EXACT APPROACHES FOR BIAS DETECTION AND AVOIDANCE WITH SMALL, SPARSE, OR CORRELATED CATEGORICAL DATA

by

Sarah E. Schwartz

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences (Statistics Specialization)

Approved:

Chris Corcoran, Sc.D.
Major Professor

D. Richard Cutler, Ph.D.
Committee Member

Daniel Coster, Ph.D.
Committee Member

Kady Schneiter, Ph.D.
Committee Member

Thomas Ledermann, Ph.D.
Committee Member

Mark McLellan, Ph.D.
Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2017
ABSTRACT

Exact Approaches for Bias Detection and Avoidance with
Small, Sparse, or Correlated Categorical Data

by

Sarah E. Schwartz, Doctor of Philosophy
Utah State University, 2017

Major Professor: Dr. Chris Corcoran
Department: Mathematics and Statistics

Every day, traditional statistical methodology are used world wide to study a variety of topics and provides insight regarding countless subjects. Each statistical model is based on assumptions which must be met to ensure valid results. Additionally, many statistical approaches rely on large sample asymptotic and may collapse or degenerate in the presence of small, spare, or correlated data. This dissertation details several advancements to detect these conditions, avoid their consequences, and analyze categorical data in a new way to yield trustworthy results.

One of the most commonly used modeling techniques for binary outcomes is logistic regression. While the problem of complete separation is widely known, the consequences of near separability are not. Many investigators are unaware that their particular data, while not completely separable, are at risk of bias since detecting near separability is not as straight forward.

We have developed a routine for determining if concern regarding standard maximum likelihood estimates (MLE) in logistic regression is warranted. This diagnostic tool signals when data are small, spare, or correlated to the degree that
the analyst should consider penalized maximum likelihood estimation (Firth’s bias correction) or an exact approach for analysis.

Correlated data may arise from common situations such as multi-site medical studies, research on family units, or investigations on students within classrooms. In these circumstance the associations between cluster members must be included in any statistical analysis testing the hypothesis of a connection between predictor and response in order for results to be valid.

Previously investigators had to choose between using a method intended for small or sparse data while assuming independence between observations (Chi Squared Test for Independence) or a method that allowed for correlation between observations, such as generalized estimating equations (GEE) and multilevel mixed effects models (MLM), but that is only applicable for larger samples. We present a new method that allows for small, clustered samples to be assessed for evidence of a relationship between a binary predictor and a multinominal outcome.
Exact Approaches for Bias Detection and Avoidance with Small, Sparse, or Correlated Categorical Data
Sarah E. Schwartz

Every day, traditional statistical methodology are used worldwide to study a variety of topics and provides insight regarding countless subjects. Each technique is based on a distinct set of assumptions to ensure valid results. Additionally, many statistical approaches rely on large sample behavior and may collapse or degenerate in the presence of small, sparse, or correlated data. This dissertation details several advancements to detect these conditions, avoid their consequences, and analyze data in a different way to yield trustworthy results.

One of the most commonly used modeling techniques for outcomes with only two possible categorical values (e.g., live/die, pass/fail, better/worse, etc.) is logistic regression. While some potential complications with this approach are widely known, many investigators are unaware that their particular data does not meet the foundational assumptions, since they are not easy to verify. We have developed a routine for determining if a researcher should be concerned about potential bias in logistic regression results, so they can take steps to mitigate the bias or use a different procedure altogether to model the data.

Correlated data may arise from common situations such as multi-site medical studies, research on family units, or investigations on student achievement within classrooms. In these circumstance the associations between cluster members must be included in any statistical analysis testing the hypothesis of a connection between two variables in order for results to be valid.
Previously investigators had to choose between using a method intended for small or sparse data while assuming independence between observations or a method that allowed for correlation between observations, while requiring large samples to be reliable. We present a new method that allows for small, clustered samples to be assessed for a relationship between a two-level predictor (e.g., treatment/control) and a categorical outcome (e.g., low/medium/high).
I dedicate this to my four daughters: Allie, Kylie, Zoey, and Megan.
ACKNOWLEDGMENTS

I would like to first thank Dr. Chris Corcoran for his instrumental role as my mentor throughout my collegiate education spanning more than a decade. I would also like to thank my committee members, Drs. Richard Cutler, Dan Coster, Kady Schneiter, and Thomas Ledermann for their support and assistance throughout the entire process.

This work would not be possible without the pioneering advancements of the co-founders of Cytel Inc., Drs. Cyrus Mehta and Nitin Patel. Senior Vice President for Research & Development, Dr. Pralay Senchaudhuri’s input has been invaluable in this work. I would also like to thank Sumit Singh and Ashwini Kulkarni, of the development team, who have provided technical support throughout.

I wish I knew how my amazing parents, Dr. Paul and Martha Tew, raised my sisters and I to believe we were brilliant, hard workers able to tackle any problem, despite evidence to the contrary. I wish to pass on the grit, determination, and self-confidence they instilled in me to my children, as it has seen me through many challenges.

Last, but far from least, I thank my husband, Nathan Schwartz, and our four daughters who have supported me through many years of long days and nights. This has been a team effort!

Sarah E. Schwartz
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.2 Multinomial Distributions with Ordered Categories</td>
<td>56</td>
</tr>
<tr>
<td>3.4.3 Exact Trend Test for Correlated Multinomial Data (ETT-CMD)</td>
<td>59</td>
</tr>
<tr>
<td>3.4.4 Additional Conditioning for the Network Algorithm</td>
<td>60</td>
</tr>
<tr>
<td>3.4.5 Specification of the Correlation Conditioning</td>
<td>62</td>
</tr>
<tr>
<td>3.5 Application</td>
<td>64</td>
</tr>
<tr>
<td>3.5.1 Clinical Trial for Pain Medication</td>
<td>64</td>
</tr>
<tr>
<td>3.5.2 Toxicology Study on Ethylene Glycol in Mice</td>
<td>65</td>
</tr>
<tr>
<td>3.6 Conclusion</td>
<td>69</td>
</tr>
<tr>
<td>4 CONDITIONING CONSIDERATIONS FOR THE EXACT TREND TEST</td>
<td>70</td>
</tr>
<tr>
<td>4.1 Background</td>
<td>70</td>
</tr>
<tr>
<td>4.2 Motivation</td>
<td>72</td>
</tr>
<tr>
<td>4.2.1 Correlated Multinomial Variate Generating Mechanisms</td>
<td>72</td>
</tr>
<tr>
<td>4.2.2 Selected Method for Sample Generation</td>
<td>75</td>
</tr>
<tr>
<td>4.3 Simulation Study</td>
<td>79</td>
</tr>
<tr>
<td>4.3.1 Model Parameterization and Data Generation</td>
<td>80</td>
</tr>
<tr>
<td>4.3.2 Results and Discussion</td>
<td>82</td>
</tr>
<tr>
<td>4.4 Conclusion</td>
<td>85</td>
</tr>
<tr>
<td>5 CONCLUSION</td>
<td>88</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>91</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>97</td>
</tr>
<tr>
<td>Appendix A: R Code for Chapter 2</td>
<td>98</td>
</tr>
<tr>
<td>Appendix B: R Code for Chapter 3</td>
<td>128</td>
</tr>
<tr>
<td>Appendix C: R Code for Chapter 4</td>
<td>134</td>
</tr>
<tr>
<td>VITA</td>
<td>148</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Illustrative Dataset: Fifth Observation with Variable Value $k$.</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Collett’s ESR Data: Effect of Near Separation on the MLE and SE.</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Generic $R \times C$ Contingency Table, without Clustering</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Simple 2x2: Observed Contingency Table</td>
<td>27</td>
</tr>
<tr>
<td>2.5 Simple 2x2: Logistic Regression MLEs for Observed Contingency Table.</td>
<td>27</td>
</tr>
<tr>
<td>2.6 Simple 2x2: MLE and PMLE Estimates with wCI and plCIs for Observed Contingency Table.</td>
<td>28</td>
</tr>
<tr>
<td>2.7 Simple 2x2: Generic Form of Tables in the Reference Set.</td>
<td>30</td>
</tr>
<tr>
<td>2.8 Simple 2x2: Ranking of the Reference Set.</td>
<td>31</td>
</tr>
<tr>
<td>2.9 Collett’s ESR data: MLE and PMLE Estimates with wCI and plCIs.</td>
<td>38</td>
</tr>
<tr>
<td>3.1 CHED: Number of Rejected Corneal Grafts.</td>
<td>45</td>
</tr>
<tr>
<td>3.2 Multicenter Pain and Spasm Study: Observed Contingency Table.</td>
<td>47</td>
</tr>
<tr>
<td>3.3 Generic $2 \times C$ Contingency Table for a Binary Outcome, with Clustering</td>
<td>52</td>
</tr>
<tr>
<td>3.4 Congenital Eye Disease and Corneal Grafts: Observed $2 \times C$ Contingency Table, with Clustering</td>
<td>53</td>
</tr>
<tr>
<td>3.5 Congenital Eye Disease and Corneal Grafts: Exact Permutation Distribution for the Test Statistic for the Exact Trend Test for Correlated Binomial Data (ETT-CBD)</td>
<td>54</td>
</tr>
<tr>
<td>3.6 Generic $2 \times C$ Contingency Table for a Multinomial Outcome, with Clustering.</td>
<td>56</td>
</tr>
<tr>
<td>3.7 Multicenter Pain and Spasm Study: Crossproducts for $\delta_{jk}$ Sufficient Statistics</td>
<td>65</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>3.8</td>
<td>Multicenter Pain and Spasm Study: Exact p-values for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)</td>
</tr>
<tr>
<td>3.9</td>
<td>Toxicology Study on Ethylene Glycol in Mice: Observed Contingency Table.</td>
</tr>
<tr>
<td>3.10</td>
<td>Toxicology Study on Ethylene Glycol in Mice: Exact p-values for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)</td>
</tr>
<tr>
<td>4.1</td>
<td>Randomly Generated Clusters of ( n = 10 ) Observations and Univariate Probabilities of ( p = (0.2, 0.3, 0.5) ) to Investigate the Behavior of Intracluster Dispersion Specification.</td>
</tr>
<tr>
<td>4.2</td>
<td>Randomly Generated Clusters of Various Cluster Sizes in Conjunction with Univariate Probabilities of ( p = (0.2, 0.3, 0.5) ) and Overdispersion Parameterization ( \theta_{12} = 0.90, \theta_{23} = 0.85, \theta_{13} = 0.75 ).</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation: Two Univariate Probability Distributions</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation: Three Sets of Dispersion Parameters</td>
</tr>
<tr>
<td>4.5</td>
<td>Simulation: An Example Dataset Generated for Ascertaining Power in the Presence of Overdispersion within Clusters of Size ( n = 8 ).</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.1</td>
<td>Illustration of Separation in Two and Three Dimensions.</td>
</tr>
<tr>
<td>2.2</td>
<td>SAS warning message: Complete separation.</td>
</tr>
<tr>
<td>2.3</td>
<td>Illustrative Dataset: Covariate Space.</td>
</tr>
<tr>
<td>2.4</td>
<td>Illustrative Dataset: Effect of Near Separation on the MLE and SE.</td>
</tr>
<tr>
<td>2.5</td>
<td>Collett’s ESR data: Covariate Space.</td>
</tr>
<tr>
<td>2.6</td>
<td>Illustration of Simulation Study: Characteristics of MLE.</td>
</tr>
<tr>
<td>2.7</td>
<td>Collett’s ESR data: Convex Hull.</td>
</tr>
<tr>
<td>2.8</td>
<td>Illustration of a Generic Profile Likelihood Confidence Interval.</td>
</tr>
<tr>
<td>2.9</td>
<td>Illustration of plCIs for Logistic Regression.</td>
</tr>
<tr>
<td>2.10</td>
<td>Simple 2x2: All Possible Tables with N = 20 and Half exposed (x = 1).</td>
</tr>
<tr>
<td>2.11</td>
<td>Simple 2x2 Crosstabulation: Probability Density of the Reference Set.</td>
</tr>
<tr>
<td>2.12</td>
<td>Simple 2x2: Simulation to Study Relationship Between plCI Asymmetry Ratio and Absolute Value of MLE Bias.</td>
</tr>
<tr>
<td>2.13</td>
<td>Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation.</td>
</tr>
<tr>
<td>2.14</td>
<td>Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation and Sample Balance, with Logarithmic Scales.</td>
</tr>
<tr>
<td>2.15</td>
<td>Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation and Sample Balance, with Restricted Range Scales.</td>
</tr>
<tr>
<td>2.16</td>
<td>Collett’s ESR data: Asymmetry of Profile Likelihood Confidence Intervals.</td>
</tr>
</tbody>
</table>
3.1 Multicenter Pain and Spasm Study: Exact Distribution of the Exact Trend Test for Correlated Multinomial Data (ETT-CMD) ........ 66

3.2 Toxicology Study on Ethylene Glycol in Mice: Exact Distribution of the Test Statistic for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD) ......................... 68

4.1 Simulation: Example Dataset Exact Distribution of the Test Statistic Under Each of the Correlation Conditions of the ETT-CMD .... 84

4.2 Simulation: Comparing the Three Correlation Conditioning Options Regarding Power and Type I Error Rate. ...................... 86
“Big data enchants us with the promise of new insights,” says Raymond Yee, a self-titled ‘data architect’ at US Berkeley School of Information (Dutcher, 2014). Yet he cautions, “let’s not forget the knowledge hidden in the small data right before us.” The same problem that has occupied statisticians and scientists for centuries remains. Insight, inference, and implementing augmentations to systems for the better. “We have a new resource here,” says Professor David Hand of Imperial College London (Harford, 2014) of the big data movement, “but nobody wants ‘data’. What they want are the answers.”

In today’s data-rich world, many big-data analyses are simply comprised of a large number of small-sample problems. Regarding these, Fisher (1925) observed that,

\[
\text{The traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small problems on their merits does it seem possible to apply accurate tests to practical data.}
\]

Likewise, increased computing power has had a profound effect on the gathering of information. Even by today’s standards, data considered moderate in size often include a large number of covariates and involve more complex designs. Many big data problems boil down to repeatedly solving small data problems. When modeling proportions, simply gathering more data does not necessarily lead to well-behaved asymptotic likelihood-based procedures. Particularly when many covariates are measured, the data may be sparse or unbalanced. While exact conditional
methods alleviate some of the drawbacks of asymptotic inference, they may be computationally infeasible. Further research is needed in extending the exact methodology to larger data and more complex designs, as well as evaluating the relative characteristics of existing exact and asymptotic methods. The present work attempts to address problems in each of these areas.

In chapter 2, we will present a diagnostic tool to detect potential risk of bias for point estimates involving binary data. Logistic regression is the most frequently used model for binary data and has widespread applicability in the health, behavioral, and physical sciences. Unconditional logistic regression is widely used to model the probability of a binary response as a function of covariates; however, inferences are based on the asymptotic properties of maximum likelihood statistics. Maximum likelihood estimates (MLE) are reliable for problems with large samples and when the proportion of responses is neither too small or too large.

It is well documented that MLE are not reliable for small, sparse or unbalanced datasets and several methods for handling such situations appear in the literature. One of the most promising methods is a first-order bias correction approach proposed originally by Firth (1993), which is now available in most statistical software packages. In addition to being computationally efficient, Firth’s method, which maximized a penalized likelihood function, also guarantees that the parameter estimates will be finite. While methods for modeling such problematic data exist, little attention has been paid to detecting such conditions. Only data that exhibits complete separability raise explicit alarm signals in most software packages (Allison, 2003). Near separation is actually more worrisome because it can easily go undetected in the course of model fitting using maximum likelihood, except for the most extreme cases, which result in absurdly large estimates.

Our approach utilizes the profile-likelihood confidence interval for point estimates. While the standard Wald-type confidence intervals (wCI) are inherently
symmetric since they are based on asymptotic approximation distribution of the log likelihood ratio test statistic, profile likelihood based confidence intervals (plCI) may exhibit asymmetry due to inverting the likelihood ratio test. We define a new ratio statistic which is the quotient of the length of the longer and the shorter sides of the interval around the estimate. While computation of plCIs is executable within many software packages, we show that the degree of asymmetry of this interval is directly related to the potential for bias.

The exact trend test for correlated binary data (ETT-CBD) developed by Corcoran et al. (2001) will be extended in Chapter 3. We have developed an exact trend test for correlated multinomial data (ETT-CMD). Correlated observations arise in a variety of settings, including those in which multiple or repeated responses are measured on the same individual, or in which responses are clustered, sharing some common underlying latent factor (e.g. children from the same family, eyes or ears from the same individual). The analysis of such designs has received much attention in the literature during recent decades. While there has been significant progress made in the development and implementation of exact testing procedures for uncorrelated data, few exact alternatives exist for investigators who require valid testing procedures when faced with small or sparse samples of correlated binary data.

While Corcoran’s trend test has already been derived from the model of Molenberghs and Ryan (1999) for clustered multivariate binary data, it only allows for two-level outcomes. Because the model is a member of the exponential family, we can readily condition on its sufficient statistics in order to eliminate the nuisance parameters, including the dispersion parameter, which models the intraclass correlation, under the null hypothesis of no covariate effect. We can therefore maintain the underlying correlation structure of the data without needing to estimate the degree of overdispersion. This method implicitly enumerates
all possible permutations of the observed data, subject to the constraints of the observed sufficient statistics, using a graphical network representation. This allows efficient recovery of the exact tail distribution. Our approach builds upon the network-based methods pioneered by Mehta, Patel and Senchaudhuri (1992).

Chapter 4 investigates three distinct correlation conditioning constraint options available for computing the exact distribution of the test statistic in the ETT-CMD developed in the second paper. Of primary concern is the sensitivity of the trend test to the assumptions pertaining to clusters arising from non-constant pairwise correlations between outcome levels. Simulations under various conditions investigate the power resulting from the three possible correlation constraints as well as the potential trade off in terms of type I error rate inflation.

No matter the size of data at hand, we echo George Box’s familiar refrain: “all models are wrong, but some are useful” (Box, 1976). Despite the usefulness of logistic regression, maximum likelihood estimates are vulnerable to bias due to small or sparse data. We hope that the tools within this document aid researchers in detecting these conditions, thus allowing them to circumvent erroneous conclusions. Notwithstanding the utility of generalized estimation equations, realized samples may constrain application. We offer an alternative for clustered data when finite sample sizes prohibit asymptotic machinery. Let us not be as Rutherford D. Roger warned, “drowning in information and starving for knowledge”.
CHAPTER 2
POTENTIAL BIAS DIAGNOSTIC TOOL

*The statistician cannot evade the responsibility for understanding the processes he applies or recommends.*
*Sir Ronald A. Fisher in The Design of Experiments, 1935*

**Background**

Categorical outcomes are ubiquitous in many areas of research, and generalized linear models (GLMs) represent the most widely applied methodology for testing associations between categorical variables and fixed investigative factors. Logistic regression in particular is the most frequently used model for binary data and has widespread applicability in the health, behavioral, and physical sciences. King and Ryan (2002) stated there were 2,770 research papers published in 1999 in which “logistic regression” was in the title of the paper or among the keywords. In fact, King and Zeng (2001) referred to the use of the maximum likelihood method in logistic regression as “the nearly universal method”.

Maximum likelihood estimates (MLE) for logistic regression are based on large sample approximations that are reliable for problems with large samples and when the proportion of responses is not too small or too large. However, it has been known for several years that MLE are not reliable for small, sparse or unbalanced datasets, with the latter referring to a considerable difference between the number of zeros and ones of the response variable (see, e.g., Jennings (1986)). Exact logistic regression is a computationally intensive method developed by Mehta...
and Patel (1995) that is often useful in such situations. Corcoran et al. (2001) discuss four published datasets that demonstrate its value in applications.

A phenomenon referred to as separation results in unbounded maximum likelihood estimates when fitting a GLM to certain types of datasets (see panel A in Figure 2.1). For example when modeling binomials, separation is said to occur in a dataset when a hyperplane can be constructed in the covariate space such that only responses ($Y = 1$) lie in one (open) half space generated by the hyperplane and only non-responses ($Y = 0$) lie in the other (open) half space (see panel B in Figure 2.1). Allison et al. (2004) provides a clear exposition of this phenomenon and an explanation of why it occurs often in practice. Heinze and Schemper (2002) examined separation when there are binary covariates and showed that the probability of separation depends on sample size, the number and balance of binary covariates, and the magnitude of the regression coefficients. Specifically, their Table I shows the probability of separation can be as large as 0.98; this number occurring when there are 10 covariates and the sample size is 30 with coefficients corresponding to odds ratios of 16 : 1.

Figure 2.1: Illustration of Separation in Two and Three Dimensions.
Motivation

Since separation can frequently occur, it follows that near or quasi separation, where the responses are almost separable, should be at least as common. Near separation has received very little attention in the literature and the extent to which near separation undermines MLE and causes it to produce poor and misleading results is not widely appreciated. Near separation is actually more worrisome because it can easily go undetected in the course of model fitting using maximum likelihood. While separation raises an explicit alarm signal in most software packages (see Figure 2.2), it leads to absurdly large estimates in others (Allison et al., 2004). Indeed, it appears as though King and Ryan (2002) may be the only journal article in which the effects of near separation have been examined or even discussed to any extent. We illustrate some of these effects through the following examples.

Illustrative Dataset

Consider fitting a logistic regression model of the form

$$logit(Y|X_1,X_2) = \beta_0 + \beta_1X_1 + \beta_2X_2$$

(2.1)

to the data shown below in Table 2.1, and consider the value $X_2 = k$ for the
Figure 2.3: Illustrative Dataset: Covariate Space.

fifth observation:

Table 2.1: Illustrative Dataset: Fifth Observation with Variable Value \( k \).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>19</td>
<td>17</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>X2</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>14</td>
<td>( k )</td>
<td>13</td>
<td>11</td>
<td>15</td>
<td>19</td>
<td>17</td>
<td>16</td>
<td>18</td>
<td>22</td>
<td>20</td>
<td>24</td>
<td>19</td>
<td>17</td>
<td>21</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2.3 shows that if \( k = 18 \) (or \( k > 18 \)) there is separation, since a hyperplane (a straight line in 2 dimensions) can be drawn through the \( Y = 1 \) points that separates the plane into open half-planes. The lower half-plane is empty while the upper half-plane contains only points for which \( Y = 0 \). Thus the MLEs are unbounded and hence do not exist. If \( k = 17.999 \), the MLEs do exist, but we must ask how good they are, especially in light of a sensitivity or data perturbation analysis.

Panel A of Figure 2.4, shows the effect of near separation on the maximum likelihood estimates \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) of \( \beta_1 \) and \( \beta_2 \), respectively, for \( k = 17.999 \), with \( \hat{\beta}_1 \)
being the larger of the two estimates. The graph illustrates that the MLE grows exponentially as the point of complete separation is approached. Specifically, very small changes in $k$ cause large changes in the MLE. Although we do not know the true underlying population values of the coefficients, since the data for this example were not generated by simulation from a known model, there should not be large changes in parameter estimates resulting from very small data changes in covariate values for any estimation method.

Panel B of Figure 2.4 illustrates that in addition to the deleterious effect on the MLE, the standard errors ($SE$) of the estimators also approach infinity near the point of separation. It may appear that this is one way of noting that there is a problem with near separation, but this ad hoc approach only works in extreme cases. Problems with both the estimates and their standard errors will not be obvious to users when both are of the same order of magnitude and are not very large. Instead, users may erroneously conclude that a covariate is of no value because the estimate is small relative to its standard error. We illustrate this misconception with a real data set in the next example.

Any method that attempts to address problems created by near separation will need a quantitative measure for it. There is not at present a good way to measure near separation. Christmann and Rousseeuw (2001) indirectly addressed near separation when they developed an algorithm that determined the number of points (say, $m$) in a dataset, which, if removed, would create separation. Ryan and Blankmeyer (2003) showed that this measurement is not effective in characterizing separation, as for a given $m$ we may or may not have near separation. This is illustrated by Figure 2.4 where $m = 1$ for all values of $k$ less than 18, and yet $k = 17.999$ implies near separation whereas $k = 5$ clearly does not.
Figure 2.4: Illustrative Dataset: Effect of Near Separation on the MLE and SE.
Erythrocyte Sedimentation Rate

Near separation is most likely to occur with many covariates and small sample sizes, but can also occur under various other conditions. For example, King and Ryan (2002), and Heinze (2006) analyzed the erythrocyte sedimentation rate (ESR) data of Collett (2002), which was originally given by Collett and Jemain (1985). ESR is a simple blood test used to detect and monitor inflammation associated with conditions such as infections and autoimmune diseases. Figure 2.5 displays the ESA binary response variable with two different style points and the two covariates, Fibrinogen and Gamma-globulin, displayed on the horizontal and vertical axes respectively.

Collett (2002) identified two outliers. Near separation is evident from Figure 2.5, constructed after the outliers have been deleted, as a vertical line can be drawn with \( Fibrinogen = 3.36 \) that separates the \( Y = 1 \) values from the \( Y = 0 \) values. Although Collett (2002) analyzed the data very carefully, he did not con-
sider near separation. This indicates a lack of recognition of the problems associated with MLE in the statistical community.

