all treatment groups as the independent variables is fit, parameter estimates \( \hat{\beta} \) and \( \hat{V} \) will not change. Therefore statistics \( S_i = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_i} \) as well as the test statistic \( \max_{i \in I}|S_i| \) for testing each subset null hypothesis \( H_{0I} : \beta_i = 0, i \in I \) will not change with \( I \) either.

For example, regardless of whether \( H_{0\{23\}} : \{\beta_2 = 0 \text{ and } \beta_3 = 0\} \) or \( H_{0\{235\}} : \{\beta_2 = 0 \text{ and } \beta_3 = 0 \text{ and } \beta_5 = 0\} \) is tested, the test statistics used in based on \( |S_2| = \frac{|\hat{\beta}_2 - 0|}{\hat{\sigma}_2}, |S_3| = \frac{|\hat{\beta}_3 - 0|}{\hat{\sigma}_3} \) and \( |S_5| = \frac{|\hat{\beta}_5 - 0|}{\hat{\sigma}_5} \), which can be readily obtained from the ‘comprehensive’ logistic regression model. Suppose \( |S_2| = \max(|S_2|, |S_3|, |S_5|) \). Therefore, the test statistic is \( |S_2| \), which does not change regardless of whether \( I = \{2, 3\} \) or \( I = \{2, 3, 5\} \) is to be tested.

4: Critical values \( c_I \) have the property that if \( J \subset I \) then \( c_J \leq c_I \).

Check: Since \( S_i = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_i}, i \in I \) follow a multivariate normal distribution under the null hypothesis with mean \( 0 \) and variances 1. If \( J \subset I \), \( S_j, j \in J \) also follows a multivariate normal distribution with a lower dimension \( |J| \). Then it
is trivially satisfied that critical values $c_{\alpha,J}$ for testing $H_{0,J}$ : $\beta_j = 0$, $j \in J$ have the property that if $J \subset I$, then $c_{\alpha,J} < c_{\alpha,I}$.

Therefore, the proposed test statistics from the ‘comprehensive’ logistic regression model satisfy one set of sufficient conditions for a step-down short-cut (Xu and Hsu, 2007).

For testing $k$ null hypotheses $H_{0,i}$ : $\beta_i = 0$, $i = 1, \ldots, k$ in the ‘comprehensive’ logistic regression setting, the step-down short-cut test procedure is as follows:

Let $[1], \ldots, [k]$ be the random indices such that $|S_{[1]}| < \ldots < |S_{[k]}|$, where $S_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i}, i = 1, \ldots, k$. Let $\beta_{[i]}$ indicate the parameter of test interest corresponding to each $S_{[i]}$.

Instead of testing up to $2^k - 1$ subspaces, the step-down short-cut version tests in following sequence:

Step 1: Test $H_{0,\{[1], \ldots, [k]\}}$ : $\beta_{[1]} = 0$ and $\beta_{[2]} = 0$ and $\ldots$ and $\beta_{[k]} = 0$.
If $|S_{[k]}| > c_{\alpha,\{[1], [2], \ldots, [k]\}}$ then infer $\beta_{[k]} \neq 0$ and go to step 2; else stop.

Step 2: Test $H_{0,\{[1], \ldots, [k-1]\}}$ : $\beta_{[1]} = 0$ and $\beta_{[2]} = 0$ and $\ldots$ and $\beta_{[k-1]} = 0$.
If $|S_{[k-1]}| > c_{\alpha,\{[1], [2], \ldots, [k-1]\}}$, then infer $\beta_{[k-1]} \neq 0$ and go to step 3; else stop.

\vdots

Step $k$: Test $H_{0,\{[1]\}}$ : $\beta_{[1]} = 0$
If $|S_{[1]}| > c_{\alpha,\{[1]\}}$, then infer $\beta_{[1]} \neq 0$ and stop; else stop.

Therefore, instead of testing up to $2^k - 1$ tests, at most $k$ tests are needed to make valid inference for $\beta_i, i = 1, \ldots, k$. 

116
3.3.7 Example analysis

Now let us illustrate how to apply the step-down shortcut of the partitioning test in logistic regression model setting using the honey bee experiment as an example.

The experiment was interested in comparing four treatment groups: untreated control, min oil, mint patty and oil patty. The odds of mites infection versus no infection can be modeled through the ‘comprehensive’ logistic regression model (3.10).

Let $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$.

$$
\pi_i = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}; \text{ for } i = 0
$$

$$
\pi_i = \frac{\exp(\beta_1 + \beta)}{1 + \exp(\beta_1 + \beta)}; \text{ for } i = 1, 2, 3
$$

The parameter estimates $\hat{\beta}_i$ as well as the variance-covariance estimate $\hat{\Sigma}$ can be obtained by fitting the ‘comprehensive’ logistic regression model through many statistical software packages. The R 2.7.0 package ‘Design’ was used for the following analysis.

We obtain the parameter estimates:

$$
\hat{\beta} = \begin{pmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\hat{\beta}_3 \\
\end{pmatrix} = \begin{pmatrix}
0.1736 \\
0.1233 \\
0.0379 \\
0.0189 \\
\end{pmatrix}
$$

with the variance-covariance matrix estimate:

$$
\hat{\Sigma} = \begin{pmatrix}
0.001423 & 0.000294 & 0.000284 & 0.000283 \\
0.000294 & 0.001168 & 0.000272 & 0.000272 \\
0.000284 & 0.000272 & 0.000983 & 0.000262 \\
0.000283 & 0.000272 & 0.000262 & 0.000976 \\
\end{pmatrix}
$$

By using the appropriate elements in the variance-covariance matrix $\hat{\Sigma}$, we obtain the statistics $(S_1, S_2, S_3)$:
\[
\begin{pmatrix}
3.607 \\ 1.207 \\ 0.604
\end{pmatrix}
\] (3.34)

The estimated correlation matrix of \( \hat{\beta} \) can be easily obtained:

\[
\hat{R} = \begin{pmatrix}
1.000000 & 0.227658 & 0.239845 & 0.240263 \\
0.227658 & 1.000000 & 0.253835 & 0.254277 \\
0.239845 & 0.253835 & 1.000000 & 0.267889 \\
0.240263 & 0.254277 & 0.267889 & 1.000000
\end{pmatrix}
\] (3.35)

According to the MLE asymptotic theory, \( S_i = \frac{\hat{\beta}_i - \beta_i}{\sigma_i}, i = 1, 2, 3 \) follow the joint distribution \( \text{MVN}(0, \mathbf{r}) \) under the null hypothesis, where \( \mathbf{r} \) consists the \( i \)th rows and columns of \( \mathbf{R}, i = 1, 2, 3 \):

\[
\hat{r} = \begin{pmatrix}
1.000000 & 0.253835 & 0.254277 \\
0.253835 & 1.000000 & 0.267889 \\
0.254277 & 0.267889 & 1.000000
\end{pmatrix}
\] (3.36)

Therefore, the step-down version of partitioning test proceeds as:

First order \( S_i \) according to their absolute values. Let \([1],[2],[3]\) be the random indices such that \( |S_{[1]}| \leq |S_{[2]}| \leq |S_{[3]}| \). Therefore

\[
S_1 = S_{[3]}, S_2 = S_{[2]}, S_3 = S_{[1]}
\] (3.37)

Let \( \beta_{[i]}, i = 1, 2, 3 \) be \( \beta_i \) that corresponds to each \( S_{[i]} \). Therefore, \( \beta_1 = \beta_{[3]}, \beta_2 = \beta_{[2]}, \beta_3 = \beta_{[1]} \).

Step 1: Test \( H_{0([1],[2],[3])} : \beta_1 = 0 \) and \( \beta_2 = 0 \) and \( \beta_3 = 0 \).

Recall that as in (3.29), the critical value can be obtained based on the joint distribution of \( S_1, S_2, S_3 \),

118
\[ 1 - \int_{-c_{1,2,3}}^{c_{1,2,3}} \int_{-c_{1,2,3}}^{c_{1,2,3}} \int_{-c_{1,2,3}}^{c_{1,2,3}} f(S_1, S_2, S_3) dS_1 dS_2 dS_3 = \alpha \] (3.38)

This could be solved by using the ‘mvtnorm’ package in R based on algorithm proposed by Genz and Bretz (Genz, 1993; Genz and Bretz, 1999).

Assuming \( \alpha = 0.05 \), \( c_{1,2,3} = 2.379 \). Since

\[ |S_3| = |S_1| = 3.607 > c_{1,2,3} = 2.379, \] (3.39)

we reject \( H_0\{[1],[2],[3]\} \), infer \( \beta_1 \neq 0 \) and go to the next step.

Step 2: Test \( H_0\{[1],[2]\} : \beta_1 = \beta_3 = 0 \) and \( \beta_2 = \beta_2 = 0 \) by comparing the second largest test statistic \( |S_2| \) to the critical value \( c_{2,3} \) which satisfies

\[ 1 - \int_{-c_{2,3}}^{c_{2,3}} \int_{-c_{2,3}}^{c_{2,3}} f(S_2, S_3) dS_2 dS_3 = \alpha \] (3.40)

Since

\[ |S_2| = |S_2| = 1.207 < c_{2,3} = 2.230, \] (3.41)

we cannot reject the null and stop at this step.

As shown in (3.28), the p-value for testing each subset null hypothesis can be obtained by integration over the area defined by the statistics \( S_i \). For example, p-value for testing \( \beta_1 = 0 \) equals

\[ 1 - \int_{-s_1}^{s_1} \int_{-s_1}^{s_1} \int_{-s_1}^{s_1} f(S_1, S_2, S_3) dS_1 dS_2 dS_3 = 0.0009 \] (3.42)

The \( 1 - \alpha \) simultaneous confidence interval for parameters \( \beta_i \) can be easily obtained by \([\hat{\beta}_i - c_i \hat{\sigma}_i, \hat{\beta}_i + c_i \hat{\sigma}_i]\).
Thus we reject the null hypothesis $H_{01} : \beta_1 = 0$ but not $H_{02} : \beta_2 = 0$ or $H_{03} : \beta_3 = 0$. There is evidence indicating that the mite infection probability of the mint oil group is significantly higher than the untreated group. However, there is not enough evidence to show the mint patty or the oil patty group had significantly different mites infection probabilities compared to the control group from the observed data.

To construct a single-step multiple testing with the control procedure, we could simply use the critical value obtained from the first step, i.e. $c_{\{1,2,3\}} = c^D = 2.379$ to test all three null hypotheses. Since

$$|S_1| = 3.607 > c^D = 2.379$$

$$|S_2| = 1.207 < c^D = 2.379$$

$$|S_3| = 0.604 < c^D = 2.379$$

we reject the hypothesis that $H_{01} : \beta_1 = 0$ but not $H_{02}$ or $H_{03}$. Therefore, for this particular data set, the extended single-step Dunnett’s method reached the same conclusion as the step-down procedure did.

