STATISTICAL METHODOLOGY FOR EVALUATING MULTIPLE RESPONSE CRITERIA FROM BIOEQUIVALENCE AND CLINICAL EQUIVALENCE TRIALS

by

Sonia M. Davis

Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2127T

May 1994
Statistical Methodology for Evaluating
Multiple Response Criteria from
Bioequivalence and Clinical Equivalence Trials

by

Sonia M. Davis

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

Chapel Hill

1994

Approved by:

[Signatures]

Advisor

Reader

Reader

Reader

Reader
Bioequivalence and clinical equivalence trials evaluate the similarity of two or more treatments. For continuous responses, equivalence is demonstrated if a 90% confidence interval for the ratio of treatment means is contained within a prespecified limit, usually 0.8-1.25. When multiple response criteria are evaluated, investigators might conclude overall treatment equivalence if equivalence is demonstrated for each outcome. The power to detect equivalence for all outcomes decreases as the number of outcomes increases, and is affected by correlation between responses. Power of the multiple 90% confidence interval method is evaluated for bivariate log-normal data from parallel and cross-over designs using numerical integration and simulation. Correlation is found to have a stronger effect on multivariate power when each univariate power is moderate to low.

Equivalence for dichotomous responses is evaluated with one-sided confidence intervals for odds ratios, ratios, and differences of proportions. Univariate and bivariate power are evaluated for each definition. Confidence intervals for differences appear to have the best power for response proportions near 1.0, but power for differences and odds ratios is similar for response proportions near 0.5.

Equivalence can alternately be defined in terms of the ratio of an individual’s response to two treatments. A method of evaluating individual equivalence is presented based on ANOVA prediction intervals for an individual’s ratio of responses from cross-over designs with log-normal data. Univariate and bivariate power are evaluated for this method.
Confidence intervals obtained univariately ignore the correlation between responses. Intervals based on multivariate models which take account of the correlation may have decreased variance, and hence better power. Applications of multivariate modeling to equivalence are described using weighted least squares, mixed linear models, and survey linear and logistic regression for log-normal, normal, and dichotomous responses. In examples, multivariate strategies generally achieved slightly smaller variance than univariate counterparts.

A concern with clinical equivalence studies is the lack of a placebo to assure efficacy of active treatments. A definition of equivalence is evaluated based on ratios of mean differences from placebo. Univariate and multivariate analysis strategies are presented for parallel designs with normally distributed and dichotomous responses.
Acknowledgements

My sincere thanks go to Gary Koch for everything, including the many years of support, encouragement, and education in the Biometric Consulting Lab. He makes everything possible.

I would also like to thank my committee members Ed Davis, Suzanne Edwards, Kerry Lee, and Al Tyroler for their time and suggestions. The data set for the clinical trial comparing treatments of hypertension was kindly provided by James A. Phillips. A special thanks to my fellow students from the BCL for their friendship and kindness.

Thanks of course to Tom, the best husband in the world, who has done so much to help me - even graciously sharing the family computer. And to my little buddy inside who had no choice but to be there every minute: now we can sleep!
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td></td>
</tr>
<tr>
<td>I.1</td>
<td></td>
</tr>
<tr>
<td>I.1.1</td>
<td></td>
</tr>
<tr>
<td>I.1.2</td>
<td></td>
</tr>
<tr>
<td>I.2</td>
<td></td>
</tr>
<tr>
<td>I.2.1</td>
<td></td>
</tr>
<tr>
<td>I.2.2</td>
<td></td>
</tr>
<tr>
<td>I.3</td>
<td></td>
</tr>
<tr>
<td>I.3.1</td>
<td></td>
</tr>
<tr>
<td>I.3.2</td>
<td></td>
</tr>
<tr>
<td>I.4</td>
<td></td>
</tr>
<tr>
<td>I.4.1</td>
<td></td>
</tr>
<tr>
<td>I.4.2</td>
<td></td>
</tr>
<tr>
<td>I.4.3</td>
<td></td>
</tr>
<tr>
<td>I.4.4</td>
<td></td>
</tr>
<tr>
<td>I.5</td>
<td></td>
</tr>
<tr>
<td>II.</td>
<td></td>
</tr>
<tr>
<td>II.1</td>
<td></td>
</tr>
<tr>
<td>II.1.1</td>
<td></td>
</tr>
<tr>
<td>II.1.2</td>
<td></td>
</tr>
</tbody>
</table>

**I. Introduction to Clinical Trials, Equivalence, and Multiple Comparison Techniques**

1.1 Introduction ........................................................................... 1

1.1.1 Introduction to Clinical Trials ........................................... 1

1.1.2 Clinical Trials in This Work ............................................. 3

1.2 Studies to Show Equivalence ............................................... 4

1.2.1 Bioequivalence Trials ....................................................... 4

1.2.2 Clinical Equivalence ....................................................... 14

1.3 Multiple Comparison Issues and Methods ............................ 18

1.3.1 Multiple Endpoints ......................................................... 18

1.3.2 Multiple Treatment Comparisons ...................................... 22

1.4 Examples ............................................................................ 23

1.4.1 Bioequivalence Meta Analysis of a Psychotropic Drug .......... 23

1.4.2 Parallel Clinical Trial of Hypertension ............................. 24

1.4.3 Parallel Clinical Trial of Muscle Soreness ........................ 25

1.4.4 Equivalence of Dermatologic Therapies ............................. 26

1.5 Research Goals ................................................................. 26

**II. Review of Standard Methods for Showing a Difference** ............. 28

2.1 Dichotomous Response ....................................................... 28

2.1.1 Direct Comparisons ....................................................... 29

2.1.2 Stratified Comparisons .................................................. 32
2.1.3 Covariate Adjusted Comparisons ........................................... 35

2.2 Ordinal Categorical Response .................................................. 38
  2.2.1 Direct Comparisons ......................................................... 39
  2.2.2 Stratified Comparisons ....................................................... 40
  2.2.3 Covariate Adjusted Comparisons ......................................... 42

2.3 Continuous Response ............................................................. 44
  2.3.1 Direct Comparisons ......................................................... 44
  2.3.2 Stratified Comparisons ....................................................... 46
  2.3.3 Covariate Adjusted Comparisons ......................................... 48

2.4 Multivariate Techniques ........................................................ 48

2.5 Conclusions ........................................................................... 52

III. Univariate Statistical Analysis for Bioequivalence and Clinical Equivalence ......................................................... 53

3.1 Analysis of the Two-period Cross-over Study ................................ 53

3.2 Assessments of Equivalence Based on ANOVA ................................ 57
  3.2.1 The Power Approach ........................................................ 57
  3.2.2 The Confidence Interval Approach ....................................... 60
  3.2.3 The Null Hypothesis of Non-equivalence ................................ 64
  3.2.4 Confidence Intervals for Ratios without Log-transforming ........ 66

3.3 Non-parametric Confidence Intervals ......................................... 68
  3.3.1 Based on Wilcoxon Signed-Rank Test .................................... 69
  3.3.2 Based on Pitman’s Permutation Test ..................................... 69
  3.3.3 Based on Wilcoxon-Mann Whitney Test ................................ 70
  3.3.4 Robust Procedures .............................................................. 71

3.4 Example .................................................................................. 72

3.5 Categorical Data Methods .......................................................... 74
  3.5.1 Parallel Designs to Test $\Delta$ from a Dichotomous Response ...... 74
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.2 Ordered Categorical Data</td>
<td>78</td>
</tr>
<tr>
<td>3.5.3 Categorical Data Example</td>
<td>80</td>
</tr>
<tr>
<td>3.6 Bayesian Methods for Bioequivalence</td>
<td>82</td>
</tr>
<tr>
<td>3.7 Individual Equivalence</td>
<td>84</td>
</tr>
<tr>
<td>3.7.1 FDA Tolerance Interval</td>
<td>85</td>
</tr>
<tr>
<td>3.7.2 The TIER Method</td>
<td>86</td>
</tr>
<tr>
<td>3.7.3 Individual Equivalence Example</td>
<td>88</td>
</tr>
<tr>
<td>3.7.4 Further Individual Equivalence Strategies</td>
<td>88</td>
</tr>
<tr>
<td>3.8 Additional Bioequivalence Topics</td>
<td>90</td>
</tr>
<tr>
<td>3.9 Power and Sample Size Calculations</td>
<td>91</td>
</tr>
<tr>
<td>3.10 Conclusion</td>
<td>94</td>
</tr>
<tr>
<td>IV. Multiple Endpoint Equivalence Techniques: Two Univariate Confidence</td>
<td>96</td>
</tr>
<tr>
<td>Intervals</td>
<td></td>
</tr>
<tr>
<td>4.1 Introduction to Multiple Endpoint Analysis</td>
<td>96</td>
</tr>
<tr>
<td>4.1.1 Current Strategy</td>
<td>96</td>
</tr>
<tr>
<td>4.1.2 Power of the Current Strategy</td>
<td>98</td>
</tr>
<tr>
<td>4.2 Cross-Over Bivariate Normal Distribution</td>
<td>99</td>
</tr>
<tr>
<td>4.2.1 Statistical Model</td>
<td>99</td>
</tr>
<tr>
<td>4.2.2 Obtaining the Test Statistic</td>
<td>101</td>
</tr>
<tr>
<td>4.2.3 Power Calculation from Numerical Integration</td>
<td>103</td>
</tr>
<tr>
<td>4.2.4 Power Estimation from Simulation Assuming Known Variance</td>
<td>106</td>
</tr>
<tr>
<td>4.2.5 Power Estimation from Simulation Assuming Unknown Variance</td>
<td>108</td>
</tr>
<tr>
<td>4.2.6 Application and Results</td>
<td>109</td>
</tr>
<tr>
<td>4.2.7 Conclusions</td>
<td>117</td>
</tr>
<tr>
<td>4.2.8 Example</td>
<td>121</td>
</tr>
<tr>
<td>4.3 Parallel Bivariate Normal Distribution</td>
<td>122</td>
</tr>
</tbody>
</table>
4.3.1 Statistical Model .................................................. 122
4.3.2 Obtaining the Test Statistic ..................................... 123
4.3.3 Power Calculation from Numerical Integration ............ 125
4.3.4 Power Estimation from Simulation Assuming Known Variance 126
4.3.5 Power Estimation from Simulation Assuming Unknown Variance ................ 126
4.3.6 Application and Results ........................................ 127
4.3.7 Example .................................................................. 134
4.3.8 Extension to More Than Two Outcomes ....................... 135
4.4 Conclusions .......................................................... 136

V. Bivariate Dichotomous Responses from Parallel Designs ......... 137
5.1 Introduction ............................................................ 137
5.1.1 Response Notation .................................................. 138
5.2 Difference of Proportions ............................................ 141
5.2.1 Model Assuming Normal Distribution ....................... 142
5.2.2 Bivariate Power Assuming Normality ....................... 144
5.2.3 Modifications for Small Samples .............................. 146
5.2.4 Bivariate Power Assuming Normal Approximation ....... 148
5.2.5 Applications and Results ........................................ 149
5.3 Ratios of Proportions ............................................... 156
5.3.1 Model Assuming Normal Distribution ....................... 157
5.3.2 Bivariate Power Assuming Normality ....................... 159
5.3.3 Modifications for Small Samples .............................. 160
5.3.4 Bivariate Power Assuming Normal Approximation ....... 161
5.3.5 Applications and Results ........................................ 161
5.4 Odds Ratios .......................................................... 165
5.4.1 Model Assuming Normal Distribution ....................... 166
5.4.2 Bivariate Power Assuming Normality.......................... 167
5.4.3 Modifications for Small Samples............................... 168
5.4.4 Bivariate Power Assuming Normal Approximation............ 169
5.4.5 Applications and Results.................................... 169
5.5 Example............................................................ 174
5.6 Conclusions....................................................... 177

VI. Individual Equivalence from Cross-Over Designs.................. 180
6.1 Introduction....................................................... 180
6.2 Statistical Model and Test Statistic............................ 181
6.3 Choosing Individual Equivalence Limits........................... 184
6.4 Power Assuming Known Variance................................ 190
6.5 Power Assuming Unknown Variance................................ 192
6.6 Application and Results......................................... 192
6.7 Example............................................................ 199
6.8 Discussion and Conclusions...................................... 200

VII. Multivariate Models for Equivalence.............................. 202
7.1 Introduction....................................................... 202
7.2 Methodology...................................................... 202
  7.2.1 Weighted Least Squares.................................. 203
  7.2.2 Survey Data Regression.................................. 206
  7.2.3 Mixed Linear Model....................................... 209
7.3 Log-Normal Responses............................................ 211
  7.3.1 Parallel Design............................................ 212
  7.3.2 Cross-Over Design....................................... 218
  7.3.3 Covariate Adjustment..................................... 219
List of Tables