Data perturbation analysis on the leftmost \( Y = 1 \) (solid square) point yields results similar to those illustrated in Figure 2.3. Hence, if we move the point very slightly to the right, so as to create severe near separation, the MLEs increase drastically, as is shown in Table 2.2. Thus, very small changes in a single Fibrinogen value results in very large change in the MLEs near the value of 3.36 at which complete separation occurs.

**Table 2.2:** Collett’s ESR Data: Effect of Near Separation on the MLE and SE.

<table>
<thead>
<tr>
<th>Permute the Single Point’s Fibrinogen Value</th>
<th>(True value is 3.34)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.33</td>
</tr>
<tr>
<td>fibrinogen</td>
<td>19.507</td>
</tr>
<tr>
<td></td>
<td>(21.694)</td>
</tr>
<tr>
<td>globulin</td>
<td>−0.079</td>
</tr>
<tr>
<td></td>
<td>(0.332)</td>
</tr>
<tr>
<td>Observations</td>
<td>30</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>−2.892</td>
</tr>
</tbody>
</table>

*Note:*  
*\( p < 0.1; \) **\( p < 0.05; ***\( p < 0.01 \)

Various other datasets that exhibit near separation are published in the literature, including the vaso constriction dataset given by Finney (1947), the cancer remission dataset of Lee (1974), the food stamp dataset of Künsch, Stefanski and Carroll (1989), and a low birth weight dataset given by Hosmer and Lemeshow (2005). These datasets were not written as examples of near separation, however. Rather, the food stamp dataset and the vaso constriction dataset are well-known in
the literature on outlier detection and robust logistic regression. For example, Künsch et al. (1989) and Carroll and Pederson (1993) applied robust logistic regression approaches to the food stamp dataset. Such datasets as prototypes of near separation and its deleterious effects would proliferate the literature if near separation were widely recognized as the problem that it is when MLE, in particular, is used.

**Simulation Study**

Several computing experiments have been conducted to examine bias in MLE for models with one and two continuous covariates in different configurations. The results show that MLE has substantial bias, not only when there is near separation, but also over a range of values where the regression coefficients are not large. Note that bias and mean squared error (MSE) are computed conditional on *not* observing response vectors that exhibit separation. In the absence of this conditioning, neither bias nor MSE exist for MLE. This is because, for any sample size, the mean of MLE does not exist as there is a non-zero probability that it is unbounded.

Here is one such illustration of the poor performance of MLE using simple logistic regression (one covariate) with 20 observations. The values of the covariate $X$ range from 5 to 100, in steps of 5. Due to space limitations we will be describing only this specific example in depth; nevertheless, it is important to keep in mind that the behavior of MLE follows the same pattern in all our simulated datasets.

All possible response vectors $y = (y_1, y_2, y_3, \ldots, y_{20})$ that did not result in separation, were enumerated. Values of $X = (5, 10, 15, \ldots, 100)$ were standardized by dividing by the standard deviation of $X$. This is useful as it enables us to compare findings across datasets with covariate values that lie in different ranges. For each value of $\beta_0$ from -4 to 4 in steps of 0.5, we varied $\beta_1$ from 0 to 5 in steps of 0.15. (There is no need to consider negative values of $\beta_1$ because of the symmetry in the
values of $X$ around its mean.) For each pair $(\beta_0, \beta_1)$ we calculated the bias and MSE of MLE conditional on separation not having occurred. This is done by computing the estimates and probability of occurrence of every possible non-separating response vector and using these values to calculate the bias conditional on non-occurrence of separation. For all values of $\beta_0$ we observed a pattern of behavior of MLE similar to that shown in Figure 2.6 for the case when $\beta_0 = 0$. (Note that the vertical axis for panel (b) is drawn on a logarithmic scale.) We can also see from panel (b) that MLE has tolerably low bias (<10%) when $3 < \beta_1 < 4$ and that there are two distinct regions where it has large bias.

As expected from the discussion of the previous example, the MLE has substantial bias when there is near separation (i.e., when the probability of separation
is high, greater than 0.7, see panel (d) of Figure 2.6). What is perhaps surprising is that panel (d) also shows that MLE has large relative bias for smaller values of $\beta_1$ also (when $0 < \beta_1 < 3$) when the probability of separation is small, less than 0.5. For instance, when $\beta_1 = 1.50$ the expected value of the MLE for $\beta_1$ is 2.14, so that the bias is 43%.

The previous three examples illustrate the long-recognized finite-sample bias of MLEs in logistic regression and among GLMs for categorical data in general. Although exact conditional regression provides a reliable alternative for small or sparse samples, this approach may often be too computationally difficult even for moderately sized samples, while at the same time such sample sizes may still be insufficient to overcome the bias of likelihood-based methods. These situations are prime candidates for bias-corrections methods, including the first order bias correction approach proposed originally by Firth (1993). A wholly new diagnostic measure (computed using a unique proprietary algorithm) that will inform analysts about potential problems with near separability and parameter estimation bias for a given GLM analysis is needed.

Methodology

Many data may be represented in the form of a $R \times C$ contingency table $x$, as displayed in Table 2.3. The entry in each cell of the table represents the number of observations falling in the corresponding row and column classifications, for either nominal or ordered variables.

The exact probability distribution of $x$ contains unknown parameters, $\pi_{ij}$, relating to the individual cells. Consider the null hypothesis of no row by column interaction. Since statistical inference is based on the distribution of $x$ under the null...
hypothesis of no row by column interaction, the number of unknown parameters is reduced ($\pi_{ij}$ being replaced by $\pi_i + \pi_j$ or $\pi_j$ depending on the sampling scheme) but not eliminated.

Asymptotic inference relies on estimating these unknown parameters by maximum likelihood and related methods. The key to exact permutation inference is getting rid of all nuisance parameters from the probability distribution of $x$. This is accomplished by restricting the sample space to the set of all $R \times C$ contingency tables that have the same marginal sums as the observed table $x$. Specifically, defined as the reference set

$$
\Gamma = \{ y : y \text{ is } R \times C; \sum_{j=1}^{c} y_{ij} = m_i; \sum_{i=1}^{r} y_{ij} = n_j; \text{ for all } i, j \}.
$$  \tag{2.2}

Then one can show that, under the null hypotheses of no row by column interaction, the probability of observing $x$ conditional on $x \in \Gamma$ is of the hypergeometric form

$$
Pr(x|x \in \Gamma) \equiv P(x) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{n_j!m_i!}{N!x_{ij}!}.
$$  \tag{2.3}

Equation 2.3, which is free of all unknown parameters, holds for categorical data whether the sampling scheme used to generate $x$ is full multinomial, product multinomial, or Poisson (See, for example, Agresti (1992)). Since Table 2.3 contains

<table>
<thead>
<tr>
<th></th>
<th>$Col_1$</th>
<th>$Col_2$</th>
<th>$\ldots$</th>
<th>$Col_c$</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Row_1$</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$\ldots$</td>
<td>$x_{1c}$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>$Row_2$</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$\ldots$</td>
<td>$x_{2c}$</td>
<td>$m_2$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ddots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$Row_r$</td>
<td>$x_{r1}$</td>
<td>$x_{r2}$</td>
<td>$\ldots$</td>
<td>$x_{rc}$</td>
<td>$m_r$</td>
</tr>
<tr>
<td>Col Total</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$\ldots$</td>
<td>$n_c$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Table 2.3: Generic $R \times C$ Contingency Table, without Clustering
tains no unknown parameters, exact inference is possible. The nuisance parameters were, however, eliminated by conditioning on the margins of the observed contingency table. Some of these margins were not fixed when the data were gathered. Thus it is reasonable to question the appropriateness of fixing them for purposes of inference. The justification for conditioning at inference time on margins that were not naturally fixed at data sampling time has a long history. Fisher (1925) first proposed this idea for exact inference on a single $2 \times 2$ contingency table. At various times since then, prominent statisticians have commented on this approach. See Mehta and Patel (1998) for further justification. It is this conditional approach which provides us with a unified way to perform exact inference and thereby compute accurate p-values and confidence intervals for $R \times C$ contingency tables, stratified $2 \times 2$ contingency tables, stratified $2 \times C$ contingency tables, and logistic regression.

Having assigned an exact probability $P(y)$ to each $y \in \Gamma$, the next step is to order each contingency table in $\Gamma$ by a test statistic or discrepancy measure that quantifies the extent to which that table deviates from the null hypothesis of no row by column interaction. Let us denote the test statistic by a real valued function $D : \Gamma \rightarrow \mathbb{R}$ mapping $R \times C$ tables from $\Gamma$ onto the real line $\mathbb{R}$. The p-value is defined as the sum of null probabilities of all the tables in $\Gamma$ which are at least as extreme as the observed table, $x$, with respect to $D$. In particular if $x$ is the observed $R \times C$ table, the exact p-value is

$$p = \sum_{D(y) \geq D(x)} P(y) = Pr \left[ D(y) \geq D(x) \right].$$ (2.4)

Classical non-parametric methods rely on the large-sample distribution of $D$ to estimate $p$. For $R \times C$ tables with large cell counts and the usual forms for the function $D$ it is possible to show that $D$ converges in distribution to a chi-square
with appropriate degrees of freedom. Thus \( p \) is usually estimated by \( \tilde{p} \), the chi-square tail area to the right of \( D(x) \). Modern algorithmic techniques have made it possible to compute \( p \) directly instead of relying on \( \tilde{p} \), its asymptotic approximation. This is achieved by powerful recursive algorithms that are capable of generating the actual permutation distribution of \( D \) instead of relying on its asymptotic chi-square approximation. This is important as \( p \) and \( \tilde{p} \) can differ considerably for contingency tables with small cell counts.

The main advantage of using \( p \) rather than \( \tilde{p} \) is that it is guaranteed to bound the type-1 error rate of the hypothesis testing procedure to any desired level. Moreover, this guarantee is provided unconditionally even though each p-value, \( p \), is calculated conditionally by restricting attention to a specific reference set \( \Gamma \). To see this, let

\[
S(\Gamma) = Pr(p \leq \alpha | \Gamma).
\]  

(2.5)

Accordingly, \( S(\Gamma) \) is the conditional type-1 error rate of a level-\( \alpha \) hypothesis testing procedure in which \( R \times C \) tables are repeatedly generated from the same reference set, \( \Gamma \), under the null hypothesis, and rejected whenever \( p \leq \alpha \) under the null hypothesis \( S(\Gamma) \leq \alpha \). Now the unconditional type-1 error rate, where \( \Gamma \) may be different each time you execute the test, is

\[
S = \sum S(\Gamma)Pr(\Gamma),
\]

(2.6)

the sum being taken over all possible reference sets, \( \Gamma \). Notice that Equation 2.6 is a weighted sum of terms of the form \( S(\Gamma) \), where each such term is less than or equal to \( \alpha \), the weights, \( Pr(\Gamma) \), are positive, and sum to 1. Thus \( S \leq \alpha \).

Hence the guaranteed protection against the type-1 error of an exact conditional hypothesis test also applies unconditionally. This guarantee does not hold if
you use \( \bar{p} \) rather than \( p \) in the decision to reject the null hypothesis, since \( Pr(\bar{p} \leq 0.05|\Gamma) \leq \alpha \) holds only asymptotically.

Exact conditional inferential methods utilize the distribution of the sufficient statistic for a given parameter, conditional on sufficient statistics for the other parameters (nuisance parameters) in the model. In 1983 Cyrus Mehta, Nitin Patel, and their colleagues at Harvard School of Public Health devised an innovative network algorithm that eliminated the need to completely enumerate all possible outcome vectors in the appropriate reference set. Referring to Table 2.3, \( \Gamma \) can be represented as a network of nodes and arcs constructed in \( c \) stages, wherein each complete path through the network, starting at the initial node to the terminal node, represents one and only one unique contingency table in the reference set.

Convex Hull and the Probability of Separability

A primary aim of a recent grant by Cytel included calculating the probability of separability, \( Pr(S) \). Their algorithm follows thus. Given a vector of outcomes \( y = (y_1, y_2, \ldots, y_n) \), a straightforward method for computing \( Pr(S) \) would be to sequentially generate every possible \( y \)-vector, then test it for separation in the covariate space and accumulate the probabilities associated with \( y \)-vectors that lead to separation.

As an example, for a dataset with 100 observations this would require enumeration of \( 2^{100} \) \( y \)-vectors at the first step followed by solving a linear program (see Santner and Duffy (1986), and Kolassa (1997)) for each \( y \)-vector to test for separability. Such a procedure would be prohibitively slow. They attempted to modify the state-of-the-art implementation of the network algorithm for exact logistic regression in LogXact [(Hirji, Mehta and Patel, 1987; Mehta, Patel and Senchaudhuri, 2000; Cytel Software Corporation Inc, 2016a) to incorporate a fast convex hull algo-
algorithm (e.g., Barber, Dobkin and Huhdanpaa (1996)) enabling rapid computation of $Pr(S)$. This approach sought to exploit the key idea underlying the success of the network algorithm, yet was still not realistically feasible.

For a logistic regression model with $p$ covariates, we can implicitly enumerate the $y$-vectors by examining the vector of sufficient statistics, $t = (t_0, t_1, \ldots, t_p)$, where $t_i$ is the sufficient statistic for $i, i = 1, 2, \ldots, p$. It is known that separation exists if and only if the observed $t$-vector is on the convex hull of the set of all possible $t$-vectors (see for example Andersen (2012)). Therefore to compute $Pr(S)$ we can sum the probabilities associated with the sufficient statistics vector on the convex hull.

We illustrate this view of the sufficient statistics vector using the ESR data (Section 2.2.2) in Figure 2.7. Since there are two covariates, $p = 2$ and $t = (t_0, t_1, t_2)$, we can display all 1,351 possible $t$-vectors as the scatter plot in panel (a). Note that due to size and plotting resolution limits, the points not separated vertically by spaces and appear as vertical bars. However, we see in panel (b) that we can obtain a thousand-fold reduction in effort, because instead of checking $2^{20}$ or approximately 1,000,000 $y$-vectors for separation, we examine only about 1,000 $t$-vectors. Conditioning on the observed value of the first sufficient statistic, $t_0$, the convex hull is a 40-sided polygon and the 40 $t$-vectors that lie on the boundary of the polygon are sufficient statistics pairs $(t_1, t_2)$ corresponding to all $y$-vectors that exhibit separation. Panel (c) zooms in on the part of panel (b) where the observed pair of sufficient statistic lies for the ESR data, demonstrating the near separability in this particular example.

Although the theory for computation of the probability of separability, $Pr(S)$, via the convex hull is reliable, the algorithm has proven too computationally expensive for a realistic dataset. Efforts to make the algorithm more efficient have largely failed. Yet there has been an important finding from this effort. Many sam-
Figure 2.7: Collett’s ESR data: Convex Hull.

...were generated under conditions not considered small or spare in practice, conditions causing most analysts not be overly concerned about the performance of MLE, though they demonstrate a marked advantage of penalized maximum likelihood estimation (PMLE) over MLE in terms of MSE and bias.

Profile Likelihood Confidence Intervals

The standard procedure for computing a confidence interval (CI) for a parameter in a GLM is by the formula: \( \text{estimate} \pm \text{percentile} \times SE_{\text{estimate}} \), where \( SE \) is the standard error. The \text{percentile} is selected according to a desired confidence level and a reference distribution (a \( t \)-distribution for regression coefficients in a linear model). The procedure is commonly referred to as a Wald-type CI (wCI) and is known to work poorly if the distribution of the parameter estimator is markedly skewed or if the standard error is a poor estimate of the standard deviation of the estimator. Since the standard errors in GLM’s are based on asymptotic variances obtained from the information matrix, wCI’s may perform poorly for small to moderate sample sizes. Profile likelihood confidence interval (plCI), also called the likelihood ratio confidence interval, do not assume normality of the estimator and ap-
pear to perform better for small sample sizes. They are, nonetheless, still based on
an asymptotic approximation - the asymptotic chi-squared distribution of the log
likelihood ratio test statistics.

The basic rationale is as follows. Suppose \( \lambda \) is a scalar parameter and we wish
to test whether \( \lambda = \lambda_0 \) where \( \lambda_0 \) is some specific value of interest. The likelihood
ratio statistic for this test is

\[
LR = 2\log \left[ \frac{L(\hat{\lambda})}{L(\lambda_0)} \right] \sim \chi^2_1
\]  

(2.7)

where \( \hat{\lambda} \) is the MLE of \( \lambda \). We reject \( \lambda = \lambda_0 \) at \( \alpha = 0.05 \) if \( LR > \chi^2_1(0.95) \) is the
.95 quantile of a chi-squared distribution with one degree of freedom. Similarly we
would fail to reject \( \lambda = \lambda_0 \) if it turns out \( LR < \chi^2_1(0.95) \). Essentially, the inequal-
ity defines the lower limit for the likelihood confidence interval for \( \lambda \), but on a log-
likelihood scale.

\[
2\log \left[ \frac{L(\hat{\lambda})}{L(\lambda_0)} \right] < \chi^2_1(0.95) \iff \log L(\hat{\lambda}) - \log L(\lambda_0) < \frac{1}{2} \chi^2_1(0.95)
\]

\[
\iff \log L(\hat{\lambda}) - \frac{1}{2} \chi^2_1(0.95) < \log L(\lambda_0)
\]  

(2.8)

Figure 2.8 shows the set of log-likelihood values that satisfy our inequality
(indicated by the vertical arrow at top of graph) and in turn define a correspond-
ing set of values for the parameter \( \lambda \) (horizontal arrow at the bottom of the graph).
These values for \( \lambda \) comprise what is called the profile likelihood confidence interval
for \( \lambda \). The boundaries of this confidence interval are defined by the places
where the long-dashed horizontal lower-limit line intersects the smooth curve of the log-
likelihood, then projected down to the \( \lambda \)-axis by the vertical dotted lines.

We could estimate the confidence limits graphically, but it is far simpler to
use numerical methods. The horizontal line in Figure 2.8 intercepts the axis at
Figure 2.8: Illustration of a Generic Profile Likelihood Confidence Interval.

the value $H$. We seek the values of $\lambda$ such that

$$
\log L(\lambda) = H \iff \log L(\lambda) - H = 0 
\iff f(\lambda) = 0
$$

(2.9)

where $f(\lambda) = \log L(\lambda) - H$, thus we need to find the roots or zeros of this function.

Suppose we have binomial data with a single covariate $x$ with $N$ levels, so that the number of successes at $x_i$ is $Y_i \sim \text{Binomial}(n_i, \pi_i), i = 1, 2, \ldots, N$. We fit the logit model

$$
\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \mu + \theta x_i.
$$

(2.10)
The log-likelihood function is then

$$
\log(\mu, \theta) = \sum_{i=1}^{N} \left[ y_i \log\left( \frac{\pi_i}{1 - \pi_i} \right) + n_i \log(1 - \pi_i) + \log\left( \frac{n_i}{y_i} \right) \right] = \sum_{i=1}^{N} \left[ y_i (\mu + \theta x_i) + n_i \log\left( \frac{1}{1 + \exp(\mu + \theta x_i)} \right) + \log\left( \frac{n_i}{y_i} \right) \right].
$$

(2.11)

While joint profile confidence regions are probably the most sensible interval estimates to consider, they are rather difficult to work with because they are difficult to solve for explicitly. A decidedly inferior but more popular alternative is to use marginal profile confidence intervals. In a marginal profile confidence interval we fix all but one parameter at specified values, typically the MLE, so that the log-likelihood becomes a function of just this one parameter.

To test the hypothesis for the mean $\mu$

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

To carry out this test we need to compare two log-likelihoods, one from a model in which $\mu$ and $\theta$ are estimated and the second from a model in which $\theta$ is estimated, but $\mu$ is held fixed at $\mu_0$.

$$LR_{\mu|\theta} = 2 \log \frac{L(\hat{\mu}, \hat{\theta})}{L(\mu_0, \hat{\theta})} \sim \chi_1^2.
$$

(2.13)

Here $\hat{\mu}$ and $\hat{\theta}$ are the MLEs. Operationally $\theta$ is fixed in both models at its MLE $\hat{\theta}$ so that it is no longer a free parameter in the model. Because the two models differ at a single parameter $\mu$, its free in one model and fixed in the other, the likelihood ratio statistic has an asymptotic chi-squared distribution with one degree of freedom. We reject the null hypotheses at $\alpha = .05$ if $LR_{\mu|\theta} < \chi_1^2(.95)$. This in-
equality defines the rejection region for the likelihood ratio statistic and all values of the statistic that would cause us to reject the null hypothesis. If we flip the inequality around we get the *fail-to-reject region*, or all values of the likelihood ratio statistic that would cause us to not reject the null hypothesis. This region can be written as follows

\[
2 \log \frac{L(\hat{\mu}, \hat{\theta})}{L(\mu_0, \hat{\theta})} \leq \chi^2_1(.95) \iff \log L(\hat{\mu}, \hat{\theta}) - \log L(\mu_0, \hat{\theta}) \leq \chi^2_1(.95) \]

(2.14)

\[
\iff \log L(\hat{\mu}, \hat{\theta}) \leq \log L(\mu_0, \hat{\theta}) - \chi^2_1(.95)
\]

**Figure 2.9** shows an example of the pICIs (red arrows) in which the *fail-to-reject region* is shaded for each MLE parameter (red square) by the inequalities in 2.14. The end points of each pCI is defined as the parameter value where the log-likelihood curve (solid black curve) intersects the LR lower bound \( H = \log L(\mu_0, \hat{\theta}) - \chi^2_1(.95) \) (horizontal dotted red line).

**Innovation**

We propose a wholly new diagnostic measure that will inform analysts about potential problems with near separability and parameter estimation bias for a given GLM analysis. This diagnostic will alert users of the need to take steps to mitigate potential bias, such as using exact tests or applying the Firth bias correction (Heinze, 2006).
Diagnostic Tool

The focus regards the degree of asymmetry in the plCI, defined as the ratio of the longer to the shorter side. To demonstrate, consider Table 2.4 as a simple example of an observed $2 \times 2$ table. Suppose the rows represent a binary outcome $y$ and the columns represent a binary exposures $x$. The cell counts represent the number of subjects observed for each combination of exposure and outcome, for a total of 20 subjects. This observed table conveys that 6 of the 10 exposed subjects experienced the outcome compared to only 4 of the 10 unexposed subjects.

Classic logistic regression models the dependency of $\pi_i = Pr(y_i = 1)$ on $x_i$ through the relationship

$$logit(y|x) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x.$$  \hspace{1cm} (2.15)
Table 2.4: Simple 2x2: Observed Contingency Table

<table>
<thead>
<tr>
<th>y</th>
<th>x</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Sum</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

Typical inference about $\beta_0$ and $\beta_1$ involves maximizing the likelihood function with respect to these regression coefficients.

$$Pr(y_1, y_2, \ldots, y_n) = \frac{\exp\left[\sum_{j=1}^{n} y_i (\beta_0 + \beta_1 x_i)\right]}{\prod_{i=1}^{n} \left[1 + \exp(\beta_0 + \beta_1 x_i)\right]}.$$  \hspace{1cm} (2.16)

Results are displayed in Table 2.5. The MLE estimate for the exposure $x$ was found to be $\hat{\beta}_1 = 0.8109$ which corresponds to an odds ratio of 2.25.

Table 2.5: Simple 2x2: Logistic Regression MLEs for Observed Contingency Table.

<table>
<thead>
<tr>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Log Likelihood</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2.6 displays results for four alternative logistic regression models, all of which model the effect of the exposure $x$ on the outcome $y$. The first two rows correspond to the MLE model shown in Table 2.5. Note the symmetry of the classic Wald confidence interval (wCI: Ratio = 1) compared to the asymmetry in the
profile likelihood confidence interval (plCI: ratio = 1.062). The plCI is also slightly wider than the wCI. The bottom two rows correspond to logistic regression utilizing the Firth bias correction or penalized maximum likelihood estimate (PMLE). The estimated parameter and odds ratio for PMLE are lower than MLE, yet the respective confidence intervals are slightly narrower.

Table 2.6: Simple 2x2: MLE and PMLE Estimates with wCI and plCIs for Observed Contingency Table.

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>CI_Type</th>
<th>Lower.95.</th>
<th>Upper.95.</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.811</td>
<td>Wald</td>
<td>-0.978</td>
<td>2.600</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>0.811</td>
<td>pl</td>
<td>-0.954</td>
<td>2.685</td>
<td>1.062</td>
</tr>
<tr>
<td>PMLE</td>
<td>0.735</td>
<td>Wald</td>
<td>-1.047</td>
<td>2.518</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>0.735</td>
<td>pl</td>
<td>-0.946</td>
<td>2.506</td>
<td>1.053</td>
</tr>
</tbody>
</table>

Figure 2.10 displays all possible 2 × 2 contingency tables with a total sample size of 20, having half exposed (x = 1). Although there are $2^{20}$ or 1,048,576 possible binary $y$ vectors with length 20, there are only 81 distinct contingency tables that are not completely separable. This set is further reduced if we constrain the tables to match the margins of the observed table shown in Table 2.4. The points shown in red represent the reference set $\Gamma$ (constrained to $t_0 = 10$) as defined in Equation 2.2 (excluding the two tables that are completely separable since their MLE for $\beta_1$ is undefined). The MLE for $\beta_1$ is represented with a dashed red horizontal line and the observed table is a circled red point on that line. For reference, the color blue is used to display the reference set if the observed table had contained only 8 cases ($t_0 = 8$) instead of 10, with the same proportion of 50% exposed (x = 1) for which $\hat{\beta}_1 = \log(6)$.