Although we could model the honey bee data by the ‘comprehensive’ logistic regression model and obtain the joint distribution of the test statistics analytically, Holm’s method is commonly used for convenience in practice. Compared to the step-down method and the single step Dunnett’s method, it ignores the joint distribution of the test statistics. Compared to the single-step Dunnett’s method, Holm’s method is a step-down procedure shortcutting partition testing, which enjoys the advantage of stepwise method. It would be interesting to see how Holm’s method performs in current example.

Since $c^H_{1,2,3} = c_{\alpha/3} = 2.394, c^H_{1,3} = c_{\alpha/2} = 2.241, c^H_1 = c_\alpha = 1.960$, the Holm’s procedure proceeds as:

120
Step 1: since $|S_{[3]}| = |S_1| = 3.607 > c_{1,2,3}^H = 2.394$, reject $H_{01}$ and infer $\beta_1 \neq 0$ and go to step 2;

Step 2: since $|S_{[2]}| = |S_2| = 1.207 < c_{1,3}^H = 2.241$, do not reject $H_{02}$ and stop here.

Therefore, Holm’s method reached the same conclusion as the other two methods did.

The parameter estimates with corresponding test statistics, critical values, p-values and confidence intervals obtained from the step-down partitioning test, the single-step Dunnett’s method, as well as the Holm’s step-down method are summarized in following tables.

Figure 3.3.7 further shows the test statistics and critical values from the step-down partition test, the single-step extended Dunnett’s method and Holm’s method.
<table>
<thead>
<tr>
<th>Dose</th>
<th>$\hat{\beta}_i$</th>
<th>Test Statistic</th>
<th>Critical Value</th>
<th>P-value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mint oil</td>
<td>0.1233</td>
<td>3.607</td>
<td>2.379</td>
<td>0.00091</td>
<td>(0.0420, 0.2046)</td>
</tr>
<tr>
<td>Mint patty</td>
<td>0.0379</td>
<td>1.207</td>
<td>2.379</td>
<td>0.523</td>
<td>(-0.0320, 0.1125)</td>
</tr>
<tr>
<td>Oil patty</td>
<td>0.0189</td>
<td>0.604</td>
<td>1.960</td>
<td>0.545</td>
<td>(-0.0423, 0.0801)</td>
</tr>
</tbody>
</table>

Table 3.1: Test statistics, critical values, P-values and confidence intervals by step-down shortcut of partitioning test for data from (Dasgupta et al., 2000)

<table>
<thead>
<tr>
<th>Dose</th>
<th>$\hat{\beta}_i$</th>
<th>Test Statistic</th>
<th>Critical Value</th>
<th>P-value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mint oil</td>
<td>0.1233</td>
<td>3.607</td>
<td>2.394</td>
<td>0.00093</td>
<td>(0.0415, 0.2051)</td>
</tr>
<tr>
<td>Mint patty</td>
<td>0.0379</td>
<td>1.207</td>
<td>2.241</td>
<td>0.405</td>
<td>(-0.0324, 0.1082)</td>
</tr>
<tr>
<td>Oil patty</td>
<td>0.0189</td>
<td>0.604</td>
<td>1.960</td>
<td>0.545</td>
<td>(-0.0423, 0.0801)</td>
</tr>
</tbody>
</table>

Table 3.2: Test statistics, critical values, P-values and confidence intervals by single-step Dunnett’s method for data from (Dasgupta et al., 2000)

<table>
<thead>
<tr>
<th>Dose</th>
<th>$\hat{\beta}_i$</th>
<th>Test Statistic</th>
<th>Critical Value</th>
<th>P-value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mint oil</td>
<td>0.1233</td>
<td>3.607</td>
<td>2.394</td>
<td>0.00093</td>
<td>(0.0415, 0.2051)</td>
</tr>
<tr>
<td>Mint patty</td>
<td>0.0379</td>
<td>1.207</td>
<td>2.241</td>
<td>0.405</td>
<td>(-0.0324, 0.1082)</td>
</tr>
<tr>
<td>Oil patty</td>
<td>0.0189</td>
<td>0.604</td>
<td>1.960</td>
<td>0.545</td>
<td>(-0.0423, 0.0801)</td>
</tr>
</tbody>
</table>

Table 3.3: Test statistics, critical values, P-values and confidence intervals by Holm’s method for data from (Dasgupta et al., 2000)
Figure 3.1: Honey bee example result: test statistics and critical values from the step-down partition test, the single-step extended Dunnett’s method and Holm’s method.
3.4 Likelihood Ratio Tests Cannot Shortcut

Recall that by the statistic ‘deviance’ \( D \), one can always compare the likelihood of one model to that of the ‘saturated’ model which fits exactly to the observed data, i.e.

\[
D = -2\ln[L_c/L_s]
\]  

(3.44)

where \( L_c \) is the likelihood of current model, \( L_s \) is the likelihood of the saturated model.

Deviance is a useful measurement for goodness-of-fit. Similarly, the deviance \( D \) for another nested model can be obtained. Consequently one can compare two nested models directly by:

\[
G = -2\ln[L_{-\theta_i}/L_{+\theta_i}]
\]  

(3.45)

with \( L_{+\theta_i} \) being the likelihood with parameter \( \theta_i \), \( L_{-\theta_i} \) being the likelihood without parameter \( \theta_i \). Under the null hypothesis that the additional parameter \( \theta_i \) in the more complex model is zero, one compares \( G \) to a chi-square distribution with 1 degree of freedom.

Even in the case that MLEs do not exist, likelihoods still can be maximized. Therefore, comparison between nested models can be always accomplished by comparing their likelihoods given current data. Specifically, let \( L(\theta) \) denote the likelihood for a model contains parameters \( \theta = \{\theta_1, \ldots, \theta_k\} \) given observed data. To test the null hypothesis

\[
H_{0\{1, \ldots, k\}} : \theta_1 = \ldots = \theta_k = 0
\]  

(3.46)
is equivalent to comparing the likelihood of the reduced model which does not contain any element of $\theta$ with that of the full model which has all $k$ elements of $\theta$.

Let $\Theta_0$ be the parameter space under the null hypothesis. Let $\Theta$ be the parameter space without restriction. The likelihood ratio statistic is:

$$\Lambda_{1,\ldots,k} = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} \quad (3.47)$$

With large sample size, likelihood ratio test compares $-2\ln \Lambda_{1,\ldots,k}$ to a chi-square distribution with $k$ degrees of freedom, i.e. $\chi^2_{k,\alpha}$. If a subset of the parameter set $\theta_{i_1}, \ldots, \theta_{i_s}, \{i_1, \ldots, i_s\} \subseteq \{1, \ldots, k\}$ are of test interest, then the likelihood ratio statistic $-2\ln \Lambda_{i_1,\ldots,i_s}$ will be compared to $\chi^2_{s,\alpha}$, and so forth. Therefore, for a multiple testing problem of whether all or some of $k$ parameters equal zero, one approach is to execute a series of likelihood ratio tests.

In theory, testing

$$H_{0i} : \theta_i = 0 \, i = 1, \ldots, k \quad (3.48)$$

by the partitioning principle (Stefansson et al., 1988; Finner and Strassburger, 2002) or the closed testing principle (Marcus et al., 1976), up to $2^k - 1$ subset hypotheses tests are required, each testing for null hypotheses on a subset of the $k$ parameters. If no short-cut is permitted, the number of total tests will increase quickly with $k$.

Let us consider the simplified honey bee example (Dasgupta et al., 2000). Recall that

$$Y_{ij} = \begin{cases} 1 & \text{infection} \\ 0 & \text{no infection} \end{cases} \quad (3.49)$$

$Y_{ij}$ are independent samples from $\text{Bernoulli}(\pi), i = 0, 1, 2, 3$, so that the probability of infection for honey bee $j$ is determined by its treatment $i$. 

125
Consider the comparison between two treatment groups versus the control, logistic regression model can be used modeling the mites infection probabilities in group $i$:

$$P(Y_{ij} = 1) = \pi_i = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)}, i = 1, 2$$

Under the null hypothesis, every group has the same infection probability regardless the treatment. That is, $H_{0i}^0: \pi_0 = \pi_i, i = 1, 2$, which is equivalent to testing:

$$H_{0i}: \beta_i = 0, i = 1, 2$$

One way to test (3.51) is to follow the closed testing principle (Marcus et al., 1976), where the null hypothesis $H_{0(12)}: \beta_1 = 0$ and $\beta_2 = 0$ as well as $H_{0(1)}: \beta_1 = 0$, $H_{0(2)}: \beta_2 = 0$ will be tested separately at the desired $\alpha$ level. Only if both $H_{0(12)}$ and $H_{0(1)}$ are rejected, $H_{01}: \beta_1 = 0$ will be rejected. Only if both $H_{0(12)}$ and $H_{0(2)}$ are rejected, $H_{02}: \beta_2 = 0$ will be rejected.

It is natural to consider the likelihood ratio test to test $H_{0(12)}, H_{0(1)}$ and $H_{0(2)}$, since the parameter estimation in logistic regression is achieved by the maximum likelihood method (Dasgupta et al., 2000). Define $\Lambda_{1,2}$ as the test statistic for testing $H_{0(12)}: \beta_1 = 0$ and $\beta_2 = 0$:

$$\Lambda_{1,2} = \frac{\sup_{\beta \in B} L(\beta)}{\sup_{\beta \in B} L(\beta)}$$

$$= \frac{\sup_{\beta_1 = 0, \beta_2 = 0} \prod_{i=0}^{2} \prod_{j=1}^{n} \pi_i^{Y_{ij}} (1 - \pi_i)^{1 - Y_{ij}}}{\prod_{i=0}^{2} \sup_{\beta_i \in \mathbb{R}} \prod_{j=1}^{n} \pi_i^{Y_{ij}} (1 - \pi_i)^{1 - Y_{ij}}}$$

To test the null hypothesis involving the single parameter $H_{0(1)}: \beta_1 = 0$ or $H_{0(2)}: \beta_2 = 0$, the test statistics are:
\[ \Lambda_i = \frac{\sup_{\beta_1=0} L(\beta)}{\sup_{\beta_1, \beta_2 \in \mathbb{R}^2} L(\beta)} \]

\[ = \frac{\sup_{\beta_1=0} \prod_{r=0}^{2} \prod_{j=1}^{n} \pi_{rj}^{Y_{rj}} (1 - \pi_{rj})^{1-Y_{rj}}}{\prod_{r=0}^{2} \sup_{\beta_1, \beta_2 \in \mathbb{R}^2} \prod_{j=1}^{n} \pi_{rj}^{Y_{rj}} (1 - \pi_{rj})^{1-Y_{rj}}} , \text{ for } i = 1 \text{ or } 2 \tag{3.53} \]

The likelihood ratio test compares \(-2ln\Lambda_{1,2}\) to the upper \(\alpha\) quantile of chi-square distribution with 2 degrees of freedom, i.e. \(\chi^2_{2,0.05}\). Similarly, to test \(H_0(1)\) and \(H_0(2)\), \(-2ln\Lambda_1\) and \(-2ln\Lambda_2\) will be compared to \(\chi^2_{1,0.05}\).