<p>| Table 3.1 | Equivalence Results for Area under the Curve (AUC) Comparing Generic and Reference Formulations of a Psychotropic Drug, Study #1 | 73 |
| Table 3.2 | Equivalence Results for Maximum Concentration ($C_{\text{max}}$) Comparing Generic and Reference Formulations of a Psychotropic Drug, Study #1 | 73 |
| Table 3.3 | Test of $H_0 , \delta &lt; -0.1$ for the Study of Delayed Onset Muscle Soreness Using the Chi-squared method and Normal Approximation, and 90% CIs for $\delta$ based on the Normal and Gart Approximations | 81 |
| Table 3.4 | 90% Confidence Intervals for the Odds Ratio of improvement for Drug A vs Drug B in the study of Delayed Onset Muscle Soreness. | 81 |
| Table 3.5 | 90% Confidence Intervals for the Odds of Improvement vs. None for Drug A vs B for Three Ordered Categories of Pain Improvement by 48 hours in the study of Delayed Onset Muscle Soreness | 82 |
| Table 3.6 | Individual Equivalence Results for Area Under the Curve (AUC) and Maximum Concentration ($C_{\text{max}}$) Comparing Generic and Reference Formulations of a Psychotropic Drug, Study #1 | 88 |
| Table 4.1 | Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 90% Confidence Intervals. Cross-over Design, N per Sequence Group=10, Equivalence Range=0.8-1.25, Obtained via Simulation | 111 |
| Table 4.2 | Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 90% Confidence Intervals. Cross-over Design, N per Sequence Group=20, Equivalence Range=0.8-1.25, Obtained via Numerical Integration | 112 |
| Table 4.3 | Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 95% Confidence Intervals. Cross-over Design, N per Sequence Group=10, Equivalence Range=0.8-1.25, Obtained via Simulation | 113 |
| Table 4.4 | Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 95% Confidence Intervals. Cross-over Design, N per Sequence Group=20, Equivalence Range=0.8-1.25, Obtained via Numerical Integration | 114 |
| Table 4.5 | Power for Testing Equivalence of a Log-Normal Response for Method of 100(1-2\alpha)% Confidence Intervals. Cross-over Design, Equivalence Range=0.8-1.25, Obtained via Simulation (n per sequence group=10) and Numerical Integration (n per sequence group=20) | 115 |</p>
<table>
<thead>
<tr>
<th>Table 4.6</th>
<th>Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 90% Confidence Intervals. Parallel Design, N per Treatment Group=25, Equivalence Range=0.8-1.25, Obtained via Numerical Integration</th>
<th>129</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.7</td>
<td>Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 90% Confidence Intervals. Parallel Design, N per Treatment Group=50, Equivalence Range=0.8-1.25, Obtained via Numerical Integration</td>
<td>130</td>
</tr>
<tr>
<td>Table 4.8</td>
<td>Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 95% Confidence Intervals. Parallel Design, N per Treatment Group=25, Equivalence Range=0.8-1.25, Obtained via Numerical Integration</td>
<td>131</td>
</tr>
<tr>
<td>Table 4.9</td>
<td>Power for Testing Equivalence of Bivariate Log-Normal Responses for Method of Two Univariate 95% Confidence Intervals. Parallel Design, N per Treatment Group=50, Equivalence Range=0.8-1.25, Obtained via Numerical Integration</td>
<td>132</td>
</tr>
<tr>
<td>Table 4.10</td>
<td>Power for Testing Equivalence of a Log-Normal Response for Method of 100(1-2α)% Confidence Intervals. Parallel Design, N per Treatment Group=50, Equivalence Range=0.8-1.25, Obtained via Numerical Integration</td>
<td>133</td>
</tr>
<tr>
<td>Table 5.1.1</td>
<td>Power for Testing Test-Reference Percent Response is no Less than $\Delta=-0.10$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, n per group=200, Obtained via Numerical Integration (Direct Method) and Simulation (Method A)</td>
<td>152</td>
</tr>
<tr>
<td>Table 5.1.2</td>
<td>Power for Testing Test-Reference Percent Response is no Less than $\Delta=-0.10$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, n per group=40, Obtained via Simulation</td>
<td>153</td>
</tr>
<tr>
<td>Table 5.2.1</td>
<td>Power for Testing Test-Reference Percent Response is no Less than $\Delta=-0.20$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, n per group=100, Obtained via Numerical Integration (Direct Method) and Simulation (Method A)</td>
<td>154</td>
</tr>
<tr>
<td>Table 5.2.2</td>
<td>Power for Testing Test-Reference Percent Response is no Less than $\Delta=-0.20$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, n per group=15, Obtained via Simulation</td>
<td>155</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Power for Testing Test-Reference Percent Response is no Less than $\Delta$. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration (Direct Method) and Simulation (Methods A, B)</td>
<td>156</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Power for Testing Test/Reference Percent Response is no Less than $\Theta=0.9$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration (n=200) and Simulation (n=40)</td>
<td>162</td>
</tr>
</tbody>
</table>
Table 5.5  Power for Testing Test/Reference Percent Response is no Less than $\Theta=0.8$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration ($n=100$) and Simulation ($n=15$) .................................................. 163

Table 5.6  Power for Testing Test/Reference Percent Response is no Less than $\Psi=0.5$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration ($n=200$) and Simulation ($n=40$) .................................................. 164

Table 5.7  Power for Testing Odds Ratio for Test vs. Reference is no Less than $\Psi=0.5$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration ($n=100$) and Simulation ($n=15$) .................................................. 171

Table 5.8  Power for Testing Odds Ratio for Test vs. Reference is no Less than $\Psi=0.5$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration ($n=100$) and Simulation ($n=15$) .................................................. 172

Table 5.9  Power for Testing Odds Ratio for Test vs. Reference is no Less than $\Psi=0.5$ for Two Dichotomous Outcomes. Parallel Design, $\alpha=0.05$, Obtained via Numerical Integration ($n=100$) and Simulation ($n=15$) .................................................. 173

Table 5.10 Univariate Analysis for Example Data .................................................. 175

Table 5.11 Univariate and Bivariate Power for Example Data .................................. 175

Table 5.12 Univariate and Bivariate Power of $\pi_{11}=0.45$ for Example Data ............ 176

Table 5.13 Power for Testing Proportion of Negative Responses when $\pi_{11}=0.55$ for Example Data .................................................. 177

Table 6.1 Percent of Population Individual Treatment Ratios Within Specified Intervals For $\Theta=1.0$ and 1.25 and Equal Test and Reference Variance .................................................. 188

Table 6.2 Power for Testing Individual Equivalence of a Log-Normal Response From a Two Period Cross-Over ANOVA Prediction Interval, $N$ per Sequence Group= 20, Obtained via Simulation .................................................. 193

Table 6.3 Power for Testing Individual Equivalence of a Log-Normal Response From a Two Period Cross-Over ANOVA Prediction Interval, $N$ per Sequence Group= 30, Obtained via Simulation .................................................. 194

Table 6.4 Power for Testing Individual Equivalence of Two Bivariate Log-Normal Responses From a Two Period Cross-Over ANOVA Prediction Interval, $N$ per Sequence Group= 20, Obtained via Simulation .................................................. 195

Table 6.5 Power for Testing Individual Equivalence of Two Bivariate Log-Normal Responses From a Two Period Cross-Over ANOVA Prediction Interval, $N$ per Sequence Group= 30, Obtained via Simulation .................................................. 196
Table 7.1  Univariate Confidence Intervals for Equivalence of AUC and $C_{max}$ from Several Univariate and Multivariate Strategies .......... 223
Table 7.2  Confidence Intervals for Global Equivalence of AUC and $C_{max}$ from Several Multivariate Strategies ........................................ 224
Table 7.3  Univariate Confidence Intervals for Equivalence of AUC and $C_{max}$ for Study #1 from Several Univariate and Multivariate Strategies ........................................ 227
Table 7.4  Confidence Intervals for Equivalence of Diastolic Blood Pressure from Several Univariate and Multivariate Strategies .......... 234
Table 7.5  Confidence Intervals for Equivalence of Systolic Blood Pressure from Several Univariate and Multivariate Strategies .......... 234
Table 7.6  Confidence Intervals for Global Equivalence of Diastolic and Systolic Blood Pressure from Several Multivariate Strategies ..... 235
Table 7.7  Confidence Intervals of Odds Ratios for Equivalence of Ulcer Treatments from Several Univariate and Multivariate Strategies . 238
Table 7.8  Confidence Intervals for Global Equivalence of Ulcer Treatments from Several Multivariate Strategies .................. 239
Table 8.1  Confidence Intervals for Placebo-Controlled Equivalence for Diastolic Blood Pressure from Several Univariate and Multivariate Strategies ........................................ 252
Table 8.2  Confidence Intervals for Placebo-Controlled Equivalence for Systolic Blood Pressure from Several Univariate and Multivariate Strategies ........................................ 252
Table 8.3  Confidence Intervals for Global Placebo-Controlled Equivalence .. 253
Table 8.4  Confidence Intervals for Placebo-Controlled Equivalence of Treatments for Endoscopy Exam from Several Univariate and Multivariate Strategies ........................................ 256
Table 8.5  Confidence Intervals for Placebo-Controlled Equivalence of Treatments for Reported Absence of Pain from Several Univariate and Multivariate Strategies ........................................ 256
Table 8.6  Confidence Intervals for Global Placebo-Controlled Equivalence .. 257
List of Figures

Figure 1.1 Drug Concentrations in the Blood Following a Single Dose of a Standard-Release Formulation ........................................... 5
Figure 1.2 Drug Concentrations in the Blood Following a Single Dose of a Modified-Release Formulation................................. 6
Figure 1.3 Treatment Groups for Parallel Clinical Trial of Hypertension ....... 24
Figure 2.1 Notation for Dichotomous Response........................................ 29
Figure 3.1 Response Notation for Dunnett and Gent's Method .................... 75
Figure 3.2 Expected Cell Counts for Dunnett and Gent's Method ................. 76
Figure 4.1 Bivariate Power Density ........................................................... 119
Figure 5.1 Notation for Bivariate Dichotomous Response ................................ 138
Figure 5.2 Notation for Univariate Dichotomous Response ......................... 140
Figure 5.3 Notation for Outcome 1 vs Outcome 2 for Treatment g .............. 140
Figure 7.1 Ulcer Study Example of Bivariate Dichotomous Response .......... 237
Figure 8.1 Ulcer Study Example of Bivariate Dichotomous Response .......... 255
## List of Symbols

- **CV**: coefficient of variation $= \sigma/\mu$
- **$\delta$**: treatment difference
- **$\Delta_1, \Delta_2$**: equivalence limits for a difference
- **N**: total sample size
- **$n_A, n_B$**: sample size in cross-over study sequence group A or B
- **$n_t, n_r$**: sample size in parallel study treatment group test or reference
- **$\pi$**: population percent from a dichotomous variable
- **$\psi_{12}$**: odds ratio of outcome 1 vs outcome 2
- **$\psi$**: treatment odds ratio of test vs reference
- **$\Psi_1, \Psi_2$**: equivalence limits for an odds ratio
- **$s$**: ANOVA root mean square error
- **$\sigma^2$**: population variance
- **$\theta$**: treatment ratio
- **$\Theta_1, \Theta_2$**: equivalence limits for a ratio
- **$\mu$**: population mean
- **$\nu$**: population median
Chapter 1
Introduction to Clinical Trials, Equivalence, and Multiple Comparison Techniques

1.1 Introduction

This chapter is a general introduction to the studies and issues that will be addressed in this work. It begins with a general overview of clinical trials, explaining what features are applicable to this research. It then turns to two specific trials which are designed to show equality between treatments: bioequivalence trials, and active control equivalence studies. The problems of multiple endpoints and comparisons involving several treatment groups are introduced, and statistical techniques which address these problems in standard efficacy trials are examined. Several equivalence trials which will be used as examples throughout the work are described. Shortcomings of multiple comparison methods as applied to equivalence studies are pointed out, and the main focus of this research is defined.

1.1.1 Introduction to Clinical Trials

Clinical trials are conducted to compare two or more treatments, one of which is often placebo, for efficacy or safety issues. Volunteer subjects are randomized to a treatment group, and follow a specific protocol of treatment regimen and measurements. Trials are most often conducted in a double blind fashion, where neither the subject nor clinician know which treatment has been assigned. Data gathered from the treatment groups are compared statistically to address the investigator’s specific research hypotheses.
Clinical trials are planned by first specifying the condition or outcome for investigation, for instance healing of an ulcer or occurrence of a heart attack. Often clinical trials have one primary outcome measure, but generally several responses related to the investigated condition are measured. The outcomes can be continuous, categorical, ordered categories, or any combination of these.

The time frame of the study and study design are also crucial planning steps. Study designs usually follow two basic patterns: a parallel design where each subject receives one treatment, and a cross-over design where each subject receives at least two different treatments with a wash-out period between them. Cross-over studies have the advantage that the subject acts as his/her own control, which reduces variability for treatment comparisons, but have the disadvantage of a more complicated protocol and analysis than parallel designs.

Next for consideration is the number of treatments to be evaluated. Trials always have at least two treatments, which consist usually of an active therapy and placebo, or a test and reference therapy. Studies with more than two treatments have one or more reference treatments, one or more test treatments, and often a placebo. In addition, reference and test formulations may be applied at several dose levels, increasing the number of treatment groups.

Volunteers are randomized in a clinical trial only if they meet certain inclusion and exclusion criteria. These criteria cover a variety of requirements including demographics, general health, and specific health conditions pertaining to the outcomes of interest or the proposed treatments. Sometimes subjects are accepted only if they demonstrate adequate compliance to the therapy or placebo in a run-in period.

During the study, drop-outs and protocol violations are identified. There are two strategies for dealing with such data at analysis time. Intent to treat is a conservative approach which analyzes available data from all randomized subjects,
regardless of protocol violations or treatment compliance. A more anti-conservative approach is to include in analysis only those subjects who followed the protocol. Missing data, resulting from missed visits or otherwise uncollected data, can cause a problem at analysis time. One method to deal with missed visits is to apply the last collected measure to future missing values, called the last observation carried forward (LOCF) approach.

Clinical trials are designed to show either a difference between treatments, or equivalence of treatments. Bioavailability studies, discussed in section 1.2.1, are the most common equivalence studies. Trials to show equivalence versus a difference require separate strategies for planning sample size and analysis. Some trials with more than two treatments may be designed to test significant equivalence between a test and reference, and a significant difference between a test and placebo. Other trials with many treatments may wish to show that a set of treatments are all jointly equivalent.

Statistical concerns for clinical trials include defining appropriate treatment comparisons of the outcome variables to answer the research questions, planning the sample size so that appropriate power is available for addressing the research questions, and applying appropriate analysis techniques. Analysis can follow a non-parametric or model-based orientation depending on assumptions concerning the data and the sample size.

1.1.2 Clinical Trials In this Work

The focus of this dissertation is primarily on clinical trials which intend to show equivalence between treatment formulations, for at least some of the study treatments. In particular, studies will be considered that have more than one possibly correlated response criteria.

The studies addressed are either parallel or cross-over, with continuous,
ordinal, or categorical outcomes. Statistical methods will cover both non-parametric and model-based strategies, and will consider sample size calculation as well as analysis. Studies involving survival analysis are not examined.

For simplicity, it is assumed there is no missing data, or that missing data has already been replaced following the last observation carried forward (LOCF) approach. Analysis will follow an intent-to-treat plan, assuming that all patients were randomized, received the appropriate treatment, took or were exposed to at least some of the treatment, and had at least one post baseline visit. Therefore, no drop-out or missing data dilemmas will be addressed.

1.2 Studies to Show Equivalence

There are two clinical trial situations where showing equivalence is of interest. One is clinical equivalence, where treatment equality is evaluated in large studies of generally unhealthy individuals, with methodology as demanding as studies attempting to show a difference between an active treatment and placebo. While clinical equivalence is of broad interest, it has been implemented on a fairly limited basis, and has received little attention in the development of statistical methodology.

Bioequivalence studies are more common. They are usually cross-over trials applied to a small number of healthy volunteers, generally measure blood concentrations over time, and cost much less than a clinical equivalence trial. The bulk of statistical methodology for equivalence has been developed for this specific trial. For this reason, bioequivalence is discussed in depth in the next section and in Chapter 3.

1.2.1 Bioequivalence Trials

Bioavailability
Bioavailability of a drug is defined as the amount of active drug which is absorbed from an administered dose into the circulatory system, and the rate at which it gets there (Drug Info J 1991 25:474). A good discussion of the topic is given in Westlake (88). Unless otherwise noted, the following overview is based on this reference. ‘Availability’ refers to the amount of drug available in the circulatory system. Drugs usually get to the site of action through the circulatory system, but the site of action is often unknown or unreachable, so availability to the site of action is evaluated by availability in the blood.

In a bioavailability study, often called a blood-level trial, a drug is administered, and concentration in the blood is measured at designated times until the concentration diminishes. Most trials make blood measurements every thirty minutes at the beginning, and then every one or two hours as blood concentrations fall (Schuirmann 90). Concentrations usually increase sharply as the drug is absorbed, and then gradually decrease as the drug is metabolized by the liver. Controlled or modified release formulations tend to have an initial rise, followed by a plateau, and then a gradual decrease. Examples of blood concentration curves over time are shown in Figures 1.1 and 1.2.