All tables in the reference set $\Gamma$ are of the form for Table 2.7. Since the margins are constrained (20 subjects total, half of whom are exposed, and ten expe-
Figure 2.10: Simple 2x2: All Possible Tables with N = 20 and Half exposed (x = 1).
Table 2.7: Simple 2x2: Generic Form of Tables in the Reference Set.

<table>
<thead>
<tr>
<th>y</th>
<th>x</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>$t_1$</td>
</tr>
<tr>
<td>Sum</td>
<td>-</td>
<td>10</td>
</tr>
</tbody>
</table>

Sum: $t_0 = 10$

Table 2.7: Simple 2x2: Generic Form of Tables in the Reference Set.

Experience the outcome), given any one of the four internal cell counts, all other cell counts may be derived, thus a single cell is sufficient to convey all the information contained in the entire table. Thus, $t_0$ is sufficient for $\beta_0$ and $t_1$ is sufficient for $\beta_1$.

Suppose we have interest in inference about $\beta_1$, and regard $\beta_0$ as a nuisance parameter. Then, instead of estimating $\beta_0$ from the unconditional likelihood function, we can eliminate it by conditioning on the observed value of its sufficient statistic, $t_0 = \sum_{i=1}^{n} y_i$, or the total number of observed outcomes. This yields the conditional likelihood function

$$Pr(y_1, y_2, \ldots, y_n | t_0) = \frac{\exp \left( \sum_{j=1}^{n} y_i \beta_1 x_i \right)}{\prod \exp \left( \sum_{j=1}^{n} y_i \beta_1 x_i \right)}.$$ (2.17)

Here we define $t_1 = \sum_{i=1}^{n} x_i y_i = \sum_{x_i=1}^{y_i}$, or the total number of observed outcomes among the exposed cases. Figure 2.11 displays the exact distribution of $t_1$ in $\Gamma$ as calculated by StatXact 11.

Development

The first approach we took was to generate small, categorical datasets such the one given in Table 2.4 and generate all possible permutations of the table in the reference set, conditioning on the sufficient statistics given in Table 2.7. Since
<table>
<thead>
<tr>
<th>Statistic $T(Y^*)$</th>
<th>Probability $Pr(T(Y^*))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.41254e-006</td>
</tr>
<tr>
<td>1</td>
<td>0.000541254</td>
</tr>
<tr>
<td>2</td>
<td>0.0109604</td>
</tr>
<tr>
<td>3</td>
<td>0.0779406</td>
</tr>
<tr>
<td>4</td>
<td>0.238693</td>
</tr>
<tr>
<td>5</td>
<td>0.343718</td>
</tr>
<tr>
<td>6</td>
<td>0.238693</td>
</tr>
<tr>
<td>7</td>
<td>0.0779406</td>
</tr>
<tr>
<td>8</td>
<td>0.0109604</td>
</tr>
<tr>
<td>9</td>
<td>0.000541254</td>
</tr>
<tr>
<td>10</td>
<td>5.41254e-006</td>
</tr>
</tbody>
</table>

**Table 2.8:** Simple 2x2: Ranking of the Reference Set.

**Figure 2.11:** Simple 2x2 Crosstabulation: Probability Density of the Reference Set.
the data was generated, we would know the true parameters and thus calculate the bias of the MLE of logistic regression to compare with against the asymmetry ratio of the plCI. The exact probability distribution, such as the one given in Table 2.8 would then be used to determine the probability weights of each bias-ratio pair.

Figure 2.12 displays the results of one such cycle of result. The data were generated from a range of true parameter combinations regarding the intercept \( \beta_0 \) and predictor effect \( \beta_1 \), which are denoted by the row and column labels. For each of the nine combined conditions, 50 repetitions included simulating a dataset of ten observations from the true logistic model, modeling the data with a logistic model to compute the MLE and plCI, and then computing both the bias and plCI ratio of asymmetry. The size of the points in the figure signify the number of times the ratio-bias pair was observed. Although this approach indicated a correlation between bias and asymmetry, it was ultimately found lacking.

A second approach focused on simulating data with two correlated continuous predictors instead of a single binary predictor. Figure 2.13 displays 16 example dataset generated from the multivariate normal distribution of two such predictor variables. In each panel, the 20 circles were generated from a mean expected value at the origin, \( \mu = (0, 0) \), with both predictors having variance of 1 to represent the observations with \( y = 0 \). Conversely, the 20 x’s were generated from a mean expected value of \( \mu = (a, a) \) and similar variance of 1. The columns vary by the separation between the means \( a \), and the rows vary by the degree of correlation between the two predictors \( r = \text{covariance} \). Moving from left to right, the two data clouds become more separated with increasing \( a \). Initial simulation results to follow set the separation \( a = 1 \), while investigating the effect of correlation.

For values of \( r = (0.05, 0.25, 0.50, 0.75) \), 500 samples were generated with 20 observations for each of three options for the count of the outcome variable. For instance, in Figure 2.14 the left most column is titled \( n: (6, 14) \) denoting that of
Figure 2.12: Simple 2x2: Simulation to Study Relationship Between plCI Asymmetry Ratio and Absolute Value of MLE Bias.
Figure 2.13: Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation.
the 20 observations in each generated dataset, six were from the $y = 0$ distribution and 14 were from the $y = 1$ distribution. Here both axes are represented on the log scale due the extreme skewness in both the bias and asymmetry ratio.

Figure 2.15 displays the same simulation results as Figure 2.14, but with restricted axes rather than log scales. Since this simulation including modeling for three parameters, the intercept as well as main effects of two continuous predictors, the axes display the maximum for all three values in each model fit, asymmetry ratio along the horizontal axis and the absolute value of the MLE bias on the vertical axis. Under all conditions there was a strong positive correlation between asymmetry in the plCI and bias in the MLE. Under all conditions there was a strong positive correlation between asymmetry in the plCI and bias in the MLE.

Application

Erythrocyte Sedimentation Rate

As shown in Figures 2.10, 2.14, and 2.14, as the absolute value of the MLE increases, so does the asymmetry of the plCI. Thus, the closer the data is to being separable, the larger the ratio of asymmetry becomes. Returning to the ESR dataset, we fit both MLE and PMLE models and calculate both types of confidence intervals, wCI and plCI. Results tabulated in Table 2.9 are plotted in Figure 2.16.
Figure 2.14: Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation and Sample Balance, with Logarithmic Scales.
Figure 2.15: Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation and Sample Balance, with Restricted Range Scales.
Table 2.9: Collett’s ESR data: MLE and PMLE Estimates with wCI and plCIs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Variable</th>
<th>Estimate</th>
<th>CI_Type</th>
<th>Lower.95.</th>
<th>Upper.95.</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>fibrinogen</td>
<td>24.877</td>
<td>Wald</td>
<td>-38.091</td>
<td>87.845</td>
<td>1.000</td>
</tr>
<tr>
<td>MLE</td>
<td>globulin</td>
<td>-0.139</td>
<td>Wald</td>
<td>-1.006</td>
<td>0.728</td>
<td>1.000</td>
</tr>
<tr>
<td>MLE</td>
<td>fibrinogen</td>
<td>24.877</td>
<td>pl</td>
<td>3.188</td>
<td>166.424</td>
<td>6.526</td>
</tr>
<tr>
<td>MLE</td>
<td>globulin</td>
<td>-0.139</td>
<td>pl</td>
<td>-1.869</td>
<td>0.532</td>
<td>2.579</td>
</tr>
<tr>
<td>PMLE</td>
<td>fibrinogen</td>
<td>4.421</td>
<td>Wald</td>
<td>0.024</td>
<td>8.818</td>
<td>1.000</td>
</tr>
<tr>
<td>PMLE</td>
<td>globulin</td>
<td>0.050</td>
<td>Wald</td>
<td>-0.246</td>
<td>0.345</td>
<td>1.000</td>
</tr>
<tr>
<td>PMLE</td>
<td>fibrinogen</td>
<td>4.421</td>
<td>pl</td>
<td>1.171</td>
<td>25.429</td>
<td>6.463</td>
</tr>
<tr>
<td>PMLE</td>
<td>globulin</td>
<td>0.050</td>
<td>pl</td>
<td>-0.230</td>
<td>0.414</td>
<td>1.305</td>
</tr>
</tbody>
</table>

Figure 2.16: Collett’s ESR data: Asymmetry of Profile Likelihood Confidence Intervals.

Conclusion

One of the most common quantitative analysis tools, logistic regression, is open to significant bias in the parameter estimates fit by typical maximum likelihood when population and sample conditions are near separable. This can include a small overall sample, a sparse outcome, or a high correlation between indepen-
dent variables and the outcome of interest, perhaps high level interaction. Though extremely large standard errors for the parameter estimates may indicate such a situation, inflated bias may occur without this warning sign. We have shown that the degree of asymmetry in the plCI (ratio of long to short side) is a simple to implement diagnostic tool will help warn researchers if their data exhibit near separation, alerting them to take steps to mitigate the potential for bias.
CHAPTER 3

EXACT TREND TEST FOR CORRELATED MULTINOMIAL DATA

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations.

George EP Box, 1978

Background

The problem of analyzing correlated categorical outcomes arises in a variety of biomedical research settings, especially in areas such as genetics, ophthalmology, and teratology. One can encounter correlated categorical data wherever multiple outcomes are measured on an individual over time, or on several different individuals who share common genetic or environmental exposures.

Another example is developmental toxicology, in which impregnated rodents are exposed to ordered levels of a potentially toxic substance and their offspring are examined for adverse outcomes. Because of shared environmental and genetic influences, offspring within the same litter tend to show greater similarity in their individual probabilities of response. This well-known litter effect is analogous to group effects observed in other areas of research, such as ophthalmology or periodontal studies. Ignoring such inherent correlation can lead to erroneous inferences (Zeger, Liang and Self, 1985; Kapper et al., 1986; Yinheng, Piedmonte and Williams, 1994), and much consideration has been given to methods that adjust for correlation within groups or clusters of observations. Pendergast et al. (1996)
reviewed many such statistical developments.

A large body of work has yielded models for analyzing correlated categorical outcomes. For example, marginal regression models using generalized estimating equations (GEE) (see Diggle et al., 2002) and nonlinear mixed models such as the beta-binomial and logistic-normal-binomial models (see chapter 2 McCulloch, Searle and Neuhaus, 2008) are readily available in most statistical software packages. There has also been some special attention paid to adapting versions of conventional contingency table tools (such as the Pearson chi-square test) to correlated data. Thomas, Singh and Roberts (1996) provides a review of some these developments. However, virtually all of these methods rely on large-sample approximate normality to justify their inferences.

Prior work indicates methods that rely on asymptotic approximations fare poorly in the presence of small or sparse samples. For example, Yinsheng et al. (1994) show that GEE confidence intervals for parameters associated with cluster-level covariates provide increasingly conservative coverage for decreasing sample sizes when the naive covariance estimator is used and increasingly anti-conservative coverage when a robust covariance estimator is used. Molenberghs, Declerck and Aerts (1998) evaluate the effect of model misspecification on dose-response testing in a fully parametric setting. They compare results assuming the Bahadur, beta-binomial, and exponential family (Molenberghs and Ryan, 1999) models and find that, even for moderately large samples, the likelihood-based tests are generally unreliable when the underlying response rate is small.

For uncorrelated data, there has been significant progress made in the development and implementation of exact testing procedures. Mehta (1994) provides a unifying overview of the theory underlying the exact analysis of contingency tables as well as a summary of recent algorithmic advances that have led to more widespread application of these methods, now available in packages such as StatX-
Few exact alternatives exist, however, for investigators who require valid testing procedures when faced with small or sparse samples of correlated binary data. Gladen and Rogan (1979) suggest some ad hoc procedures that employ summary response measures. For example, one could make group comparisons using the sample proportions for each cluster. A Wilcoxon or generalized rank test could then be used. Gladen shows such methods break down quickly when cluster sizes are unequal. Luta et al. (1996) propose using exact conditional logistic regression, treating the clusters as fixed effects and conditioning out their associated nuisance parameters. This approach may result in over conditioning as the number of clusters grows large.

Unlike many models for uncorrelated data, parametric models that adjust for intracluster correlation do not allow straightforward conditioning that can eliminate nuisance parameters. The model of Molenberghs and Ryan (1999) and Ryan and Molenberghs (1999) for clustered multivariate binary data provides a useful framework for exact inference in the presence of correlated binary data. This model includes a two-parameter multiplicative generalization of the binomial distribution proposed by Altham (1979) as a special case. Though, unlike Altham, Molenberghs and Ryan use a parameterization that directly lends itself to modeling covariates. Because the model is a member of the exponential family, we can readily condition on its sufficient statistics in order to eliminate the nuisance parameters, including the dispersion parameter, under the null hypothesis of no covariate effect. We can therefore maintain the underlying correlation structure of the data without needing to estimate the degree of overdispersion. Unlike the method of Luta et al. (1996), regardless of the number of clusters, we need only condition out a single dispersion parameter that captures the cluster effects. Moreover, the formulation of this exact test naturally incorporates unequal cluster sizes.
Conditioning on a multidimensional sufficient statistic introduces additional computational complexity in obtaining an exact p-value. We can implicitly enumerate all possible permutations of the observed data, subject to the constraints of the observed sufficient statistics, by using a graphical network representation. This allows efficient recovery of the exact tail distribution. This approach builds on the network-based methods developed by Mehta et al. (1992).

While conditioning provides a useful way to eliminate nuisance parameters, the resulting exact methods are computationally intensive. Exact inference therefore did not gain widespread use until the advent of modern computing over the past two or three decades. The development of fast algorithms during this period has expanded the feasibility of exact inference. One class of algorithms called network algorithms were developed by Dr. Cyrus Mehta, Dr. Nitin Patel, and their colleagues at the Harvard School of Public Health. A full list of references to this body of work is given in chapter 17 of the StatXact User’s Manual (Cytel Software Corporation Inc, 2016b). Many of these algorithms have been implemented in statistical software packages such as StatXact, SAS, SPSS, and Splus. Oster (2002) compares the implementation and use of exact methods across these and other products.

For correlated categorical data, however, eliminating nuisance parameters is made more difficult by the need to consider parameters that model the correlation structure of the data. Conditioning is easiest in the context of exponential family models, in which sufficient statistics for nuisance canonical parameters are more readily available. The quadratic exponential model proposed by Gourieroux and Trognon (1984) and Zhao and Prentice (1990) provides a basis for this sort of exact conditional inference in the context of correlated binary data. This model has been adapted in several ways to a variety of study designs. Molenberghs and Ryan (1999), for example, reformulated it to include multivariate outcomes. Heagerty and
Zeger (1996) extended it to ordered multinomials. Corcoran et al. (2001) used it as a basis for an exact trend test applied to cluster-correlated ordered binary populations. An exponential family model of this type allows us to condition away the nuisance correlation parameter, thus making exact inference possible when observations are dependent.

When faced with a small or sparse sample of categorical data, an investigator has relatively few widely available analytic techniques available and we seek to bridge this gap by extending and implementing recent developments in exact inference for correlated ordinal data. We focus on extending the exact trend test for exchangeable correlated binary data (Corcoran et al., 2001) to groups of correlated observations on which the categorical outcome has more than two ordered levels (binary extended to multinomial).

**Motivation**

Based on our broad experience in the analysis of categorical data, and our extensive interactions with biostatisticians and other scientists in public health and medicine, we are likewise convinced of the practical relevance of this project. The following examples demonstrate the potential utility of exact methods for correlated data. These cases involve actual small or sparse data set for which asymptotic methods are insufficient. The first example can now be solved but is included here for illustration of the ETT-CBD. The remaining example will be addressed by the developments proposed here. These examples are typical of the kinds of small-sample problems encountered by Drs. Corcoran and Coull and their colleagues in the course of their collaborations.
A study of nine children diagnosed with the ophthalmic dysfunction congenital hereditary endothelial dystrophy (CHED) was made by Schaumberg et al. (1999), in part to assess the impact of potential risk factors on the success of corneal implants to correct the loss of visual acuity resulting from CHED. Table 3.1 displays the number of these children who did and did not reject corneal grafts. Seven of the children received implants in both eyes, and two received implants in only a single eye. It would be inappropriate in this case to treat individual eyes of the same child as independent. All of the children were diagnosed before the age of six years, and all received implants before the age of 12 years. Early intervention in cases of CHED may be critical to prevent amblyopia (commonly known as “lazy eye”) due to opacification of the corneas. However, the necessary surgery carries greater risk when performed on young children as opposed to older children and adults. The investigators were therefore interested in comparing the results between two age groups: those younger than three years versus those aged at least three years at the time of surgery. The small sample sizes reported in this and other studies of CHED reflect the extreme rarity of the condition.

**Table 3.1:** CHED: Number of Rejected Corneal Grafts.

<table>
<thead>
<tr>
<th>Age at Diagnosis</th>
<th>&lt; 3 yrs</th>
<th>&gt;= 3 yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/2</td>
<td>0/2</td>
<td>0/2</td>
</tr>
<tr>
<td>0/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>0/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>0/2</td>
<td>1/1</td>
<td>1/1</td>
</tr>
</tbody>
</table>

Of the eight eyes among the younger children, there were no rejections,
whereas four of the eight implants were rejected among the older children. Does this imply a higher average rejection rate among children over three years of age compared to those who are younger? To answer this question while accounting for the correlation between the eyes of a given child, one might use a random-effects or marginal model as described in benchmark texts such as Diggle et al. (2002) or Fitzmaurice, Laird and Ware (2010). However, such modeling methods are infeasible in this case. Note that all four rejections occurred in older children, making estimation of an age-related regression parameter impossible. Even if parameter estimates were obtainable, the small sample size would render the corresponding asymptotic inference for these estimates suspect. In their published comments, these investigators expressed their disappointment at the failure of conventional large-sample methods to draw reliable inferences from their data.

We are now able to analyze data such as this via the ETT-CBD developed by Corcoran et al. (2001). This non-parametric method conditions on nuisance parameters to derive an exact distribution for the trend statistic. This test is detailed and applied to this data in the upcoming Methodology section to set the stage for the current innovation which is the focus of the bulk of this chapter.

**Clinical Trial for Pain Medication**

Whereas the previous example possessed an outcome of a binary nature, the next example exhibits an outcome with more than two levels. Miller and Landis (1991) detail a randomized study to compare an investigative drug to placebo in relieving symptoms characterized by pain and spasms. The study involved nine different investigators who treated and evaluated patients on three levels of response: *Worse or No Change*, *Slight Improvement*, or *More Improvement or Cured*. The cross-tabulation of partial data is reproduced in Table 3.2. With small observed
Table 3.2: Multicenter Pain and Spasm Study: Observed Contingency Table.

<table>
<thead>
<tr>
<th>Dose</th>
<th>Site (i)</th>
<th>Response (j)</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Same</td>
<td>Better</td>
</tr>
<tr>
<td>Placebo</td>
<td>x_i = 1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Drug</td>
<td>x_i = 0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>20</td>
<td>21</td>
</tr>
</tbody>
</table>

Counts in some cells, these data may benefit from the exact methods proposed for ordered response categories and will provide tools for an exact comparison of two ordinal multinomial populations with correlated data.

Methodology

The methodological core of our approach is the use of an exponential family model for correlated categorical data. The form of such a model allows straightforward conditioning on sufficient statistics in order to eliminate nuisance parameters, hence yielding a distribution for hypothesis testing and estimation that is entirely specified. Although this approach is widely used to derive exact tests for uncorrelated data, only recently have the appropriate models and algorithms been developed to allow similar applications to dependent data (Corcoran et al., 2001). We will employ the quadratic exponential models (QEM) suggested by Gourieroux and Trognon (1984) and Zhao and Prentice (1990), and further adapted by Molenberghs and Ryan (1999) for simultaneously clustered and multivariate outcomes. This approach allows for generalization to arbitrary two-way contingency tables of
correlated data.

** Exact Trend Test for Correlated Binomial Data (ETT-CBD) **

This section lays the ground work for the extensions to the exact trend test for correlated binary data (ETT-CBD) developed by Corcoran et al. (2001). Suppose that we have \( N \) clusters, where the \( i \)th cluster has \( n_i \) subjects. Let \( Z_{ij} \) represent the binary response of the \( j \)th observation, in the \( i \)th cluster, such that

\[
Z_{ij} = \begin{cases} 
1, & \text{if observation } j, \text{ in cluster } i, \text{ is a SUCCESS} \\
0, & \text{if observation } j, \text{ in cluster } i, \text{ is a failure.}
\end{cases}
\]

Therefore, if the probability of success in the \( i \)th cluster is \( \theta_i \), then \( Z_{ij} \sim Bernoulli(\theta_i) \). So the density is given by

\[
Pr(Z_{ij} = z_{ij}) = \begin{cases} 
\theta_i, & \text{if } z_{ij} = 1, \\
1 - \theta_i, & \text{if } z_{ij} = 0,
\end{cases}
\]

or more compactly

\[
f(z_{ij}) = Pr(Z_{ij} = z_{ij} | \theta_i) = \theta_i^{z_{ij}} (1 - \theta_i)^{1-z_{ij}}. \tag{3.1}
\]

This single-cluster density of an independent response \( Z_{ij} \) may be rearranged to
show it is a member of the exponential family

\[
Pr(Z_{ij} = z_{ij} | \theta_i) = \theta_i^{z_{ij}} (1 - \theta_i)^{1-z_{ij}}
\]

\[
= \exp \left\{ \log \left[ \theta_i^{z_{ij}} (1 - \theta_i)^{1-z_{ij}} \right] \right\}
\]

\[
= \exp \left\{ z_{ij} \log \theta_i + (1 - z_{ij}) \log (1 - \theta_i) \right\}
\]

\[
= \exp \left\{ z_{ij} \log \theta_i + \log(1 - \theta_i) - y_{ij} \log(1 - \theta_i) \right\}
\]

\[
= \exp \left\{ z_{ij} \log \left( \frac{\theta_i}{1 - \theta_i} \right) + \log(1 - \theta_i) \right\}
\]

\[
= \exp \left\{ \eta(\theta_i) T(z_{ij}) - A(\theta_i) \right\},
\]

where the natural parameter is \(\eta(\theta_i) = \log \frac{\theta_i}{1 - \theta_i}\) (logit function), the sufficient statistic is \(T(z_{ij}) = z_{ij}\) (identity function), and the normalizing constant is \(A(\theta_i) = -\log(1 - \theta_i) = \log \frac{1}{1-\theta_i}\).

The Molenberghs and Ryan (1999) model specifies the density of the vector of all binary responses in the \(i\)th cluster \(Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{im_i})'\) as

\[
Pr(Z_i = z_i) = \exp \left\{ \theta_i y_i - \delta_i y_i(n_i - y_i) + A_i(\theta_i, \delta_i) \right\},
\]

where the total number of responses within the \(i\)th cluster is \(Y_i = \sum_{j=1}^{n_i} z_{ij}\), \(\delta_i\) represents the dispersion parameter, and \(A_i(\theta_i, \delta_i)\) is the normalizing constant, summing over all possible realizations of \(Z_i\). The parameter \(\delta_i\) reflects intracluster correlation. It is easy to show that the model reduces to a product of independent binary probabilities when \(\delta_i = 0\). Using the logit link for \(\theta_i = \alpha + \beta x_i\), where \(x_i\) is constant for all observations in the \(i\)th cluster, and assuming equal dispersion among clusters
(δ_i = δ for each i), the density is expressed as

\[ Pr(Z_i = z_i|x_i) = \exp\left\{ (\alpha + \beta x_i)y_i - \delta y_i(n_i - y_i) + A_i(\alpha, \beta, \delta) \right\}. \]  

(3.4)

Assuming exchangeability of the binary responses within a given cluster, the density for each cluster depends on the individual responses, Z_i, only through their sum, Y_i. Any permutation of the binary responses within Z_i yields the same probability. The density of Y_i can therefore be expressed as

\[ Pr(Y_i = y_i|x_i, n_i, \alpha, \beta) = \binom{n_i}{y_i} \exp\left\{ (\alpha + \beta x_i)y_i - \delta y_i(n_i + y_i) + A_i(\alpha, \beta, \delta) \right\}, \quad y_i = 0, \ldots, n_i. \]  

(3.5)

Assuming cluster independence, \( Y = (Y_1, Y_2, \cdots, Y_N)' \), and constant covariate level \( x = (x_{i1}, \cdots, x_{im_i})' \) within each cluster, the joint distribution of the set of sums, Z_i,
is obtained as

\[
Pr(Y|x, n, \alpha, \beta, \delta) = \prod_{i=1}^{N} \left( \frac{n_i}{y_i} \right) \exp \left\{ (\alpha + \beta x_i) y_i - \delta y_i (n_i - y_i) + A_i(\alpha, \beta, \delta) \right\}
\]

\[
= \left\{ \prod_{i=1}^{N} \left( \frac{n_i}{y_i} \right) \right\} \prod_{i=1}^{N} \exp \left\{ (\alpha + \beta x_i) y_i - \delta y_i (n_i - y_i) + A_i(\alpha, \beta, \delta) \right\}
\]

\[
= \left\{ \prod_{i=1}^{N} \left( \frac{n_i}{y_i} \right) \right\} \exp \left\{ \sum_{i=1}^{N} (\alpha + \beta x_i) y_i - \delta y_i (n_i - y_i) + \sum_{i=1}^{N} A_i(\alpha, \beta, \delta) \right\}
\]

\[
= \left\{ \prod_{i=1}^{N} \left( \frac{n_i}{y_i} \right) \right\} \exp \left\{ \alpha \sum_{i=1}^{N} y_i + \beta \sum_{i=1}^{N} x_i y_i - \delta \sum_{i=1}^{N} y_i (n_i - z_i) + \sum_{i=1}^{N} A_i(\alpha, \beta, \delta) \right\}
\]

\[
= \left\{ \prod_{i=1}^{N} \left( \frac{n_i}{y_i} \right) \right\} \exp \left\{ \alpha s^\alpha + \beta t - \delta s^\delta + \sum_{i=1}^{N} A_i(\alpha, \beta, \delta) \right\}.
\]

Because this density is also of the exponential family, then the sufficient statistics are

\[
\mathbf{s} = \begin{cases} 
  s^\alpha = \sum_{i=1}^{N} y_i & \text{for } \alpha \\
  t = \sum_{i=1}^{N} x_i y_i & \text{for } \beta \\
  s^\delta = \sum_{i=1}^{N} y_i (n_i - y_i) & \text{for } \delta.
\end{cases}
\]

(3.7)

To formulate an exact test for the hypothesis of no dose-response effect with respect to \(x_i\), we confine ourselves to tables of the form \(Y\), with fixed cluster sizes \(n\), and condition on the sufficient statistics \(s^\alpha\) and \(s^\delta\) in order to eliminate the nuisance parameters \(\alpha\) and \(\delta\) in Equation 3.6.