Therefore, to make inference for two parameters, at most three tests are required. As the number of hypotheses increases, this would soon be infeasible if no short-cut path can be taken in the testing procedure.

Although the likelihood ratio test seems a good choice for testing multiple hypotheses, through a simple example we will show that the test procedures using likelihood ratio statistics can not have a short-cut.

We can construct a data set with \(n = 100\) binary observations recorded for each of three treatment groups \(i, i = 0, 1, 2\), with the probabilities of infection \(\hat{\pi}_0 = 0.35, \hat{\pi}_1 = 0.48, \hat{\pi}_2 = 0.32\), respectively. This can be easily achieved by creating three vectors of length 100 consisting binary outcomes with exactly 35, 48 and 32 observations to be 1.

By using the ‘rlm’ function in R, one could fit a series of logistic regression models with either one or two group indicator variables in addition to the intercept \(\beta_0\). Then the likelihood ratio statistics for testing \(H_0(1,2), H_0(1)\) and \(H_0(2)\) are obtained by comparing the likelihoods of the reduced models corresponding to each of above null hypothesis with that from the saturated models.
\[-2\ln\Lambda_{1,2} = 6.14\]
\[-2\ln\Lambda_1 = 0.50\]
\[-2\ln\Lambda_2 = 2.91\] (3.54)

Therefore, to test $H_{0(12)}$, $-2\ln\Lambda_{1,2}$ is compared to $\chi^2_{2,0.05} = 5.99$. Since $-2\ln\Lambda_{1,2} = 6.14 > 5.99$, we reject $H_{0(12)} : \theta_1 = 0$ and $\theta_2 = 0$. To test $H_{0(1)}$ and $H_{0(2)}$, $-2\ln\Lambda_1$ and $-2\ln\Lambda_2$ are compared to $\chi^2_{1,0.05} = 3.84$. Since both $-2\ln\Lambda_1 = 0.50 < 3.84$ and $-2\ln\Lambda_2 = 2.91 < 3.84$, we do not reject either $H_{0(1)}$ or $H_{0(2)}$. That is, rejection of the intersection hypothesis $H_{0(12)}$ does not lead to automatic rejection of either $H_{0(1)}$ or $H_{0(2)}$. To make valid inference about $\theta_1$ and $\theta_2$, all three null hypotheses $H_{0(12)}, H_{0(1)}, H_{0(2)}$ need to be tested. Therefore, no short-cut pathway can be taken by skipping any of the null hypothesis testing.

The underlying reason for likelihood ratio statistics based method not being able to take short-cut is the ‘shape’ of the rejection regions in each step of subset hypothesis testing.

We illustrate this by considering the simplest possible example, where two parameters of interest in the test are the means of two independent standard normal variables. Let $X = (X_1, \ldots, X_n)$ be a random sample from $N(\theta_1, \sigma^2)$, $Y = (Y_1, \ldots, Y_n)$ be a random sample from $N(\theta_2, \sigma^2)$. $X$ and $Y$ are independent. Therefore,

$$f(x_i|\theta_1) = \frac{1}{\sqrt{2\pi}} e^{\frac{(x_i - \theta_1)^2}{2\sigma^2}}, \ i = 1, \ldots, n$$

$$f(y_j|\theta_2) = \frac{1}{\sqrt{2\pi}} e^{\frac{(y_j - \theta_2)^2}{2\sigma^2}}, \ j = 1, \ldots, n$$ (3.55)

We are interested in testing if $\theta_1$ and/or $\theta_2$ equal zero, i.e.
By maximum likelihood method, \( \theta_1 \) is estimated by \( \bar{X} \) and \( \theta_2 \) is estimated by \( \bar{Y} \).

Recall that the likelihood ratio statistic is the ratio of two likelihoods: the likelihood from parameters under the null hypothesis restriction, and the likelihood from parameters under no constraints. \( H_{0(12)} : \theta_1 = 0 \) and \( \theta_2 = 0 \) are tested by using the likelihood ratio statistic

\[
-2 \ln \Lambda_{1,2}
\]

where

\[
\Lambda_{1,2} = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}
\]

\[
= \frac{\prod_{i=1}^{n} f(x_i | \theta_1) f(y_i | \theta_2)}{\prod_{i=1}^{n} \sup_{\theta_1 \in \Theta} f(x_i | \theta_1) \prod_{i=1}^{n} \sup_{\theta_2 \in \Theta} f(y_i | \theta_2)}
\]

By simple algebra,

\[
\Lambda_{1,2} = \sup_{\theta_1, \theta_2 \in \Theta_0} \{ \exp \left[ -n(\bar{x} - \theta_1)^2 / 2\sigma^2 \right] \exp \left[ -n(\bar{y} - \theta_2)^2 / 2\sigma^2 \right] \}
\]

\[
= \sup_{\theta_1, \theta_2 \in \Theta_0} \{ \exp \left[ -\frac{n}{2\sigma^2} [(\bar{x} - \theta_1)^2 + (\bar{y} - \theta_2)^2] \right] \}
\]

One rejects \( H_{0(12)} \) if \(-2 \ln \Lambda_{1,2} > \chi^2_2\) where \( \chi^2_2 \) is the chi-square statistic with 2 degrees of freedom, that is,

\[
-2 \ln \Lambda_{1,2} > \chi^2_2
\]

\[
\Rightarrow \sup_{\theta_1, \theta_2 \in \Theta_0} \{ \exp \left[ -\frac{n}{2\sigma^2} [(\bar{x} - \theta_1)^2 + (\bar{y} - \theta_2)^2] \right] \} < \lambda
\]

\[
\Rightarrow \inf_{\theta_1, \theta_2 \in \Theta_0} \{ (\bar{x} - \theta_1)^2 + (\bar{y} - \theta_2)^2 \} > \lambda'
\]
where $\lambda, \lambda'$ are constant.

Therefore, if the hypothesized values for $\theta_1$ and $\theta_2$ are 0, the rejection region for $H_{0\{1\}}$ is outside of a circle whose origin is (0, 0), i.e. $\hat{\theta}_1^2 + \hat{\theta}_2^2 > c$ (shown as the circle in Figure 3.4). The ‘points’ on the circle satisfy the equation $\hat{\theta}_1^2 + \hat{\theta}_2^2 = c$. In general, if $\theta_1$ and $\theta_2$ are correlated, the rejection region has an elliptical shape.

To test $H_{0\{1\}} : \theta_1 = 0$ and $H_{0\{2\}} : \theta_2 = 0$, $-2\ln\Lambda_1$ and $-2\ln\Lambda_2$ will be calculated, where

$$
\Lambda_1 = \sup_{\theta_1 = 0} \exp\left[-\frac{n}{2\sigma^2} (\bar{x} - \theta_1)^2\right] \\
\Lambda_2 = \sup_{\theta_2 = 0} \exp\left[-\frac{n}{2\sigma^2} (\bar{y} - \theta_2)^2\right] 
$$

(3.60)

Furthermore, $H_{0\{1\}}$ is rejected if $-2\ln\Lambda_1 > \chi_1^2$ and $H_{0\{2\}}$ is rejected if $-2\ln\Lambda_2 > \chi_1^2$.

$$
-2\ln\Lambda_i > \chi_1^2 \\
\rightarrow sup_{\theta_i = 0} \{\exp[-\frac{n}{2\sigma^2} (\bar{x} - \theta_i)^2]\} < \lambda_1 \\
\rightarrow inf_{\theta_i = 0} (\bar{x} - \theta_i)^2 > \lambda_1' 
$$

(3.61)

where $\lambda_1, \lambda_1'$ are constant. This is equivalent to rejecting $H_{0\{1\}}$ if $\hat{\theta}_1^2 > c^*$ and rejecting $H_{0\{2\}}$ if $\hat{\theta}_2^2 > c^*$, where $c^*$ is constant. Therefore, the rejection region for $H_{0\{i\}}, i = 1, 2$ is

$$
\hat{\theta}_i > \sqrt{c^*} \text{ or } \hat{\theta}_i < -\sqrt{c^*}, i = 1, 2, 
$$

(3.62)

which is defined by four straight lines perpendicular to the axes and going through points $(0, \sqrt{c^*}), (0, -\sqrt{c^*}), (\sqrt{c^*}, 0), (-\sqrt{c^*}, 0)$, respectively (shown as the four dashed lines in Figure 3.4).
Clearly, if the size of the test for $H_{0\{12\}}$, $H_{0\{1\}}$ and $H_{0\{2\}}$ are all $\alpha$, neither rejection region completely dominates the other.

It is easy to see that any $\hat{\theta}_1, \hat{\theta}_2$ values within the four ‘corners’ where the red dots are will lead to rejection of $H_{0\{12\}}$, since they are within the rejection region of $H_{0\{12\}}$, i.e. $\hat{\theta}_1^2 + \hat{\theta}_2^2 > c$, where $c$ is constant. However, neither $H_{0\{1\}}$ or $H_{0\{2\}}$ will be further rejected since the areas where the red dots locate are within the acceptance regions of the two single null hypotheses $H_{0\{1\}}$, $H_{0\{2\}}$, which are defined by the four dashed lines, i.e. $|\hat{\theta}_i| \leq |\sqrt{c^*}|$. Therefore, rejection by $H_{0\{12\}}$ does not guarantee rejection of either $H_{0\{1\}}$ or $H_{0\{2\}}$ by using likelihood ratio statistics. Consequently, no short-cut can be taken to rejecting $H_{0i}$, $i \in I$ even if one knows for sure the ‘bigger’ intersection hypothesis set $H_{0I}$ is rejected.