**Figure 1.1** Drug Concentrations in the Blood
Following a Single Dose of a Standard-Release Formulation

![Graph showing drug concentration over time](image)

Bioavailability studies are usually conducted on healthy male volunteers. There is some concern that the population who will receive the drug is a better
study group, but trials are rarely conducted on patients (Hauck and Anderson 91). Sometimes a regimen of blood samples can not be drawn (i.e., for very young or ill patients). In such cases, metabolite concentrations in the urine are used as a surrogate measure. Bioequivalence trials of topical treatments generally evaluate specific skin reactions, such as vasoconstriction.

Figure 1.2 Drug Concentrations in the Blood
Following a Single Dose of a Modified-Release Formulation

Different formulations of the same drug (e.g., intravenous, tablets, capsule) can yield very different blood level curves. The same formulation produced by different manufacturers, or even different plants of the same manufacturer can also vary widely, due to factors such as the degree of compression of a tablet.

Variability in blood concentration curves also stems from the subject. The absorption and metabolism of a drug can vary widely between people, and even between different administrations to the same subject. Many factors, including food intake, circulation, and metabolism rates lead to this variability.

Bioequivalence

It is often of interest to show that two different formulations of the same drug have the same therapeutic effect. But therapeutic effects must be compared in a full-scale clinical trial where the effects can be measured and compared. An alternative approach is to conduct a comparative bioavailability trial of the blood
levels of each formulation. If the two formulations lead to essentially the same blood level curves, they are said to be bioequivalent. The important underlying assumption is that two bioequivalent formulations of a drug have equivalent therapeutic effects. This can not be proven, but is considered scientifically sound by regulatory agencies, including the United States Food and Drug Administration (FDA) (Clin Pharm 1988 7:334-335).

Bioequivalence is used for regulatory approval of generic drugs. When a drug patent expires, other pharmaceutical companies may manufacture and market the drug under a generic name. To gain approval from the FDA, the new manufacturer submits an Abbreviated New Drug Application, in which they show bioequivalence between their formulation and the original product. Efficacy of the new formulation is assumed, based on bioequivalence to the original, and expensive efficacy clinical trials are not required for the new manufacturer. This leads to the dramatically cheaper shelf price for generic drugs.

A test drug found to have significantly increased bioavailability over the reference drug is termed suprabioavailable. Such a drug is considered a new dosage form by the FDA, and requires clinical efficacy trials before approval (Drug Info J 1991 25: 471-482).

Bioequivalence Trials

Typically bioequivalence trials involve administration of a single dose of two drugs: the test (generic) drug and the reference drug. A placebo is also of interest for some trials. The trial design can follow two patterns: a parallel groups design, or a crossover design. Due to the large variability in drug metabolism from person to person, bioequivalence studies are much more suited to crossover designs. Each person acts as his own control, which leads to much smaller variation than a parallel groups design.
This basic study design can be modified to allow more than one test formulation, or repeated treatment administration (for instance a four period crossover study with two treatments).

Steady-State Bioequivalence

Another type of bioavailability is called steady-state bioavailability. Multiple doses of each drug are given to a subject at equally spaced intervals so that the blood concentration maintains a constant level over a prolonged period of time (usually 12 hours). The plot of a blood concentration curve from a multidose study resembles that of a modified release formulation (Figure 1.2), but may stay at a plateau for a longer period of time, depending on the length of the study.

Steinijans et al (89) suggest several advantages of multiple dose equivalence studies over their single dose counterparts. Switch-over between test and reference drugs is done at steady state, so a wash-out period between treatments is not necessary. Since concentrations fluctuate little, blood can be drawn and analyzed as little as once during each dosing interval. Also, concentrations are measured at levels well above the detection limit (which may not be the case for some time points in single dose studies), and the study design directly reflects the recommended use of the drug. Disadvantages include the added logistic problems of subject compliance and food interactions over a longer time.

Pharmacokinetic Measures

Bioequivalence trials are essentially repeated measures experiments, and could be analyzed using repeated measures techniques or principal components. This was done in initial stages of bioequivalence development (for instance Westlake 1974, Wallenstein and Fisher 1977). However, analysis has turned away from this approach, sparked by Westlake’s suggestion that blood level
comparisons should focus only on those characteristics which have a meaningful relationship to the therapeutic effect (Westlake 1979). Current strategies are based on univariate analyses of a few pharmacokinetic parameters derived from the concentration curve. Substantial literature has been published concerning the appropriate parameters to analyze. All authors agree that there are two important bioavailability characteristics: the amount or extent of drug absorption, and the rate of drug absorption.

Extent of Absorption

Let time be designated by $T$, and blood concentration by $C$. The amount of drug absorbed is estimated by $\int_0^T C(T) dT / [k_e V]$, where $V$ is the volume of the blood, and $k_e$ is a constant describing the elimination rate of the drug. The integral is referred to as the area under the blood level curve (AUC). Since $k_e V$, referred to as the clearance, is unknown and unmeasurable, AUC is accepted as the best measure of amount of drug available to the circulatory system.

Blood concentrations are measured at only a few discrete time points, so AUC must be estimated. The estimate is calculated in two steps. The area under the observed data is calculated by the trapezoidal rule, which states:

$$\int_0^T C(T) dT \approx \sum_{i=1}^{n-1} \left\{ \frac{1}{2} [C(T_i) + C(T_{i+1})][T_{i+1} - T_i] \right\},$$

(1.1)

where $T_n$ is the time of the last measurement, and $n$ is the number of measurements made, including time zero (Swokowski 79). The extrapolation to infinity is essentially calculated by estimating the slope at the tail of the concentration curve with log-linear regression, dividing the last concentration by this slope, and adding this to the trapezoidal rule value (Sauter 92). Computer software packages which calculate AUC as well as measures of rate of absorption are available. (Johnston, Wollard 83).
Rate of Absorption

While AUC is widely accepted as a reasonable measure of the extent of absorption for both standard and modified release drugs, there is no clear favorite measure of the rate of absorption. Different measures may be more appropriate for some drugs than others, but they all have flaws.

The most commonly used measure of absorption rate is the maximum concentration reached, $C_{\text{max}}$. $C_{\text{max}}$ is of importance because some drugs may need to reach a certain level to achieve a therapeutic effect, and may need to stay below another level to avoid toxic effects. A drawback of $C_{\text{max}}$ is that it is highly correlated with AUC, and thus is not a pure measure of absorption rate (Kaniwa 89). It is also more variable, since it is based on only one extreme measure per person. Endrenyi et al (91) suggested evaluating $C_{\text{max}}$/AUC instead of $C_{\text{max}}$, because the ratio measures the absorption rate independent of the amount of drug absorbed. This suggestion does not appear to have come into common practice.

Another frequently used measure is the time it takes to reach the peak blood level, referred to as $T_{\text{max}}$. $T_{\text{max}}$ is important for drugs that need to reach peak concentration as soon as possible, such as analgesics and antibiotics, but is less important for drugs requiring multiple dosing before a therapeutic effect is observed, for instance some psychotropic drugs. $T_{\text{max}}$ is independent of extent of absorption, but power from this measure is low because of its discrete nature, which especially causes problems in products with short $T_{\text{max}}$ values (Kaniwa 89).

The majority of bioequivalence studies are performed on modified release products (Sauter et al 92). For these formulations, neither $C_{\text{max}}$ nor $T_{\text{max}}$ are reasonable measures, since the concentration curve maintains a plateau, and the maximum concentration is the result of variability. Plateau time (the amount of time the concentration is at least 75% of $C_{\text{max}}$, labeled T75%$C_{\text{max}}$) has been proposed as the least flawed measure of rate of absorption for these studies.
(Hauschke et al 90, Schulz and Steinijans 91, Sauter 92). The mean residence time (MRT), which can be interpreted as the mean of the concentration curve is another suggestion, but Sauter et al (92) point out that adequate calculation of MRT may require measurement at many time points.

For multiple dose studies, several measures of the steady-state rate of absorption have been suggested. Schulz and Steinijans (91) and Sauter et al (92) recommend the percent peak-trough fluctuation, \(\%\text{PTF} = 100(C_{\text{max}} - C_{\text{min}})/C_{\text{avg}}\). Other measures proposed include the half-value duration (HVD), plateau time, time above \(C_{\text{avg}}\), and percent AUC fluctuation.

In summary, for standard drugs, bioequivalence analyses are based on univariate analyses of \(C_{\text{max}}, T_{\text{max}},\) and AUC. \(C_{\text{max}}\) and \(T_{\text{max}}\) do not measure rate of absorption well, but they continue to be used for lack of more suitable parameters. For modified release forms, plateau time is the preferred rate measure, and for steady state studies, the percent peak trough fluctuation is recommended. The literature recommends picking one rate measure a priori, but also does not discourage examining several measures as secondary characteristics. They do not resolve how to interpret results when conclusions regarding equivalence based on these measures are conflicting (Sauter 92), and rarely acknowledge that this is a multiple comparison issue (Steinijans, Hauschke et al 92, Westlake 88).

Additive versus Multiplicative Models

In order to compare bioavailability measures between treatments, assumptions about their distributions must be made. In particular, if measures fit an additive model, comparisons should be made on differences, whereas measures which follow a multiplicative model are compared using ratios. The natural log \((\ln)\) transform is useful for multiplicative parameters, since \(\ln(x/y) = \ln(x) - \ln(y)\).
and multiplicative comparisons can be made using differences on the log scale, then exponentiating to return to ratios on the original scale. This is helpful statistically since many statistical procedures, in particular analysis of variance (ANOVA), assume an additive model. Schuirmann (87) points out that often bioavailability variables are skewed, with a long tail of higher values. A log transformation of the data is more likely to follow a normal distribution.

For many bioequivalence responses, the log transform is justified by a multiplicative pharmacokinetic model. Area under the curve follows a multiplicative model because AUC equals the amount of drug actually absorbed multiplied by the clearance. The log transform makes this additive, and the clearance is taken into account as part of the subject effect in an ANOVA model (see Chapter 3). One can argue that $C_{\text{max}}$ is also multiplicative, because the volume of blood distribution in the body follows a multiplicative model. Multiplicative models can also be assumed for mean residence time, percent peak trough fluctuation and percent AUC fluctuation (Diletti et al 91).

No justification exists for log transforming $T_{\text{max}}$, and in fact, it makes more sense to base bioequivalence of $T_{\text{max}}$ on absolute differences instead of relative percents (Steinijans and Hauschke 90). Additive models are also assumed for half value duration, plateau time, and time above $C_{\text{avg}}$ (Diletti et al 91).

Definition of Equivalence for Bioavailability Studies

Equivalence of a bioavailability parameter is most often assessed via 90% confidence intervals for the ratio of the treatment medians for multiplicative parameters, and for the difference of means or medians for additive parameters. In 1986, the FDA defined equivalence for AUC and $C_{\text{max}}$ acceptable when a 90% confidence interval of ratio of treatment medians is contained within 0.80 - 1.20 (Clin Pharm 88 7(5):334). The FDA Generic Drugs Advisory Committee changed
the bioequivalence range to 0.8-1.25 in September 1991 (Steinijans, Hauck et al 92), after many authors suggested that 0.8-1.25 is preferable because it is symmetrical on the log scale and leads to the same conclusion regardless of which treatment is in the denominator of the ratio (Kirkwood 81, Steinijans, Hauschke et al 92).

Sauter et al (92) suggest that measures following an additive model be considered equivalent if the CI of the test-reference difference is within $\pm 20\%$ of the mean for the reference population, but Westlake (88) and others criticize the use of relative percent criteria for parameters following additive models.

There is an interest in developing consistent definitions of equivalence between regulatory agencies from different countries. Currently, the European Committee for Proprietary Medicinal Products (CPMP) supports equivalence limits of 0.8-1.25 for AUC, and 0.7-1.43 for the more variable $C_{max}$. For drugs with a narrow therapeutic range, tighter limits are recommended: 0.9-1.11 for AUC and 0.8-1.25 for $C_{max}$ (Schulz and Steinijans 91). In Canada, a confidence interval for $C_{max}$ is not required: only the estimated $C_{max}$ ratio must be within $0.80-1.25$. While 90% confidence intervals are accepted in the US and Europe, Canada requires 95% CIs for certain toxic drugs and drugs with a narrow therapeutic range (McGilvery 92). Japan also requires 95% CIs (Kaniwa 90).

Other Applications of Bioequivalence

Bioequivalence studies are used in the initial development of a drug to compare bioavailability of different oral forms with intravenous injection to determine the most desirable delivery formulation (Westlake 88). They are also used to evaluate pharmacokinetic interaction between two drugs (Steinijans et al 91). A drug is tested for interaction with a reference by giving the reference drug alone, and then giving the two drugs simultaneously. Drug interaction is assumed
not to occur if the reference drug blood concentrations at the two times are equivalent. Food intake is sometimes used instead of a second drug to evaluate the effect of food on drug absorption (Steinijans, Hauschke et al 92).

1.2.2 Clinical Equivalence

Introduction

Sometimes a newly developed treatment or therapy can not be tested for efficacy against a placebo in a controlled clinical trial. This is usually due to ethical problems with randomizing patients to placebo in studies of serious diseases where an active drug has been proven effective, such as cancer or AIDS. Often the goal of the new treatment is not to improve efficacy over the existing treatment, but to reduce side effects, decrease medical costs, or maintain a higher quality of life. In other instances, different dosing regimens or therapies already proven effective in previous studies may be compared for equivalence to show that they are interchangeable, allowing the safest or easiest therapy to be prescribed for each patient.

These situations require a clinical equivalence trial between the new and existing therapy. Makuch et al (1990) labeled such a trial an Active Control Equivalence Study (ACES). They point out that while ACES are necessary, they are not scientifically ideal studies, due to several issues. Except where noted, Makuch is the reference for the following discussion.

Examining Efficacy through Historical Controls

There is one major theoretical difference between bioequivalence studies and ACES involving new therapies not yet proven effective. Bioequivalence examines a new manufacturing of an already approved therapy, and if equivalence is concluded, it is readily accepted that the new formulation is effective. Clinical
studies, however, often compare two completely separate treatments, and efficacy of a new therapy is not so easily accepted by equivalence alone.

If the two treatments are found equivalent, this does not automatically imply that the test drug is effective. Additional information is needed to examine whether in the current trial neither treatment was actually effective, or whether the study was not sensitive enough to detect important differences between the two treatments.

One way to evaluate treatment efficacy versus a placebo is to examine historical controls. Historical controls are the placebo group from an efficacy clinical trial of the reference therapy. Confidence intervals of the treatment effect are required for the placebo, the new therapy, and the reference therapy from both trials. If the two reference drug confidence intervals overlap somewhat, and if the lower bound of the ACES reference CI is larger than the upper bound of the historical placebo CI, then it may be concluded that the current study did find the reference drug effective. Similarly, if the lower bound of the CI for the new drug exceeds the upper bound of the placebo CI, then indirect efficacy can be cautiously accepted. On the other hand, if the ACES treatment CIs overlap with the historical placebo CI to a large extent, then efficacy of the new treatment is questionable. Such comparisons may also be made on confidence intervals of treatment differences. Even if indirect efficacy seems to have been met, it can not be fully accepted, due to factors such as 'positive-control study upward bias', where responses may be exaggerated in both ACES treatment groups because both patient and clinician know that an active drug has definitely been administered.