If we denote the set of tables of the form of Table 3.3 subject to these con-
Table 3.3: Generic $2 \times C$ Contingency Table for a Binary Outcome, with Clustering

<table>
<thead>
<tr>
<th>Cluster $(i)$</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$N$</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose $(x_i)$</td>
<td>$x_1$</td>
<td>$x_2$</td>
<td>...</td>
<td>$x_N$</td>
<td></td>
</tr>
<tr>
<td>$y_i$</td>
<td>$y_1$</td>
<td>$y_2$</td>
<td>...</td>
<td>$y_N$</td>
<td>$s^\alpha$</td>
</tr>
<tr>
<td>$n_i - y_i$</td>
<td>$n_1 - y_1$</td>
<td>$n_2 - y_2$</td>
<td>...</td>
<td>$n_N - y_N$</td>
<td></td>
</tr>
<tr>
<td>$n_i$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>...</td>
<td>$n_N$</td>
<td>$N$</td>
</tr>
<tr>
<td>$x_i y_i$</td>
<td>$x_1 y_1$</td>
<td>$x_2 y_2$</td>
<td>...</td>
<td>$x_N y_N$</td>
<td>$t$</td>
</tr>
<tr>
<td>$y_i (n_i - y_i)$</td>
<td>$y_1 (n_1 - y_1)$</td>
<td>$y_2 (n_2 - y_2)$</td>
<td>...</td>
<td>$y_N (n_N - y_N)$</td>
<td>$s^\delta$</td>
</tr>
</tbody>
</table>

Constraints by $s^\alpha$ and $s^\delta$

$$\Gamma(s^\alpha, s^\delta) = \left\{ Y^* : \sum_{k=1}^N y_k^* = s^\alpha, \sum_{k=1}^N y_k^*(n_k - y_k^*) = s^\delta \right\} \quad (3.8)$$

where $Y^*$ is any generic table of the form $Y = (y_1^*, y_2^*, \ldots, y_N^*)'$, holding the cluster sizes $n_i$ and sufficient statistics for the nuisance parameters $s = (s^\alpha, s^\delta)$ fixed. Then the conditional density of $Y$ under the hypothesis of no dose-response effect ($\beta = 0$) can be expressed as

$$Pr\left(Y|n, s^\alpha, s^\delta, \alpha, \beta, \delta\right) = \frac{\left[ \prod_{i=1}^N \binom{n_i}{y_i} \right] \exp\left\{ \alpha s^\alpha + \beta t - \delta s^\delta + \sum_{i=1}^N A_i(\alpha, \beta, \delta) \right\} \sum_{Y^* \in \Gamma(s^\alpha, s^\delta)} \left[ \prod_{k=1}^N \binom{n_k}{y_k^*} \right] \exp\left\{ \alpha s^\alpha + \beta t - \delta s^\delta + \sum_{k=1}^N A_k(\alpha, \beta, \delta) \right\}}\right). \quad (3.9)$$

Under the null hypothesis ($H_0 : \beta = 0$) this further reduces further to

$$Pr\left(Y|n, s^\alpha, s^\delta, H_0\right) = \frac{\prod_{i=1}^N \binom{n_i}{y_i}}{\sum_{Y^* \in \Gamma(s^\alpha, s^\delta)} \left[ \prod_{k=1}^N \binom{n_k}{y_k^*} \right]}, \quad (3.10)$$

which is free of all unknown parameters. Since this distribution is completely specified under the hypothesis of no dose-response effect, we can order the tables in $\Gamma(s^\alpha, s^\delta)$ according to their respective values of the linear rank statistic $T(Y) =$
Table 3.4: Congenital Eye Disease and Corneal Grafts: Observed $2 \times C$ Contingency Table, with Clustering

<table>
<thead>
<tr>
<th>Child ($i$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>$\sum_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age $\geq 3$ ($x_i$)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$\sum_i$</td>
</tr>
<tr>
<td>$y_i$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>s$^\alpha$ = 4</td>
<td></td>
</tr>
<tr>
<td>$n_i - y_i$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_i$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>N = 16</td>
<td></td>
</tr>
<tr>
<td>$x_i y_i$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>t = 4</td>
<td></td>
</tr>
<tr>
<td>$y_i (n_i - y_i)$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

$\sum_i x_i y_i$, and compute an exact p-value $p$ as

$$p = Pr\left(T > t_{obs} \mid n, s^\alpha, s^\delta, H_0\right) = \sum_{\gamma \in \Gamma\left(s^\alpha, s^\delta\right)} \left( \sum_{T(\gamma^*) \geq t_{obs}} \left[ \prod_{i=1}^{N} \left( \frac{n_i}{y_i} \right) \right] \right), \quad (3.11)$$

where $t_{obs}$ represents the observed test statistic. A one-sided $\alpha$-level test then rejects $H_0$ when $t_{obs} > t_{\alpha}$, where $t_{\alpha}$ is defined as the smallest value such that

$$Pr\left(T > t_{\alpha} \mid n, s^\alpha, s^\delta, H_0\right) \leq \alpha. \quad (3.12)$$

Returning to the first example detailed in the prior Motivation section, we will show the ETT-CBD at work. The data that were originally presented in Table 3.1 is now reformatted to match the layout of Table 3.3 to facilitate calculations pertinent to the test.

Referring to the sufficient statistics for the exponential family given in Equation 3.7, we define the reference set given in Equation 3.8 to be $\Gamma = \{\gamma^* \mid s^\alpha = 4, \ s^\delta = 2\}$. Table 3.5 displays the exact permutation distribution of all tables in this reference set, which conditions on the margins of the observed table and the clustering set forth in Table 3.4. The exact test of the association between age
Table 3.5: Congenital Eye Disease and Corneal Grafts: Exact Permutation Distribution for the Test Statistic for the Exact Trend Test for Correlated Binomial Data (ETT-CBD)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T(Y^*)$</td>
<td>$Pr(T(Y^*))$</td>
</tr>
<tr>
<td>4</td>
<td>0.047619</td>
</tr>
<tr>
<td>5</td>
<td>0.285714</td>
</tr>
<tr>
<td>6</td>
<td>0.285714</td>
</tr>
<tr>
<td>7</td>
<td>0.285714</td>
</tr>
<tr>
<td>8</td>
<td>0.0952381</td>
</tr>
</tbody>
</table>

Group $(X)$ and graft rejection $(Y)$ yields linear rank statistic of $t = 4$ which corresponds to a two-sided $p$-value of $0.04762$. Where unconditional asymptotic methods fail to produce any inference, the exact test provides some evidence of an age effect on the average rejection probability.

**Network Algorithm for Conditional Reference Set**

Identifying the conditional reference set $\Gamma(s^\alpha, s^\delta)$ associated with the proposed procedures is conceptually straightforward, but computationally daunting. The denominator in the expressions 3.11 often includes far too many elements to enumerate explicitly, even for relatively small data sets. A network algorithm approach to efficiently identify the tail area of the appropriate distribution has been developed. This is accomplished by representing the reference set with a network or directed graph. The network approach has been used successfully by Dr. Senchaudhuri and his colleagues at Cytel Software Corporation for solving a variety of exact inference problems for independent categorical data, as well as for exact conditional logistic regression (for an extensive list of related references, see chapter 17 of the StatXact User’s Manual, (Cytel Software Corporation Inc, 2016b)).

Building network algorithms in the correlated data setting requires significant
computational innovation, as the conditioning has a higher dimension than conventional independent data applications. Fortunately, Dr. Cyrus Mehta is the purveyor of the network algorithm and the co-founder of Cytel Inc., along with his colleague Dr. Nitin Patel, has made substantial progress in this area.

Innovation

The ETT-CBD (Corcoran et al., 2001) described in the preceding section is a major advancement, but is restricted to outcomes with only two levels such as the congenital eye disease and corneal grafts study. In this section we will develop the Exact Trend Test for Correlated Multinomial Data which is able to consider a possible association between a binary covariate (population) an outcome with three or more levels, in which the observations are clustered within population, such as the data gathered for the clinical trial for pain medication. The tabulation and layout of such data will be addressed, followed by comparing two multinomial distributions with ordered categories in settings where the data are clustered. Once we have defined the test statistic we will construct its associated one and two-sided exact p-values. In order to calculate the p-values, we will need to delineate the impact on the network algorithm and the options for specifying the correlation conditioning.

Tabulation and Layout of Correlated Multicategory Outcomes

We wish to assess the association between two discrete, ordinal variables denoted generically by $X$ and $Y$, measured on a sample of clustered data. Specifically, we want to test whether the distribution of $Y$ is independent of $X$. We denote the number of categories of $X$ (binary) and $Y$ (ordinal) by $2$ and $C$, respectively. Each
**Table 3.6:** Generic $2 \times C$ Contingency Table for a Multinomial Outcome, with Clustering.

<table>
<thead>
<tr>
<th>Category of $X$ ($i$)</th>
<th>Cluster ($i$)</th>
<th>Category of $Y$ ($j$)</th>
<th>Cluster Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>$y_{1,1}$ $\ldots$ $y_{1,C}$</td>
<td>$n_1$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$y_{2,1}$ $\ldots$ $y_{2,C}$</td>
<td>$n_2$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>0</td>
<td>$N_1$</td>
<td>$y_{N_1,1}$ $\ldots$ $y_{N_1,C}$</td>
<td>$n_{N_1}$</td>
</tr>
<tr>
<td>$1$</td>
<td>$N_1 + 1$</td>
<td>$y_{N_1+1,1}$ $\ldots$ $y_{N_1+1,C}$</td>
<td>$n_{N_1+1}$</td>
</tr>
<tr>
<td>$N_1 + 2$</td>
<td>$y_{N_1+2,1}$ $\ldots$ $y_{N_1+2,C}$</td>
<td>$n_{N_1+2}$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$N_1 + N_2$</td>
<td>$y_{N_1+N_2,1}$ $\ldots$ $y_{N_1+N_2,C}$</td>
<td>$n_{N_1+N_2}$</td>
<td></td>
</tr>
</tbody>
</table>

Observation within a given cluster can be classified in one of $C$ categories. Individual clusters are represented by $Y_1, Y_2, \ldots, Y_N$, where $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iC})'$ and $y_{ij}$ is the number of subjects in the $i$th cluster classified in category $j$, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, C$. Let $X_i$ represent the category of $X$ associated with the $i$th cluster, and let $n_i$ be the cluster size, so that $\sum_j y_{ij} = n_i$. Letting $N_1, N_2, \ldots, N_R$ represent the numbers of clusters within each category of $X$ ($N = \sum_{i=1}^R N_i$), the data can be viewed as an $2 \times C$ table with clustering, as illustrated in Table 3.6.

**Multinomial Distributions with Ordered Categories**

We can generalize the approach used for several ordered binomials described in the Methodology section by combining the extension of the exponential model to ordered multinomials suggested by Heagerty and Zeger (1996) with the conditioning argument made by Corcoran et al. (2001). Assuming exchangeability between observations within cluster, and setting three-way and higher association parameters...
ters to zero, for a single cluster we have

\[
Pr(Y_i|\theta_i, \delta, n_i) = \left(\frac{n_i}{y_{i1} \ldots y_{iC}}\right) \exp\left\{\sum_{j=1}^{C} \theta_{ij} y_{ij} + \sum_{j<k} \delta_{ij} y_{ij} y_{jk} + A_i(\theta_i, \delta, n_i)\right\},
\]

(3.13)

where \(\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iC})\), \(\delta_i = \{\delta_{jk}\}\) for all \(j, k = 1, \ldots, C\) and \(j < k\), and \(A_i(\theta_i, \delta, n_i)\) is the normalization constant summing over all possible permutations of counts within each category. Assuming cluster independent and common values of \(\delta_{ij}\) across clusters, the joint density of \(Y = (Y_1, \ldots, Y_N)\) is given by

\[
Pr(Y|\theta, \delta, n) = \prod_{i=1}^{N} \left(\frac{n_i}{y_{i1} \ldots y_{iC}}\right) \exp\left\{\sum_{i=1}^{N} \sum_{j=1}^{C} \theta_{ij} y_{ij} + \sum_{i=1}^{N} \sum_{j<k} \delta_{ij} \sum_{i=1}^{N} y_{ij} y_{jk} + \sum_{i=1}^{N} A_i(\theta_i, \delta, n_i)\right\},
\]

(3.14)

where \(n = (n_1, \ldots, n_N)\). One of the attractive features of this formulation is that if \(C = 2\) it reduces readily to Equation 3.6.

If the distribution of a given \(Y_i\) depends only on its associated level of \(X\), we can alternatively express \(\theta_{ij}\) as \(\beta_{j|x_i}\). However, under the hypothesis that the distribution of \(Y\) does not depend on \(X\) we have \(\beta_{j|x_i} = \beta_j\) for \(j = 1, \ldots, C\) and \(i = 1, \ldots, N\). This reduces Equation 3.14 to

\[
Pr(Y|\beta, \delta, n) = \prod_{i=1}^{N} \left(\frac{n_i}{y_{i1} \ldots y_{iC}}\right) \exp\left\{\sum_{j=1}^{C} \beta_j \sum_{i=1}^{N} y_{ij} + \sum_{i=1}^{N} \delta_{ij} \sum_{i=1}^{N} y_{ij} y_{jk} + \sum_{i=1}^{N} A_i(\beta_j, \delta, n_i)\right\},
\]

(3.15)
where $\beta = (\beta_1, \ldots, \beta_C)$. From Equation 3.15 it is clear the sufficient statistics are

$$ s = \begin{cases} 
  s_j^\beta = \sum_i y_{ij} & \text{for } \beta_j \\
  s_{jk}^\delta = \sum_{i<j} y_{ij}y_{ik} & \text{for } \delta_{jk}
\end{cases} \quad (3.16) $$

Note that $s_j^\beta$ represents the total number of observations, across all clusters, that fall into category $j$ of $Y$. The $s_{jk}^\delta$ represent the $\binom{C}{2}$ sums of cross-products between unordered $(y_{ij}, y_{jk})$ pairs of counts across clusters.

Treating $\beta$ and $\delta$ as nuisance parameters, we can condition on their corresponding sufficient statistics for the betas $s^\beta = (s_1^\beta, s_2^\beta, \ldots, s_C^\beta)$ and the deltas $s^\delta = (s_{1,2}^\delta, s_{1,3}^\delta, \ldots, s_{1,C}^\delta, s_{2,3}^\delta, s_{2,C}^\delta, \ldots, s_{C-1,C}^\delta)$ in order to eliminate $\beta$ and $\delta$ from the density of $Y$.

Let $\Gamma(s^\beta, s^\delta)$ represent the set of all $R \times C$ tables of the form seen in Table 3.6, where the cluster sizes $n$ are fixed and the sufficient statistics $s^\beta$ and $s^\delta$ are equal to the values of the observed table. Under the hypothesis of no association between $X$ and $Y$, the conditional distribution is then computed as

$$ P(Y|s^\beta, s^\delta, n) = \frac{\prod_{i=1}^C \left( \sum_{y_{i1} \cdots y_{iC}} \right) \exp \left\{ \sum_{j=1}^C \beta_j s_j^\beta + \sum_{j<k} \delta_{jk} s_{jk}^\delta + \sum_{i=1}^N A_i(\beta, \delta, n) \right\}}{\sum_{Y^* \in (s^\beta, s^\delta)} \prod_{i=1}^C \left( \sum_{y_{i1}^* \cdots y_{iC}^*} \right) \exp \left\{ \sum_{j=1}^C \beta_j s_j^\beta + \sum_{j<k} \delta_{jk} s_{jk}^\delta + \sum_{i=1}^N A_i(\beta, \delta, n) \right\}} $$

$$ = \frac{\prod_{i=1}^N \left( \sum_{y_{i1}^* \cdots y_{iC}^*} \right)}{\prod_{i=1}^N \left( \sum_{y_{i1}^* \cdots y_{iC}^*} \right)}, \quad (3.17) $$

where $Y^*$ is any table of the generic form shown in Table 3.6, with $Y^* = (Y_1^*, \ldots, Y_N^*)$ and $Y_i^* = (y_{i1}^*, \ldots, y_{iC}^*)'$ for the $i$th cluster.
Exact Trend Test for Correlated Multinomial Data (ETT-CMD)

Note that Equation 3.17 contains no unknown parameters. An exact test of no association can be performed by ordering the tables \( Y \in \Gamma(s^\beta, s^\delta) \) according to their corresponding values of an appropriately chosen test statistic \( T(Y) \). Let \( Y_{obs} \) represent the observed table. Then a conditional exact p-value \( p \) of the hypotheses of no association between \( X \) and \( Y \) is computed as

\[
p = Pr(T \geq T(Y_{obs})|s^\beta, s^\delta, n) = \sum_{Y^* \in \Gamma(s^\alpha, s^\delta)} Pr(Y^*|s^\beta, s^\delta, n)
\]

We now will describe the choice of \( T(Y) \) to test for association between \( X \) and \( Y \). For ordered tables with clustered data, we will test the hypothesis of no association specifically against positive or negative association. Suppose that we represent levels of \( T(Y) \) with ordinal scores \( u_1, \ldots, u_C \), where the \( u_i \) are appropriately chosen scores that are strictly increasing over the \( C \) increasing levels of \( Y \). Likewise, let \( v_i \) represent the score for a cluster associated with level \( X_i \) of \( X \), where the \( v_i \) are also strictly increasing for increasing levels of \( X \). Since \( X \) is binary, we will use the simplest scores of \( u_1 = 0 \) and \( u_2 = 1 \) as this choice simplifies the computation of the test statistic. We can generalize the trend test from correlated binomial data to a association test between \( X \) and \( Y \) by considering the linear rank statistic

\[
T(Y) = \sum_{ij} v_i u_j y_{ij} = \sum_{x_i=1} u_j y_{ij} .
\]

Thus, the test statistic is the weighted number of outcomes in one population (the one assigned \( x_i = 0 \)). This is analogue to the linear-by-linear association test statistic for the case of independent data, where the summation is across clusters.
as opposed to levels of $X$. Note that $Y$ is binary, this reduces to the trend test for correlated binary data (ETT-CBD).

**Additional Conditioning for the Network Algorithm**

Consider first the conditional reference set $\Gamma(s^\alpha, s^\delta)$ defined in Equation 3.8 and used to compute the denominator in Equation 3.11. The ETT-CBD of Corcoran et al. (2001) details how to identify this set, by first identifying the set $\Gamma(s^\alpha)$, for which $\Gamma(s^\alpha, s^\delta) \in \Gamma(s^\alpha)$. We then recursively derive $\Gamma(s^\alpha, s^\delta)$ from $\Gamma(s^\alpha)$. The advantage of this method is that both the problem of deriving $\Gamma(s^\alpha)$ and then processing $\Gamma(s^\alpha, s^\delta)$ to obtain the critical region can be handled by adapting the work of Mehta et al. (1992).

We can generalize this approach to $2 \times C$ tables as defined in Table 3.6. In this case, we need to identify the reference set $\Gamma(s^\beta, s^\delta)$ used in Equation 3.18 to compute the denominator. We will accomplish this first by recursively identifying the larger set $\Gamma(s^\beta)$ using the network approach, and then pruning out the elements of $\Gamma(s^\beta)$ that do not meet the additional constraints imposed by $s^\delta$, hence yielding $\Gamma(s^\beta, s^\delta)$ from $\Gamma(s^\beta)$.

We begin by representing the reference sets $\Gamma(s^\beta)$ and $\Gamma(s^\beta, s^\delta)$ individually as directed networks of nodes and arcs. Each network is divided into $N + 1$ stages, indexed over $0, \ldots, N$. At each stage $k$ there is a set of nodes. For $\Gamma(s^\beta)$), each node is indexed by two elements, denoted by $(k, s_k^\beta)$. The first element $k$ represents the $k^{th}$ cluster of correlated observations. The second component has the form $s_k^\beta = (s_{k1}, \ldots, s_{kC})$, where $s_{kl}^\beta = \sum_{i=1}^{k} y_{kl}$ is one possible value of the partial sum of observations from the first $k$ clusters that fall into category $l$ of $Y$, for $l = 1, \ldots, C$. For $\Gamma(s^\beta, s^\delta)$, the nodes are indexed by three elements, with each node denoted as $(k, s_k^\beta, s_k^\delta)$. The elements $k$ and $s_k^\beta$ are as defined for $\Gamma(s^\beta)$, while the additional ele-
ment has the $s_k^\delta = \{s_{kpq}^\delta\}$, where $s_{kpq}^\delta = \sum_{i=1}^k y_{ip}y_{iq}$ represents one possible value of the partial sum of $\binom{c}{2}$ products between the counts within categories $p$ and $q$ of $Y$, for $p, q = 1, \ldots, C$ and $p < q$.

For each network, there is a single initial node and a single terminal node. The initial nodes for $\Gamma(s^\beta)$ and $\Gamma(s^\beta, s^\delta)$ respectively are defined as $(0, 0^\beta)$ and $(0, 0^\beta, 0^\delta)$, where $0^\beta$ and $0^\delta$ are vectors of zeroes with the same respective lengths as $s^\beta$ and $s^\delta$. The terminal nodes for $\Gamma(s^\beta)$ and $\Gamma(s^\beta, s^\delta)$ respectively are defined as $(N, s^\beta)$ and $(N, s^\beta, s^\delta)$.

Hence, for either of the two networks, each path which begins at the initial node and ends at the terminal node corresponds to exactly one element or table within the corresponding conditional reference set. For example, for the $l^{th}$ category of $Y$ we can obtain the number of responses across the $N$ clusters as

$$
\begin{bmatrix}
s_1^\beta - 0 \\
s_2^\beta - s_1^\beta \\
\vdots \\
s_l^\beta - s_{N-1,l}^\beta
\end{bmatrix}.
$$

For such a path in either $\Gamma(s^\beta)$ or $\Gamma(s^\beta, s^\delta)$, $\sum_{i=1}^N (s_i^\beta - s_{i-1,l}^\beta) = s_l^\beta$. For a path in $\Gamma(s^\beta, s^\delta)$ we are further guaranteed that $\sum_{i=1}^N (s_{ip}^\beta - s_{i-1,p}^\beta)(s_{iq}^\beta - s_{i-1,q}^\beta) = s_{pq}^\delta$, with $p$ and $q$ previously defined.

A key difference between the two networks is that for any given node in $\Gamma(s^\beta)$ we can easily enumerate its successor nodes using the closed-form solution to a linear programming problem identified by Mehta and Patel (1983). They show that the set of successor nodes $R(k - 1, s_{k-1}^\beta)$ to a given node $(k - 1, s_{k-1}^\beta)$ contains all nodes of the form $(k, s_k^\beta)$ such that
max \left( s_{k-1,l}^\beta, s_l^\beta - \sum_{i=k}^{N} n_i \right) \leq \min \left( s_l^\beta, s_{k-1}^\beta + n_k \right) \text{ simultaneously for each } l = 1, \ldots, C, \text{ and } k = 1, \ldots, N.