Obviously, it is easy for one to see from the plot that there are also areas where $\hat{\theta}_1, \hat{\theta}_2$ values will lead to failure to reject $H_{0\{12\}}$ while $H_{0\{1\}}$ or $H_{0\{2\}}$ are to be rejected, i.e. $|\hat{\theta}_i| > |\sqrt{c^*}|$ but $\hat{\theta}_1^2 + \hat{\theta}_2^2 \leq c$. 
Figure 3.2: Rejection regions by likelihood ratio statistics: \( H_{0(12)} : \theta_1 = 0 \) and \( \theta_2 = 0 \) (outside of the circle), \( H_{0(1)} : \theta_1 = 0 \) or \( H_{0(2)} : \theta_2 = 0 \) (outside of the dashed lines)
From (3.59), in fact, $H_{0\{12\}}$ is rejected if the smallest distance between the MLE vector $\hat{\theta}$ and the parameter vector under the null hypothesis $\theta \in \Theta^0$ is larger than a critical value $\lambda'$. Since the data are following multivariate normal distribution, the ‘distance’ between the two vectors $\hat{\theta}$ and $\theta$ is measured by the Mahalanobis distance (Mahalanobis, 1936), which is a metric in the quadratic difference form normalized by its variance-covariance matrix $S$, i.e.

$$d = \sqrt{(\hat{\theta} - \theta)'S^{-1}(\hat{\theta} - \theta)}$$  \hspace{1cm} (3.63)

Therefore, rejection of $H_{0\{12\}}$ by $-2\Lambda_{1,2} > \chi^2_2$ is equivalent to rejecting $H_{0\{12\}}$ by $\inf_{\theta \in \Theta_0} d > c'$, which is equivalent to rejecting $H_{0\{12\}}$ if

$$\inf_{\theta \in \Theta_0} d^2 = \inf_{\theta \in \Theta_0}\{(\hat{\theta} - \theta)'S^{-1}(\hat{\theta} - \theta)\} > c''$$  \hspace{1cm} (3.64)

Note that not only likelihood ratio statistics based tests cannot shortcut in constructing multiple testing procedure, other methods based on the test statistics in the form of quadratic difference, such as the chi-square test, F-test and the quadratic form Wald test, do not have short-cut version testing procedures due to similar reasons.

In contrast, the partitioning test using maxT or minp type of test statistics based on the asymptotic multivariate normal distribution of MLEs from the ‘comprehensive’ logistic regression model satisfy a set of sufficient conditions for a step-down shortcut procedure.

If a maxT type of test is used, rejection of an intersection null hypothesis $H_{0I}$ is due to $|T_I| = \max_{i \in I}|T_i| > c_I, i = 1, \ldots, k$, where $c_I$ is the critical value properly adjusted for $k$ null hypotheses.
Assume $k = 2$. Let $T_i = \theta_i, i = 1, 2$. Then one rejects $H_{0\{12\}} : \theta_1 = 0$ and $\theta_2 = 0$ if the observed test statistic $|T_{\{1,2\}}| = \max(|T_1|, |T_2|) > c_2$, where $c_2$ is the critical value. The rejection region is outside of the square which is defined by four straight lines perpendicular to the axes and going through points $(0, c_2), (0, -c_2), (c_2, 0), (-c_2, 0)$ respectively (shown as the four straight lines in Figure 3.4).

Further, the null hypothesis involving single parameter $H_{0\{i\}}$ is rejected if $|T_i| > c_1, i = 1, 2$. The union of the rejection regions of $H_{0\{i\}}, i = 1, 2$ is a square defined by four straight lines perpendicular to the axes and going through points $(0, c_1), (0, -c_1), (c_1, 0), (-c_1, 0)$, respectively (shown as the four dashed lines in Figure 3.4). Since the monotonicity criterion of the sufficient conditions (Xu and Hsu, 2007) is satisfied, the critical value $c_1$ is not larger than the critical value used in testing the intersection hypothesis involving two parameters, i.e. $c_1 \leq c_2$. Therefore, the rejection region of $H_{0I}$ is completely within the union of the rejection regions of $H_{0\{i\}}, i = 1, 2$. Obviously, if $|T_{\{1,2\}}| = \max_{i=1,2}|T_i| > c_2, |T_{\{1,2\}}| > c_1$ is guaranteed to be true. That is, if $H_{0I}$ is rejected, at least one of $H_{0\{i\}}, i \in I$ is automatically rejected. This indicates one subset hypothesis test can be skipped without further testing.

Thus for testing $k$ null hypotheses simultaneously, the number of tests is significantly reduced from $2^k - 1$ to at most $k$ while valid inferences are still guaranteed. This is one major achievement of the short-cut procedures.

In addition, compared to the single-step multiple testing procedures such as the Dunnett’s method, the step-down procedure we proposed in this chapter is advantageous in terms of being more powerful in finding more positive discoveries by conditioning on the observed data, which will be discussed in greater details in the next chapter.
Figure 3.3: Rejection regions by maxT statistics: $H_{0\{12\}} : \theta_1 = 0$ and $\theta_2 = 0$ (outside of the straight lines), $H_{0\{1\}} : \theta_1 = 0$ or $H_{0\{2\}} : \theta_2 = 0$ (outside of the dashed lines)
CHAPTER 4

MULTIPLE TESTING IN MARGINAL MODELS

To construct multiple testing procedures that properly control false rejection rates, such as FWER, one needs to define the appropriate test statistics and find critical values according to their reference distributions under the null hypothesis. Recall that one important technique to construct the reference distributions is to model the data. Through modeling, the joint distribution of the test statistics can sometimes be derived analytically, so that exact computation of the critical value is possible.

4.1 Marginal Models and GEE

It is quite common that observations or samples are collected in groups. For example, fishes caught from the same lake, or students from the same school district. In biomedical studies and clinical trials, sometimes measurements are made repeatedly over a period of time on the same subject or over a group of individuals that are associated with each other. The group in which the observations belong to is a ‘cluster’. Observations from the same cluster are often correlated since they have something in common.

Let $Y_{st}$ denote the $t$th observation from the $s$th subject, $s = 1, \ldots, n; t = 1, \ldots, m_s$, where $s$ in the subscript indicates different numbers of observations are allowed for
each subject. Suppose each subject belongs to one of $k + 1$ ‘treatment’ groups $i, i = 0, \ldots, k$.

This set-up allows one to compare the changes in the response variable across both treatment groups and time or space. Due to subject-specific effect, it is expected to see associations between observations from the same subject.

For most members in the exponential family, especially discrete data, it is not easy to define the multivariate joint distribution, which is required for likelihood based modeling. More importantly, if the scientific question that the researchers try to address is in the marginal or population level, marginal models using Generalized Estimating Equations (GEE) by Liang and Zeger are widely used (Liang and Zeger, 1986; Zeger and Liang, 1986).

Link functions are utilized in marginal models to link the ‘marginal mean’ $E[Y_s]$ to the ‘linear predictor’ $\eta_s = X_s \beta$, i.e.

$$h(E[Y_s]) = X_s \beta.$$  \hspace{1cm} (4.1)

For example, the logit link function can be utilized for modeling binary outcome variables.

Estimation in GEE models is not based on the likelihood or the joint density functions. In contrast, a series of equations called the ‘quasi score equations’ are solved. That is, the GEE estimates $\hat{\beta}$ for parameters $\beta$ are the solution of the ‘Generalized Estimating Equations’:

$$U(\beta) = \sum_{s=1}^{n} D_s V_s^{-1} (Y_s - \mu_s) = 0$$  \hspace{1cm} (4.2)
where

\[ D_s = \frac{\partial \mu_s}{\partial \beta} = \frac{\partial h^{-1}(X_s\beta)}{\partial \beta} \quad (4.3) \]

and

\[ V_s = A_s^{1/2} R_s(\alpha) A_s^{1/2} \quad (4.4) \]

where \( V_s \) is the so called ‘working’ correlation matrix. \( R_s(\alpha) \) is the assumed correlation structure of \( Y_s \) defined by parameters \( \alpha \). By specification of the working correlation, the dependence between repeated measurements is taken into account.

Although the GEE method only requires first-order moment estimation, robust and consistent estimates of marginal effects as well as their variance-covariance matrix are guaranteed by solving the generalized estimating equations and using the working correlation matrix in the empirical variance estimator.

By the method of moments, Liang and Zeger (Liang and Zeger, 1986) showed that under mild conditions, when the marginal mean is correctly specified, GEE estimates of \( \beta \) are asymptotically normally distributed with mean \( \beta \) and variance-covariance matrix \( V_G \), i.e.

\[ n^{\frac{1}{2}}(\hat{\beta} - \beta) \rightarrow_D \text{MVN}(0, V_G). \quad (4.5) \]

Liang and Zeger suggest to estimate \( V_G \) by the empirical variance estimator, which is also called the ‘sandwich’ estimator, as introduced in Chapter 1.

The conditions for asymptotic normality include: 1) the existence of a limit \( R(\alpha)^* \) for \( R(\alpha) \), (i.e. \( R(\alpha) \) converges to a fixed correlation matrix); 2) While the number of ‘clusters’ \( n \) goes to infinity, the maximum cluster size \( m_s \) and dimension of \( \alpha^* \) is finite. Under these conditions, if \( \hat{\alpha}(Y, \beta, \phi) \) and \( \hat{\phi}(Y, \beta) \) are \( n^{\frac{1}{2}} \) consistent estimators.
for $\alpha$ and $\phi$; (i.e. $n^\frac{1}{2}(\hat{\alpha} - \alpha) = 0_p(1)$, $n^\frac{1}{2}(\hat{\phi} - \phi) = 0_p(1)$), when the number of subjects is large compared to the number of measurements within each subject, then the GEE estimates are asymptotically consistent.

The advantage of the GEE method is that the working correlation assumption $R(\alpha)$ does not need to be correct for GEE to obtain consistent estimates. The asymptotic normality of GEE estimates will also not be affected even if the working correlation structure differs from the true structure. For this reason, the ‘sandwich’ estimator is also called the ‘robust’ variance estimator.

In constructing a multiple testing procedure, one has to obtain the reference distribution of the test statistics under the null hypothesis. Whether an observation is ‘unusual’ will be determined by comparing it to its reference distribution. Obtaining the joint distribution and calculating the critical values exactly is the key to more powerful procedures. Marginal models using GEE estimation are powerful tools for correlated discrete data modeling. They provide opportunities for one to obtain the joint distribution of the test statistics analytically. Based on the joint distribution, further computation of the critical values to find the appropriate rejection region is possible.

Orelien et al. generalized the single-step Dunnett’s method for multiple comparisons with the control problem in the correlated binary data setting in GEE models (Orelien et al., 2002). Their results showed validity of the proposed method with false rejection rates close to the nominal $\alpha$ level.

In this chapter, we show that when GEE estimation is possible, and the sufficient conditions for short-cut the partitioning test are satisfied, by applying the partition principle, as well as the step-down shortcut version of it, more powerful multiple
testing procedures can be constructed based on the asymptotic distribution of the GEE parameter estimates.