This indirect efficacy issue requires that the new study protocol be as similar to the historical study as possible, and equivalence studies be undertaken with reference treatments that have been found to consistently provide an easily
detectable therapeutic effect. Circumstances will arise when these requirements can not be met, due to the nature of the therapies, cost, specific study objectives, lack of information regarding previous studies, new measurement techniques, or other factors. In such situations, efficacy of the new treatment may remain uncertain, but equivalence between ACES treatments can still be evaluated.

Study Design Concerns

In addition to the historical control problem, ACES suffer from other criticisms. If the objective of the study is to show equivalence, there is less incentive to conduct a high quality study. Trials with imprecise measurements, inappropriate patient populations, inappropriate diagnostic procedures, subpotent doses, etc. are more likely to find an equivalence between treatments. Some of these issues can be combated with appropriate statistical analysis, but others can not. For this reason, ACES must be conducted under the highest scientific standards. Makuch et al suggest private companies insure quality by contracting the work to an independent firm.

Statistical Analysis Issues

There are several concerns regarding analysis that arise in ACES but are not mentioned in the bioequivalence literature. One of these is using the intent-to-treat approach, where data from protocol violations such as noncompliance are included in analysis. While this strategy is conservative in placebo-controlled trials, it can be the opposite for ACES, making it easier to show treatment equivalence. A more appropriate strategy for ACES may be the 'efficacy' approach, where noisy data due to such violations is removed.

Another concern is that the sample size be large enough, and that conservative approaches such as sub-group analysis and adjusting for multiple
comparisons not underpower studies so they are more likely to find treatment equivalence.

It seems these concerns are founded on the fact that many ACES are not analyzed correctly. A test of the traditional null hypothesis is not appropriate. The bioequivalence analysis techniques discussed in Chapter 3 are easily modified for parallel clinical studies, but do not appear to be in wide use for ACES. Makuch et al highly encourage the use of confidence intervals on the test-reference treatment difference, or a test of the appropriate null hypothesis of non-equivalence.

Definition of Clinical Equivalence

While bioequivalence is accepted if a confidence interval of the treatment ratio is within 0.8-1.25, the clinical equivalence literature appears to prefer treatment differences. Equivalence is accepted if a CI of the test-reference difference is contained within a specified range, ±\( \Delta \). The range may be based on clinical significance, but Makuch et al point out that clinical significance may be irrelevant to treatment effects assessed under the highly controlled environment of a clinical trial. Choosing an appropriate value of \( \Delta \) thus becomes a problem. Ng (92) suggests defining \( \Delta \) as a fraction of the observed reference-placebo difference from the historical trial. A fraction of 0.20 is recommended, due to its acceptance in the bioequivalence literature.

Conclusions

While active control equivalence studies have unavoidable flaws, most of them can be minimized with proper planning and statistical analysis. Confidence intervals are usually examined for treatment differences, but could also be applied to treatment ratios, particularly when there is no clear choice of \( \Delta \). When
considering multiple endpoints, CIs of ratios may be more useful for several reasons. Using differences, Δ would have to be specified for each endpoint. Ratio equivalence limits may be easier to define, and ratios are more easily combined for overall assessments of equivalence.

1.3 Multiple Comparison Issues and Methods

Often there are many comparisons of interest in clinical trials, but conducting multiple comparisons can cause problems with both an inflated type I error rate, and interpretation of results. Multiple comparisons discussed in this work fall into two classes: comparing two treatments for many response criteria, or comparing a set of several treatments for one response criterion. A combination of these types is frequent in studies of more than two treatments. Multiple responses as a result of repeated measures of the same outcome over time or space are not considered.

Methods which control the type I error rate for each multiple comparison class are briefly described below. Discussion is primarily focused on studies designed to show a treatment difference, since little has been done in the area of equivalence. Methods which are potentially extendible to equivalence situations are emphasized.

1.3.1 Multiple Endpoints

Often in clinical trials, more than one response outcome is of interest, usually from the assessment of several potentially correlated measurements of treatment effectiveness. An example is the assessment of both rate and extent of drug absorption in bioequivalence trials. It is common in medical literature to evaluate all endpoints separately, and report multiple p-values. A trial is often interpreted as having 'positive' results if one or more of several clinically important treatment comparisons are found significantly different at the $p \leq 0.05$ level. Yet as each
additional outcome is evaluated, the overall probability of finding a significant result by chance alone is markedly increased, leading to an inflated risk of false-positive findings. Similar problems are present in the interpretation of multiple confidence intervals.

One solution to this problem is to apply an adjustment for multiple comparisons. There are many multiple comparison procedures available (Dunnett and Goldsmith 81, Bauer 91, D'Agostino and Heeren 91), but the Bonferroni method is perhaps the most common. The false-positive error rate of $0.05/k$ is applied to each of $k$ endpoints, to maintain the overall error rate at 0.05. Unfortunately, this strategy lacks power to identify true treatment differences when the outcome measures are correlated with each other, as is often the case (O’Brien 84). A less conservative approach based on the union-intersection principle is the Bonferroni-Holm method (Bauer 91). The p-values from the $k$ endpoints are ordered from smallest to largest. Significance is concluded for the smallest p-value if it is $\leq \alpha/k$. If significant, the next p-value is evaluated relative to $\alpha/(k-1)$, and so on. No further p-values are evaluated after one is found non-significant.

Another modification of the Bonferroni correction which may be more appropriate for correlated outcomes is the Hailperin-Rüger method. It requires $t$ of the $k$ endpoints to have $p \leq 0.05t/k$ to insure an overall error rate at the 0.05 level, with $t$ specified in advance (Bauer 1991). For an example with four outcomes, significance would be concluded with $t=3$ if any three measures had $p \leq 0.0375$. This method is not in common use and has not been extensively studied for assessment of multiple endpoints. Its power depends on the choice of $t$ as well as the pattern of treatment differences and correlation among the $k$ responses.

An alternate strategy used frequently is to choose a single primary outcome
whose result will constitute drug efficacy or equivalence. All other endpoints are secondary, and interpretation of their results are supplementary or descriptive in nature. Problems arise when the choice of primary endpoint is controversial or subjective. For example, one investigator studying arthritis pain may choose the physician's measure of global pain status as the primary endpoint, while another may identify the number of swollen joints as primary. As a result, p-values for both primary and secondary endpoints are reported in journals, often leading the reader to draw conclusions without regard to the false-positive error rate (O'Brien 84). In bioequivalence trials, both rate and extent of drug absorption are of concern to the FDA, so choosing one primary outcome is not possible.

Perhaps a more appropriate strategy for multiple correlated outcomes is to combine the information from all endpoints, and obtain one global p-value for determining treatment efficacy. If the global test is found significant, separate evaluation of each endpoint is of descriptive interest. The goal a global procedure is both to control the over-all error rate, and to maintain power to detect true differences. It achieves the goal by being applied at the usual $p \leq 0.05$ level and by taking into account the correlation of the outcome measures.

Several global methods have been suggested for trials where showing a treatment difference is of interest. Hotelling's $T^2$ is often used as a global test of multiple responses. However, this method tests for general treatment differences: there is potential detection of alternatives which involve one treatment favored for some endpoints, and another treatment favored for other endpoints. This alternative is of little interest for multiple response criteria from clinical trials, since consistent patterns favoring one therapy are generally expected. Less power is available for the specific alternative of interest, making Hotelling's $T^2$ a very insensitive test for scenarios discussed in this work (O'Brien 84).

O'Brien (84) proposed a rank-sum method in which each separate outcome is
ranked over all study participants from worst to best, where 1 = worst, and so on. The ranks are summed over the endpoints for each subject, and the resulting rank-sums are compared between the treatment groups with a t-test or analysis of variance. The author showed this method is relatively efficient for normally distributed as well as skewed continuous responses. It performs well for situations where outcome measures tend to agree with one another, sample sizes are moderately large, and the number of endpoints is at least moderate. This method is usually considered for continuous responses but is applicable in principle to dichotomies or ordered categories. Since the outcomes are ranked, the fact that they may have different units is not a problem. Its performance when there is a mixture of dichotomous, ordered categorical, and continuous responses is of interest for investigation.

A limitation of the O'Brien rank-sum method is that it does not readily convert to a meaningful confidence interval. Intervals are more informative than p-values, because they allow an assessment of the extent of equivalence. Also, O'Brien's method does not provide a way to evaluate the homogeneity of the patterns of treatment differences across the multiple endpoints. Although expected to be correlated a priori, outcomes with substantially different response patterns should not be combined in a global test of this nature.

A class of comparisons of two treatments for multiple correlated responses which does allow for homogeneity assessment and confidence interval estimation (in some cases) is based on defining asymptotically normal statistics for each outcome. Treatment comparisons and response homogeneity are assessed using weighted least squares. Such methods are applicable to dichotomies, ordered categories, continuous responses, survival data, and are potentially extendible to combinations of these. The statistics of interest include mean differences, Mann-Whitney and related rank association statistics, and log-odds ratios. Sample sizes
must be large enough for the normal approximation to apply. Variations of this method are described by Koch et al (85), Carr et al (89), O'Brien (84), Pocock et al (87), Wei and Lachin (84), Wei and Johnson (85), and Lachin (92).

Tang et al (93) suggest an asymptotic likelihood ratio test of the one-sided alternative that one treatment is worse than another. It has the undesirable property that treatment differences in the unexpected direction are ignored, and thus appears inferior to weighted least squares methods.

Strategies for combining responses into a global test for the evaluation of treatment equivalence is of interest in this research. Important characteristics of such strategies include maintaining reasonable power, combining measures from different metrics, and assessing homogeneity of the responses. Methods discussed above are examined in later chapters for their application to multivariate equivalence assessment.

1.3.2 Multiple Treatment Comparisons

For studies comparing many treatments, multiple pairwise treatment comparisons can be adjusted with Bonferroni or other multiple comparison procedures (for example D'Agostino and Heeren 91, Dunnett and Tamhane 92). Overall tests of multiple treatment differences can be conducted using analysis of variance, extended Mantel-Haenszel, or other strategies.

Methods for simultaneous evaluation of multiple treatment equivalence are less available. One strategy is to impose treatment equivalence within the design matrix of a statistical model. Examples of this are models for dose-response and response surface methodology (J. Phillips et al 92). If the goodness of fit of the model is supported, this is partial evidence that treatments may be equivalent. This is not, however, a legitimate test for simultaneous treatment equivalence. An underpowered study may appear to fit a model imposing treatment
equivalence when in fact there are true treatment differences.

One appropriate strategy for multiple treatment equivalence is to form a global confidence interval of treatment ratios, and evaluate this against a predetermined equivalence range. Such methods are of interest for investigation.

1.4 Examples

Described in this section are several trials designed to show treatment equivalence. Some are used as examples throughout this work, others are used to demonstrate that equivalence study designs have a variety of complex treatment regimens and multiple outcome measures - designs with more breadth than are currently addressed by statistical equivalence methods.

1.4.1 Bioequivalence Meta Analysis of a Psychotropic Drug

A series of seven bioavailability studies were conducted to compare a new generic formulation of a treatment for a psychological disorder with the name-brand reference. In each study, a cross-over design was performed, where healthy volunteers received both treatments with a washout period between them. Blood concentrations of the drug and two of its major metabolites were monitored. Nine outcome measures were collected, consisting of three bioavailability parameters derived from the concentration curves of the drug and each metabolite. The measured parameters are area under the blood concentration curve (AUC), maximum blood concentration ($C_{\text{max}}$), and time to maximum concentration ($T_{\text{max}}$). The studies were conducted on separate individuals, possibly by separate clinicians, but followed similar protocols. Sample size varied from 19 to 26 subjects, for a total of 155 subjects from the combined studies. The goal was to show that the test and reference formulations have equivalent bioavailability parameters.
1.4.2 Parallel Clinical Trial of Hypertension

This example is a randomized double-blind parallel-group multi-center study of several treatments for hypertension. There are twelve treatment groups, based on a factorial structure of different doses of two drugs: ramipril, an angiotensin converting enzyme, and hydrochlorothiazide (HCTZ), a diuretic. The lowest dose of each drug is placebo. Treatment groups are displayed in Figure 1.3.

After a 2-4 week placebo run-in phase, 529 subjects were randomized to the treatment groups, with between 42 and 48 subjects per group. The subjects were followed for 6 weeks. At each visit, standing and supine blood pressure were recorded as the average of three successive measurements. The outcome variables are standing and supine diastolic and systolic blood pressure at each follow-up visit. Of primary interest for this example is the equivalence between treatment combinations for blood pressure change from baseline to the final visit, for all four measurement types. In particular, it is hoped that several subsets of treatment combinations can be shown equivalent: therapies 2 and 5; therapies 3, 4, 6, 9, 10; and therapies 7, 8, 11, 12. An analysis based on response surface methodology found that models assuming equivalence between these treatment groups had reasonable goodness-of-fit, but equivalence was not specifically tested (Phillips et al 92).

Figure 1.3
Treatment Groups for Parallel Clinical Trial of Hypertension

<table>
<thead>
<tr>
<th></th>
<th>Ramipril placebo</th>
<th>Ramipril 2.5 mg</th>
<th>Ramipril 5 mg</th>
<th>Ramipril 10 mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HCTZ placebo</td>
<td>HCTZ 2.5 mg</td>
<td>HCTZ placebo</td>
<td>HCTZ placebo</td>
</tr>
<tr>
<td>5</td>
<td>Ramipril placebo</td>
<td>Ramipril 2.5 mg</td>
<td>HCTZ 2.5 mg</td>
<td>HCTZ 12.5 mg</td>
</tr>
<tr>
<td>9</td>
<td>HCTZ 12.5 mg</td>
<td>Ramipril 5 mg</td>
<td>HCTZ 12.5 mg</td>
<td>HCTZ 12.5 mg</td>
</tr>
<tr>
<td>10</td>
<td>Ramipril placebo</td>
<td>HCTZ 25 mg</td>
<td>Ramipril 5 mg</td>
<td>Ramipril 10 mg</td>
</tr>
<tr>
<td>12</td>
<td>HCTZ 25 mg</td>
<td>HCTZ 25 mg</td>
<td>HCTZ 25 mg</td>
<td>HCTZ 25 mg</td>
</tr>
</tbody>
</table>
1.4.3 Parallel Clinical Trial of Muscle Soreness

In this randomized double-blind study, two treatments were compared for their ability to decrease delayed onset muscle soreness, generally the result of unaccustomed exercise or exertion. After meeting entry criteria, a subject's dominant bicep muscle was challenged with an exercise regimen. 129 subjects were randomized to one of two treatment groups (A and B), and treatments were applied topically at regular intervals over three days. Muscle pain was assessed by two scales: an ordered category scale consisting of 5 categories, (0) no soreness, (1) a little soreness, (2) some soreness, (3) a lot of soreness, and (4) as sore as can be; and a visual analog scale rated from 0 to 100, where 0 is no muscle soreness and 100 is as sore as can be. Three types of muscle soreness were measured: global pain, soreness on active range of motion, and soreness on palpation. Evaluation of both scales was made at baseline and 12 time points post randomization.