On the other hand, identifying successor nodes to a given triple in \( \Gamma(s^\beta, s^\delta) \) involves simultaneously solving both the linear and quadratic programming problems introduced by further conditioning on \( s^\delta \). This solution is not closed-form, but can be obtained through a series of \( \binom{C}{2} \) recursions. For each, we first recursively calculate the largest and smallest possible values of the secondary sufficient statistic over all paths that begin at the initial node and end at each of the remaining given nodes. This information enables us to eliminate paths that do not belong in \( \Gamma(s^\beta, s^\delta) \). Having identified the conditional reference set we can process the network in the manner of Mehta and Patel (1983) in order to compute the desired tail probability in Equation 3.18.

**Specification of the Correlation Conditioning**

Analogous to the tests of ordered independent multinomials, we compute an exact p-value using a conditional reference set of tables with row (cluster) totals and column totals equal to those of the observed table. Adapting the approach of Corcoran et al. (2001) to the clustered multinomial formulation of Heagerty and Zeger (1996), we can account for the correlation within cluster by conditioning further on the set of sufficient statistics defined by

\[
U_{ij}(Y) = \sum_{i=1}^{N} \sum_{j<k} y_{ij} y_{jk}, \quad j, k = 1, 2, \ldots, C. \tag{3.20}
\]

The exponential family model used assumes exchangeability between observations within cluster, cluster independence, and that all three way and higher associ-
ation parameters are zero.

The $U_{jk}$ represent the $\binom{C}{2}$ sums of cross-products between unordered $(y_{ij}, y_{ik})$ pairs of counts across clusters. This additional conditioning ensures a distribution of the test statistic that is fully specified under the null hypothesis of no association between the group and response variables, while accounting for cluster correlation.

Additionally, the conditional reference set must satisfy that the sufficient statistic, based on a specified correlation condition, is equal to that of the observed table. Three correlation conditions are possible:

1. **Intra-Cluster Correlation (Unequal)**: correlation for every column-pair is equal to that of the observed table:

   $s_{ij}^\delta = \sum_{i=1}^{N} y_{ij} y_{jk} = \sum_{i=1}^{N} y_{ij}^* y_{jk}^*$, for $\begin{cases} j = 1, 2, \ldots, (C - 1) \\ k = (j + 1), (j + 2), \ldots, C \end{cases}$ (3.21)

2. **Total Correlation (Equal)**: the sum of intra-cluster correlation for all column-pairs is equal to that of the observed table:

   $s^\delta = \sum_{j=1}^{C-1} \sum_{k=(j+1)}^{C} \sum_{i=1}^{N} y_{ij} y_{jk} = \sum_{j=1}^{C-1} \sum_{k=(j+1)}^{C} \sum_{i=1}^{N} y_{ij}^* y_{jk}^*$ (3.22)

3. **Adjacent Correlation**: the sum of intra cluster correlation for all $k$-adjacent column-pairs is equal to that of the observed table for $k = 1, 2, \ldots, (C - 1)$.

   $s^\delta = \sum_{j=1}^{C-k} \left( \sum_{i=1}^{N} y_{ij} y_{jk} \right) = \sum_{j=1}^{C-k} \left( \sum_{i=1}^{N} y_{ij}^* y_{jk}^* \right)$, for $k = 1, 2, \ldots, (C - 1)$ (3.23)

If columns are indexed as $1, 2, \ldots, C$ then $k$ adjacent column pairs means two
columns with index difference of \(k\).

These three options will be studied more in the next chapter.

Application

Clinical Trial for Pain Medication

Referring back to Table 3.2, there are \(C = 3\) ordinal levels to the outcome \(Y\) on the pain scale. Applying the definitions from Equation 3.16, the three sufficient statistics for the three parameters for the marginal means (\(\beta_1\), \(\beta_2\), and \(\beta_3\)) are

\[
s_1^\beta = 20, \quad s_2^\beta = 21, \quad s_3^\beta = 43,
\]

total number of observations that fall into each category \(j\) of \(Y\), across all clusters.

The sufficient statistics for the \(\binom{C}{2} = 3\) parameters for dispersion (\(\delta_{12}\), \(\delta_{23}\), and \(\delta_{13}\)) are

\[
s_{12}^\delta = 51, \quad s_{23}^\delta = 149, \quad s_{13}^\delta = 128,
\]
as calculated as the sums of cross-products between unordered pairs of counts across clusters, displayed in Table 3.7.

The exact distribution for the test statistic, under each of the three possible correlation conditioning specifications, is displayed in Figure 3.1. Note that for computing the test statistics according to Equation 3.19, the scores were set such that the data was coded with \(v_i = 1\) for control and \(v_i = 0\) for drug and the outcomes levels as \(u_j = 0\) for same, \(u_j = 1\) for better, and \(u_j = 2\) for cured. The test
Table 3.7: Multicenter Pain and Spasm Study: Crossproducts for $\delta_{jk}$ Sufficient Statistics

<table>
<thead>
<tr>
<th>Site $(i)$</th>
<th>Levels 1 vs. 2</th>
<th>Levels 2 vs. 3</th>
<th>Levels 2 vs. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same Better</td>
<td>$\Pi$</td>
<td>Better Cured</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$\sum_i$</td>
<td>20</td>
<td>21</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 3.8: Multicenter Pain and Spasm Study: Exact p-values for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)

<table>
<thead>
<tr>
<th>Correlation Conditioning</th>
<th>Point Probability</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-sided</td>
</tr>
<tr>
<td>Equal</td>
<td>.03338</td>
<td>.1960</td>
</tr>
<tr>
<td>Adjacent</td>
<td>.03241</td>
<td>.1965</td>
</tr>
<tr>
<td>Unequal</td>
<td>.04337</td>
<td>.1572</td>
</tr>
</tbody>
</table>

statistic hence observed as 38. The p-values are given in Table 3.8. In all three cases the test did not show any evidence of an association between the drug and outcome.

Toxicology Study on Ethylene Glycol in Mice

Coull and Agresti (2000) published an experiment that was carried out under the U.S. National Toxicology Program involving litters of mice, which were exposed to varying doses for ethylene glycol (EG) during gestation. Investigators were interested in the dose-response effect for either death, malformation, or both. For purposes of this example, data on placebo and dose (1.50 g/kg) for 10 litters will
Figure 3.1: Multicenter Pain and Spasm Study: Exact Distribution of the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)
Table 3.9: Toxicology Study on Ethylene Glycol in Mice: Observed Contingency Table.

<table>
<thead>
<tr>
<th>Dose</th>
<th>Litter (i)</th>
<th>Response (j)</th>
<th></th>
<th></th>
<th></th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Placebo</td>
<td>Death</td>
<td>Malformed</td>
<td>Normal</td>
<td>Sum</td>
<td></td>
</tr>
<tr>
<td>$x_i = 1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td>$x_i = 0$</td>
<td>6</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>20</td>
<td>21</td>
<td>43</td>
<td>84</td>
<td></td>
</tr>
</tbody>
</table>

Correlation Conditioning  | Point Probability  | p-Value 1-sided | p-Value 2-sided |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal</td>
<td>.006201</td>
<td>.01016</td>
<td>.021710</td>
</tr>
<tr>
<td>Adjacent</td>
<td>.004871</td>
<td>.006484</td>
<td>.016280</td>
</tr>
<tr>
<td>Unequal</td>
<td>.004065</td>
<td>.004891</td>
<td>.006891</td>
</tr>
</tbody>
</table>

Table 3.10: Toxicology Study on Ethylene Glycol in Mice: Exact p-values for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)

be examined. Three outcomes are considered: death, malformation, and normality. Each litter forms a natural cluster. A summary of counts is given in Table 3.9.

The exact distribution for the test statistic, under each of the three possible correlation conditioning specifications, is displayed in Figure 3.2. Note that for computing the test statistics according to Equation 3.19, the scores were set such that the data was coded with $v_i = 1$ for control and $v_i = 0$ for dose and the outcomes levels as $u_j = 0$ for death, $u_j = 1$ for malformed, and $u_j = 2$ for normal. The test statistic hence observed as 113. The p-values are given in Table 3.10. In all three cases the test did show evidence of an association between the drug and outcome.
Figure 3.2: Toxicology Study on Ethylene Glycol in Mice: Exact Distribution of the Test Statistic for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)
Conclusion

The Exact Trend Test for Correlated Multinomial Data (ETT-CMD) set forth in this chapter is a significant advancement as a non-parametric approach for ascertaining if the distribution of an ordinal outcome with at least three levels is independent of a binary independent variable assigned at the cluster level. The innovation of the quadratic exponential model allows for conditioning on a multidimensional sufficient statistic for nuisance parameters that include for the marginal proportions, as well as the pairwise odds ratios which capture the intracluster correlation. This is especially useful in studies with samples too small or sparse to allow for traditional multilevel or marginal regression models.

With the foundation of the ETT-CMD now developed, there are several potential extensions. The test could be expanded to allow for predictor covariates with more than two ordinal levels, in order to test for a linear association. Conversely, to analyze the full dataset gathered during the clinical trial for pain medication, we will need to allow for the covariate level to vary within cluster.
CHAPTER 4

CONDITIONING CONSIDERATIONS FOR THE EXACT TREND TEST

While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will be up to, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician.

Arthur Conan Doyle

Background

Some explanation is warranted regarding the three correlation conditions given in Innovations Section of Chapter 3 especially with regard to making the most reasonable choice when analyzing a given dataset. The underlying issue in each of the three cases has to do with what assumption can be made about the nature of the correlation between pairs of observations within a cluster.

Note that the \( \binom{C}{2} \) sufficient statistics correspond to \( \binom{C}{2} \) correlation parameters in the \( C \)-level multinomial modeling approach of Heagerty and Zeger (1996). The “Unequal” condition assumes these underlying parameters assume distinct values. Therefore, this can be considered the most ubiquitous condition, in the sense that it will yield a valid result even in the most general case. However, for some datasets the underlying \( \binom{C}{2} \) correlation parameters may follow a simplified pattern. For example, they may all have the same value, which is referred to as the “Equal” case. Alternatively, they may follow a structure such that observations from any two “Adjacent” or consecutive multinomial categories have the same correlation;
any two observations two categories apart have the same correlation, and so on.

The choice of one of these three assumptions may have important practical implications on a given data analysis. Each case results in a different degree of conditioning to obtain the exact distribution of the trend test statistic. The “Unequal” condition requires the most conditioning, with $\binom{C}{2}$ additional constraints imposed to yield the reference set $\Gamma^M$. On the other hand, the “Equal” condition requires only a single additional constraint, and the “Adjacent” requires $C - 1$ constraints. All else being equal, more conditioning yields a reference set $\Gamma^M$ containing fewer elements, and hence a more discrete permutation distribution.

In practice, this means that if one chooses the “Unequal” condition when either “Adjacent” or “Equal” could be correctly assumed, then there will be a loss of statistical power in detecting an existing association. The resulting p-values will likely be conservative. On the other hand, incorrectly assuming a simpler correlation structure may yield invalid results as well. Such is the case when choosing “Equal” when either “Adjacent” or “Unequal” is correct, or choosing “Adjacent” when “Unequal” is more appropriate. In particular, the Type I error rate may tend to be slightly inflated.

In addition, there are critical trade-offs relative to computational as well as statistical efficiency. More conditioning increases the computational complexity, sometimes significantly. Under either “Unequal” or “Adjacent”, the memory burden will quickly grow prohibitively large for increasing $C$. In other words, even where “Unequal” or “Adjacent” may represent more theoretically ideal choices, they may require too much computation time to obtain the exact p-value. In practice, as with any data analysis involving correlated data, the choice of correlation condition can be informed by an exploratory analysis that examines the strength and nature of within-cluster correlation.
Motivation

To determine more fully the impact of a wrong assumption about the correlation condition for this exact test, we set out to perform a simulation study to address the robustness of the ETT-CMD set forth in Chapter 3. Since the primary factors influencing this question are the strength and structure of the intrachuster correlations, the process of generating correlated data is of utmost importance. As we postulate that real world processes are more likely to give rise to data with unequal degrees of correlation, we focus on generating data with differential pair-wise dispersion between levels of ordinal outcome. This will help gauge risk of inflation to the type I error rate we hypothesize will be the price for more simple nurtured correlation structure.

As presented in the previous chapter, the formulation of the ETT-CMD hinges on the Quadratic Exponential Model (QEM) given by Heagerty and Zeger (1996) utilized in Equation 3.14. Whereas our test leverages the specification of pair-wise associations between levels of the ordinal outcome, they proceed to use a single global odds ratio (Dale, 1986) as a “natural” measure for the association between cumulative indices for the ordinal levels instead of category indicators as used by Miller, Davis and Landis (1993). Generating data in such a way served their purpose in testing the relative efficiency of marginal models, but will not meet our current objective.

Correlated Multinomial Variate Generating Mechanisms

Generation of correlated continuous variates has long been studied and different options are routinely available in common software packages. In the late 20th
century there was an outburst of publications regarding correlated binary data simulation, mostly based on the multivariate normal distribution, (Emrich and Piedmonte, 1991) generating multivariate binary data with fixed marginal distributions and specified degrees association. Yet multivariate ordinal data turns out to be extremely nontrivial to generation and has seen far fewer advancements. Over twenty years ago Gange et al. (1995) bemoaned that “unlike for continuous multivariate data, where the multivariate normal distribution plays a central role, no convenient multivariate distribution for correlated discrete data is readily available.” Since then several methods for generating correlated ordinal data have been proposed.

Some of the earliest publications regarding simulation of correlated ordinal data were penned by Gange et al. (1995), who suggested a simple but computationally intensive (for the time) procedure. He derived his method from the log-linear representation (Bishop, Fienberg and Holland, 1975) of multivariate categorical data and local association measures for all possible $2 \times 2$ sub-tables within the $C \times C$ bivariate joint distribution. Using this Iterative Proportional Fitting (IPF) to derive the full joint probability distribution allows for third and higher order interactions to be studied. We graciously received Gange’s R program via personal correspondence and were able to simulate new data. Though different correlations may be specified for each pair of observations in the cluster, we were unable to use this method to specify differential associations between outcome level pairs.

Perhaps the most common approach for generating independent ordinal data involves first generating variates from a continuous distribution (uniform or Gaussian) and then converting to an ordinal scale via prespecified thresholds or cut-points. This approach is problematic for several reasons. It has been well documented that the correlation between the underlying latent scale does not transfer in a meaningful way to the derived ordinal variates. Additionally, the marginal probabilities of the categorical variates constrains the correlation parameters, be they
odds-ratios, correlations, or other associations (see Prentice, 1988; Lipsitz, Laird and Harrington, 1991; and Liang, Zeger and Qaqish, 1992).

Lee (1997) described a variety of ways to generate correlated random variates, both binary and categorical. His “simple” method uses linear programming to find the solution of a joint distribution, given the association between variates in terms of a single Goodman-Kruskal’s $\gamma$ uncertainty coefficient, (ratio of concordance to discordance) and then computes convex combinations of the most extreme such tables. Although, no programming was included in his publication, AkmaIbrahim and Sulidi (2011) circulated macros for Lee’s algorithm using R and SAS. Lee (1997) also presented a second method for generating the joint distributions of ordinal variates based on the threshold method via the Archimedian copulas.

More recent publications include the Biswas (2004) algorithm which is specific to autoregressive-type correlations among identical distributed ordinal variates in direct contrast to the prevalent latent variable approach. In like manner, Choi (2008) started by simplifying the convex combination approach of Lee (1997) by using Pearson’s chi-squared statistic as the single measure of association between two ordinal variates, instead of the uncertainty coefficient, and applied the greedy algorithm instead of the linear programming method to find the maximal distributions leading to the two-dimensional joint distribution. Further generation of the multidimensional distribution follows the IPF algorithm detailed by Gange et al. (1995).

Demirtas (2006) responded with a more flexible algorithm allowing for any type of correlation structure between variates and no assumption of identical distributions. The algorithm starts by collapsing the ordinal categories into binary levels (0 and 1) and proceeds with the generation of correlated binary variates based on the associated marginal exceptions and the correlation structure. Resulting binary variates are then converted to the ordinal levels by reversing the process. While the algorithm generates data that match the targeted marginal means and associations,
the binary correlations are not free to range between -1 and 1. As implemented in the \texttt{MultiOrd} R package, it relies on a fixed but common marginal probabilities across all subjects and a predefined correlation matrix for the correlated categorical responses.

Barbiero and Ferrari (2017) created the \texttt{GenOrd} R package for stochastic simulation of discrete variables in direct competition with \texttt{MultiOrd}. Based on the Gaussian copula (also known as “\textit{NOR}mal To \textit{A}nything” or the NORTA method), it links the discrete distributions together and uses an iterative scheme to recover the correlation matrix for the copula that ensure the desired correlations among the discrete variables. Implementation allows for association in terms of Spearman’s correlation in addition to the standard Pearson’s correlation.

A third R package available through the Comprehensive R Network (CRAN) is \texttt{SimCorMultRes} Touloumis (2016). It implements a modified latent variable threshold approach and relies on the NORTA principle to allow for the marginal probabilities to carry across subjects based on categorical and/or continuous covariates.

The more novel approach of Traylor (2017) invokes a distinctly computer science strategy of constructing a tree structure to represent the generalized multinomial distribution in which all categorical variables are identically distributed, by equally correlated based on the inclusion of a single dependency coefficient.

\textit{Selected Method for Sample Generation}

Although together these advancements in the realm of discrete data generation are remarkable, none of the previously summarized methods meet the current need to simulate ordinal data with distinct level-pairwise correlations. Ideally we would accomplish this with the QEM, but the difficulty precluded even Heagerty
and Zeger (1996). Ultimately, we determined the applicable approach was the multivariate generalization of the multiplicative binomial distribution, as implemented by Altham, Hankin and others (2012) in the \texttt{MM} R package.

The model developed from the two-parameter exponential family generalization of the binomial distribution first introduced by Altham (1978) allows for over- or under-dispersion by introduction of the \textit{shape} parameter, $\theta$. This is further generalized to the Multivariate Multinomial distribution which maps directly to a single cluster in the methodology of Chapter 3. The counts of observations exhibiting each ordinal level is given by $y_1, \ldots, y_C \geq 0$ with a corresponding cluster size of $\sum_j y_j = n$; a fixed integer. The probability mass function of $Y_1, \ldots, Y_C$ is defined as

$$P(y_1, \ldots, y_C) = A^{-1}(y, p, \theta) \left( \frac{n}{y_1 \cdots y_C} \right)^C \prod_{j=1}^C p_j^{y_j} \prod_{1 \leq j < k \leq C} \theta_{jk}^{y_j y_k}$$

where $A^{-1}(y, p, \theta)$ is a normalizing constant.

The free parameters are the \textit{univariate ordinal level probabilities} $p = (p_1, \ldots, p_C)$, for which $p_j \geq 0$ and $\sum_{j=1}^C p_j = 1$, and the $\binom{C}{2}$ pair-wise dispersion parameters $\theta = \theta_{ij}$, for which $\theta_{jk} > 0$. Thus the standard multinomial is recovered in the case where all the shape parameters $\theta_{jk} = 1$.

A Poisson distribution with a log link function may be used to generate the counts $(y_j)$ due to the simple exponential form and the ingenious result of Lindsey and Mersch (1992) without the need to compute the normalizing constants. In this way the Multivariate Multinomial distribution may be seen as a discrete analogy of the multivariate normal distribution.

Altham et al. (2012) show that the set of $\theta_{jk}$ may be interpreted in terms of the \textit{conditional} cross-ratios for a cluster of $n = 4$ interchangeable members may be
computed as

\[ \theta_{jk}^2 = \frac{P_{12hl} \times P_{21hl}}{P_{11hl} \times P_{22hl}} \]  

(4.2)

for which \( \theta_{rs} = \theta_{sr}, \forall r, s \in r, s, h, l \) and

\[ P_{jkh} = \frac{1}{A} \times p_{j}p_{k}p_{h}p_{l} \times \theta_{jk} \theta_{kh} \theta_{hl} \]  

(4.3)

where \( P_{jkh} \) would be the cell probability for the \( C^4 \) contingency table for the multinomial joint distribution with \( 1 \leq i, j, k, l \leq C \).

Although it is clear that when all the dispersion parameters \( \theta_{jk} = 1 \), the cells of the \( P_{jkh} \) contingency table are equivalent to the product of the marginal univariate probabilities, and thus the independent case (no correlation), it is less clear the exact connection the \( \delta_{jk} \) given in Equation 3.15. This is made even less transparent as the \( \delta_{jk} \) are never actually computed in the QEM, only treated as nuisance parameters and conditioned out via their sufficient statistics, \( s_{jk}^\delta \).

To investigate the role of the \( \theta_{jk} \) on the dispersion of clustered multinomial data, Table 4.1 displays \( N = 30 \) clusters of \( n = 10 \) observations each for a \( C = 3 \) level ordinal outcome. The univariate probabilities for each level of \( Y \) are \( p = (0.2, 0.3, 0.5) \). For a single outcome, \( p_1 = P(y_j = 1) = 0.2, p_2 = P(y_j = 2) = 0.3 \), and \( p_3 = P(y_j = 3) = 0.5 \).

The left most set of columns display the ten clusters \( Y = (y_1, y_2, y_3) \) frequency counts for the three levels of the ordinal outcome under the standard multinomial distribution where all \( \theta_{jk} = 1 \), thus all members of each cluster are generated from the univariate distribution independent of each other. The marginal probabilities are roughly 20/30/50 and the set of clusters exhibit a fair degree of variation. Now compare these data with the rest of the table.

The middle set of columns in Table 4.1 correspond to the ten clusters gener-
ated under the condition of overdispersion by setting all the $\theta_{jk} = 0.8$. With only three levels, in order to exhibit strong correlation and homogeneity within a cluster, the outcomes must fall into the same category most of the time. Usually this leads to an overload of observations in the category with the largest univariate probability, however, several of the clusters in this set of ten end up with the majority of observations failing in a low probability category. For instance, the fourth cluster results in all but two observations in the middle level. Therefore, the marginal distribution is somewhat consistent with the univariate distribution, if not exhibiting an enhanced trend. It is somewhat counter intuitive that the correlation between members of a cluster results in this *lumpy-ness*, which turns the homogeneity within clusters into heterogeneity between clusters. This consequence of overdispersion on multinomial variates will be crucial to understanding the simulation results presented in the following section.

Conversely, the last set of columns on the right side of Table 4.1 were generated with all the $\theta_{jk} = 2.0$ and exhibit underdispersion. Although rare in most real world data, investigating this opposite condition helps us understand overdispersion. With a limited number of levels, the only way for observations within a cluster to not be associated is to spread out. In this way, observations within a cluster heterogeneity result in a consistent spread across clusters and a breakdown of any marginal trend that was initial in the univariate probabilities.

A second set of $N = 30$ clusters are displayed in Table 4.2. For all clusters, the univariate probabilities are $p = (.2, .3, .5)$ and overdispersion constant for all samples, but differential for each pair of levels, such that $\theta_{12} = 0.90$, $\theta_{23} = 0.85$, $\theta_{13} = 0.75$. Here the effect of sample size was investigated. Recall that overdispersion results in *lumpy* clusters. It is important to note that the larger the cluster size, the more extremely the univariate probabilities are violated in a cluster where the *lump* falls in a minority category. Again, this consequence will be crucial to un-
Table 4.1: Randomly Generated Clusters of \( n = 10 \) Observations and Univariate Probabilities of \( p = (0.2, 0.3, 0.5) \) to Investigate the Behavior of Intrachluster Dispersion Specification.

<table>
<thead>
<tr>
<th>Cluster ( (i) )</th>
<th>A. Standard Multinomial ( \theta_{jk} = 1.0 )</th>
<th>B. (Lumpy) Overdispersion ( \theta_{jk} = 0.8 )</th>
<th>C. (Smooth) Underdispersion ( \theta_{jk} = 2.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3 2 5</td>
<td>5 1 4</td>
<td>2 4 4</td>
</tr>
<tr>
<td>2</td>
<td>1 3 6</td>
<td>0 3 7</td>
<td>2 4 4</td>
</tr>
<tr>
<td>3</td>
<td>3 4 3</td>
<td>1 2 7</td>
<td>3 4 3</td>
</tr>
<tr>
<td>4</td>
<td>3 2 5</td>
<td>1 8 1</td>
<td>3 3 4</td>
</tr>
<tr>
<td>5</td>
<td>4 2 4</td>
<td>1 6 3</td>
<td>3 3 4</td>
</tr>
<tr>
<td>6</td>
<td>2 3 5</td>
<td>0 1 9</td>
<td>3 3 4</td>
</tr>
<tr>
<td>7</td>
<td>1 6 3</td>
<td>0 2 8</td>
<td>4 3 3</td>
</tr>
<tr>
<td>8</td>
<td>3 2 5</td>
<td>0 0 10</td>
<td>4 3 3</td>
</tr>
<tr>
<td>9</td>
<td>6 2 2</td>
<td>0 4 6</td>
<td>3 4 3</td>
</tr>
<tr>
<td>10</td>
<td>1 3 6</td>
<td>1 1 8</td>
<td>2 3 5</td>
</tr>
<tr>
<td>( \sum_i )</td>
<td>27 29 44</td>
<td>9 28 63</td>
<td>29 34 37</td>
</tr>
<tr>
<td>%</td>
<td>27 29 44</td>
<td>9 28 63</td>
<td>29 34 37</td>
</tr>
</tbody>
</table>

Understanding the simulation results presented in the following section.