4.2 Motivating Example

The motivating example is from the Shell toxicology laboratory (Paul, 1982), where an experiment was conducted to study the toxic effects of a chemical. The experiment follows a typical setup for toxicology studies, where pregnant female animals were randomly administered different doses of chemicals. The toxic effects were observed on the offspring.

The outcome variable is the binary presence or absence of malformations. Besides one ‘control’ group, where no toxic chemical was fed, there are three other dose groups, i.e. ‘low’, ‘medium’ and ‘high’ dose groups. There are 27 litters in the control group and 19, 21 and 17 litters in the low, medium and high dose groups, respectively. The data is summarized in Table 1 (Paul, 1982).

Specifically, the question of interest is the relative dose effects measured by malformation probabilities compared to the control group.
Table 4.1: Summary of Toxicology Data, Paul S.R. Analysis of Proportions of Affected fetuses in Teratological Experiments, Biometrics 1982, 38:2, 361-70. ‘M’-malformed fetuses; ‘T’- total number of fetuses.

<table>
<thead>
<tr>
<th>Dose Group</th>
<th>M</th>
<th>1</th>
<th>1</th>
<th>4</th>
<th>0</th>
<th>0</th>
<th>4</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>4</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>11</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Control</td>
<td>M</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Low</td>
<td>M</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>5</td>
<td>11</td>
<td>7</td>
<td>9</td>
<td>12</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Medium</td>
<td>M</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>13</td>
<td>6</td>
<td>8</td>
<td>13</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>M</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>
4.3 Problem Formulation

Let \( Y_{st} \) indicate the malformation status of the \( t \)th fetus in the \( s \)th litter, \( s = 1, \ldots, n; t = 1, \ldots, m_s \). Each litter \( s \) belongs to one out of four dose groups: control, low, medium or high dose, denoted by \( i = 0, 1, 2, 3 \), respectively.

Obviously, the multiple fetuses within the same litter are genetically related. Also, they shared the same environmental factors during the pregnancy procedure. Therefore fetuses within each litter are associated. That is, there is a litter effect in addition to the dose effect. This data can be viewed as clustered data where litter serves as the cluster.

The research goal is to compare the proportion of malformation in each dose group with that of the control group. For \( i = 0, \ldots, 3 \), let \( \pi_i \) indicate the proportions of malformation in dose group \( i \). The two-sided null hypotheses of interest are:

\[
H_{0i}^\pi : \pi_i - \pi_0 = 0 \text{ vs. } H_{ai}^\pi : \pi_i - \pi_0 \neq 0; \ i = 1, \ldots, k
\]

This is a multiple comparisons with the control problem (MCC) in a correlated binary data setting.

For \( k \) treatment groups, hypotheses of \( k \) differences in malformation probabilities between the treatment and the control group will be tested. It is well known that the probability of making false positive findings by chance increases with the number of tests. To draw a valid conclusion, multiplicity adjustment for the total number of null hypotheses being tested is required. Relatively simpler multiplicity adjustment methods exist, such as the Bonferroni’s adjustment, which ignores the fact that in comparing each dose group with the same control, the test statistics are positively correlated. Hence Bonferroni’s method is overly conservative in this case.
If the partitioning principle (Stefansson et al., 1988; Finner and Strassburger, 2002) or the closed testing principle (Marcus et al., 1976) is used to test for $k$ null hypotheses simultaneously, the total number of tests needed is at most $2^k - 1$, which increases quickly with $k$. If no short-cut version can be applied, it will be too time consuming to test all $2^k - 1$ hypotheses even for medium sized $k$. Therefore the question is how to construct more powerful and efficient multiple testing procedures.

4.4 GEE modeling

4.4.1 The ‘Comprehensive’ Model

For the $t$th fetus of the $s$th litter and $i$th dose, $s = 1, \ldots, n, t = 1, \ldots, m_s$, define:

$$Y_{ist} = \begin{cases} 
1 & \text{if the fetus is malformed} \\
0 & \text{if the fetus is normal}
\end{cases} \quad (4.7)$$

It is reasonable to assume marginally

$$Y_{ist} \sim Bernoulli(\pi_i), \quad (4.8)$$

so that on average fetuses from litters in dose group $i$ have the same probability $\pi_i$ of experiencing the event \{$Y_{ist} = 1$\} (malformation).

The outcome vector for the $s$th litter is a $m_s \times 1$ column vector:

$$Y_{is} = \begin{pmatrix} 
Y_{is1} \\
Y_{is2} \\
\vdots \\
Y_{ism_s}
\end{pmatrix} \quad (4.9)$$
In addition, the inference one is interested in is whether toxic effects from different dose groups differ from that of the control group, but not for any individual litter. Thus, a marginal model for correlated binary data analysis is useful. With the logit link function employed, the proportion of malformation for dose group $i$ could be modeled by:

$$
\text{logit}(\pi_i) = \text{logit}\{P(Y_{ist} = 1)\} = \log\frac{P(Y_{ist} = 1)}{1 - P(Y_{ist} = 1)} = X_{ist}\beta 
$$

Let $\beta$ be the vector of the parameters for the fixed dose group effects:

$$
\beta = \begin{pmatrix} 
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3 
\end{pmatrix} 
$$

$x_{is}$, the design matrix for litter $s$ in dose group $i$, where each row is $x_{ist}$, could be designed in such a way so that $\beta_i, i = 2, 3, 4$ represent the relative dose effects by ‘low’, ‘medium’ or ‘high’ dose group compared to the control group, where $i = 0$. For example, if $s$ belongs to the control group, each row $x_{ist}$ is a $1 \times 4$ vector with the first element being equal to 1 and all the rest elements being 0, i.e.

$$
x_{is} = \begin{pmatrix} 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
... & ... & ... & ...
\end{pmatrix} 
$$

For $s$ from other dose groups where $i = 1, 2, 3$, besides the first element, let the $(i + 1)$th element of each row of $x_s$ also be 1. For example, litter $s$ is from the ‘medium’ dose group where $i = 2$, $x_s$ has each row $x_{ist} = (1, 0, 1, 0)$; for $t = 1, \ldots, m_s$. 

144
By formula (4.10),

\[
P(Y_{st} = 1) = \pi_i = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \text{ for } i = 0;
\]

\[
P(Y_{st} = 1) = \pi_i = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)} \text{ for } i = 1, 2, 3.
\]

We name this model the ‘comprehensive’ GEE model since it estimates all group effects in one model.

Notice that

\[
\frac{P(Y_{st} = 1)}{1 - P(Y_{st} = 1)} = \frac{P(Y_{st} = 1)}{P(Y_{st} = 0)} = \frac{\pi_i}{1 - \pi_i}
\]

are the ‘odds’ of experiencing an event. In this example, \( x_{st} \) indicates dosage. \( \beta_i \) estimates the difference between the ‘log odds’, i.e. the log odds ratio of having malformation in dose group \( i \) versus the control group, where \( i = 0 \).

\[
\beta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right) - \log\left(\frac{\pi_0}{1 - \pi_0}\right) = \log\left\{\frac{\pi_i/(1 - \pi_i)}{\pi_0/(1 - \pi_0)}\right\}
\]

4.4.2 Comparison on the means equals comparison on comprehensive model parameters

To compare the malformation probabilities of each dose group with the control group, \( H_{0i} : \pi_i - \pi_0 = 0 \) versus \( \pi_i - \pi_0 \neq 0 \) for \( i = 1, \ldots, k \) are to be tested.

Notice that by formula (4.16), it follows that \( \pi_i = \pi_0, i = 1, \ldots, k \), if and only if \( \beta_i = 0 \). Also, \( \pi_i - \pi_0 \) monotonically increases with \( \beta_i \). This is simply because

\[
\pi_i - \pi_0 = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}
\]
Thus,

$$\frac{\partial (\pi_i - \pi_0)}{\partial \beta_i} = \frac{\partial}{\partial \beta_i} \left\{ \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right\} = \frac{\exp(\beta_0 + \beta_i)}{[1 + \exp(\beta_0 + \beta_i)]^2} > 0.$$  

(4.18)

Therefore testing $\pi_i - \pi_0 = 0$ versus $\pi_i - \pi_0 \neq 0, i = 1, \ldots, k$ is equivalent to testing

$$H_{0i} : \beta_i = 0 \text{ vs. } H_{ai} : \beta_i \neq 0, i = 1, \ldots, k$$  

(4.19)

That is, by testing null hypotheses for GEE ‘comprehensive’ model-based parameters $\beta_i, i = 1, \ldots, k$, one can make inference for marginal differences in probabilities of a binary event.

### 4.4.3 Test Statistics and Critical Value

To test $\beta_i = 0$ for $i = 1, \ldots, k$, one can form test statistics according to the asymptotic properties of the GEE parameter estimates $\hat{\beta}_i$. By Theorem 2 by Zeger and Liang (Zeger and Liang, 1986; Liang and Zeger, 1986), under mild conditions, the large sample property of the GEE estimates is

$$n^{\frac{1}{2}} (\hat{\beta} - \beta) \rightarrow_D MVN(0, V_G)$$  

(4.20)

Thus by Slutsky’s theorem,

$$\text{diag}(V_G)^{-\frac{1}{2}} n^{\frac{1}{2}} (\hat{\beta} - \beta) = \text{diag}(n(I_0^{-1}I_I^{-1}))^{-\frac{1}{2}} n^{\frac{1}{2}} (\hat{\beta} - \beta) \rightarrow_D MVN(0, R_G)$$  

(4.21)

where $R_G$ is the correlation matrix of GEE estimates $\hat{\beta}$. 

146
Define the test statistic for testing the two-sided single null hypothesis $H_{0i}: \beta_i = 0$ versus $H_{ai}: \beta_i \neq 0$ as:

$$|S_i| = \left|\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i}\right|, i = 1, \ldots, k$$ (4.22)

where $\hat{\sigma}_i$ is the square root of the $i$th diagonal element of the GEE variance-covariance matrix estimate $\hat{\mathbf{V}}_G$. This test statistic can be readily computed from the ‘comprehensive’ GEE model output for $i = 1, \ldots, k$.

For testing multiple hypotheses simultaneously, i.e. $H_{0\{I\}}: \beta_i = 0, i \in I, I = \{1, \ldots, k\}$, the maxT type of test for each subset hypothesis $H_{0\{I\}}$ can be constructed using the statistics $\max_{i \in I} |S_i|$. That is, one rejects the two-sided null hypothesis $H_{0\{I\}}$ if

$$\max_{i \in I} |S_i| > c_{\alpha,I}$$ (4.23)

Under the null hypothesis $\beta_i = 0, i = 1, \ldots, k$, the vector $\mathbf{S} = (S_1, \ldots, S_k)$ follows a multivariate normal distribution $\text{MVN}(\mathbf{0}, \mathbf{R}_G)$, where $\mathbf{R}_G$ is the correlation matrix of $\mathbf{\hat{\beta}}$.