The primary outcome of interest for each of the three soreness evaluations is the weighted average change in soreness from baseline, averaging over all twelve time points, for both the categorical and visual analog scales - interpreted as the reduction in soreness halfway through a subject's follow-up period. Secondary categorical measures include reporting of no pain on the category scale after 48 hours, and reporting a 50% reduction in pain on the visual analog scale after 48 hours. Also of interest is an ordinal variable, defined as change in pain from baseline to 48 hours. It has four categories: worse/no change, some improvement, moderate improvement, and marked improvement. The goal of the study was to show that the two treatment regimens, A and B, resulted in equivalent pain reduction.
1.4.4 Equivalence of Dermatologic Therapies

In this randomized double-blind trial, a test and reference formulation of a topical corticosteroid are compared for equivalence. Each drug is prepared in three different ointment formulations, making a total of six unique treatments. The ointments are applied to healthy skin and evaluated for blanching, the visual effect of vasoconstriction.

The study was conducted at two centers, on 24 healthy volunteer subjects per center. The treatments were applied to one of 6 skin sites on the arms and legs of both the left and right sides of the body, so that subjects received each treatment twice, once on each side of the body. Skin site randomization was balanced across the study so that each treatment was applied to each site an equal number of times. The twelve skin sites were then covered for sixteen hours, after which they were evaluated for blanching.

Vasoconstriction was evaluated on an ordered category scale: (0) no blanching, (1) barely detectable, (2) definite but not maximum, (3) marked maximal blanching. A second evaluation was made allowing a half step between the 0-1, 1-2, and 2-3 categories. Vasoconstriction was evaluated on both scales by two independent observers, resulting in four blanching scores per site.

The goal of the study is to compare the six drug-ointment formulations for equivalence of vasoconstriction, by examining both measurement scales. The factors of center, side of the body, sites of the treatments, and observer must all be taken into account.

1.5 Research Goals

This chapter has examined clinical trial issues, particularly for trials designed to show treatment equivalence. Statistical issues for the evaluation of multiple endpoints or multiple treatment comparisons were examined, generally for trials
attempting to show a treatment difference. There are no multiple endpoint or multiple treatment strategies specifically designed for equivalence studies.

The goal of this research is to develop statistical methods for achieving a reasonably powerful analysis of the combined information from multiple response criteria in an equivalence clinical trial. While no new theoretical results are derived, new techniques are extended and developed from existing statistical methodology. Emphasis is placed on ratios of treatment measures, since they are easily combinable due to their unitless nature, but some methods may also be applicable on the additive scale.

The application of methods is focused on clinical trials as described in Section 1.1.2. While bioequivalence studies are an immediate candidate for such methods and receive considerable attention in several chapters, applications to the broader class of clinical equivalence trials may prove more beneficial. Clinical trials may be more likely to have multiple endpoints and multiple groups, whereas bioequivalence studies have multiple endpoints, but are generally limited to two treatments.

A review of standard techniques used for analyzing traditional difference studies, with possibly multiple endpoints or multiple treatment groups, is presented in Chapter 2. Univariate techniques designed for bioequivalence and clinical equivalence are presented in Chapter 3. The remaining chapters evaluate strategies for combining multiple outcomes from equivalence studies.
Chapter 2
Review of Standard Methods for Showing a Difference

Often a common strategy for examining multiple endpoints is to conduct an analysis of each measurement type, time point, or outcome response separately. In such scenarios, the analyses are based on univariate methods. In other situations, multivariate methods are available to test some endpoints jointly. The following is a brief review of the major univariate and multivariate methods most commonly used in analyses comparing two or more treatments. This review covers dichotomous, ordinal categorical, and continuous responses and strategies for direct comparisons, stratification, and covariate adjustment. Both non-parametric and parametric methods are presented. Except where noted, reference to these methods is from Koch and Edwards (1988).

Univariate methods comparing two treatments, generally for a difference, are discussed in detail. Some multivariate methods and strategies which combine endpoints or evaluate several treatments simultaneously are mentioned briefly. Nearly all methods discussed are designed to show a difference between treatments, as opposed to an equivalence. Multivariate methods applicable to equivalence studies are discussed in further detail in later chapters.

2.1 Dichotomous Response

This section is limited to the comparisons of two treatments for outcomes in the form of presence or absence of a condition or response. In the following discussion, the notation is based on a $2 \times 2$ contingency table with the structure as described in Figure 2.1:
Figure 2.1 Notation for Dichotomous Response

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Favorable</th>
<th>Unfavorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
</tr>
<tr>
<td>B</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
</tr>
<tr>
<td></td>
<td>( n_1 )</td>
<td>( n_2 )</td>
</tr>
</tbody>
</table>

2.1.1 Direct Comparisons

Fisher's Exact Test

Under the assumption of randomization to treatments, the treatment marginal totals can be considered fixed, and under the null hypothesis of no treatment group difference in response, the column totals can also be considered fixed. The univariate hypergeometric distribution describes the probability of observing a count in any cell from the \( i \)th row or \( j \)th column:

\[
Pr(n_{ij} | H_0) = \frac{n_1! \cdot n_2! \cdot n_1! \cdot n_2!}{n! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!}
\]  

(2.1)

The null hypothesis of no treatment difference can be examined using Fisher’s exact test. A p-value is found by calculating and then summing the probabilities of the observed table plus any possible table with the same fixed marginals more extreme than observed in the direction of the alternative hypothesis. A two-sided alternative of treatment differences involves tabulating both tails of the distribution, while for a one-sided hypothesis only one tail is summed. Fisher's test can not be converted to a confidence interval, and since it is randomization based, inferences can be made only to the study population. This exact method is valid for any sample size but is best applied to small studies since it can be computationally intensive.
**Pearson Chi-square Test**

If the sample size is large and marginals can be assumed fixed, then the cell counts have an approximate normal distribution by the Central Limit Theorem. The expected count for cell \( n_{ij} \) under the null hypothesis \( E(n_{ij}|H_0) = m_{ij} = (n_i .)(n .j)/n \) and \( Var(n_{ij}|H_0) = v_{ij} = (n_1 .n_2 .n_3 .)/n^2 (n-1) \). To test the null hypothesis of no treatment difference against the 2-sided alternative of treatment differences, the Pearson chi-square statistic is calculated as

\[
Q_p = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}
\]  

(2.2)

If all expected cell counts are greater than 5 or 10, \( Q_p \) has an approximate \( \chi^2 \) distribution with 1 degree of freedom. The less-used randomization chi-square statistic is

\[
Q = Q_p (n-1)/n = \frac{(n_{ij} - m_{ij})^2}{v_{ij}}, \text{ for any } n_{ij}.
\]  

(2.3)

For large samples the two are essentially equivalent. The chi-square test is always two-sided, and like Fisher's test, conclusions from it are applicable only to the study population. It is also not directly invertible to a confidence interval. Yate's correction to \( Q_p \), obtained by adding 0.5 to each expected count, yields a \( p \)-value potentially more similar to Fisher's exact test (Fleiss 81).

**Odds Ratio**

While chi-square and Fisher's tests evaluate the association between treatment and response, they do not give a quantitative measure of the association. Many such measures of association have been suggested in the literature. The most common of these is the odds ratio. Odds ratios are most appropriate when the response occurs in less than 15% or greater than 85% of the population. It is calculated as

\[
\hat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}}
\]  

(2.4)

and is interpreted as the ratio of the odds of a favorable response of treatment A vs treatment B. If treatment is not associated with response, the odds ratio is 1.0.

For reasonably large samples, \( \log_e(\hat{\psi}) \) approximately has a normal
distribution with mean \( \log_e \psi \) and variance consistently estimated by \( \text{var}(\log_e \psi) = \left( \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right) \). The null hypothesis that \( \psi = 1 \) is traditionally tested against the general alternative using Pearson's chi-square or Fisher's test. However, a 100(1 - \( \alpha \))% confidence interval (CI) for the odds ratio is obtained by:

\[
\exp\left[ \log_e \psi \pm z_{1 - \alpha / 2} \sqrt{\text{var}(\log_e \psi)} \right]
\]

One-sided CIs can be similarly constructed.

The normal approximation is less biased when 0.5 is first added to each cell count, but results are essentially identical for large samples. Without this continuity correction, the odds ratio can be undefined when zero cell counts are encountered. Typically zero cells arise in small sample situations when inferences concerning the odds ratio should be viewed with caution. In small samples, a confidence interval based on Mantel-Haenszel methods [equations (2.12) and (2.15), below] with the number of strata = 1 may be more appropriate.

**Difference Between Proportions**

When the population proportion of the response is less extreme (i.e., between 15% and 85%), an appropriate measure of association is the treatment difference of proportions with a favorable response, calculated as \( d = (p_1 - p_2) = n_{11}/n_1 - n_{21}/n_2 \). Under the assumption that the study population is a simple random sample, for large sample size the difference has an approximate normal distribution with mean \( E(d) \) and variance consistently estimated by \( \text{var}(d) = (p_1)(1-p_1)/n_1 + (p_2)(1-p_2)/n_2 \). A 100(1 - \( \alpha \))% CI is:

\[
d \pm \left[ z_{1 - \alpha / 2} \sqrt{\text{var}(d)} + 1/2(1/n_1 + 1/n_2) \right].
\]

The last term is a continuity correction factor. A test of the null hypothesis of no difference is carried out with the continuity corrected Wald chi-square statistic:

\[
Q_{Wc} = \frac{[|p_1 - p_2| - 1/2(1/n_1 + 1/n_2)]^2}{(p_1)(1-p_1)/n_1 + (p_2)(1-p_2)/n_2}.
\]

For large samples \( Q_{Wc} \) approximately has a \( \chi^2 \) distribution with 1 df (Fleiss 81).
2.1.2 Stratified Comparisons

Mantel-Haenszel Chi-square Test

Often it is of interest to evaluate the association between treatment and outcome by combining the results from a set of strata, for instance gender, categorical baseline status, or clinic in a multi-center trial. With the assumption of fixed row and column marginals, the probability distribution of each stratum under the null hypothesis is hypergeometric, as described previously. When randomization has occurred separately in each stratum (as in a multi-center trial) or when the distributions are examined conditional on the stratification variable, the strata are independent and the overall probability is described by the product hypergeometric:

\[
Pr(n_{11j}, n_{21j}, ..., n_{h1j} | H_0) = \prod_{h=1}^{H} \frac{n_{h1!} n_{h2!} n_{h1.1} n_{h1.2}}{n_h! n_{h11!} n_{h12!} n_{h21!} n_{h22!}}
\]  

(2.8)

Exact p-value computation for the test of no treatment differences controlling for the stratification variable is available from StatXact (Cytel Software 1991).

The Mantel-Haenszel chi-square statistic, used more frequently, is defined as:

\[
Q_{MH} = \frac{\left( \sum_{h=1}^{H} n_{h11}^{-1} \right) \left( \sum_{h=1}^{H} m_{h11} \right)^2}{\sum_{h=1}^{H} V_{h11}}
\]  

(2.9)

where \( h \) refers to the stratum (1, 2, ... H), \( m_{h11} \) is \( E(n_{h11} | H_0) = n_{h1}. n_{h1.1}/n_h \), and \( V_{h11} \) is \( \text{Var}(n_{h11} | H_0) = \frac{n_{h1} n_{h2} n_{h1.1} n_{h1.2}}{n_h^2 (n_h - 1)}. \) For reasonably large samples, \( Q_{MH} \) approximately has a \( \chi^2 \) distribution with 1 d.f. Fleiss suggests applying a continuity correction to \( Q_{MH} \) to improve the \( \chi^2 \) approximation. A rule of thumb used to determine adequacy of the sample size is called the Mantel-Fleiss criterion:

\[
\min \left\{ \left[ \sum_{h=1}^{H} m_{h11} - \sum_{h=1}^{H} (n_{h11})_L \right], \left[ \sum_{h=1}^{H} (n_{h11})_U - \sum_{h=1}^{H} m_{h11} \right] \right\} \geq 5
\]  

(2.10)

where \((n_{h11})_L = \max(0, n_{h1} - n_{h.2})\) and \((n_{h11})_U = \min(n_{h1.1}, n_{h.1})\) are the smallest
and largest possible values for \( n_{h11} \) across all possible randomizations with the marginal frequencies fixed. If the Mantel-Fleiss criterion is not met, \( \chi^2 \) is a poor approximation and only exact methods should be used. The criterion is generally met when the combined strata treatment sample sizes are large (i.e., all \( n_{i} > 20 \) for several strata).

Due to sample size constraints, this method has a limited ability to adjust for many strata. As with all randomization-based methods, statistical inference is limited to the study population.

The Mantel-Haenszel statistic is a test of average partial association, and is effective for detecting patterns of treatment group differences across the respective strata when these patterns are similar, or homogeneous. It is not sensitive when patterns are dissimilar, and should be interpreted with caution when patterns are strongly dissimilar. Homogeneity should be evaluated with the Breslow-Day test, described below.

**Common Odds Ratio**

When there is sufficient homogeneity of association across strata, the Mantel-Haenszel estimate of the common odds ratio is an appropriate measure of the odds ratio for the strata considered jointly. It is calculated as

\[
\hat{\psi}_{MH} = \frac{\sum_{h=1}^{H} n_{h11} n_{h22}}{\sum_{h=1}^{H} n_{h12} n_{h21}/n_h}.
\]  

(2.11)

The null hypothesis \( H_0: \psi_{MH} = 1 \) is tested against \( H_A: \psi_{MH} \neq 1 \) by the Mantel-Haenszel chi-square test. Under \( H_0 \), \( \text{Var} (\log_e (\hat{\psi}_{MH})) = Q_{MH} \). A test-based 100(1-\( \alpha \))% Confidence Interval can thus be constructed by:

\[
\exp \left[ (\log_e \hat{\psi}_{MH}) \times (1 \pm z_{1-\alpha/2} / \sqrt{Q_{MH}}) \right].
\]

(2.12)

Another frequently used form of the common odds ratio is the weighted
regression, or logit, estimate $\hat{\psi}_W$:

$$\hat{\psi}_W = \exp \left( \frac{\sum_{h=1}^{H} w_h (\log_e \hat{\psi}_h)}{\sum_{h=1}^{H} w_h} \right) ,$$

where $w_h = \left[ \frac{1}{n_{h11}} + \frac{1}{n_{h12}} + \frac{1}{n_{h21}} + \frac{1}{n_{h22}} \right]^{-1}$. A 100(1-$\alpha$)% CI for $\hat{\psi}_W$ is

$$\exp \left[ \log_e (\hat{\psi}_W) \pm z_{1-\alpha/2} \left( \sum_{h=1}^{H} w_h \right)^{-1/2} \right] .$$

(2.14)

The Mantel-Haenszel formula is appropriate whenever the sample size meets the requirements for the $Q_{MH}$ test, but the weighted regression estimate requires all cell counts > 5, and is thus limited to large sample situations.