Simulation Study

Now that we have established a viable mechanism to generate ordinal data with distinct pair-wise correlation conditions on the \( \binom{C}{2} \) pairs of levels, we are poised to address the very difficult question of determining the robustness of the ETT-CMD to misspecification. It is important to note that this is a novel question being posed of a cutting edge tool concerning abstract quantities. Determining an appropriate data generating mechanism alone is a milestone in many regards and the initial results presented here are the first of their kind. Plainly all concerns surrounding the three options for correlation constraints for the ETT-CMD will not be answered in this chapter, but this simulation is impactful nonetheless.
Table 4.2: Randomly Generated Clusters of Various Cluster Sizes in Conjunction with Univariate Probabilities of \( p = (0.2, 0.3, 0.5) \) and Overdispersion Parameterization \( \theta_{12} = 0.90, \theta_{23} = 0.85, \theta_{13} = 0.75. \)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Small Clusters ( n = 8 )</th>
<th>Moderate Clusters ( n = 12 )</th>
<th>Large Clusters ( n = 16 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((i))</td>
<td>1 2 3</td>
<td>1 2 3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>1</td>
<td>1 0 7</td>
<td>0 1 11</td>
<td>0 2 14</td>
</tr>
<tr>
<td>2</td>
<td>0 0 8</td>
<td>0 3 9</td>
<td>0 1 15</td>
</tr>
<tr>
<td>3</td>
<td>0 4 4</td>
<td>0 0 12</td>
<td>0 3 13</td>
</tr>
<tr>
<td>4</td>
<td>1 4 3</td>
<td>0 2 10</td>
<td>0 2 14</td>
</tr>
<tr>
<td>5</td>
<td>0 1 7</td>
<td>1 1 10</td>
<td>0 0 16</td>
</tr>
<tr>
<td>6</td>
<td>1 4 3</td>
<td>0 2 10</td>
<td>0 3 13</td>
</tr>
<tr>
<td>7</td>
<td>1 3 4</td>
<td>0 2 10</td>
<td>0 0 16</td>
</tr>
<tr>
<td>8</td>
<td>0 3 5</td>
<td>0 0 12</td>
<td>0 6 10</td>
</tr>
<tr>
<td>9</td>
<td>2 1 5</td>
<td>0 3 9</td>
<td>2 11 3</td>
</tr>
<tr>
<td>10</td>
<td>0 0 8</td>
<td>0 4 8</td>
<td>1 8 7</td>
</tr>
<tr>
<td>(\sum_i)</td>
<td>6 20 54</td>
<td>1 18 101</td>
<td>3 36 121</td>
</tr>
<tr>
<td>%</td>
<td>7.5 25.0 67.5</td>
<td>0.8 15.0 84.2</td>
<td>1.9 22.5 75.6</td>
</tr>
</tbody>
</table>

Model Parameterization and Data Generation

For this simulation, samples were generated with four clusters from each of two different populations \( (N = 8) \), distinguished by a binary covariate \( X \) coded such that the first four clusters were assigned the value of one and the second four the value of two. Cluster size within each sample was held constant \( (n = n_i, \forall i = 1, \ldots, 8) \), but sets of samples of various sizes \( (n = 4, n = 8, n = 12) \) were generated in turn. Note that samples of 8 clusters of size 16 each were prohibitive due to memory constraints.

To address the paired responses of power and type 1 error rate, two univariate probability distributions were utilized. For comparisons regarding power, data were generated under the mean distribution given in the top portion of Table 4.3. Significant results (p-value < \( \alpha = 0.05 \)) under these conditions indicate the trend test
was able to correctly reject the null hypothesis of no association and conclude the distribution of $Y$ is stochastically larger in one population compared to the other.

Conversely, data generated under the probability distributions given in the bottom portion Table 4.3 were employed to ascertain type I error rates. Significant results ($p$-value < $\alpha = 0.05$) under these conditions indicate the test incorrectly rejected the null hypothesis of no association and erroneously found evidence that the distribution of $Y$ was stochastically larger in one population compared to the other, even though all clusters were generated from the same mean distribution.

To address the open research question regarding the impact the within-cluster correlations have on the trend test, data were also generated under the three different sets of dispersion parameters give in Table 4.4, in terms of the $\theta_{jk}$ presented by Altham et al. (2012).

The rMM() function from the MM package for R (Altham et al., 2012) was then used to generate values of a three-level ($C = 3$) multinomial outcome $Y_{ij}$ based on
Table 4.5: Simulation: An Example Dataset Generated for Ascertaining Power in the Presence of Overdispersion within Clusters of Size \( n = 8 \).

<table>
<thead>
<tr>
<th>( X )</th>
<th>Cluster ((i))</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_i = 0 )</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>( x_i = 1 )</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>13</td>
<td>29</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>

The multiplicative multinomial distribution given a combination of the population specific univariate probabilities and a set of \( \binom{C}{2} = 3 \) dispersion parameters. For each of the 18 combinations of mean structure (trend, no trend), dispersion structure (standard, over, under), and cluster size \((n = 4, 8, or 12)\), 200 samples were generated, for a grand total of 3600 datasets. The standard Metropolis-Hastings simulation was used for each sample of four clusters from a given population. The first \( 4 \times n \) initial observations were ignored (e.g. burnin = 32 for clusters size \( n = 8 \)). After the initial cluster was recorded, only every \( 4n^{th} \) cluster generated from the Markov chain was added to the sample.

Results and Discussion

The ETT-CMD, described in Chapter 3, was applied to each of the 3600 datasets in order to calculate exact p-values (one- and two-sided) under all three possible correlation conditions: Equal, Adjacent, and Unequal. Figure 4.1 displays the exact distribution of the test statistics under each of the three correlation conditions for the example dataset given in Table 4.5. Although the observed test
statistic is the same ($t = 56$) for the given scores $u_i = (0,1,2)'$, the reference set is decidedly different depending on the specification of the correlation conditioning.

The aggregated results for the total 3200 simulated datasets are displayed in Figure 4.2. The nine panels correspond to the various combinations of dispersion parameterization and cluster size. The horizontal axis separates the three correlation conditioning options. Within each panel there are two types of points. The circles denote the proportion of 200 samples with significant exact two-side p-values ($\alpha = .05$) when the two univariate probabilities display opposite trends and thus $X$ and $Y$ are associated. The squares represent the same proportion, but for a different set of 200 samples which were all clusters were generated from the same univariate distribution, thus $X$ and $Y$ have no association.

At first glance, the center bottom panel seems out of place and to not follow the patterns displayed in the other panels; however, we will discuss this case after detailing the general patterns elsewhere and explain why it is so different.

As expected, data in which overdispersion was present perform better in terms of power while underdispersed data perform more poorly. Also, power was highest when the reference set was only conditioned in terms of the total intracluster correlation and power is lowest when the conditioning includes separate correlations for each column-pair being equal to the observed table. The “Equal” conditioning that results in the highest power, also exhibits inflated type I error rates. Though this was postulated prior to the simulation, the increase was not as extreme as feared (except for the middle central panel).

Now back to the exceptional bottom center panel. Of highest concern is the drastically high type I error rate of 20 percent. Recall from the discussion around Tables 4.1 and 4.2 and the observation that overdispersion within clusters results in lumpy clusters. In turn, larger clusters will have larger lumps in the non-dominant category level. If a sample happens to contain one of these exceptional
Figure 4.1: Simulation: Example Dataset Exact Distribution of the Test Statistic Under Each of the Correlation Conditions of the ETT-CMD.
clusters, the population trend, and thus association with population membership, will not be supported. Hence, highly overdispersed large clusters are problematic in much the same way having a high inter-class correlation (ICC) in a generalized estimating equations (GEE) model results in a lower effective sample size. So, though this panel seems divergent at first glance, it is to be expected. Whether data in the real world would exhibit this intensive dispersion for clusters of this size has yet to be determined.

Conclusion

When we set out to study the performance of the ETT-CMD, the proposed simulation seemed straightforward enough; however, determining a mechanism for generating multinomial data with intracluster correlation induced by distinct odds ratios for each pair of outcome levels was not trivial. Most existing research has included assuming a common global odds ratio or single measure of association, unless the members of a cluster were not exchangeable. The Multiplicative Multinomial distribution was the only method we found in the literature.

As postulated, conditioning the reference set only on the total correlation resulted in higher power at the expense of inflation of the type I error rate, when compared to further restricting the exact distribution to additionally require the adjacent, or even all pairwise, intracluster correlations be equivalent to the observed data. We were pleasantly surprised that the preliminary estimate of the cost of over simplification of the relationship did not come at a higher cost, at least within the constraints of this initial simulation study. The seemingly robust nature of the test leaves us optimistic regarding the validity of assuming a global odds ratio even when the data are generated from a more complex phenomenon.
Figure 4.2: Simulation: Comparing the Three Correlation Conditioning Options Regarding Power and Type I Error Rate.
These promising results will now springboard our future study of the performance of this test under a wider array of conditions. This will include simulating samples with a wider range of distinct pairwise odds ratios between outcome levels to verify that the robustness holds. Also of interest is simulating data under conditions of a single global odds ratio between all pairs of outcome levels to see if the more fully constrained reference set yields better performance.
CHAPTER 5

CONCLUSION

Every day, traditional statistical methodology are used world wide to study a variety of topics and provides insight regarding countless subjects. Each statistical model is based on assumptions which must be met to ensure valid results. Additionally, many statistical approaches rely on large sample asymptotic and may collapse or degenerate in the presence of small, sparse, or correlated data. This dissertation contains several advancements for detecting these situations, avoid their consequences, and analyze data in such a way to yield trustworthy results.

While logistic regression is one of the most commonly used modeling techniques for binary outcomes, the results are not always trustworthy. While the problem of complete separation is widely known, the consequences of near separability are not. We hope investigators will become more unaware that their particular data, while not completely separable, may be at risk of bias due to near separability.

We have developed a routine for determining if concern regarding standard maximum likelihood estimates (MLE) in logistic regression is warranted. This diagnostic tool signals when data are small, sparse, or correlated to the degree that the analyst should consider penalized maximum likelihood estimation (Firth’s bias correction) or an exact approach for analysis.

This advancement leverages profile likelihood confidence intervals, which are usually used as a measure of effect size, to assess the likelihood surface for lack of symmetry in a straightforward way. It is both simple to implement and takes almost no additional computation time when performing typical logistic regression,
yet it can reveal concern regarding potential bias in a method that is extremely common in everyday research.

Correlated data may arise from common situations such as multi-site medical studies, research on family units, or investigations on students within classrooms. In these circumstances, the associations between cluster members must be included in any statistical analysis testing the hypothesis of a connection between predictor and response in order for results to be valid.

Previously, investigators had to choose between using a method intended for small or sparse data while assuming independence between observations (Chi Squared Test for Independence) or a method that allowed for correlation between observations, such as generalized estimating equations (GEE) and multilevel mixed effects models (MLM), but that is only applicable for larger samples.

We set out a new method that allows for small, clustered samples to be assessed for evidence of a relationship between a binary predictor and a multinomial outcome. This innovation conditions on a multidimensional sufficient statistic during enumeration of all possible permutations of the data to construct the exact distribution of a test for the hypothesis of no association.

As the test of association allows for three different options to condition on the dispersion parameters, we conducted a salutation to investigate the repercussions of each choice when the data is generated for differential pair-wise associations between levels of the outcome.

Although we confirmed that conditioning only on the total association (Equal deltas) does inflate the type I error rate despite increasing the power of the test, the increase is not as detrimental as previously feared. Although these initial investigations into the performance of this new test are promising, more inspection is planned.

Together these advancements allow researchers to study small, sparse, and
correlated data where other methods fall short or are known to exhibit bias.
REFERENCES


Dutcher, J. (2014) What is big data?


John Wiley; Sons Ltd.


Traylor, R. (2017) A generalized multinomial distribution from dependent categorical


APPENDICES
Appendix A: R Code for Chapter 2

Required: R Packages

```r
packages = c("tidyverse",
             "broom",
             "purrrlyr",
             "forcats",
             "pander",
             "xtable",
             "HSAUR",
             "reshape",
             "stargazer",
             "texreg",
             "logistf")

package.check <- lapply(packages, FUN = function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
})
```

Table 2.1 on page 8

Illustrative Dataset: Fifth Observation with Variable Value $k$.

```r
x <- c(10, 12, 16, 14, 18, 13, 11, 15, 19, 17)
x1 <- rep(x, times = 2)
x2 <- c(x, x + 6)
y <- rep(c(1, 0, 1, 0), times = c(4, 1, 5, 10))
data <- rbind("X1" = x1,
               "X2" = x2,
               "Y" = y)
data[1, 5] <- 18
data[2, 5] <- "k"
data[3, 5] <- 0
colnames(data) <- 1:ncol(data)
colnames(data)[5] <- 5
data \%>%
  xtable(comment = FALSE,
          align = paste(c("l|", rep("c", 20)), collapse = ""),
          caption = "Illustrative Dataset: Fifth Observation with Variable Value $k$.",
          label = "tab:ex1data") \%>
print(hline.after = c(-1, 0, 2, 3),
      booktabs = TRUE,
      sanitize.text.function = function(x) {x},
      caption.placement = "top",
      scalebox='0.85')
```
**Figure 2.3** on page 8

Illustrative Dataset: Covariate Space.

data <- rbind("X1" = rep(x, 2),
   "X2" = c(x, x + 6),
   "Y" = rep(c(1, 0, 1, 0), times = c(4, 1, 5, 10)))
t(data) %>%
data.frame %>%
dplyr::mutate(Y = Y %>% factor) %>%
ggplot(aes(x = X1,
   y = X2,
   shape = Y)) +
  annotate("text", x = 18.6, y = 15, size = 7, label = "k") +
  annotate("text", x = 18, y = 22, size = 4, label = "separable") +
  annotate("text", x = 17.5, y = 13, size = 4, label = "not\nseparable") +
  geom_segment(aes(x = 17.5, y = 18, xend = 18.55, yend = 18)) +
  geom_segment(aes(x = 18.15, y = 0, xend = 18.55, yend = 0)) +
  geom_segment(aes(x = 18.35, y = 0, xend = 18.35, yend = 18),
    arrow = arrow(length = unit(.15, "inches"))) +
  geom_segment(aes(x = 18, y = 18, xend = 18, yend = 15),
    color = "blue",
    arrow = arrow(type = "closed",
      length = unit(.12, "inches"))) +
  geom_segment(aes(x = 18, y = 18, xend = 18, yend = 21),
    arrow = arrow(type = "closed",
      length = unit(.12, "inches")),
    color = "red") +
  geom_point(size = 4, fill = "white") +
  theme_bw() +
  scale_shape_manual(values = c(21, 15)) +
  scale_x_continuous(breaks = seq(from = 10, to = 20, by = 2)) +
  theme(legend.position = c(.135, .115),
    legend.direction = "horizontal",
    legend.background = element_rect(color = "black"))
log_coef <- function(data, k) {
  df <- t(data) %>%
    data.frame %>%
    dplyr::mutate(X2 = ifelse(X1 == 18 & Y == 0, k, X2))
  fit <- glm(Y ~ ., data = df, family = binomial)
  return(fit)
}

data.frame(k = c(seq(from = 5, to = 17.99, by = 0.01))) %>%
purrr::invoke_rows(.f = log_coef, data = data) %>%
dplyr::mutate(fit = map(.out, tidy)) %>%
tidyr::unnest(fit) %>%
dplyr::filter(term != "(Intercept)") %>%
tidyr::gather(key = term, value = est, estimate, std.error) %>%
dplyr::mutate(term = term %>% factor) %>%
dplyr::mutate(type = type %>% factor) %>%
dplyr::mutate(type_text = case_when(type == "estimate" ~
  "A. Estimate (MLE)",
  type == "std.error" ~
  "B. Standard Error (SE)")) %>

ggplot(aes(x = k, y = est,
    linetype = term,
    color = term)) +
facet_wrap(~ type_text,
    ncol = 1,
    scale = "free_y") +
geom_line(size = 1) +
theme_bw() +
labs(y = NULL,
    linetype = "Parameter",
    color = "Parameter") +
theme(legend.position = c(.138, .33),
    legend.key.width = unit(1.25, "cm"),
    legend.background = element_rect(color = "black")) +
scale_x_continuous(breaks = seq(from = 4, to = 20, by = 2)) +
scale_color_manual(values = c("black", "blue")) +
scale_linetype_manual(values = c("solid", "dashed")) +
geom_vline(xintercept = 18,
    linetype = "dotted")
**Figure 2.5** on page 11
Collett’s ESR data: Covariate Space.

```r
data("plasma")

df_2 <- data.frame(new = c(3.32, 3.33, 3.35, 3.359, 3.36),
                    const = c(32, 32, 32, 32, 32))

plasma %>%
dplyr::filter(ESR == "ESR < 20" | fibrinogen > 3) %>%
dplyr::mutate(ESR = ESR %>%
              forcats::fct_recode("No (Y = 0)" = "ESR < 20",
                                   "Yes (Y = 1)" = "ESR > 20")) %>%
ggplot(aes(x = fibrinogen,
               y = globulin,
               shape = ESR)) +
geom_segment(aes(x = 3.34, y = 30.5, xend = 3.34, yend = 33.5)) +
geom_segment(aes(x = 3.34, y = 32, xend = 3.0, yend = 32),
               color = "blue",
               arrow = arrow(type = "closed",
                              length = unit(.12, "inches"))) +
geom_segment(aes(x = 3.34, y = 32, xend = 3.68, yend = 32),
               arrow = arrow(type = "closed",
                              length = unit(.12, "inches")),
               color = "red") +
geom_vline(xintercept = 3.36,
               color = "Red",
               linetype = "dashed") +
geom_point(size = 3,
            fill = "white") +
theme_bw() +
labs(x = "Fibrinogen",
     y = "Gamma Globulin",
     shape = "ESR > 20 mm/hr:" ) +
scale_shape_manual(values = c(21, 15)) +
theme(legend.position = c(.82, .8),
      legend.background = element_rect(color = "black")) +
annotate("text", x = 3.7, y = 33.5, size = 4, label = "separable")
```
Table 2.2 on page 12

Collett's ESR Data: Effect of Near Separation on the MLE and SE.

```r
plasma_reduced <- plasma %>%
dplyr::filter(ESR == "ESR < 20" | fibrinogen > 3) %>%
dplyr::mutate(ESR = as.numeric(ESR == "ESR > 20"))

fit_0 <- plasma_reduced %>% glm(ESR ~ ., data = ., family=binomial)

fit_1 <- plasma_reduced %>%
dplyr::mutate(fibrinogen = ifelse(globulin == 32 & fibrinogen == 3.34, 3.35, fibrinogen)) %>%
  glm(ESR ~ ., data = ., family=binomial)

fit_2 <- plasma_reduced %>%
dplyr::mutate(fibrinogen = ifelse(globulin == 32 & fibrinogen == 3.34, 3.359, fibrinogen)) %>%
  glm(ESR ~ ., data = ., family=binomial)

fit_3 <- plasma_reduced %>%
dplyr::mutate(fibrinogen = ifelse(globulin == 32 & fibrinogen == 3.34, 3.359999, fibrinogen)) %>%
  glm(ESR ~ ., data = ., family=binomial)

fit_4 <- plasma_reduced %>%
dplyr::mutate(fibrinogen = ifelse(globulin == 32 & fibrinogen == 3.34, 3.33, fibrinogen)) %>%
  glm(ESR ~ ., data = ., family=binomial)

fit_5 <- plasma_reduced %>%
dplyr::mutate(fibrinogen = ifelse(globulin == 32 & fibrinogen == 3.34, 3.32, fibrinogen)) %>%
  glm(ESR ~ ., data = ., family=binomial)

stargazer(fit_4, fit_0, fit_1, fit_2, fit_3,
  keep = c("fibrinogen", "globulin"),
  dep.var.caption = "Permute the Single Point's Fibrinogen Value",
  dep.var.labels = "(True value is 3.34)",
  column.labels = c("3.33", "3.34", "3.35", "3.359", "3.359999"),
  model.numbers = FALSE,
  header = FALSE,
  label = "tab:esrBeta",
  title = "Collett's ESR Data: Effect of Near Separation on the MLE and SE."
)
```
Figure 2.8 on page 23

Illustration of a Generic Profile Likelihood Confidence Interval.

num.stems <- c(6, 8, 9, 6, 6, 2, 5, 3, 1, 4)

# generate raw data from tabulated values
aphid.data <- rep(0:9, num.stems)

# ML estimation for Poisson model
poisson.LL <- function(lambda){ sum(log(dpois(aphid.data, lambda)))}
poisson.negloglik <- function(lambda){ -poisson.LL(lambda) }

# Non-linear minimization
out <- nlm(poisson.negloglik, 3, hessian = TRUE)

# Lower Limit for the Likelihood
lower.limit <- -out$minimum - 0.5 * qchisq(0.95, 1)

# Profile Likelihood Confidence Interval Endpoints
plCI_lower <- out$estimate - qnorm(.975)*sqrt(1/out$hessian)
plCI_upper <- out$estimate + qnorm(.975)*sqrt(1/out$hessian)

data.frame(lambda = seq(from = 2.3, to = 4.7, by = 0.1)) %>%
dplyr::rowwise() %>%
dplyr::mutate(loglik = poisson.LL(lambda)) %>%
ggplot(aes(x = lambda, y = loglik)) +
geom_line(size = 2.25, alpha = .5) +
theme_bw() +
geom_hline(yintercept = lower.limit, linetype = "longdash", size = 1) +
geom_segment(aes(x = plCI_lower, xend = plCI_lower, y = lower.limit, yend = -137), linetype = "dotted", size = 1, arrow = arrow(length = unit(0.4, "cm"), type = "closed")) +
geom_segment(aes(x = plCI_upper, xend = plCI_upper, y = lower.limit, yend = -137), linetype = "dotted", size = 1,
```
arrow = arrow(length = unit(0.4, "cm"),
    type = "closed")
+ geom_point(aes(x = out$estimate,
                y = -out$minimum),
    shape = 15,
    size = 2.5) +
geom_segment(aes(x = out$estimate, xend = out$estimate,
                y = -out$minimum, yend = lower.limit),
    arrow = arrow(length = unit(0.25, "cm"),
                  ends = "both") +
geom_segment(aes(x = plCI_lower, xend = plCI_upper,
                y = -135, yend = -135),
    arrow = arrow(length = unit(0.25, "cm"),
                  ends = "both",
                  type = "closed")) +
labs(x = "Parameter, lambda",
     y = "Log Likelihood") +
annotate("text", x = 2.55, y = -124.5,
         label = "Log Likelihood\nLower Limit") +
annotate("text", x = out$estimate, y = -134.5,
         label = "plCI") +
annotate("text", x = out$estimate, y = -122.7,
         label = "Maximum Likelihood")
```
**Figure 2.9** on page 26

Illustration of plCIs for Logistic Regression.

```r
NB.LL <- function(mu, theta){
    sum(log(dnbinom(aphid.data,
    mu = mu,
    size = theta)))
}

# for nlm we need negative LL and a function of a vector
NB.LL.p <- function(p){ -NB.LL(p[1], p[2]) }

# obtain MLEs
out.NB <- nlm(NB.LL.p, c(4,4))

# generate grid of values
mu <- seq(2,5, 0.1)
theta <- seq(1, 9, 0.1)
g <- expand.grid(mu, theta)

g$z <- apply(g, 1, NB.LL.p)

# reorganize z as a matrix same shape as grid
zmat <- matrix(g$z, nrow = length(seq(2, 5, 0.1)))

# boundary of acceptance region
lower.L <- -out.NB$minimum - 0.5 * qchisq(.95, df = 1)

# Fig. 2a

test.func <- function(x){ NB.LL(x, out.NB$estimate[2]) - lower.L }
left.mu <- unroot(test.func, c(2.5, 3))
right.mu <- unroot(test.func, c(4, 5))

par(mfrow = c(1, 2))

plot(seq(2.2, 5.2, 0.01), sapply(seq(2.2, 5.2, 0.01),
    function(x) NB.LL(x, out.NB$estimate[2])),
    xlab = expression(mu),
    ylab = 'log-likelihood',
    type = 'l')
```
inc <- seq(left.mu$root, right.mu$root, 0.01)

sapply(inc,
    function(x) segments(x, lower.L,
        x, NB.LL(x, out.NB$estimate[2]),
        col = 'grey70')) -> junk

abline(h = - out.NB$minimum-.5 * qchisq(0.95, df=1),
    col = 2,
    lty = 2)

points(out.NB$estimate[1],
    -out.NB$minimum - 0.5 * qchisq(0.95, df = 1),
    col = 2,
    pch = 15)

arrows(left.mu$root,
    - out.NB$minimum - 0.5 * qchisq(0.95, df = 1) - 0.5,
    right.mu$root,
    - out.NB$minimum - 0.5 * qchisq(0.95, df = 1) - 0.5,
    code = 3,
    angle = 45,
    length = 0.09,
    col = 2)