Consider the above joint distribution of the test statistics. The p-value of the test for the two-sided null hypothesis involving $k$ parameters

$$H_{0\{1,\ldots,k\}}: \beta_i = 0 \text{ vs. } H_{a\{1,\ldots,k\}}: \beta_i \neq 0, i = 1, \ldots, k$$ (4.24)

is given by:

$$\text{p-value} = 1 - Pr(|S_1| < Z_1, |S_2| < Z_2, \ldots, |S_k| < Z_k)$$

$$= 1 - \int_{-Z_k}^{Z_k} \cdots \int_{-Z_1}^{Z_1} f(S_1, \ldots, S_k) \, dS_1 \cdots dS_k$$ (4.25)

where $\mathbf{Z} = (Z_1, \ldots, Z_k), Z_i = \frac{|\hat{\beta}_i|}{\hat{\sigma}_i}$. 

147
For a given $\alpha$ level, the critical value $c_\alpha$ for rejecting $H_{0(1,\ldots,k)}$ can be obtained from multivariate normal distribution with dimension $k$, where it satisfies the equation:

\[ 1 - \int_{-c_\alpha}^{c_\alpha} \cdots \int_{-c_\alpha}^{c_\alpha} f(S_1, \ldots, S_k) dS_1 \cdots dS_k = \alpha \]

(4.26)

where $f(S_1, \ldots, S_k)$ is the multivariate normal joint distribution of $S$ under the null hypotheses, i.e. MVN(0, $R_G$).

### 4.4.4 Applying Partition Testing in GEE Setting

The partitioning principle (Stefansson et al., 1988; Finner and Strassburger, 2002) is a fundamental principle that can be used to construct valid multiple testing procedures that control FWER strongly.

Consider testing $k$ null hypotheses $H_{0i} : \theta \in \Theta_i, i = 1, \ldots, k$ simultaneously. Very briefly, the partition testing proceeds as follows:

P1: Partition the entire parameter space $\Theta_1 \times \cdots \times \Theta_k$ into $2^k$ disjoint subspaces so that each subspace contains exactly one subset of true null hypotheses.

P2: Test each of above null hypotheses subset at level-$\alpha$. There are at most $2^k - 1$ subsets to test.

P3: Infer $\theta \notin \Theta_I, I \subseteq \{1, \ldots, k\}$ if and only if all $H_{0(I)}^*$ with $I \subseteq J$, are rejected.

The partition testing procedure for $H_{0i} : \beta_i = 0, i = 1, \ldots, k$ in the GEE ‘comprehensive’ model setting could proceed as follows:

P1: Partition the parameter space $B_1 \times \cdots \times B_k$ into $2^k$ disjoint subspaces: $H_{0(1,\ldots,k)}^* : \beta_1 = 0$ and $\beta_2 = 0$ and $\ldots$ and $\beta_k = 0$
\(H_{0\{1,\ldots,k-1\}}^* : \beta_1 = 0\) and \(\beta_2 = 0\) and \(\ldots\) and \(\beta_{k-1} = 0\) and \(\beta_k \neq 0\)

\(\vdots\)

\(H_{0\{1\}}^* : \beta_1 = 0\) and \(\beta_2 \neq 0\) \(\ldots\) and \(\beta_k \neq 0\)

P2: Test each \(H_{0\{I\}}^*\) by \(H_{0\{I\}} : \beta_i = 0, i \in I\) at level-\(\alpha\). There are at most \(2^k - 1\) hypotheses to be tested.

P3: For each \(i, i = 1, \ldots, k\), infer \(\beta_i \neq 0\) if and only if all \(H_{0\{I\}}\) with \(i \in I, I \subseteq \{1, \ldots, k\}\) are rejected.

Recall that each partitioning hypothesis \(H_{0\{I\}}^*\) can be tested by \(H_{0\{I\}} : \beta_i = 0, i \in I\). The maxT type of test statistic \(max_{i \in I} |S_i| = max_{i \in I} \frac{|\hat{\beta}_i - 0|}{\hat{\sigma}_i}\) is proposed. Let \(S_i = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_i}, i \in \{1, \ldots, k\}\). Order \(|S_i|\) so that \(|S_{[1]}| \leq \ldots \leq |S_{[k]}|\). The test is in the form of rejecting each \(H_{0\{I\}}\) if \(max_{i \in I} |S_i| > c_{\alpha,I}\), where \(c_{\alpha,I}\) satisfies equation (4.26).

If \(c_{\alpha,I}\) is chosen so that each \(H_{0\{I\}}\) can be tested at desired level \(\alpha\), FWER is strongly controlled for testing \(k\) null hypotheses.

If all intersection null hypotheses \(H_{0\{i\}, i \in I}\) that imply \(H_{0\{i\}}\) are rejected, then the null hypothesis \(H_{0\{i\}} : \beta_i = 0\) is rejected. At most \(2^k - 1\) subset hypotheses are required to be tested to draw final conclusions for the \(k\) parameters.

Notice that the critical value \(c_{\alpha,\{1,\ldots,k\}}\) which satisfies equation (4.26) is the largest one among all critical values for \(2^k - 1\) subset tests, which corresponds to testing the intersection hypothesis involving all \(k\) parameters simultaneously, i.e. \(H_{0\{1,\ldots,k\}}\). If it is used to reject every single null hypothesis \(H_{0\{i\}} : \beta_i = 0\) when \(|S_i| > c_{\alpha,\{1,\ldots,k\}}\), a single-step multiple comparison with the control procedure can be constructed. This is indeed an extension of Dunnett’s method in the GEE model setting (Orelien et al., 2002).
4.4.5 Step-down The Partition Testing in GEE Setting

The number of partitioning hypotheses increases quickly with the number of parameters of interest \( k \). For instance, if \( k = 3 \), up to 7 subset hypothesis tests may be required. When \( k = 10 \), 1023 tests may be required. Even for moderate \( k \), some shortcut pathway must be taken.

Calian et al. established a set of sufficient conditions for a step-down shortcut version of the partition test (Calian et al., 2008). Now we check if the statistics

\[
|S_i| = \frac{|\hat{\beta}_i - 0|}{\hat{\sigma}_i}, i = 1, \ldots, k
\]

obtained from fitting the ‘comprehensive’ GEE model satisfy these sufficient conditions. Upon satisfaction of the conditions, we further show how to construct a valid step-down multiple comparison procedure.

D1: The test for \( H_{0I}^* \) is of the maxT form.

Check: To test \( H_0 : \beta_i = 0, i \in I \), the test statistics \( |S_i|, i \in I \) are ordered so that \( |S_{[1]}| \leq \ldots \leq |S_{[I]}| \). \( H_{0I}^* \) is rejected if \( \max_{i \in I} |S_i| > c_{\alpha,I} \). Therefore, the test statistic is in the ‘maxT’ form.

D2: \( \sup_{H_{0I}} P\{\max_{i \in I} |S_i| > c_I\} \leq \alpha \).

Check: The critical value for rejecting \( H_0 : \beta_i = 0, i \in I \) if \( \max_{i \in I} |S_i| > c_{\alpha,I} \) satisfies the equation

\[
1 - \int_{-c_{\alpha,I}}^{c_{\alpha,I}} \cdots \int_{-c_{\alpha,I}}^{c_{\alpha,I}} f(S_{i_1}, \ldots, S_{i_p}) dS_{i_1} \cdots dS_{i_p} = \alpha, \quad \{i_1, \ldots, i_p\} \subseteq I \quad (4.27)
\]

According to GEE theorem and Slutsky’s theorem, under mild conditions, asymptotically

\[
\text{diag}(\hat{V}_G)^{-\frac{1}{2}} n^\frac{1}{2}(\hat{\beta} - \beta) = \text{diag}(nI_0^{-1}I_1I_0^{-1})^{-\frac{1}{2}} n^\frac{1}{2}(\hat{\beta} - \beta) \rightarrow_d \text{MVN}(0, R_G) \quad (4.28)
\]
Therefore, $\sup_{H_{0I}^{*}} P\{\max_{i \in I} |S_i| > c_I\} \leq \alpha$ as long as the asymptotic distributional results hold.

D3: The values of the test statistics $S_i$, $i = 1, \ldots, k$, are not re-computed for different $H_{0I}^{*}$.

Check: Since we only consider one model, the so called ‘comprehensive’ GEE model is fit only once, and parameter estimates $\hat{\beta}$ and $\hat{V}_G$ do not change with $I$. Therefore the test statistic $\max_{i \in I} |S_i|$ for testing subset null hypotheses $H_{0I} : \beta_i = 0, i \in I$ does not change with $I$ either.

For example, no matter if $H_{0(23)} : \beta_2 = 0$ and $\beta_3 = 0$ or $H_{0(235)} : \beta_2 = 0$ and $\beta_3 = 0$ and $\beta_5 = 0$ is tested, the test statistics is the maximum of $|S_2| = \frac{\hat{\beta}_2}{\hat{\sigma}_2}, |S_3| = \frac{\hat{\beta}_3}{\hat{\sigma}_3}, |S_5| = \frac{\hat{\beta}_5}{\hat{\sigma}_5}$ obtained from the ‘comprehensive’ GEE model output. Therefore they are not re-computed no matter if $I = \{2, 3\}$ or $I = \{2, 3, 5\}$ is in the null hypothesis.

D4: Critical values $c_I$ have the property that if $J \subset I$ then $c_J \leq c_I$.

Check: Since $S_i, i \in I$ follow a multivariate normal distribution under the null hypotheses, it is easy to see that if $J \subset I$, $S_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j}, j \in J$ also follows a multivariate normal distribution with lower dimension $|J|$. Then it is trivially satisfied that if $J \subset I$, then $c_{\alpha,J} < c_{\alpha,I}$, as long as they satisfy equation (4.26).

Therefore, one set of sufficient conditions for shortcutting the partition testing (Calian et al., 2008; Huang and Hsu, 2007; Xu and Hsu, 2007) is satisfied by the proposed test statistics in the ‘comprehensive’ GEE model setting. We may construct a step-down shortcut version of multiple testing procedure based on partition testing in this model setting.
In general, for testing $k$ null hypotheses $H_{0i} : \beta_i = 0, i = 1, \ldots, k$. Let $[1], \ldots, [k]$ be the random indices such that $|S_{[1]}| < \ldots < |S_{[k]}|$. Let $\beta_{[i]}$ indicate the parameter of test interest corresponding to $S_{[i]}$. The step-down shortcut version tests in following steps:

Step 1: Test $H_{0\{[1],\ldots,[k]\}} : \beta_{[1]} = 0$ and $\beta_{[2]} = 0$ and $\ldots$ and $\beta_{[k]} = 0$.