An alternative to the Mantel-Haenszel test-based interval was recently proposed by Sato (90). It is based on the approximate $\chi^2$ distribution of the function $\frac{[R - \psi S - c^2]}{\psi T}$, where $c$ is a continuity correction equal to either zero or $(1+\psi)/4$,

$$R = \sum_{h=1}^{H} \frac{n_{h11} n_{h22}}{n_h} ,$$

$$S = \sum_{h=1}^{H} \frac{n_{h12} n_{h21}}{n_h} \quad \text{and}$$

$$T = \sum_{h=1}^{H} \left\{ \left[ \frac{n_{h12}}{n_h} + \frac{1}{n_h} \right] \frac{n_{h11} n_{h22}}{n_h} + \left[ \frac{n_{h11}}{n_h} + \frac{n_{h22}}{n_h} + \frac{1}{n_h} \right] \frac{n_{h12} n_{h21}}{n_h} \right\} .$$

Limits of the 100(1-$\alpha$)% confidence interval for $\psi$, solved using the quadratic formula, are:

$$\frac{(2RS + \chi^2_{1-\alpha} T) \pm \sqrt{(4RS + \chi^2_{1-\alpha} T)\chi^2_{1-\alpha} T}}{2S^2}$$

(2.15)

The continuity correction is applied by substituting $R - \frac{1}{4}$ and $S + \frac{1}{4}$ for the lower bound, and $R + \frac{1}{4}$ and $S - \frac{1}{4}$ for the upper bound. Advantages of this estimate are that a log transformation is not required, it is applicable for small and large sample sizes, and performs better than the Mantel-Haenszel estimate when the true odds ratio is one.

Generally the three methods lead to very similar intervals. The weighted regression estimate assumes the data are a stratified simple random sample from a larger population. When this is met, conclusions can be drawn to that population.
The Mantel-Haenszel estimates are test based and presume no association, so inference can be drawn only in this situation. Neither method should be used when there is significant heterogeneity, especially when the sign of the association differs between strata.

**Breslow-Day Test of Homogeneity**

For sets of two by two tables, the Breslow-Day test is used to evaluate homogeneity of the association between treatment and response across strata. With reasonably large samples, the Breslow-Day statistic is approximately distributed $\chi^2$ with $(H - 1)$ degrees of freedom.

$$Q_{BD} = \sum_{h=1}^{H} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{hij} - m_{hij})^2}{m_{hij}},$$

(2.16)

where $m_{hij} = E(n_{hij} | \hat{\psi}_{MH})$ are obtained by calculating $\hat{\psi}_{MH}$ and solving for each $m_{hij}$ as follows:

$$\frac{[m_{h11}(n_{h1} - n_{h1} + m_{h11})]}{(n_{h1} - m_{h11})(n_{h1} - m_{h11})} = \hat{\psi}_{MH},$$

$$m_{h12} = (n_{h1} - m_{h11}) \quad m_{h21} = (n_{h1} - m_{h11}) \quad m_{h22} = (n_{h} - n_{h1} - n_{h1} + m_{h11}).$$

Breslow-Day tests the null hypothesis that all stratum odds ratios are equal versus the alternative that at least two are unequal. This test requires expected cell counts in each cell of each stratum to be $\geq 5$ or 10, which is more stringent than the Mantel-Haenszel test, so sometimes homogeneity can not be tested for a valid Mantel-Haenszel test. However, problems with interpreting the Breslow-Day test in small samples are usually limited to apparent heterogeneity when in fact there is reasonable homogeneity.

### 2.1.3 Covariate Adjusted Comparisons

Logistic Regression is a statistical model for a dichotomous response which allows treatment comparisons to be adjusted for a set of explanatory variables, including continuous covariates and categorical stratifications. Treatment
comparisons can be conducted on more than two treatments at once. Homogeneity of treatment effects across strata and covariates can be assessed in a straightforward manner through interactions.

Compared with randomization-based methods, it has greater flexibility in adjustment of demographic variables which are not equally distributed among the treatments, and leads to a more powerful analysis by reducing the variability through adjustment of explanatory variables strongly associated with the response. The assumptions concerning the study population, however, are often not applicable to clinical trials. The study population is assumed to be a stratified simple random sample of some larger target population. Under this assumption, the overall stratum \( x \) treatment \( x \) response cross-classification has the product binomial distribution:

\[
P(n_{hij}) = \prod_{h=1}^{H} \prod_{i=1}^{I} \frac{n_{hi1}! n_{hi2}!}{n_{hi1} n_{hi2}} (1 - \theta_{hi})^{n_{hi1}} \theta_{hi}^{n_{hi2}}
\]  

(2.17)

where \( \theta_{hi} \) is the probability that a subject in stratum \( h \) with treatment \( i \) will have a favorable response. Each value of a continuous covariate is considered a stratum, so models with several continuous predictors are likely to have \( H \) be quite large. In such situations, the notation of \( h \) strata and \( i \) treatment groups is changed to \( s \), where \( s \) is the total number of subpopulations with at least one subject for the cross-classification of explanatory variables. Specification of the logistic model is:

\[
\theta_{hi} = \left[ 1 + \exp(-x_{hi} \beta) \right]^{-1}
\]  

(2.18)

The modeled response function is the log odds, or logit, of favorable response:

\[
\log_{e} \left( \frac{\theta_{hi}}{1 - \theta_{hi}} \right) = x_{hi} \beta
\]  

(2.19)

where \( x_{hi} \) is a \( (1 \times p) \) row vector of explanatory variables corresponding to the \( h \)th stratum and the \( i \)th treatment, and \( \beta \) is a \( (p \times 1) \) column vector of parameters. Reference cell coding is often used, in which case the first element of \( x_{hi} \) is a 1 for all observations, and the first element of \( \beta \) is interpreted as the log odds for the reference cell with all covariates at the value zero. All other elements of \( \beta \) are
interpreted as increment in log odds due to one increment of the corresponding element of the $x_{hi}$ vector. The $(n \times p)$ design matrix $X$ is a concatenation of the $x_{hi}'$ rows for all observations.

$\beta$ is estimated by maximum likelihood as follows: in the product binomial formula, or likelihood function, $\theta_{hi}$ is replaced by its logistic model specification. Then the log of the likelihood is differentiated with respect to $\beta$, and set to zero to obtain the maximum likelihood equation:

$$
\sum_{h=1}^{H} \sum_{i=1}^{2} (n_{hi1} - (n_{hi})\hat{\theta}_{hi})[x_{hi}'] = 0
$$

(2.20) where $\hat{\theta}_{hi} = [1 + \exp(-x_{hi}'\hat{\beta})]^{-1}$ is the maximum likelihood estimate of $\theta_{hi}$, and $\hat{\beta}$ is the maximum likelihood estimate of $\beta$. The estimates are calculated iteratively using the Newton-Raphson algorithm. When $\sum_{h=1}^{H} \sum_{i=1}^{2} n_{hi}x_{hi}$ is approximately normally distributed, then $\hat{\beta}$ is likewise approximately normal, with $E(\hat{\beta}) = \beta$ and estimated variance-covariance matrix

$$
v(\hat{\beta}) = \left[ \sum_{h=1}^{H} \sum_{i=1}^{2} (n_{hi})\hat{\theta}_{hi}(1 - \hat{\theta}_{hi})x_{hi}x_{hi}' \right]^{-1}
$$

(2.21) Standard errors of the estimates are calculated as $\text{diag}\{(v(\hat{\beta}))^{1/2}\}$. The null hypothesis that $\beta_k = 0$ is tested with the Wald statistic

$$
Q_W = [\hat{\beta}_k / \text{s.e.}(\hat{\beta}_k)]^2
$$

(2.22) which approximately has a $\chi^2$ distribution with 1 d.f.

Predicted probabilities of favorable response for stratum $h$ and treatment $i$ are calculated as $\exp(x_{hi}'\hat{\beta})$. A predicted odds ratio of favorable response for treatment $i$ vs. the reference treatment is obtained by $\exp(\hat{\beta}_k)$, and a 100(1-$\alpha$)% confidence interval for the odds ratio is obtained by

$$
\exp[\hat{\beta}_k \pm z_{1-\alpha/2} \times \text{SE}(\hat{\beta}_k)]
$$

(2.23) where $\beta_k$ is the coefficient corresponding to $x_{hi} = 1$ if treatment $i$, 0 otherwise.

Goodness of fit of the logistic model can be assessed by several methods. The log-likelihood chi-square statistic is defined as:

$$
Q_L = \sum_{h=1}^{H} \sum_{i=1}^{2} \sum_{j=1}^{2} 2n_{hij}\log_e(n_{hij}/\hat{m}_{hij}),
$$

(2.24)
[defined as zero when \( n_{hij} = 0 \)], where \( \tilde{m}_{hi} = n_{hi}(1-\hat{\theta}_{hi}) \), and \( \tilde{m}_{hi} = n_{hi} \hat{\theta}_{hi} \). When all \( m_{hij} \geq 5 \), \( Q_L \) approximately has a \( \chi^2 \) distribution with degrees of freedom equal to \([\text{number of (treatment x background variable) strata} - \text{rank}(X)]\). The null hypothesis is that the \( n_{hij} \) are reasonably close to the \( \tilde{m}_{hij} \). Non-significance of the test supports the model as a good fit to the data. Due to the sample size requirement, this method can usually not be used in the presence of continuous covariates.

An alternate strategy for testing goodness of fit is to specify an expanded model with the design matrix \([X \ W]\) which includes interactions between the predictor variables. Non-significance of the interactions is verified by the Wald chi-square statistic on an individual basis, or by non-significance of the log-likelihood ratio chi-square statistic which compares the fit of the expanded model to that of the original model:

\[
Q_{LR} = \sum_{i=1}^{S} \sum_{j=1}^{2} 2n_{ij} \log_e(\tilde{m}_{ij,\text{expanded}}/\tilde{m}_{ij}).
\]  

\(Q_{LR}\) is approximately distributed \( \chi^2 \) with degrees of freedom = \( \text{rank}(W) \) when the sample size is sufficiently large. This sample size is less strict than that required for \( Q_L \). An asymptotically equivalent statistic which does not require fitting the expanded model is the Rao score statistic, \( Q_{RS} \). It assesses the association of the residuals with \( W \), and is also approximately distributed \( \chi^2 \) with d.f. = \( \text{rank}(W) \).

### 2.2 Ordinal Categorical Response

Often an outcome measure has several ordinal categories such as a ranking of mild, moderate, severe. Following is a discussion of the most common analysis strategies used to evaluate treatment comparisons, as described by Koch and Edwards (1988). For the following discussion, it is assumed there are a total of \( J \) ordered outcomes.
2.2.1 Direct Comparisons

By the same arguments provided for dichotomous outcomes, the row and column totals can be considered fixed under the null hypothesis of no treatment difference, with the assumption of randomization of treatments. The $2 \times J$ cross tabulation thus has a multivariate hypergeometric distribution:

$$
Pr(n_{ij}|H_0) = \frac{\prod_{i=1}^{2} n_{i.}! \prod_{j=1}^{J} n_{..j}!}{n! \prod_{i=1}^{2} \prod_{j=1}^{J} n_{ij}!}
$$  \hspace{1cm} (2.26)

If each outcome category is given a numeric value, or score, then the mean score chi-square statistic $Q_x$ is used to provide power against the alternative of a larger mean response in one of the treatment groups.

$$
Q_x = \frac{(n-1)(\bar{f}_1 - \bar{f}_2)^2}{\frac{n}{n_1 + 1/n_2} V_a},
$$  \hspace{1cm} (2.27)

where $\bar{f}_i = \frac{\sum_{j=1}^{J} a_j n_{ij}}{n_i}$ is the mean score for treatment $i$, $V_a = \sum_{j=1}^{J} (a_j - \mu_a)^2 n_{ij}/n$ = $\text{Var}(\bar{f}_1|H_0)$ $\times \frac{n_1(n-1)}{n}$, and $\mu_a = E(\bar{f}_1|H_0) = \sum_{j=1}^{J} a_j n_{ij}/n$. The $a_j$ are a set of scores reflecting the response levels.

There are several common choices for the scores. Integer scores assign a simple count to the response categories, i.e., $a_j = 1, 2, 3 \ldots J$. This scoring is appropriate for discrete counts, such as number of occurrences of an event, or for equally spaced categories, such as hours to an event.

Modified ridit or standardized midrank scores account for the ordinality of the response categories, but impose no assumptions about the quantitative relationship between the response levels. They are calculated as

$$
a_j = \left[2 \sum_{k=1}^{j} n_{k \cdot} - n_{..j} + 1\right] / 2(n+1).
$$  \hspace{1cm} (2.28)

The mean score statistic with modified ridit scoring is the contingency table counterpart of the Wilcoxon-Mann-Whitney rank sum statistic, described below.

Other scoring strategies include logrank scores for data with L-shaped distributions for which more interest is placed on higher values of the response
than lower values. Binary scores collapse the table to a $2 \times 2$ and $Q_*$ becomes identical to the randomization chi-square discussed previously.

In large samples, $Q_*$ approximately has a chi-square distribution with 1 degree of freedom. The approximation is reasonable provided all row totals are $\geq 20$. Exact p-values for smaller samples are available from StatXact (Cytel Software 91). Limitations of this test are that it cannot be converted to a confidence interval, and one can draw conclusions only to the study population.

2.2.2 Stratified Comparisons

Extended Mantel-Haenszel Statistic

Often the association between treatment and an ordinal categorical response is stratified over a categorical descriptor. If treatment assignment can be considered independent in each stratum, then the underlying distribution is the product of the multivariate hypergeometric distribution for each strata:

$$
Pr(n_{1ij}, n_{2ij}, \ldots n_{hij} | H_0) = \prod_{h=1}^{H} \frac{n_{hi1}!}{n_{hih}!} \frac{\prod_{j=1}^{J} n_{hij}!}{\prod_{i=1}^{I} n_{hij}!}
$$

An extension of the Mantel-Haenszel mean score chi-square statistic is effective for detecting treatment differences across strata when the stratum-specific associations are similar. The extended Mantel-Haenszel chi-square is calculated as:

$$
Q_{EMH} = \left[ \sum_{h=1}^{H} \left( \frac{n_{h1} n_{h2} / n_h}{(\bar{f}_{h1} - \bar{f}_{h2})} \right)^2 \right] / \sum_{h=1}^{H} \left( \frac{n_{h1} n_{h2} / n_h}{V_h} \right)^2
$$

where $\bar{V}_h = (1/n_{h1} + 1/n_{h2})[n_h/(n_h - 1)]V_h$, $\bar{f}_{hi}$ is the mean score for the $i$th treatment in the $h$th stratum, and $V_h$ is the variance from the mean score statistic, as described above, for each stratum. When sample sizes are large, $Q_{EMH}$ approximately has a $\chi^2$ distribution with 1 d.f. All scores mentioned for
the mean score statistic also apply to this test. When modified ridits are used, $Q_{EMH}$ is the categorical data counterpart of the Van Elteren statistic, which combines Wilcoxon rank sum tests across strata.