#Fig. 2b

test.func2 <- function(x) NB.LL(out.NB$estimate[1],x) - lower.L
left.theta <- uniroot(test.func2, c(1, 2))
right.theta <- uniroot(test.func2, c(6, 7))

plot(seq(1, 7, 0.01),
    sapply(seq(1, 7, 0.01),
        function(x) NB.LL(out.NB$estimate[1], x)),
    xlab = expression(theta),
    ylab = 'log-likelihood',
    type = 'l')

inc <- seq(left.theta$root, right.theta$root, 0.05)

sapply(inc,function(x) segments(x,lower.L,
    x, NB.LL(out.NB$estimate[1], x),
    col = 'grey70')) -> junk
abline(h = - out.NB$minimum - 0.5 * qchisq(0.95, df = 1),
       col = 2,
       lty = 2)

points(out.NB$estimate[2],
       - out.NB$minimum - 0.5 * qchisq(0.95, df = 1),
       col = 2,
       pch = 15)

arrows(left.theta$root,
       - out.NB$minimum - 0.5 * qchisq(0.95, df = 1) - 0.25,
       right.theta$root,
       - out.NB$minimum - 0.5 * qchisq(0.95, df = 1) - 0.25,
       code = 3,
       angle = 45,
       length = 0.09,
       col = 2)

par(mfrow = c(1, 1))
**Table 2.4 on page 27**

Simple 2x2: Observed Contingency Table

```r
data2 <- expand.grid(x = 0:1, y = 0:1) %>%
dplyr::mutate(count = c(6, 4, 4, 6)) %>%
reshape::untable(num = .$count) %>%
dplyr::select(x, y)

addtorow <- list()
addtorow$pos <- list(0, 0)
addtorow$command <- c("& \multicolumn{2}{c}{x} \\
 & 0 & 1 & Sum \\
 & y & 0 & 1 & Sum \n")

table(data2) %>%
  addmargins() %>%
  xtable(digits = 0,
         align = "c|cc|c",
         comment = FALSE,
         caption = "Simple 2x2:
Observed Contingency Table",
         label = "tab:d1") %>%
  print(hline.after = c(-1, 0, 2, 3),
        add.to.row = addtorow,
        include.colnames = FALSE,
        booktabs = TRUE)
```

---

**Table 2.4 on page 27**

Simple 2x2: Observed Contingency Table

```r
data2 <- expand.grid(x = 0:1,
                     y = 0:1) %>%
dplyr::mutate(count = c(6, 4, 4, 6)) %>%
reshape::untable(num = .$count) %>%
dplyr::select(x, y)

addtorow <- list()
addtorow$pos <- list(0, 0)
addtorow$command <- c("& \multicolumn{2}{c}{x} \\
 & 0 & 1 & Sum \\
 & y & 0 & 1 & Sum \n")

table(data2) %>%
  addmargins() %>%
  xtable(digits = 0,
         align = "c|cc|c",
         comment = FALSE,
         caption = "Simple 2x2:
Observed Contingency Table",
         label = "tab:d1") %>%
  print(hline.after = c(-1, 0, 2, 3),
        add.to.row = addtorow,
        include.colnames = FALSE,
        booktabs = TRUE)
```
**Table 2.5** on page 27

Simple 2x2: Logistic Regression MLEs for Observed Contingency Table.

```r
fit_mle <- glm(y ~ x, data = data2, family = binomial)
b_x <- fit_mle$coefficients["x"]
```

```r
fit_mle %>%
  stargazer(title = "Simple 2x2: Logistic Regression MLEs for Observed Contingency Table.",
             label = "tab:data1mle",
             header = FALSE,
             single.row = TRUE,
             ci = TRUE)
```
Function for Logistic Regression: plCI Ratio & Bias.

```r
log4 <- function(df, #data.frame
  f, #logical: TRUE = firth vs. MLE
  p){ #logical: TRUE = plCI vs. Wald

  fit <- logistf::logistf(y ~ x,
    data = df,
    firth = f,
    pl = p)

  est <- coefficients(fit) %>%
    broom::tidy() %>%
    dplyr::filter(names == "x") %>%
    dplyr::select(-names) %>%
    unlist(use.names = FALSE)

  ci <- confint(fit) %>%
    broom::tidy() %>%
    dplyr::filter(.rownames == "x") %>%
    dplyr::select(-.rownames)

  out <- cbind(Firth = f, pl = p, Estimate = est, ci)

  return(out)
}
```

```markdown
results <- rbind(log4(data2, FALSE, FALSE),
  log4(data2, FALSE, TRUE),
  log4(data2, TRUE, FALSE),
  log4(data2, TRUE, TRUE)) %>%
  dplyr::mutate(low = Estimate - Lower.95.) %>%
  dplyr::mutate(upp = Upper.95. - Estimate) %>%
  dplyr::group_by(Firth, pl) %>%
  dplyr::mutate(Ratio = max(low, upp)/min(low, upp)) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(Method = case_when(.Firth == TRUE ~ "PMLE",
                                  TRUE ~ "MLE")) %>%
  dplyr::mutate(CI_Type = case_when(.pl == TRUE ~ "pl",
                                   TRUE ~ "Wald")) %>%
  dplyr::mutate(OR = exp(Estimate)) %>%
  dplyr::mutate(CI_Width = Upper.95. - Lower.95.) %>%
  dplyr::select(Method, Estimate, OR,
                 CI_Type, Lower.95., Upper.95.,
```
\text{CI\_Width, Ratio}
Table 2.6 on page 28

Simple 2x2: MLE and PMLE Estimates with wCI and plCIs for Observed Contingency Table.

```r
results %>%
dplyr::select(-OR, -CI_Width) %>%
xtable(digits = 3,  
    align = "ccc|ccc|c",  
    caption = "Simple 2x2:
    MLE and PMLE Estimates with wCI and
    plCIs for Observed Contingency Table.",
    label = "tab:d1est") %>%
print(include.rownames = FALSE,  
    hline.after = c(-1,0,2,4),  
    booktabs = TRUE,  
    caption.placement = "top")
```
Figure 2.11 on page 31

Simple 2x2 Crosstabulation: Probability Density of the Reference Set.

\[ P_{\text{obs}} = 0.238693 \]

```r
data.frame(t1 = 0:10, prob = c(5.41254e-006, 0.000541254, 0.0109604, 0.0779406, 0.238693, 0.343718, 0.238693, 0.0779406, 0.0109604, 0.000541254, 5.41254e-006)) %>%
dplyr::mutate(obs = case_when(t1 == 6 ~ 1, prob <= P_obs ~ 2, TRUE ~ 3) %>%
  factor(levels = 1:3, labels = c("Observed", "More Extreme", "Less Extreme"))) %>%
  ggplot(aes(x = t1, y = prob, fill = forcats::fct_rev(obs))) +
  geom_hline(yintercept = P_obs, size = 2, linetype = "dashed", alpha = .3) +
  geom_segment(aes(xend = t1, y = 0, yend = prob)) +
  geom_point(size = 3, shape = 21) +
  theme_bw() +
  labs(x = "Test Statistic", y = "Probability", fill = NULL, size = NULL) +
  scale_fill_manual(values = c("white", "gray", "black")) +
  theme(legend.position = c(.828, .78), legend.background = element_rect(color = "black"))
```
Figure 2.12 on page 33

Simple 2x2: Simulation to Study Relationship Between plCI Asymmetry Ratio and Absolute Value of MLE Bias.

```r
sim_log <- function(n_obs, beta_0, beta_1) {
  data.frame(id = 1:n_obs,
             x = rep(0:1, each = n_obs/2)) %>%
  dplyr::mutate(eta = beta_0 + beta_1 * x) %>%
  dplyr::mutate(pi = exp(eta)/(1 + exp(eta))) %>%
  dplyr::rowwise() %>%
  dplyr::mutate(y = rbinom(n = 1, size = 1, prob = pi)) %>%
  dplyr::select(x, y)
}

log4 <- function(data, f, p){
  fit <- logistf(y ~ x,
                 data = data,
                 firth = f,
                 pl = p)

  est <- coefficients(fit) %>%
         tidy %>%
         filter(names == "x") %>%
         dplyr::select(-names) %>%
         unlist(use.names = FALSE)

  ci <- confint(fit) %>%
       tidy %>%
       filter(.rownames == "x") %>%
       dplyr::select(-.rownames)

  out <- cbind(Firth = f,
               pl = p,
               Estimate = est,
               ci)
  return(out)
}

run_log2 <- function(n_obs, beta_0, beta_1){
  sim_log(n_obs, beta_0, beta_1) %>%
  log4(TRUE, TRUE) %>%
  dplyr::mutate(lower = Estimate - Lower.95.) %>%
```
```r
dplyr::mutate(upper = Upper.95. - Estimate) %>%
dplyr::mutate(plCI_ratio = max(lower, upper)/min(lower, upper)) %>%
dplyr::mutate(beta_bias = Estimate - beta_1) %>%
dplyr::select(plCI_ratio, beta_bias)
}

reps = 50
set.seed(1234)

cycle1 <- data.frame(rep = 1:reps,
  n_obs = 10) %>%
tidyrr::crossing(data.frame(beta_0 = c(-1, 0, +1))) %>%
tidyrr::crossing(data.frame(beta_1 = c(0, 1, 3))) %>%
dplyr::select(-rep) %>%
purrrlyr::invoke_rows(.f = run_log2) %>%
tidyrr::unnest(.out)

cycle1 %>%
ggplot(aes(x = log(plCI_ratio),
  y = abs(beta_bias))) +
  geom_hline(yintercept = 0,
    alpha = .5) +
  geom_vline(xintercept = 0,
    alpha = .5) +
  geom_count() +
  theme_bw() +
  facet_grid(beta_1 ~ beta_0,
    labeller = "label_both") +
  theme(legend.position = "none")
```
**Function: data_log** Simulate data for logistic outcome

* `n_0` number of observations with $y = 0$
* `n_1` number of observations with $y = 1$
* `a` distance between centers
* `r` covariance
* `v` variance

Generate two sets of data, one for $y = 0$ and the other for $y = 1$ with two multivariate normal covariates.

data_log <- function(n_0, n_1, a, r, v){
  rbind(MASS::mvrnorm(n = n_0,
                      mu = c(x_1 = 0, x_2 = 0),
                      Sigma = matrix(c(v, r, r, v), nrow = 2)) %>%
         cbind(y = 0),
       MASS::mvrnorm(n = n_1,
                     mu = c(x_1 = a, x_2 = a),
                     Sigma = matrix(c(v, r, r, v), nrow = 2)) %>%
         cbind(y = 1)) %>%
  data.frame
}


Simulated Data for Logistic Regression: Two Multivariate Normally Distributed Predictors with Various Degrees of Correlation.

```r
expand.grid(a = c(0, 0.5, 1, 2),
             r = c(.05, .25, .5, .75)) %>%
dplyr::mutate(n_0 = 20,
               n_1 = 20,
               v = 1) %>%
purrrlyr::invoke_rows(.f = data_log, .d = .) %>%
tidyrr::unnest(.out) %>%
dplyr::mutate(y = factor(y)) %>%
ggplot(aes(x = x_1, y = x_2, shape = y)) +
geom_point() +
theme_bw() +
facet_grid(r ~ a, labeller = "label_both") +
scale_shape_manual(values = c(1, 4)) +
theme(legend.position = "bottom")
```
Function: MLE_plCI_ratio

Compute the MLE and plCI ratio for logistic regression data from the data_long() function

```
MLE_plCI_ratio <- function(df) {
  # df = data frame
  logistf(y ~ x_1 + x_2,
    data = df,
    pl = TRUE,
    firth = FALSE)[c(1, 17, 19:21)] >>
  data.frame >>
  tibble::rownames_to_column("variable") >>
  dplyr::mutate(left = coefficients - ci.lower) >>
  dplyr::mutate(right = ci.upper - coefficients) >>
  dplyr::rowwise() >>
  dplyr::mutate(ratio = max(left, right)/min(left, right)) >>
  dplyr::rename(var = variable,
    est = coefficients,
    ratio = ratio) >>
  dplyr::select(var, est, ratio)
}
```
Function: sim_fit

```r
sim_fit <- function(n_0, n_1, r, a = 1, v = 1, reps = 1000) {

  dat <- data.frame(id = 1:reps) %>%
      dplyr::mutate(r = r,
                  a = a,
                  n_0 = n_0,
                  n_1 = n_1,
                  v = v) %>%
      dplyr::select(-id) %>%
      purrrlyr::invoke_rows(.f = data_log, .d = )

  dat_fit <- data.frame(id = 1) %>%
      dplyr::mutate(r = r,
                   a = a,
                   n_0 = 10000,
                   n_1 = 10000,
                   v = v) %>%
      dplyr::select(-id) %>%
      purrrlyr::invoke_rows(.f = data_log, .d = ) %>%
      dplyr::mutate(fit = map(.out, MLE_plCI_ratio)) %>%
      dplyr::mutate(id = row_number()) %>%
      tidyr::unnest(fit) %>%
      dplyr::mutate(est_true = est) %>%
      dplyr::select(var, est_true)

  fit <- dat %>%
      dplyr::mutate(fit = map(.out, MLE_plCI_ratio)) %>%
      dplyr::mutate(id = row_number()) %>%
      tidyr::unnest(fit) %>%
      dplyr::left_join(dat_fit, by = "var") %>%
      dplyr::mutate(bias = abs(est - est_true))

  return(fit)
}
```
Run Simulation

```r
fit_10_05 <- sim_fit(n_0 = 10, n_1 = 10, r = 0.05)
fit_10_25 <- sim_fit(n_0 = 10, n_1 = 10, r = 0.25)
fit_10_50 <- sim_fit(n_0 = 10, n_1 = 10, r = 0.50)
fit_10_75 <- sim_fit(n_0 = 10, n_1 = 10, r = 0.75)

fit_12_05 <- sim_fit(n_0 = 8, n_1 = 12, r = 0.05)
fit_12_25 <- sim_fit(n_0 = 8, n_1 = 12, r = 0.25)
fit_12_50 <- sim_fit(n_0 = 8, n_1 = 12, r = 0.50)
fit_12_75 <- sim_fit(n_0 = 8, n_1 = 12, r = 0.75)

fit_14_05 <- sim_fit(n_0 = 6, n_1 = 14, r = 0.05)
fit_14_25 <- sim_fit(n_0 = 6, n_1 = 14, r = 0.25)
fit_14_50 <- sim_fit(n_0 = 6, n_1 = 14, r = 0.50)
fit_14_75 <- sim_fit(n_0 = 6, n_1 = 14, r = 0.75)

fits <- bind_rows(fit_10_05, fit_10_25, fit_10_50, fit_10_75,
                  fit_12_05, fit_12_25, fit_12_50, fit_12_75,
                  fit_14_05, fit_14_25, fit_14_50, fit_14_75)
```
Figures 2.14 and 2.15 on pages 36 and 37

```r
fits_plot <- fits %>%
dplyr::mutate(n = paste0("(\", n_0, ", \", n_1, ")") %>%
  factor(labels = c("(6, 14)",
    "(8, 12)",
    "(10, 10)"))) %>%
dplyr::group_by(n, r, id) %>%
dplyr::summarise(bias_max = max(bias),
  bias_mean = mean(bias),
  ratio_max = max(ratio),
  ratio_mean = mean(ratio)) %>%
ggplot(aes(x = ratio_max,
  geom_point(shape = 21,
    alpha = .3) +
  theme_bw() +
  labs(x = "Maximum plCI Ratio, Among All Parameters",
    y = "Maximum Absolute Bias, Among All Parameters") +
  facet_grid(r ~ n, labeller = "label_both")

fits_plot +
  coord_trans(x = "log",
    y = "log")

fits_plot +
  coord_cartesian(xlim = c(1, 6),
    ylim = c(0, 15))
```
Table 2.9 on page 38

Collett's ESR data: MLE and PMLE Estimates with wCI and plCIs.

```r
my_log <- function(f, p, df){
  fit <- logistf(ESR ~ fibrinogen + globulin,
               data = df,
               firth = f,
               pl = p)

  est <- coefficients(fit) %>%
  broom::tidy() %>%
  dplyr::filter(names %in% c("fibrinogen","globulin")) %>%
  dplyr::select(-names) %>%
  unlist(use.names = FALSE)

  ci <- confint(fit) %>%
  broom::tidy() %>%
  dplyr::filter(.rownames %in% c("fibrinogen","globulin")) %>%
  dplyr::select(-.rownames)

  out <- cbind(Firth = f,
               pl = p,
               var = c("fibrinogen","globulin"),
               Estimate = est,
               ci)

  return(out)
}

plasma_results <- rbind(my_log(FALSE, FALSE, plasma_reduced),
                         my_log(FALSE, TRUE, plasma_reduced),
                         my_log(TRUE, FALSE, plasma_reduced),
                         my_log(TRUE, TRUE, plasma_reduced)) %>%
  dplyr::mutate(low = Estimate - Lower.95.) %>%
  dplyr::mutate(upp = Upper.95. - Estimate) %>%
  dplyr::group_by(Firth, pl, var) %>%
  dplyr::mutate(Ratio = max(low, upp)/min(low, upp)) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(Method = case_when(.Firth == TRUE ~ "PMLE",
                                   TRUE ~ "MLE")) %>%
  dplyr::mutate(CI_Type = case_when(.Fpl == TRUE ~ "Profile Likelihood",
                                    TRUE ~ "Classic Wald")) %>%
  dplyr::mutate(OR = exp(Estimate)) %>%
  dplyr::mutate(CI_Width = Upper.95. - Lower.95.) %>%
  dplyr::select(Method, var,Estimate, OR, CI_Type,
```

Lower.95., Upper.95., CI_Width, Ratio) %>%
dplyr::mutate(CI_Type = CI_Type %>%
  forcats::fct_recode("Wald" = "Classic Wald",
  "pl" = "Profile Likelihood"))

plasma_results %>%
dplyr::rename("Variable" = var) %>%
dplyr::select(-OR, -CI_Width) %>%
  xtable(digits = 3,
    align = "cccc|ccc|c",
    caption = "Collett's ESR data:
    MLE and PMLE Estimates with wCI and plCIs.",
    label = "tab:ESRest") %>%
  print(include.rownames = FALSE,
    hline.after = c(-1, 0, 4, 8),
    booktabs = TRUE,
    caption.placement = "top")
Figure 2.16 on page 38
Collett’s ESR data: Asymmetry of Profile Likelihood Confidence Intervals.

plasma_results %>%
  ggplot(aes(x = CI_Type,
            y = Estimate)) +
  geom_errorbar(aes(ymin = Lower.95.,
                     ymax = Upper.95.,
                     color = CI_Type),
                size = 1.5,
                width = 0.4) +
  geom_hline(aes(yintercept = Estimate)) +
  geom_text(aes(label = Ratio %>% round(2)), nudge_x = .25, vjust = -.2) +
  geom_point(size = 2) +
  facet_grid(var ~ Method, scale = "free") +
  theme_bw() +
  theme(legend.position = "none") +
  labs(x = NULL,
       y = "Parameter Estimate, logOR") +
  scale_color_manual(values = c("darkgray", "black"))
Appendix B: R Code for Chapter 3

R Packages

```r
packages = c("tidyverse",
        "forcats",
        "pander",
        "xtable",
        "readxl")

package.check <- lapply(packages, FUN = function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
})
```

Table 3.1 on page 45

CHED: Number of Rejected Corneal Grafts.

```r
eye <- data.frame(id = 1:9,
  age = rep(0:1, times = c(4, 5)),
  rej = rep(0:1, times = c(5, 4)),
  num = rep(2:1, times = c(7, 2)))

# Mutate age
age <- factor(age, levels = 0:1, labels = c("< 3 yrs", ">= 3 yrs"))

dplyr::mutate(eye, age = age %>%
  factor(levels = 0:1, labels = c("< 3 yrs", ">= 3 yrs")))

addtorow <- list()
addtorow$pos <- list(-1)
addtorow$command <- c("\\multicolumn{2}{c}{Age at Diagnosis} \\\n"
)

# Mutate frac
frac <- paste0(rej, "/", num)

# Select age, frac
select(eye, age, frac)

# Mutate order
order <- c(1:4, 1:5)

# Spread frac by age
spread(eye, key = age, value = frac)

# Select order, exclude
select(-order)

# Print table
print(eye, include.rownames = FALSE, hline.after = c(-1, 0, 5),
  booktabs = TRUE, sanitize.text.function = function(x) {x},
  include.colnames = FALSE, caption.placement = "top",
  scalebox='0.85',
  add.to.row = addtorow)
```
**Figure 3.1** on page 66  
Multicenter Pain and Spasm Study: Exact Distribution of the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)

```r
equal <- read.table("data/painpartial_2mult_equal.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Equal")

unequal <- read.table("data/painpartial_2mult_unequal.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Unequal")

adjacent <- read.table("data/painpartial_2mult_adjacent.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Adjacent")

data <- rbind(equal, adjacent, unequal)

t_obs = 38

P_obs <- data %>%
dplyr::filter(Statistic == t_obs) %>%
dplyr::rename(point_prob = Probability) %>%
dplyr::select(point_prob, corr_type)

data %>%
dplyr::left_join(P_obs, by = "corr_type") %>%
dplyr::mutate(corr_type = corr_type %>%
  factor(levels = c("Equal", "Adjacent", "Unequal"),
        labels = c("A. Equal (Total Correlation)", "B. Adjacent (Adjacent Correlations)", "C. Unequal (Intra-Cluster Correlation)")) %>%
dplyr::mutate(obs = case_when(Statistic == t_obs ~ 1,
                           Probability <= point_prob ~ 2,
                           TRUE ~ 3) %>%
  factor(levels = 1:3,
         labels = c("Observed", "More Extreme", "Less Extreme"))) %>%
ggplot(aes(x = Statistic,
               y = Probability,
               color = corr_type),
       geom_point() +
       labs(title = "Figure 3.1 Multicenter Pain and Spasm Study: Exact Distribution of the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)",
            x = "Statistic",
            y = "Probability") %>%
ggtitle("Figure 3.1 Multicenter Pain and Spasm Study: Exact Distribution of the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)"),
```

```
fill = obs %>% fct_rev()) +
facet_wrap(~ corr_type, ncol = 1) +
geom_segment(aes(xend = Statistic,
    y = 0,
    yend = Probability,
    color = obs %>% fct_rev()),
    size = 4,
    alpha = .5) +
geom_segment(aes(xend = Statistic,
    y = 0,
    yend = Probability)) +
geom_point(size = 2.5,
    shape = 21) +
theme_bw() +
labs(x = "Test Statistic",
    y = "Probability",
    fill = NULL) +
scale_color_manual(values = c("white", "gray", "black")) +
scale_fill_manual(values = c("white", "gray", "black")) +
theme(legend.position = c(.125, .915),
    legend.background = element_rect(color = "black")) +
guides(color = FALSE)
Figure 3.2 on page 68
Toxicology Study on Ethylene Glycol in Mice: Exact Distribution of the Test Statistic for the Exact Trend Test for Correlated Multinomial Data (ETT-CMD)

```r
equal <- read.table("data/Toxicpartial_2mult_equal.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Equal")

unequal <- read.table("data/Toxicpartial_2mult_unequal.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Unequal")

adjacent <- read.table("data/Toxicpartial_2mult_adjacent.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Adjacent")

data <- rbind(equal, adjacent, unequal)
t_obs = 113

P_obs <- data %>%
dplyr::filter(Statistic == t_obs) %>%
dplyr::rename(point_prob = Probability) %>%
dplyr::select(point_prob, corr_type)

data %>%
dplyr::left_join(P_obs, by = "corr_type") %>%
dplyr::mutate(corr_type = corr_type %>%
  factor(levels = c("Equal", "Adjacent", "Unequal"),
         labels = c("A. Equal (Total Correlation)", "B. Adjacent (Adjacent Correlations)", "C. Unequal (Intra-Cluster Correlation)"))) %>%
dplyr::mutate(obs = case_when(Statistic == t_obs ~ 1,
                         Probability <= point_prob ~ 2,
                         TRUE ~ 3) %>%
  factor(levels = 1:3,
         labels = c("Observed", "More Extreme", "Less Extreme"))) %>%
ggplot(aes(x = Statistic, y = Probability,
            fill = obs))
```

```r
```
fill = obs %>% fct_rev()) +
facet_wrap(~ corr_type, ncol = 1) +
geom_segment(aes(xend = Statistic,
    y = 0,
    yend = Probability,
    color = obs %>% fct_rev()),
    size = 4,
    alpha = .5) +
geom_segment(aes(xend = Statistic,
    y = 0,
    yend = Probability)) +
geom_point(size = 2.5,
    shape = 21) +
theme_bw() +
labs(x = "Test Statistic",
    y = "Probability",
    fill = NULL) +
scale_color_manual(values = c("white", "gray", "black")) +
scale_fill_manual(values = c("white", "gray", "black")) +
theme(legend.position = c(.125, .915),
    legend.background = element_rect(color = "black")) +
guides(color = FALSE)
Appendix C: R Code for Chapter 4

R Packages