If $|S_{[k]}| > c_{\alpha,\{1,\ldots,[k]\}}$ then infer $\beta_{[k]} \neq 0$ and go to step 2; else stop.

Step 2: Test $H_{0\{[1],\ldots,[k-1]\}} : \beta_{[1]} = 0$ and $\beta_{[2]} = 0$ and $\ldots$ and $\beta_{[k-1]} = 0$.

If $|S_{[k-1]}| > c_{\alpha,\{1,\ldots,[k-1]\}}$, then infer $\beta_{[k-1]} \neq 0$ and go to step 3; else stop.

\ldots

Step $k$ : Test $H_{0\{[1]\}} : \beta_{[1]} = 0$

If $|S_{[1]}| > c_{\alpha,\{[1]\}}$, then infer $\beta_{[1]} \neq 0$ and stop; else stop.

Therefore, once the hypothesis of parameter $\beta_{[k]}$ which is related to the largest test statistic is rejected due to $|S_{[k]}| > c_{\alpha,\{1,\ldots,[k]\}}$, it is guaranteed that $|S_{[k]}| > c_{\alpha,\{[i_1],\ldots,[i_r]\}}$ as long as $\{[i_1], \ldots, [i_r]\} \subseteq \{1, \ldots, k\}$. That is, $\beta_{[k]}$ can be removed from future hypothesis testing. Thus instead of testing up to $2^k - 1$ null hypotheses, at most $k$ null hypotheses are needed to make valid inference for $H_{0i} : \beta_i = 0, i = 1, \ldots, k$.

4.5 Example analysis

Now let us illustrate how to apply the step-down shortcut of the partition testing in GEE models using the toxicology experiment example. Recall the experiment had four treatment groups: control, low, medium and high dose. The odds of malformation versus normality can be modeled through the ‘comprehensive’ GEE model (4.10).
Through the logit link function, the GEE ‘comprehensive’ model is formulated as,

\[
\text{logit}(\pi_i) = \log \frac{P(Y_{ist} = 1)}{1 - P(Y_{ist} = 1)} = X_{st} \beta
\]  

(4.29)

Let \( \beta = (\beta_0, \beta_1, \beta_2, \beta_3)' \).

\[
\pi_i = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}; \text{ for } i = 0
\]

\[
\pi_i = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)}; \text{ for } i = 1, 2, 3
\]  

(4.30)

The parameter estimates \( \hat{\beta} \) as well as the variance-covariance estimate \( \hat{\mathbf{V}}_G \) can be obtained by fitting the ‘comprehensive’ GEE model through many statistical packages, such as SAS, or R, etc. R 2.7.0 package “gee” was used for following analysis. The ‘compound symmetric’ working correlation structure was used since similar extent of correlations between any pair of fetuses within the same litter was assumed.

We obtain the parameter estimates:

\[
\hat{\beta} = \begin{pmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\hat{\beta}_3
\end{pmatrix} = \begin{pmatrix}
-1.8171816 \\
-0.1265536 \\
1.1784401 \\
0.6603800
\end{pmatrix}
\]  

(4.31)

With the variance-covariance matrix estimate:

\[
\hat{\mathbf{V}}_G = \begin{pmatrix}
0.09641035 & -0.09641035 & -0.09641035 & -0.09641035 \\
-0.09641035 & 0.21495051 & 0.09641035 & 0.09641035 \\
-0.09641035 & 0.09641035 & 0.17798483 & 0.09641035 \\
-0.09641035 & 0.09641035 & 0.09641035 & 0.18753627
\end{pmatrix}
\]  

(4.32)

By using appropriate elements in the variance-covariance matrix, we compute the statistics \( s_1, s_2, s_3 \):
The estimated correlation matrix of $\hat{\beta}$ can be easily obtained:

$$
\hat{R}_G = \begin{pmatrix}
1.0000000 & -0.6697189 & -0.7359873 & -0.7170001 \\
-0.6697189 & 1.0000000 & 0.4929047 & 0.4801885 \\
-0.7359873 & 0.4929047 & 1.0000000 & 0.5277029 \\
-0.7170001 & 0.4801885 & 0.5277029 & 1.0000000 \\
\end{pmatrix}
$$

(4.34)

According to GEE theorem (Zeger and Liang, 1986; Liang and Zeger, 1986), $S_i = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_i}, i = 1, 2, 3$ follow the joint distribution MVN(0, $r_G$), where $r_G$ is part of $R_G$ matrix that consists the $i$th rows and columns of $R_G, i = 1, 2, 3$,

$$
\hat{r}_G = \begin{pmatrix}
1.0000000 & 0.4929047 & 0.4801885 \\
0.4929047 & 1.0000000 & 0.5277029 \\
0.4801885 & 0.5277029 & 1.0000000 \\
\end{pmatrix}
$$

(4.35)

The step-down version of the partition testing proceeds as follows:

We first order $S_i$ according to their absolute values. Let $[1], [2], [3]$ be the random indices such that $|S_{[1]}| \leq |S_{[2]}| \leq |S_{[3]}|$. Therefore

$$
S_{[1]} = S_1, S_{[3]} = S_2, S_{[2]} = S_3
$$

(4.36)

Let $\beta_i, i = 1, 2, 3$ be $\beta_i$ that corresponds to each $S_{[i]}$. Therefore, $\beta_1 = \beta_{[1]}, \beta_2 = \beta_{[3]}, \beta_3 = \beta_{[2]}$.

Step 1: Test $H_0([1],[2],[3]): \beta_1 = 0$ and $\beta_2 = 0$ and $\beta_3 = 0$.

Recall that as in equation (4.26), the critical value can be obtained based on the joint distribution of $S_1, S_2, S_3$,

$$
1 - \int_{-c_{(1,2,3)}}^{c_{(1,2,3)}} \int_{-c_{(1,2,3)}}^{c_{(1,2,3)}} \int_{-c_{(1,2,3)}}^{c_{(1,2,3)}} f(S_1, S_2, S_3) dS_1 dS_2 dS_3 = \alpha
$$

(4.37)
Assuming $\alpha = 0.05$, $c_{\{1,2,3\}} = 2.348$. Since

$$\max_{i \in \{1,2,3\}} |S_1| = |S_3| = |S_2| = 2.79 > c_{\{1,2,3\}} = 2.348, \quad (4.38)$$

we reject $H_{0\{\{1\},\{2\},\{3\}\}}$, infer $\beta_2 \neq 0$ and go to the next step.

**Step 2:** Test $H_{0\{\{1\},\{2\}\}} : \beta_1 = 0$ and $\beta_2 = 0$ by comparing the second largest test statistic $S_{[2]}$ to the critical value $c_{\{1,3\}}$, which satisfies

$$1 - \int_{-c_{\{1,3\}}}^{c_{\{1,3\}}} \int_{-c_{\{1,3\}}}^{c_{\{1,3\}}} f(S_1, S_3) dS_1 dS_3 = \alpha \quad (4.39)$$

Since

$$|S_{[2]}| = |S_3| = 1.525 < c_{\{1,3\}} = 2.214 \quad (4.40)$$

we cannot reject the null and stop therefore at this step.

As shown in formula (4.25), a p-value for testing each subset null hypothesis can be obtained by integration over the area defined by the test statistics. For example, the p-value for testing $\beta_2 = 0$ equals

$$1 - \int_{-S_2}^{S_2} \int_{-S_2}^{S_2} \int_{-S_2}^{S_2} f(S_1, S_2, S_3) dS_1 dS_2 dS_3 = 0.0144 \quad (4.41)$$

The $1 - \alpha$ confidence interval for parameters of interest can be easily obtained by $[\hat{\beta}_i - c_i \hat{\sigma}_i, \hat{\beta}_i + c_i \hat{\sigma}_i]$.

Collating the above test results, the final conclusion is we reject the null hypothesis $H_{02} : \beta_2 = 0$ but not $H_{01} : \beta_1 = 0$ or $H_{03} : \beta_3 = 0$. Therefore, there is evidence indicating that the malformation probability of the medium dose group is significantly higher than the control group. However, there is not enough evidence to show the low
dosage or the high dosage group had significantly different malformation probabilities compared to the control group.

To construct a single-step multiple test with the control procedure, we could simply use the critical value obtained from the first step, i.e. \( c_{\{1,2,3\}} = c^D = 2.348 \) for testing all three hypotheses. Since

\[
|S_2| = 2.79 > c^D = 2.348 \\
|S_3| = 1.525 < c^D = 2.348 \\
|S_1| = 0.273 < c^D = 2.348
\]

we reject the hypothesis that \( H_{02} : \beta_2 = 0 \) but not \( H_{01} \) or \( H_{03} \). Therefore, for this particular data set, the extended single-step Dunnett’s method reached the same conclusion as the step-down procedure.

Finally, recall that Holm proposed the well-known and widely used step-down testing procedure which is based on the Bonferroni inequality (Holm, 1979). This method is valid regardless of the joint distribution of the test statistics.

Although we could model the toxicity data by GEE and obtain the joint distribution of the test statistics analytically, we could still apply Holm’s method. Compared to the step-down method and the single step Dunnett’s method, Holm’s method ignores the estimable joint distribution of the test statistics. On the other hand, compared to the single-step Dunnett’s method, Holm’s method is a step-down shortcut procedure for partition testing. It would be interesting to explore how Holm’s method performs in current example.

Let \( c_{1,2,3}^H = c_{\alpha/3} \) satisfy

\[
1 - \int_{-c_{1,2,3}^H}^{c_{1,2,3}^H} \int_{-c_{1,2,3}^H}^{c_{1,2,3}^H} \int_{-c_{1,2,3}^H}^{c_{1,2,3}^H} f(S_1, S_2, S_3) dS_1 dS_2 dS_3 = \alpha/3
\]

(4.43)
and so forth for $c^H_{1,3} = c_{a/2}, c^H_1 = c_a$.

Since $c^H_{1,2,3} = c_{a/3} = 2.394, c^H_{1,3} = c_{a/2} = 2.241, c^H_1 = c_a = 1.960$, the Holm’s procedure proceeds as:

Step 1: Since $|S_2| = |S_3| = 2.79 > c^H_{1,2,3} = 2.394$, reject $H_{02}$ and infer $\beta_2 \neq 0$ and go to step 2;

Step 2: Since $|S_2| = |S_3| = 1.53 < c^H_{1,3} = 2.241$, do not reject $H_{03}$ and stop here.