$Q_{EMH}$ detects consistent patterns of treatment differences when the strata $(\bar{f}_{h1} - \bar{f}_{h2})$ have the same sign. It is a test of average partial association, and is not sensitive when patterns are dissimilar. It is therefore important to evaluate the homogeneity of the relationship across strata.

Extended Mantel Haenszel statistics require the combined strata sample sizes to be sufficiently large. No formula is available to check size, although categories could be collapsed to a dichotomy and checked via the Mantel-Fleiss criterion. Certainly there is no problem when sample sizes are sufficient for evaluation of pseudo-homogeneity, described below. In small sample situations, exact $p$-values can be obtained from StatXact (Cytel Software 1991).

**Pseudohomogeneity statistic**

Homogeneity of association across strata can not be exactly tested for ordinal categorical responses using randomization based methods. The closest alternative is called the test of pseudo-homogeneity, which actually tests no association between treatment and response, rather than homogeneous association across strata. Another strategy is to collapse ordinal variables into dichotomies and proceed with the Breslow-Day test.

The pseudo-homogeneity statistic is calculated as $Q_{PH} = Q_T - Q_{EMH}$, where $Q_T$ is the sum of the chi-square mean score statistics over each stratum:

$$Q_T = \sum_{h=1}^{H} Q_h = \sum_{h=1}^{H} \left[ \frac{(n_h - 1)(\bar{f}_{h1} - \bar{f}_{h2})^2}{n_h(1/n_{h1} + 1/n_{h2})V_h} \right].$$  \hspace{1cm} (2.31)

With large sample size, $Q_{PH}$ is approximately distributed $\chi^2$ with d.f. = $(H - 1)$. It should be viewed with caution when sample sizes are small, particularly when one or more stratum row totals are less than 20.
Rank Measures of Association

A disadvantage of extended Mantel-Haenszel tests is they have a limited ability to evaluate homogeneity across strata. The proportional odds model, described below, can evaluate homogeneity, but it may not be a good fit of the data. Another option which evaluates homogeneity of association across strata is the rank measure of association. The rank measure can be defined for each stratum as an estimate of the probability that a randomly selected test patient has a more favorable response than a placebo patient:

\[
g_h = \sum_{j=1}^{J} \frac{n_{h1j}}{n_{h1}} \left[ \left( \sum_{k=1}^{K} \frac{n_{h2k}}{n_{h2}} \right) - 0.5 \frac{n_{h2j}}{n_{h2}} \right] \tag{2.32}
\]

The covariance matrix of the \( g_h \) is constructed with linear Taylor series approximation, and variation among the strata is analyzed by weighted least squares. The advantage of homogeneity assessment is counterweighted by a more stringent sample size requirement than the Mantel-Haenszel test and computational difficulty.

2.2.3 Covariate Adjusted Comparisons

Parametric covariate adjusted comparisons of ordinal data require imposing a structure of pairwise odds ratios amongst the categories. Two models used most often are the proportional odds and the equal adjacent odds models. Equal adjacent odds makes the assumption that odds ratios between a pair of adjacent categories is equal to odds ratios formed by any set of adjacent pairs. For instance, if there are three categories of mild, moderate, and severe, the equal adjacent odds model assumes the odds ratio of treatment A vs. B for mild vs. moderate response is the same as the odds ratio for moderate vs. severe. It is computationally less available than the more used proportional odds model, and so only proportional odds is discussed in detail.
Proportional Odds Regression

Proportional odds is an extension of logistic regression for an ordered categorical response. Like logistic regression, treatment comparisons can be adjusted for a set of explanatory variables, including continuous covariates and categorical stratifications. It has the same advantages, disadvantages, and assumptions as logistic regression. One additional assumption, called the proportional odds assumption, requires that odds ratios comparing treatments be the same for each increment in the ordinal scale. For example, the odds ratio of treatment A vs treatment B is assumed to be the same for mild vs. moderate/severe, and mild/moderate vs. severe.

If there are $J$ ordered categories, define $J-1$ sets of dichotomized categories as described in the example. The modeled response function for all $J-1$ dichotomies simultaneously is the log odds, or logit, of each favorable response:

$$\log\left(\frac{\theta_{hi}^k}{1-\theta_{hi}^k}\right) = \alpha_k + x'_{hi}\beta$$  (2.33)

where $\alpha_k$ is the reference log odds for each of the $J-1$ category dichotomies, $x'_{hi}$ is a $(1 \times p)$ row vector of explanatory variables corresponding to the $h$th stratum and the $i$th treatment, and $\beta$ is a $(p \times 1)$ column vector of parameters, interpreted as the increment in log odds due to 1 increment of the corresponding element of the $x'_{hi}$ vector, regardless of category dichotomy. The $(n \times p)$ design matrix $X$ is a concatenation of the $x'_{hi}$ rows for all observations. Parameter estimation and goodness of fit of the model are identical to standard logistic regression. The proportional odds assumption can be tested with a score statistic.

Randomization Covariance Analysis

Covariate adjustment can be accomplished in a non-parametric environment using a combination of regression and Mantel-Haenszel methods. The residual
from the regression of the response variable onto the covariate is calculated using least squares. The treatment groups are compared via a version of the mean score statistic of the residuals. If stratification is also to take place, then a version of the extended Mantel-Haenszel statistic is available. The only assumption is that subjects are randomized to treatments. Sample size necessary is equivalent to that needed for the Mantel-Haenszel statistics as described above. This method could also be applied to continuous data. When both the response and covariate are in the form of ranks, the score statistic is equivalent to the rank analysis of covariance as discussed by Quade (1967) (Koch et al 1989).

2.3 Continuous Response

Following are the most common univariate tests used to compare the effectiveness of two treatments for continuous outcomes.

2.3.1 Direct Comparisons

Wilcoxon Rank Sums (Mann Whitney)

This non-parametric technique can be applied to data that is continuous, but not necessarily normally distributed, or has small sample sizes. The Wilcoxon test assumes treatment groups are formed from two independent random samples, the distributions differ only in location, but not variability, and there are few tied values. To test $H_0$: median $\nu_1 = \nu_2$ vs. $H_A$: $\nu_1 \neq \nu_2$, the Wilcoxon statistic is calculated by ranking the response outcomes for both treatment groups combined, and then summing the ranks for one treatment group:

$$W = S - n_1(n_1 + 1)/2,$$

(2.34)

where $S = \text{sum of jointly-ranked data for treatment group 1}$, and $n_1$ is the sample size from group 1. Ties in the data are assigned midranks. The exact p-value is obtained from a table of critical values for the distribution. Exact confidence
intervals can also be similarly obtained. Due to the discrete nature of the
distribution, 100(1-\(\alpha\))% CIs often have a coverage probability greater than (1-\(\alpha\)).

When \(n_1\) and \(n_2\) are both large, a normal approximation can be calculated:

\[
z = \frac{W - n_1n_2/2}{\sqrt{n_1n_2(n_1 + n_2 + 1)/12}}
\]  

(2.35)

Under the null hypothesis \(z\) is approximately normally distributed with mean 0
and variance 1. A correction can be used when there are many across-group ties.
A disadvantage is that the normal approximation can not be expressed as a
meaningful confidence interval (Daniel 90).

**Kruskal-Wallis Statistic**

This test is an extension of the Wilcoxon rank-sum statistic when there are
more than two treatment groups. It tests \(H_0\): the \(k\) treatment medians are
equivalent vs. \(H_A\): at least two treatment medians are different. The outcome is
ranked over the \(k\) treatments, and then the ranks are summed for each treatment.
The test statistic is:

\[
H = \frac{12}{n(n+1)} \left[ \sum_{i=1}^{k} \frac{S_i^2}{n_i^2} \right] - 3(n+1)
\]  

(2.36)

For small samples, significance is assessed through tabulated values of the test
statistic. If each treatment has more than 5 observations, \(H\) approximately has a
\(\chi^2\) distribution with \(k-1\) degrees of freedom. The Kruskal-Wallis statistic can also
be approximated by an analysis of variance on the ranked data. (Daniel 90)

**Two Sample T-test**

The t-test can be applied to normally distributed data, or to any continuous
data if the sample size is sufficiently large. This test assumes the treatment
groups are independent random samples. The statistic can be computed assuming
equal or unequal group variances. To test the null hypothesis of equal treatment
means \(H_0\): \(\mu_1=\mu_2\) vs. \(H_A\): \(\mu_1 \neq \mu_2\) under the assumption of equal variances,

\[
T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(1/n_1 + 1/n_2)}}
\]  

(2.37)
where $\bar{x}_1$ and $\bar{x}_2$ are the respective treatment group means, and $s_p^2$ is the pooled sample variance $= \frac{[(n_1-1)s_1^2 + (n_2-1)s_2^2]}{(n_1+n_2-2)}$. $T$ follows a $t$ distribution with $(n_1 + n_2 - 2)$ degrees of freedom. When the group variances are unequal, $T' = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$. $T'$ approximately follows a $t$ distribution with degrees of freedom equal to $\text{df} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$. (2.38)

A 100(1-α)% confidence Interval for the difference in treatment means is obtained by

$$\bar{x}_1 - \bar{x}_2 \pm t_{1-(α/2)}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$ (2.39)

for equal variances, and

$$\bar{x}_1 - \bar{x}_2 \pm t_{1-(α/2)}\sqrt{\frac{s_1^2/n_1 + s_2^2/n_2}{n_1 + n_2}}$$ (2.40)

for unequal variances. Equality of variances can be tested with the variance ratio test. The $t$-test is robust against moderate violations of its assumptions and is relatively powerful at detecting true differences (Remington and Schork 85).

2.3.2 Stratified Comparisons

Analysis of Variance

Analysis of Variance (ANOVA) can be used to compare continuous outcomes between two or more treatments while adjusting for a number of categorical stratification variables. Assumptions are that the variance of the response is the same inside each treatment-stratum cell, sample size is about the same in each cell, and the residuals are normally distributed. ANOVA is fairly robust to departures from assumptions when there are 30 or more observations.

ANOVA fits a linear model to the data, defined $Y = X\beta + \varepsilon$, where $Y$ is an $n \times 1$ vector of the response for the respective observations, $X$ is an $n \times p$ matrix of fixed known constants representing treatments and strata, $\beta$ is a $p \times 1$ vector of fixed unknown parameters, and $e$ is an $n \times 1$ vector of unobserved errors, assumed to be distributed $N(0, \sigma^2I)$. $\hat{\beta}$ is solved by $(X'X)^{-1}X'Y$, and predicted values are calculated as $\hat{Y} = X\hat{\beta}$. The residual, or estimated error is $\hat{\varepsilon} = Y - \hat{Y}$. $\hat{\beta}$ is distributed $N[\beta, \sigma^2(X'X)^{-1}]$, and $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-p)$.
In analysis of variance, $X$ is made of ones and zeros. For reference cell coding, the first column contains only ones. The column corresponding to treatment has 0 for the reference (usually placebo) and 1 for the test drug. One of the $h$ levels of the stratification variable is chosen as the reference cell. Each other stratum is represented by a column in the $X$ matrix coded as 1 for that stratum, 0 for all others. It is possible to define the interaction between treatment and stratum by multiplying the elements of the treatment column with the corresponding elements in each stratum column, resulting in $h-1$ additional columns corresponding to the interaction.

The total variation about the average response can be partitioned into sums of squares due to classification and sums of squares due to error, written $\text{SSTOT} = \text{SSREG} + \text{SSE}$, where $\text{SSTOT} = (\bar{Y}-\bar{\bar{Y}})'(\bar{Y}-\bar{\bar{Y}})$, $\text{SSE} = \mathbf{e}'\mathbf{e}$, $\text{SSREG} = (Y-\bar{Y})'(Y-\bar{Y}) - (Y-\bar{Y})'(\bar{Y}-\bar{\bar{Y}})$, and $\bar{Y}$ is an $n \times 1$ vector of the mean response. The ANOVA $F$ statistic is calculated as $\frac{\text{SSREG}/p-1}{\text{SSE}/n-p}$, and is distributed $F_{(p-1,n-p)}$. SSREG can be broken down into components due to each stratification or treatment variable. Those that are not found significant with an $F$ test can sometimes be removed from the model.

Hypotheses of treatment comparisons can be tested by creating contrast matrices of the $\beta$ estimates with a $1 \times p$ matrix $C$. Sums of squares due to hypothesis is defined as $\text{SSH} = (C\hat{\beta})'[C(X'X)^{-1}C]^{-1}C\hat{\beta}$. The null hypothesis that $C\beta=0$ is tested via $\frac{\text{SSH}/\text{rank}(C)}{\text{SSE}/n-p}$, which is distributed $F_{(\text{rank}(C),n-p)}$. The estimate $C\hat{\beta}$ is distributed $N[C\beta, \sigma^2C(X'X)^{-1}C']$.

An advantage of analysis of variance is that it can test homogeneity of treatment comparisons within strata with a test of interaction between treatment and stratum. It is also a very flexible tool which allows for the estimation of stratum-adjusted treatment means (referred to as Least Square means), and their standard errors, obtained by multiplying $\hat{\beta}$ by an appropriate adjustment matrix.
100(1 - \alpha)\% confidence intervals for contrasts of treatment means are obtained by 
\[ C\hat{\beta} \pm z_{1 - \alpha/2} \times SE(C\hat{\beta}) \] (Kleinbaum and Kupper 1978, Rawlings 1988).

For large non-normal samples, a non-parametric analysis can be conducted by ranking the response over the n subjects, and performing an analysis of variance on the ranks. In this scenario, however, meaningful confidence intervals of treatment differences are unavailable.

2.3.3 Covariate Adjusted Comparisons

General Linear Model

Regression, analysis of variance, and analysis of covariance are three special cases of the general linear model (GLM). Linear models allow treatment comparisons of continuous outcomes while adjusting for strata, continuous covariates, or a combination. Assumptions are that each observation is independent, and the model residuals are normally distributed with common variance \( \sigma^2 \). Formulas previously discussed for ANOVA also apply to the GLM, except that the X matrix generally consists of columns of ones and zeros for treatments and strata, and observed values for covariates (Kleinbaum and Kupper 78, Rawlings 88).

If sample sizes are fairly large, non-parametric covariate adjusted treatment comparisons (but not confidence intervals) can be obtained by first ranking the response variable over the n subjects, and then fitting a linear model to the ranks.