```r
packages = c("tidyverse",
             "broom",
             "purrrlyr",
             "forcats",
             "pander",
             "xtable",
             "HSAUR",
             "reshape",
             "stargazer",
             "texreg",
             "logistf",
             "MM")

package.check <- lapply(packages, FUN = function(x) {
    if (!require(x, character.only = TRUE)) {
        install.packages(x, dependencies = TRUE)
        library(x, character.only = TRUE)
    }
})
```
Table 4.1 on page 79
Randomly Generated Clusters of \( n = 10 \) Observations and Univariate Probabilities of \( p = (0.2, 0.3, 0.5) \) to Investigate the Behavior of Intracluster Dispersion Specification.

Table 4.2 on page 80
Randomly Generated Clusters of Various Cluster Sizes in Conjunction with Univariate Probabilities of \( p = (0.2, 0.3, 0.5) \) and Overdispersion Parameterization \( \theta_{12} = 0.90, \theta_{23} = 0.85, \theta_{13} = 0.75 \).

# Code generates UNTIDY table, manually formatted

```r
sample <- function(d, 
  C = 3, 
  N = 10, 
  n = 10, 
  pr = c(0.2, 0.3, 0.5)){ 
  # number of clusters 
  # number of obs per cluster 
  # univariate probabilities 

  m <- matrix(c(NA, d[["t_12"]], d[["t_13"]], 
               NA, NA, d[["t_23"]], 
               NA, NA, NA), byrow = TRUE, 
             nrow = C)

  pm <- MM::paras(p = pr[1:2], 
                  pnames = c("y_1", "y_2", "y_3"), 
                  theta = m)

  MM::rMM(n = N, Y = n, paras = pm)
}

# Vary the correlations
cbind(sample(d = c(t_12 = 1.0, t_23 = 1.0, t_13 = 1.0)), 
       sample(d = c(t_12 = 0.8, t_23 = 0.8, t_13 = 0.8)), 
       sample(d = c(t_12 = 2.0, t_23 = 2.0, t_13 = 2.0))) %>%
  data.frame %>%
  xtable(digits = 0)

# Vary the cluster size
cbind(sample(n = 8, d = c(t_12 = 0.9, t_23 = 0.85, t_13 = 0.75)), 
       sample(n = 12, d = c(t_12 = 0.9, t_23 = 0.85, t_13 = 0.75)), 
       sample(n = 16, d = c(t_12 = 0.9, t_23 = 0.85, t_13 = 0.75))) %>%
  data.frame %>%
  xtable(digits = 0)
```
Function: simMMy3sx
Simulate Correlated Multinomial Clustered Data

```r
library(tidyverse)
library(stringr)
library(MM)

## FUNCTION:
## generate clustered samples with the MM package
## specifically for a three-level dependent variable
## specifically for a two-level independent variable
## prepared for analysis by StatXact (sx)
## with one line per cluster x y-level combo
## data columns = "cluster", "x", "y", & "freq"

simMMy3sx <- function(N, n, N_x1, pdf, d) {

    ### ----- PARAMETERS ----- ###
    # N = number of clusters total (macro units)
    # n = number of observations per cluster (micro units)
    # N_x1 = number of clusters with x = 1 (exposed)
    # C = the number of outcome levels (hard coded to 3)
    # d = parameters between pairs of levels of the outcome

    C <- 3  # hard coded for now...may change later

    m <- matrix(c(NA, d["d_12"], d["d_13"],
                  NA,     NA, d["d_23"],
                  NA,   NA,     NA),
                byrow = TRUE,
                nrow = C)

    pm_x0 <- paras(p = pdf["x0", 1:2],
                    pnames = c("y_1", "y_2", "y_3"),
                    theta = m)

    pm_x1 <- paras(p = pdf["x1", 1:2],
                    pnames = c("y_1", "y_2", "y_3"),
                    theta = m)

    rbind(rMM(n = N - N_x1, Y = n, paras = pm_x0),
           rMM(n = N_x1, Y = n, paras = pm_x1)) %>%
    data.frame %>%
    tibble::rowid_to_column(var = "cluster") %>%
```

```r
dplyr::mutate(x = rep(0:1, 
    times = c(N - N_x1, N_x1))) %>%
tidyr::gather(key = y, 
    value = freq, 
    X1, X2, X3) %>%
dplyr::mutate(y = substr(y, 2, 2) %>% as.numeric)
```
Function: run_CS_reps

Write Script File for Directing StatXact to Analyse a Dataset and Save the Exact Distribution for the Test Statistic, for all Three Correlation Conditioning Settings

```r
run_CS_reps <- function(folder, set_num, rep_start, rep_end, digits){

  ins <- data.frame(rep = rep_start:rep_end) %>%
    dplyr::rowwise() %>%
    dplyr::mutate(instructions = paste0("import(C:\SxSim\", folder,"\\",
                                        set_num,"\\MData\data", pad(rep, digits), ".csv,
                                        type = ascii,
                                        header = yes,
                                        delimiter = comma);

  ll_association ( corr = unequal, row = x, col = y, clust_id = cluster, frequency = freq, method = exact, time_limit = none, dist_file = "c:\sxsim\", folder,"\\", set_num,"\\SXout\out_", pad(rep, digits), "_unequal.prn" );

  ll_association ( corr = equal, row = x, col = y, clust_id = cluster, frequency = freq, method = exact, time_limit = none, dist_file = "c:\sxsim\", folder,"\\", set_num,"\\SXout\out_", pad(rep, digits), "_equal.prn" );

  ll_association ( corr = adjacent, row = x, col = y, clust_id = cluster, frequency = freq, method = exact,
```
time_limit = none,
dist_file = "c:sxsim\", folder,\",
    set_num,\\SXout\out\",
    pad(rep, digits), "_adjacent.prn"");
})

file_name = paste0(set_num, "/instruct_rep",
    rep_start, "-", rep_end, ".cyb"
)

ins %>%
    dplyr::select(instructions) %>%
    base::unlist(use.names = FALSE) %>%
    paste0(collapse = "") %>%
    paste0("\nquilt;", collapse = "") %>%
    write(file = file_name)

system2(command = "C:/Program Files (x86)/Cytel Studio 11/cbf.exe",
        args = paste(getwd(), file_name, sep = "/"))

return(paste0("CytexStudio: silumation ",
    set_num,"- completed rep ",
    rep_start, "-", rep_end))}
**Function: test_stat**

Compute the Observed Test Statistic Given in Equation 3.19

```r
test_stat <- function(set_num, rep)
{
  read.csv(paste0(set_num, "/MMdata/data", pad(rep, digits), ".csv")) %>%
  dplyr::rowwise() %>%
  dplyr::mutate(xyfreq = x * y * freq) %>%
  dplyr::ungroup() %>%
  dplyr::summarise(sum(xyfreq)) %>%
  base::unlist(use.names = FALSE)
}
```
Function: p_val
Read in the Exact Distribution of the Test Statistic and Calculate the p-Value, for all Three Correlation Conditioning Settings

\[
p_{val} \leftarrow \text{function}(\text{set\_num}, \text{rep}, \text{t}, \text{corr\_type})\{
\]

\[
\text{out\_csv} \leftarrow \text{read.csv(paste0(set\_num, "/SXout/out\_", pad(rep, digits), ", ", corr\_type, ".prn"), sep = "")}
\]

\[
\text{t\_mean} \leftarrow \text{mean}(\text{out\_csv}\$\text{Statistic})
\]

\[
\text{point\_prob} \leftarrow \text{out\_csv} \mathbin{\%\%} \text{dplyr::filter(Statistic == t) \mathbin{\%\%} dplyr::select(Probability) \mathbin{\%\%} base::unlist(use\_names = FALSE)}
\]

\[
\text{p\_2} \leftarrow \text{out\_csv} \mathbin{\%\%} \text{dplyr::filter(Probability <= point\_prob) \mathbin{\%\%} dplyr::summarise(sum(Probability)) \mathbin{\%\%} base::unlist(use\_names = FALSE)}
\]

\[
\text{p\_both} \leftarrow \text{out\_csv} \mathbin{\%\%} \text{dplyr::mutate(pval = Probability \mathbin{\%\%} cut(breaks = c(0, point\_prob, 1), labels = c("yes", "no"))) \mathbin{\%\%} dplyr::mutate(tail = Statistic \mathbin{\%\%} cut(breaks = c(-Inf, t\_mean, Inf), labels = c("lower", "upper"), include\_lowest = TRUE)) \mathbin{\%\%} dplyr::filter(pval == "yes") \mathbin{\%\%} dplyr::group\_by(tail) \mathbin{\%\%} dplyr::summarise(prob = sum(Probability)) \mathbin{\%\%} tidyr::spread(key = sum, value = prob)}
\]

\[
\text{p\_1} \leftarrow \text{ifelse(t < t\_mean, p\_both}\$\text{lower, p\_both}\$\text{upper)}
\]

\[
\text{return(data.frame(p\_point = point\_prob, p\_1\_side = p\_1, p\_2\_side = p\_2)})
\]
Example Synax File: Generate and Test 200 Datasets

I. Declare the Parameters of the Simulation

   A. Procedural Quantities:

   * `C` = number of levels in the outcome ordinal variable
   * `N` = number of clusters (macro units)
   * `n` = number of observations per cluster (micro units)
   * `N_x1` = number of clusters out of `N` for which the predictor $x = 1$
   * `reps` = number of replications for each parameter specification

   ```
   C <- 3
   N <- 8
   n <- 4
   N_x1 <- 4
   reps <- 200
   ```

   B. Definitions

   ```
   # identify files with at least 3 digits for padding
digits <- max(5, nchar(reps))

   # list out the 3 types of correlation modeling constraints
corr_type_list = data.frame(corr_type = c("equal", "unequal", "adjacent"))

   # significance level
alpha <- 0.05
   ```

   C. Set Fixed Parameters

   1. `set_pdf` = marginal means (univariate PMF) for both levels of the predictor
      - there is one row in the PMF matrix for each level of the predictor $x$
      - each row in the PMF should have $C$ elements, one per level

   2. `set_d` = pairwise deltas for the
      - there will be “$C$ choose 2” pairs of the $C$ levels

   ```
   folder <- "simMMp2"
sset_num <- paste0("p2", "N", N, "n", n)

   set_pdf <- rbind(x0 = c(.2, .3, .5),
                    x1 = c(.5, .3, .2))

   set_d = c(d_12 = 0.90,
             d_13 = 0.75,
             ...)
   ```
d_{23} = 0.85

II. Generate the Separate Datasets
for 500 reps this took about 1 min

```r
set.seed(1234)

for (i in seq_along(1:reps)){
  simMMy3sx(N = N,
    n = n,
    N_x1 = N_x1,
    pdf = set_pdf,
    d = set_d) %>%
  write.csv(paste0(set_num, "/MMdata/data", pad(i, digits), ".csv"),
    row.names = FALSE)
}

III. Analyze each dataset separately with StatXact

For each dataset:

1. import the CSV dataset into CytelStudio
2. run the test & save the results in a CSV file

   Only run 100 reps at a time...or else bad things happen

```r
run_CS_reps(folder, set_num, 1, 100, digits)
run_CS_reps(folder, set_num, 101, 200, digits)

IV. Tally the Results

```r
simulate <- data.frame(rep = 1:200) %>%
dplyr::rowwise() %>%
dplyr::mutate(t = test_stat(set_num, rep)) %>%
tidyr::crossing(corr_type_list) %>%
dplyr::mutate(corr_type = corr_type %>% as.character) %>%
purrrlyr::invoke_rows(.f = p_val,
  set_num = set_num) %>%
unnest(.out) %>%
dplyr::mutate(param = 1,
  N = N,
  n = n) %>%
dplyr::select(param, N, n,
  rep, t, corr_type,
  p_point, p_1side, p_2side)

write.csv(simulate, file = paste0("pvals", set_num,".csv"))
IV. Repeat steps I-III for each combination of parameters and merge results

```r
simulate <- read.csv("simMMp1/pvals_p1N8n4.csv") %>%
  rbind(read.csv("simMMp1/pvals_p1N8n8.csv")) %>%
  rbind(read.csv("simMMp1/pvals_p1N8n12.csv")) %>%
  rbind(read.csv("simMMp2/pvals_p2N8n4.csv")) %>%
  rbind(read.csv("simMMp2/pvals_p2N8n8.csv")) %>%
  rbind(read.csv("simMMp2/pvals_p2N8n12.csv")) %>%
  rbind(read.csv("simMMp3/pvals_p3N8n4.csv")) %>%
  rbind(read.csv("simMMp3/pvals_p3N8n8.csv")) %>%
  rbind(read.csv("simMMp3/pvals_p3N8n12.csv")) %>%
  rbind(read.csv("simMMp11/pvals_p11N8n4.csv")) %>%
  rbind(read.csv("simMMp11/pvals_p11N8n8.csv")) %>%
  rbind(read.csv("simMMp11/pvals_p11N8n12.csv")) %>%
  rbind(read.csv("simMMp12/pvals_p12N8n4.csv")) %>%
  rbind(read.csv("simMMp12/pvals_p12N8n8.csv")) %>%
  rbind(read.csv("simMMp12/pvals_p12N8n12.csv")) %>%
  rbind(read.csv("simMMp13/pvals_p13N8n4.csv")) %>%
  rbind(read.csv("simMMp13/pvals_p13N8n8.csv")) %>%
  rbind(read.csv("simMMp13/pvals_p13N8n12.csv")) %>%
write.csv(simulate, "Phase_1_MMsx.csv")
```
**Figure 4.1** on page 84
Simulation: Example Dataset Exact Distribution of the Test Statistic Under Each of the Correlation Conditions of the ETT-CMD.

```r
equal <- read.table("data/p2N8n8data00002equal.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Equal")

unequal <- read.table("data/p2N8n8data00002unequal.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Unequal")

adjacent <- read.table("data/p2N8n8data00002adjacent.prn", header = TRUE) %>%
dplyr::mutate(corr_type = "Adjacent")

rbind(equal, adjacent, unequal) %>%
dplyr::mutate(corr_type = corr_type %>%
  factor(levels = c("Equal", "Adjacent", "Unequal"),
          labels = c("A. Equal (Total Correlation)",
                      "B. Adjacent (Adjacent Correlations)",
                      "C. Unequal (Intra-Cluster Correlation)"))) %>%
ggplot(aes(x = Statistic,
          y = Probability,
          shape = (Statistic == 56) %>%
                   factor(labels = c("Reference Set", "Observed")))) +
  facet_wrap(~ corr_type, ncol = 1) +
  geom_segment(aes(xend = Statistic,
                   y = 0, yend = Probability)) +
  geom_segment(aes(xend = Statistic,
                   y = 0, yend = Probability)) +
  geom_point(size = 2,
             fill = "white") +
  theme_bw() +
  labs(x = "Test Statistic",
       shape = NULL) +
  scale_shape_manual(values = c(21, 16)) +
  theme(legend.position = c(.86, .9),
        legend.background = element_rect(color = "black"))
```
Simulation: Comparing the Three Correlation Conditioning Options Regarding Power and Type I Error Rate.

```r
simulate <- read.csv("data/Phase_1_MMsx.csv")
alpha = .05

dplyr::mutate(sig = p_2side < alpha)

dplyr::group_by(param, corr_type, n)

dplyr::summarise(sig_percent = sum(sig)/n())

dplyr::ungroup()

dplyr::mutate(corr_type = corr_type %>% fct_reorder(-sig_percent))

dplyr::mutate(theta = case_when(param %in% c(1, 11) ~ "C. Underdispersion\n(Smooth Clusters)",
                                param %in% c(2, 12) ~ "B. Overdispersion\n(Lumpy Clusters)",
                                param %in% c(3, 13) ~ "A. Standard \nMultinomial"))

dplyr::mutate(measure = case_when(param %in% c(1, 2, 3) ~ "Power",
                                   param %in% c(11, 12, 13) ~ "Type I Error Rate"))

dplyr::mutate(n = n %>%
              factor(levels = c(4, 8, 12),
                      labels = c("Clusters size n = 4",
                                  "Clusters size n = 8",
                                  "Clusters size n = 12")))

dplyr::arrange(theta, corr_type, n, measure)

ggplot(aes(x = corr_type,
       y = sig_percent,
       fill = factor(corr_type),
       shape = factor(measure)))+
 facet_grid(n ~ theta) +
 geom_hline(yintercept = c(.05, .80),
           size = 1.5,
           alpha = .25) +
 geom_line(aes(group = factor(measure)),
           color = "black") +
 geom_point(size = 3.5) +
 theme_bw() +
labs(y = "Proportion of Samples with 2-Sided Exact p-value < .05",
     x = NULL,
     fill = NULL,
     title = "Figure 4.2 on page 86")
```

Figure 4.2 on page 86
Simulation: Comparing the Three Correlation Conditioning Options Regarding Power and Type I Error Rate.
shape = NULL) +
theme(legend.position = c(.85, .92),
    legend.box.margin = margin(6, 6, 6, 6),
    legend.background = element_rect(color = "black")) +
scale_shape_manual(values = c(21, 22)) +
scale_fill_manual(values = c("black", "gray", "white")) +
guides(fill = FALSE)
VITA
Sarah E Schwartz  
sarah.schwartz@usu.edu  
(435) 797-0169

WEBSITES

StatStudio  www.cehs.usu.edu/StatStudio
Open Science Framework  www.osf.io/3f87g
ResearcherID  E-2050-2016
LinkedIn  in/sarah-schwartz-63739010b
ORCID  0000-0001-9980-7493
GitHub  SarBearSchwartz

APPOINTMENTS

Assistant Research Professor  2015-present
Psychology Department, Utah State University
- Director, The Statistical Consulting Studio
- Instructor, graduate quantitative methods and statistics

Statistician  2013-2015
Office of Research Services, Utah State University
- Acting Director, Office of Methodological and Data Sciences
- Instructor, graduate quantitative methods and statistics

Data Manager and Statistician  2005-2013
Center for Epidemiology, Utah State University
- Managed databases, clean data, and prepare custom datasets
- Performed statistical analyses and prepared publications, posters, presentations, and grant submissions
- Worked under three main grants and many co-investigators: University of Utah, BYU, Duke, John Hopkins, University of Maryland, etc.
Data Manager 2012
Keoni Genetics Lab, Brigham Young University
- Managed databases and prepare custom datasets

Adjunct Lecturer 2006-2008
Mathematics and Statistics Department, Utah State University

High School Teacher, Math and Science 2000-2004
Sky View High School, Smithfield UT
Logan River Academy, Logan UT

EDUCATION

PhD Statistics 2013-2017
Mathematics and Statistics Department, Utah State University
Exact Approaches for Bias Detection and Avoidance with Small, Sparse, or Correlated Categorical Data
Thesis advisor: Chris Corcoran, Sc.D., Department Head
- TA: Lecturer and grader
- RA: Simulation programmer, Cytel Software Inc.
- 2015: Department Academic Excellence Award
- 2014: Department Industrious Graduate Student Award

MS Statistics 2004-2006
Mathematics and Statistics Department, Utah State University
Linear Mixed Effects Models - The Longitudinal Course of Neuropsychiatric Symptoms in Dementia
Thesis advisor: Chris Corcoran, Sc.D.
- TA: Lecturer and grader
- RA: Statistician, Center for Epidemiology Studies
- 2005: Department Graduate Student Teacher of the Year Award

Mathematics and Statistics Department, Utah State University
Magna Cum Laude
LICENCES

Secondary Education Teaching License  
*State of Utah*  
Endorsements: mathematics level 4, and chemistry  
2001-2009

SOFTWARE

R, R Studio, *and* R Markdown (LaTeX)  
*Data cleaning and wrangling, statistical analysis, and reproducible reporting*

SPSS, MPlus, G*Power, *and* SAS  
*Data cleaning and wrangling, and statistical analysis*

REDCap *and* Qualitrics  
*Creating forms and collecting data*


2016


2015


2014


2013


2011


2010


2009

SCHOLARLY PRESENTATIONS

2017


2015


2014


2013


Presented at the Society of Behavioral Medicine Conference. March 2013; San Francisco, CA.

2012


2011


Study”. *Poster presentation at the International Conference on Alzheimer’s Disease*. **July 2011**; Paris, France.

### 2010


### 2009


### 2006

TEACHING

UTAH STATE UNIVERSITY - GRADUATE

PSY 7650 Multilevel and Marginal Models for the Social Sciences
Coverage of multilevel (ie, mixed-effects, hierarchical linear) and marginal (ie, GEE) models for both continuous and categorical outcomes. Includes application of these methods to many types of cross-sectional and longitudinal research designs (eg, experimental, case-control, cohort, cross-over, complex sample, randomized controlled trials).

- 2017 Fall, using R
- 2016 Fall, using R

PSY 6600 Research Design and Analysis I
Research design and statistical concepts for research in education, human services, and psychology, with emphasis on the selection and interpretation of statistical analyses

- 2018 Summer, using R (7 week format)
- 2018 Spring, using R
- 2017 Summer, using SPSS (14 week format)
- 2017 Spring, using SPSS
- 2016 Summer, using SPSS (7 week format)
- 2015 Summer, using SPSS (7 week format)

UTAH STATE UNIVERSITY - UNDERGRADUATE

STAT 2000 Statistical Methods
Introduction to statistical concepts, graphical techniques, probability, distributions, estimation, one and two sample testing, chi-square tests, and simple linear regression, one-way ANOVA.

- 2014 Fall
- 2014 Spring
- 2005 Fall
- 2005 Spring
**STAT 1040** Introduction to Statistics  
Descriptive and inferential statistical methods. Emphasis on conceptual understanding and statistical thinking. Examples presented from many different areas.

- 2008 Spring, *included distance sites*
- 2007 Fall
- 2007 Spring
- 2006 Fall

**MATH 1050** College Algebra  

- 2004 Fall

**MATH 1010** Intermediate Algebra  
Review of introductory algebra concepts. Topics include manipulating and simplifying expressions, solving equations and inequalities, graphing equations, and inequalities. Real world applications including linear, quadratic, polynomial, rational, exponential, and radical functions.

- 2006 Summer
## COLLEGE-WIDE WORKSHOPS

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effect Size and Power Analysis Using G*Power</strong></td>
<td>10/10/17</td>
</tr>
<tr>
<td>2-hours: Lecture introducing concepts, followed with hands-on activities</td>
<td></td>
</tr>
<tr>
<td>G*Power introduced as well as conversion excel files</td>
<td></td>
</tr>
<tr>
<td><strong>Random Forests: Classification and Regression Trees</strong></td>
<td>11/22/17</td>
</tr>
<tr>
<td>2-hours: Introduction CART trees and random forests</td>
<td></td>
</tr>
<tr>
<td>focus on variable screening and examples in the literature</td>
<td></td>
</tr>
<tr>
<td><strong>Logistic Regression</strong></td>
<td>10/27/17</td>
</tr>
<tr>
<td>2-hours: Introduction CART trees and random forests</td>
<td></td>
</tr>
<tr>
<td>focus on variable screening and examples in the literature</td>
<td></td>
</tr>
<tr>
<td><strong>Random Forests: Classification and Regression Trees</strong></td>
<td>3/17/17</td>
</tr>
<tr>
<td>2-hours: Introduction CART trees and random forests</td>
<td></td>
</tr>
<tr>
<td>focus on variable screening and examples in the literature</td>
<td></td>
</tr>
<tr>
<td><strong>GEE: Generalized Estimating Equations</strong></td>
<td>2/10/17</td>
</tr>
<tr>
<td>2-hours: Introduction to GEEs for clustered and nested data (cross-sectional) and repeated measures (longitudinal)</td>
<td></td>
</tr>
<tr>
<td><strong>Intro to R: the Tidyverse way</strong></td>
<td>10/14/16</td>
</tr>
<tr>
<td>2-hours: Introduction to R via R Studio using the tidyverse packages</td>
<td></td>
</tr>
<tr>
<td>import, tidy, transform, visualization, models, and results</td>
<td></td>
</tr>
<tr>
<td><strong>Powerful Publishable Plots: The Grammar of Graphics</strong></td>
<td>4/8/16</td>
</tr>
<tr>
<td>2-hours: Focus on the do’s and don’t of good graphics</td>
<td></td>
</tr>
<tr>
<td>not specifically to any one software package</td>
<td></td>
</tr>
<tr>
<td><strong>Explore Your Data with ggplot in R</strong></td>
<td>2/19/16</td>
</tr>
<tr>
<td>2-hours: Interactive, hands-on workshop</td>
<td></td>
</tr>
<tr>
<td>assumes some familiarity with R</td>
<td></td>
</tr>
<tr>
<td><strong>Effect Size and Power Analysis Using G*Power</strong></td>
<td>11/13/15</td>
</tr>
<tr>
<td>3-hours: Lecture introducing concepts, followed with hands-on activities</td>
<td></td>
</tr>
<tr>
<td>G*Power introduced as well as conversion excel files</td>
<td></td>
</tr>
<tr>
<td><strong>Introduction to R &amp; R Studio, For Absolute Beginners</strong></td>
<td>10/2/15</td>
</tr>
<tr>
<td>3-hours: Lecture and Interactive, hands-on workshop</td>
<td></td>
</tr>
<tr>
<td>suggest coming with a laptop preloaded with R and R Studio</td>
<td></td>
</tr>
</tbody>
</table>