Therefore, Holm’s method reached the same conclusion as the other two methods.

The parameter estimates with corresponding test statistics, critical values, p-values and confidence intervals obtained from the step-down partition testing, the single-step Dunnett’s method for GEE model, as well as the Holm’s step-down method are summarized in tables 4.2, 4.3 and table 4.5.

Figure 4.1 further shows the test statistics and critical values from the step-down partition test, the single-step extended Dunnett’s method and Holm’s method graphically.

By comparing results from the step-down partition testing and the single-step Dunnett’s method in the GEE model setting, notice that the step-down method gives rise to critical values that are not larger than the ones from the single-step method. Consequently the confidence intervals $\beta_i \in [\hat{\beta}_i - c_i\hat{\sigma}_i, \hat{\beta}_i - c_i\hat{\sigma}_i]$ obtained by the step-down method are not wider than those from the single-step Dunnett’s method. This potentially indicates that the step-down method is more powerful in terms of making more positive findings compared to the single-step method.
<table>
<thead>
<tr>
<th>Dose</th>
<th>$\hat{\beta}_i$</th>
<th>Test Statistic</th>
<th>Critical Value</th>
<th>P-value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Control</td>
<td>-0.1266</td>
<td>0.2730</td>
<td>1.960</td>
<td>0.7849</td>
<td>(-1.035, 0.782)</td>
</tr>
<tr>
<td>Medium-Control</td>
<td>1.1784</td>
<td>2.7933</td>
<td>2.348</td>
<td>0.0144</td>
<td>(0.187, 2.169)</td>
</tr>
<tr>
<td>High-Control</td>
<td>0.6604</td>
<td>1.5249</td>
<td>2.214</td>
<td>0.2215</td>
<td>(-0.299, 1.619)</td>
</tr>
</tbody>
</table>

Table 4.2: Test statistics, critical values, P-values and confidence intervals by step-down shortcut of partition testing for data from (Paul, 1982)

<table>
<thead>
<tr>
<th>Dose</th>
<th>$\hat{\beta}_i$</th>
<th>Test Statistic</th>
<th>Critical Value</th>
<th>P-value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Control</td>
<td>-0.1266</td>
<td>0.2730</td>
<td>1.960</td>
<td>0.9861</td>
<td>(-1.216, 0.963)</td>
</tr>
<tr>
<td>Medium-Control</td>
<td>1.1784</td>
<td>2.7933</td>
<td>2.348</td>
<td>0.0144</td>
<td>(0.187, 2.169)</td>
</tr>
<tr>
<td>High-Control</td>
<td>0.6604</td>
<td>1.5249</td>
<td>2.348</td>
<td>0.2928</td>
<td>(-0.357, 1.678)</td>
</tr>
</tbody>
</table>

Table 4.3: Test statistics, critical values, P-values and confidence intervals by single-step Dunnett’s method for data from (Paul, 1982)

<table>
<thead>
<tr>
<th>Dose</th>
<th>$\hat{\beta}_i$</th>
<th>Test Statistic</th>
<th>Critical Value</th>
<th>P-value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Control</td>
<td>-0.1266</td>
<td>0.2730</td>
<td>1.960</td>
<td>0.7849</td>
<td>(-1.035, 0.782)</td>
</tr>
<tr>
<td>Medium-Control</td>
<td>1.1784</td>
<td>2.7933</td>
<td>2.394</td>
<td>0.0157</td>
<td>(0.106, 2.188)</td>
</tr>
<tr>
<td>High-Control</td>
<td>0.6604</td>
<td>1.5249</td>
<td>2.241</td>
<td>0.2545</td>
<td>(-0.510, 1.631)</td>
</tr>
</tbody>
</table>

Table 4.4: Test statistics, critical values, P-values and confidence intervals by Holm’s method for data from (Paul, 1982)
Figure 4.1: Toxicity example result: test statistics and critical values from the step-down partition test, the single-step extended Dunnett’s method and Holm’s method.
The reason the step-down method is able to make more positive findings is that in each step, it is guarding against the possibility that fewer and fewer null hypotheses may be true. Recall that the step-down method starts with testing the intersection null hypothesis that involves all \( k \) parameters, i.e. \( H_{0\{1,\ldots,k\}} : \beta_1 = 0 \) and \( \beta_2 = 0 \) and \ldots and \( \beta_k = 0 \). In this step, it uses the same critical value as the single-step Dunnett’s method does, \( c_{\{1,\ldots,k\}} \), which is the most stringent one.

If the sufficient conditions for short-cutting are satisfied, if the largest test statistic \( |S_{\{k\}}| > c_{\{1,\ldots,k\}} \), not only \( H_{0\{1,\ldots,k\}} \) can be rejected, but all subset null hypotheses implying \( \beta_{\{k\}} = 0 \) will also be automatically rejected. Since the parameter space of \( \beta_{\{k\}} = 0 \) is the union of all such subsets, it can be rejected and removed from future testing procedures. Therefore, the step-down procedure only tests for \( k - 1 \) parameters, \( \beta_{\{1\}}, \ldots, \beta_{\{k-1\}} \) in the second step, and so forth. As a result, fewer and fewer null hypotheses need to be guarded for erroneous rejection in each step. This is why it is called a ‘step-down’ procedure.

The critical values for guarding against fewer possibly true null hypotheses are less conservative than ones adjusting for all \( k \) null hypotheses. The latter is exactly what is used for testing every single null hypothesis by the single-step Dunnett’s method. The step-down method can be thought of controlling the false rejection rate conditioning on the real data. That is, it adjusts for critical values conditional on the data. Thus it is less stringent but still controls FWER.

Comparing Holm’s step-down procedure with the above two methods, one can see the critical value \( c^H_{\{1,2,3\}} = 2.394 \) is larger than the largest critical value used by step-down partitioning method or single-step Dunnett’s method (\( c_{\{1,2,3\}} = c^D = 2.348 \)).
The second largest critical value by Holm’s method $c_{(1,3)}^H = 2.241$ is also larger than the second largest critical value by the step-down method, where $c_{(1,3)} = 2.214$.

The reason is that Holm’s method is based on the Bonferroni’s inequality, which does not take the joint distributional of the test statistics into account. The critical values are obtained by dividing the $\alpha$ level by the number of null hypotheses in the subset. Therefore, for test statistics that are positively correlated, Holm’s method will be more conservative compared to the methods that take the positive correlation into account. However, since Holm’s method is a step-down method, it is still less conservative in testing the second largest test statistic compared to Dunnett’s method, because it makes adjustment for two null hypotheses only. Therefore $c_{(1,3)}^H < c^D$.

In this example, we take the approach of modeling and computing the critical values based on the joint distribution of the test statistics analytically. The critical values are computed in an exact way in the sense that one could find the sharpest possible critical values to control FWER at level-$\alpha$. The benefit of doing this is the test constructed based on these critical values is the most powerful one in terms of making positive discoveries compared to other tests not making use of the joint distribution information.

We discussed in detail how to construct a step-down partitioning test procedure in the GEE model setting. For other modeling tools, if the parameter estimates follow a multivariate normal distribution asymptotically, it is also possible to apply the current method in those modeling settings. For example, in logistic regression analysis, we can use similar step-down short-cut procedure, which is discussed in Chapter 3. For GLMMs and LMMs, similar procedures may be developed as well.
However, notice that the modeling approach is based on the asymptotic property of the parameter estimates. This requires that the model is able to be fit on the observed data and all the model assumptions are satisfied. For GEE models, the conditions are very mild. It is only required that 1) The correlation matrix of parameter estimates converges to a fixed correlation matrix; and 2) While the number of ‘clusters’ \( n \) goes to infinity, the maximum cluster size \( m_s \) and the number of correlation related parameters is finite. This requires the number of subjects to be large compared to the number of measurements within each subject. This could be problematic some times.

For example, when thousands of genes are measured on dozens of patients, fitting a comprehensive model including all genes is infeasible. In this case, one model could be fit for each gene. The test statistics or p-values are obtained for all genes separately. But it may not be plausible to obtain the correlation structure between all test statistics analytically due to large number of parameters. In this case, Holm’s method can be used. The result is guaranteed to be conservative.
CHAPTER 5

CONCLUSION AND FUTURE WORK

Multiple comparisons are common practice in statistical analysis. How to construct valid and powerful multiple testing procedures within the discrete data analysis setting are of special interest. In this thesis work, we discuss the limitations of permutation tests in controlling FWER for testing multiple null hypotheses on the marginal differences. Also, we discuss how to construct the step-down short-cut version of the partitioning test based on modeling in the discrete data setting once certain sufficient conditions are satisfied.

Multiple testing for significant associations between predictors and responses has a wide array of applications. One such application is pharmacogenomics, where testing for association between responses such as phenotype and a large number of genetic markers is of interest. Permuting response group labels to generate a reference distribution is often thought of as a convenient thresholding technique that automatically captures dependence in the data. For continuous predictors, it has been shown in recent publications that non-trivial model assumptions are required for permutation testing to control multiple testing error rates (Huang et al., 2006).

This work considers the case of binary predictors (such as mutation status of SNPs). By utilizing simplified examples where 2 SNPs are of comparison interest,
it is shown that, similar to the continuous predictor case, without non-trivial model assumptions (the MDJ condition (Xu and Hsu, 2007)), permutation tests based multiple testing procedures may not control unconditional multiple testing error rates. In addition, since permutation testing is explicitly conditional in nature in the binary predictors case, we also consider the conditional error rate and show lack of control of it by permutation tests if the MDJ condition is not satisfied.

Whenever modeling is possible, analytical derivation of the joint distributions of the test statistics may be feasible. To model independent binary outcome variables, the logistic regression model is a useful tool. To model correlated binary outcome variables, marginal models using GEE are often considered. Examples of multiple comparisons with the control are used for both independent and correlated binary outcome variables. We show that upon satisfaction of certain sufficient conditions, compared to single step method such as the extension of Dunnett’s method, and simpler step-wise multiple testing procedure, such as Holm’s procedure, one can construct more powerful step-down short-cut versions of the partitioning test based on the asymptotic joint distributions of the test statistics from modeling. We also discuss why testing procedures based on certain types of test statistics may not satisfy the sufficient conditions and thus are not able to short-cut the partitioning test to reduce number of tests.

The work in the permutation test also introduces challenging questions for the future research. One question is how to check the satisfaction of the MDJ condition in the discrete data setting, where the concept of residuals is not as clear as in the continuous data setting. The other interesting direction is to develop new modeling techniques for high dimensional discrete outcome variables. Further studies to explore
whether control of other types of multiple testing error rates, such as the FDR, will be of great interest as well.