2.4 Multivariate Techniques

This section contains a brief discussion of the major statistical tools for simultaneous tests of multiple comparisons. Methods used in this paper are further described in later chapters.
General Linear Multivariate Model

Treatment comparisons can be conducted for multiple continuous response criteria simultaneously in a parametric framework via the general linear multivariate model (GLMM). Resulting test statistics include Hotelling's $T^2$, Roy's largest root, and Pillai's trace. Unfortunately, when testing for significant differences from equivalence, these tests do not distinguish whether treatment differences are in the same or opposite directions across the responses. For this reason, the multivariate linear model receives little attention in this work (Koch, Elashoff, Amara 88).

Generalized Estimating Equations

Generalized estimating equations (GEE) provide a framework for conducting multivariate treatment comparisons when the responses are either dichotomous, ordered categories, or continuous. In spirit, GEE is similar to a mixed linear model (a multivariate linear model allowing fixed and random effects and the modeling of correlation between outcomes) without requiring the specification of the appropriate correlation structure between the multiple outcomes, and with relaxed assumptions about the distribution of the dependent variables.

GEE is a multivariate version of quasi-likelihood, a relaxed-assumption extension of generalized linear models. For quasi-likelihood, only the link function between $E(Y)$ and $X\beta$ and corresponding relationship between $E(Y)$ and $\text{Var}(Y)$ up to a scale parameter $\phi$ must be specified. A full parametric distribution assumption of the dependent variable is not required. A system of score-like equations are solved with iteratively reweighted least squares. Beta estimates are consistent and have asymptotically normal distributions (McCullagh, Nelder 89).

In GEE, $i$ multiple response outcomes for each of $n$ individuals are vertically concatenated to form a strung-out $(ni \times 1)$ vector $Y$. The $(n \times p)$ $X$ matrix
specifies the relationship of covariates with the outcomes, with possibly different
covariates for different outcomes. $\beta$ is a $(p \times 1)$ vector of estimates corresponding
to the columns of $X$. The link function which relates the expectation of $Y (=\mu)$ to
$X\beta$ is generally specified as the identity for continuous responses, and logit for
categorical data. Correspondingly, one can assume $\text{Var}(Y) = \mu/\phi$, $1/\phi$, or other
functions $g(\mu)/\phi$ for continuous variables, and generally $\mu(1-\mu)/\phi$ for binary
responses. The structure of a working correlation matrix between responses from
the same individual may be assumed, or estimated in its entirety. Vector
equivalents of quasi-likelihood score-like functions are solved with iterative re-
weighted least squares. If the link function is correctly specified, consistent
estimates of $\beta$ and $\text{Var}(\beta)$ are obtained, even if the correlation structure is mis-
specified. However, the efficiency of $\text{Var}(\hat{\beta})$ increases when the assumed
correlation matrix is closer to the true correlation (Zeger and Liang, 86).

Advantages of GEE are that it allows for different covariate adjustments for
each outcome, and does not require a correct specification of the correlation
matrix to obtain consistent estimates of $\beta$. GEE was formulated for repeated
measures where multiple outcomes have the same units, but is potentially
applicable to more general situations. Applications of GEE to multivariate
equivalence methods by fitting appropriate $X$ matrices and obtaining confidence
intervals based on $\hat{\beta}$ and $\text{Var}(\hat{\beta})$ are of interest.

Another modeling technique which can be applied to multivariate responses
is survey data regression (Research Triangle Institute 91). While survey
regression fits a standard least squares model to estimate $\beta$, it accounts for the
correlation of outcomes from the same individual in the calculation of $\text{Var}(\hat{\beta})$.

**Weighted Least Squares**

Functions of continuous, ordinal, or categorical outcomes can be modeled
using weighted least squares (WLS), if sample sizes are reasonably large. The covariance matrix between treatments and outcomes is estimated by Taylor series expansion. Measures which are highly correlated have less weight in the analysis than independent measures. Assumptions for WLS are that the subjects are from an independent simple random sample. Stratification can be conducted within the WLS framework, provided that sample sizes within strata are reasonably large. Covariate adjustment can be conducted to a limited degree by specifying the covariate as an additional outcome, and fitting a model that allows no group to group variation for the covariate. Often the assumptions for WLS may be met when ANOVA assumptions are not. Evaluation of homogeneity across responses, estimation of treatment differences, and goodness of fit of the model are all available in WLS, and will be explained in future chapters (Koch et al 1985).

As mentioned in Chapter 1, weighted least squares is implemented in several methods used to evaluate multivariate treatment comparisons when outcomes are expected to be correlated. Carr et al (89) apply WLS to Mann-Whitney type rank measures of association between ordinal outcomes and treatments. They suggest a modification which also allows for stratification over categorical covariates. O'Brien (84) applies WLS to continuous responses to compare the weighted average response between treatment groups. Pocock (87) suggests that WLS can be applied to any asymptotically normal statistic. Lachin (92) performs WLS on mean treatment group differences for each response, Hodges-Lehmann location shift parameters, and log-odds ratios. Pocock (87) considered combining statistics from outcomes with different metrics for the special case of two treatment groups.

Lachin (92) points out that multivariate statistics can be evaluated through frameworks other than weighted least squares. Other strategies have power for different alternative types. A multivariate $\chi^2$ test can be applied to the above statistics. This leads to an omnibus test similar to Hotelling's $T^2$. For
alternatives of stochastic ordering of treatments, a z-score can be applied. These methods do not address homogeneity, although a contrast $\chi^2$ can be applied to test homogeneity. When outcomes are expected to be fairly correlated, Lachin (92) suggests the WLS method is most powerful.

2.5 Conclusions

This chapter has examined some of the common strategies used to compare treatments for studies designed to show a treatment difference. Univariate methods were discussed in detail, and some multivariate strategies were outlined.

Many methods in this chapter can be modified to evaluate treatment equivalence, rather than a difference. Chapter 3 presents a review of the available univariate equivalence techniques. Multivariate methods for equivalence, some of which are extensions of those described above, are examined in Chapters 7 and 8.
Chapter 3
Univariate Statistical Analysis for Bioequivalence and Clinical Equivalence

This chapter is a literature review of the statistical methods developed for testing equivalence. Since most research has been done in the context of bioavailability, this topic receives considerable attention. A few methods have been developed for data obtained from clinical situations, and these are also included. Many of the bioequivalence topics discussed in this chapter have recently been summarized by Rodda (90) and Chow and Liu (92).

3.1 Analysis of the Two Period Cross-Over Study

Nearly all bioequivalence trials with a test and reference formulation follow a two period cross-over design. Pharmacokinetic parameters tend to be variable, but they vary much more between people than they do within a person. A cross-over study where each person acts as his own control has decreased variability for treatment comparisons than would be obtained with a parallel design. In a two period cross-over study, subjects receive both treatments with a suitable washout period between administrations. Subjects are randomly assigned to one of two treatment sequences: (A) test formulation in period one followed by reference formulation in period two, or (B) reference followed by test.

Analysis of data from this design requires consideration of several factors. A carry-over effect can result when the drug in period one affects the response to the drug in period two. This can occur if the washout period is not long enough. Equality of carry-over can be tested by comparing the average responses for the
two sequence groups. A period effect is a systematic difference in the response between the periods. This can arise from study logistics, such as blood assays being analyzed separately for the two periods, environmental differences of the two days, or patient conditioning leading to an altered response on period two (Schuirmann 90).

Analysis of Variance

Most parametric methods for assessing bioequivalence from a two period cross over design require an analysis of variance (ANOVA) to evaluate period and carry-over effects, and to estimate the random error variance. The appropriate ANOVA for two period cross-over studies is described by Grizzle (1965).

An initial ANOVA is fit with factors for sequence (df=1), subject within sequence (df=N-2, where N is the number of subjects), period (df=1), and treatment (df=1). The test of carry-over is conducted using the subject within sequence mean square as the error term. The test statistic (sequence mean square/subject nested within sequence mean square) has an F distribution with 1 and N-2 degrees of freedom. When carry-over effects are significant, the period effect is not interpretable using this model, because the estimate of the period 1 - period 2 difference depends on the sequence effect. Also, the treatment effect can not be evaluated by the ANOVA in this situation. When there is a significant carry-over effect, Grizzle (65) showed that an appropriate estimate for the error variance for treatment is obtained from the sums of squares for subjects within sequence for period one. Treatment comparisons can be based on data collected in period one only. The data structure for period 1 follows a parallel design with two treatment groups, which can be analyzed by a two sample t-test with df=N-2.

If carry-over effects are not significantly different (p > .10 or .15), sequence can be ignored and incorporated with the subject effect. In this situation, the
period effect becomes estimable. The new ANOVA model has factors for subject (df=N-1), period (df=1), and treatment (df=1). The error term has N-2 degrees of freedom, for a total of 2N-1 df. The mean square error is a valid estimate of the error variance for treatments. If the period effect is not significant, it may be removed from the model, and the analysis reduces to a paired t-test (with N-1 df) for the treatment differences within each of the respective subjects.

Most bioequivalence studies do not have carryover effects because they are designed with a reasonable wash-out period, use healthy subjects, and often administer single doses that will have minimal residual effect. Period effects are not common, but tend to appear in about 10% of the studies submitted to the FDA (Schuirmann 90). Since period effects do take place, appropriate equivalence methods should be designed to account for period differences.

If there is missing data, for instance a subject’s response from one of the periods is unavailable, then treatment comparisons are not available for this subject. In such cases, the entire subject is usually deleted from the analysis. To include such subjects, essentially a random effects model is needed.

There is disagreement in the bioequivalence literature about whether the subject effects should be fixed or random. Grieve (91) argues you can treat them as random, and many authors do so. Pabst and Jaeger (90) suggest that fixed is more appropriate. Most analyses in this work assume fixed subject effects.

Tests of Normality

Analysis of variance assumes that the residuals are normally distributed with equal variance. Tests of normality can help identify whether analysis should be conducted on the regular or log scale, or with non-parametric methods instead of ANOVA. Pabst and Jaeger (90) point out that often bioequivalence studies do not have enough sample size to properly test normality. The Shapiro-Wilk test
has little power in small samples (Steyn et al 92). However, least squares means are the best unbiased estimates of population means, and ANOVA is robust to slight non-normality. The authors suggest normality tests in small samples are more helpful for identifying outliers.

Steyn et al (91) recommend a bootstrapping technique to test for normal or symmetrical data with small sample sizes. They received criticism from Pabst and Plettenberg (92), and reply that the method is best used in pilot studies to determine what transformation to use in the final analysis.

Outlier Detection

The FDA Bioequivalence Task Force hearing in September 1986 suggested that more study is needed regarding outliers (Clin Pharm 1988 7:334-335). Outliers may provide critical information regarding the study formulations and may be of vital interest to the investigators. The FDA does not approve of removal of outlying subjects from analysis, especially when the origin of the outlying data is unknown.

Chow and Tse (90) recommend two methods useful for detecting outlying subjects which are based on the difference in likelihood with and without the subject included. Liu and Weng (91) suggest a further test for outlier subjects, based on a two-sample Hotelling $T^2$. They also describe a method for identifying outlying observations using residuals from formulation means. Distributions for the proposed tests are not well defined. This area requires continued research.

Models with More Than Two Formulations

Crossover trials can be conducted with more than two treatments, although this is not typical in bioequivalence studies. Multi-treatment cross-over studies can follow a variety of designs. One is an $f$ period cross-over, where $f$ is the
number of formulations. The effect of carry-over of treatments is balanced by having each formulation precede each other formulation an equal number of times (for \( f \) even, \( \# \) sequences = \( f \); for \( f \) odd, \( \# \) sequences = \( 2f \)). Such multi-period trials can be analyzed with appropriate ANOVA methods, but for equivalence purposes it may also be practical to conduct several 2-period analyses on selected treatment pairs. If many treatments are being compared, it may be impractical to give all drugs to all subjects. In such cases, a balanced incomplete blocks design can be used, in which each subject receives two or more of the treatments in a randomized block fashion, although it is a good idea to give the reference therapy to all subjects (Schuirmann 90).

3.2 Assessments of Equivalence Based on ANOVA

3.2.1 The Power Approach

Traditional clinical trials are based on disproving the null hypothesis of no treatment differences, in favor of the alternative that one therapy is better. Historically, bioequivalence trials were analyzed in this manner, and it was common practice to assume that if no significant difference was found, then equivalence could be accepted. However, the statistical null hypothesis is never proven; rather, evidence is weighed against it and it is disproved with a certain probability of error, the p-value. With this approach, the power for being able to detect a difference if one truly exists becomes very important - smaller studies with less power may not be able to detect moderate differences, while large studies with more power may identify a small, clinically unimportant difference. When trying to show equivalence, the traditional roles of false positive error rate (\( \alpha \)) and false negative error rate (\( \beta \)) are basically reversed. The traditional null hypothesis is not valid, and must be modified (Blackwelder 82).

The FDA initially responded to this problem by requiring post-hoc power
analyses in addition to a test of the traditional null hypothesis that \((\mu_t - \mu_r) = 0\). Power to detect a test-reference treatment difference of 20% was required to be at least 80%. Twenty percent was chosen as the minimum difference that would be considered therapeutically important. This procedure, labeled the power approach by Schuirmann (87), was implemented as follows: If the ANOVA F-test for treatment is significant, then equivalence is rejected. If the F-test is not significant, then examine the post-hoc power, using the ANOVA root mean square error (denoted s) to estimate \(\sigma\). The power is calculated as \(1 - \tilde{\beta} = \Phi(z)\), where \(\Phi\) is the cumulative normal distribution and
\[
z = \frac{\Delta \sqrt{N/2}}{s} - z_{0.075},
\] (3.1)
where \(\Delta\) is the desired detectable treatment difference. Westlake (88) uses \(t\) instead of \(z\), which is appropriate for small samples. The FDA's required difference of 20% is obtained by conducting the ANOVA on the log scale, and setting \(\Delta\) equal to \(\ln(1.20)\) or \(\ln(0.80)\). Since \(\ln(0.80)\) is negative, the absolute value must be used, which equals \(\ln(1.25)\). The FDA apparently did not specify which direction required power of 0.80; the most conservative strategy would be to require 0.80 power for \(\ln(1.2)\), since power will always be greater for \(\ln(1.25)\). Setting the detectable difference at 20% is not directly applicable to parameters which follow an additive model and are not log transformed. For such variables, an absolute difference on the original scale must be specified.

If the F-test null hypothesis is not rejected, and power is sufficient, then equivalence is concluded. However, if an overpowered study found a significant difference that was therapeutically unimportant, the result might be considered clinically non-significant, and equivalence might be accepted (Westlake 88).

Since the treatment F-test is identical to a t-test, non-significance of the F-test at the \(p=0.05\) level can alternatively be assessed by determining if the 95% confidence interval for \(\Delta = (\mu_t - \mu_r)\), the test-reference mean difference, contains
zero. The confidence interval for a balanced study has the form:

\[
\left[ (\bar{x}_t - \bar{x}_r) \pm s \sqrt{2/N} \times t_{0.975(\nu)} \right]
\]

(3.2)

where \(\bar{x}_t\) and \(\bar{x}_r\) are the test and reference means, s is the ANOVA root mean square error, N is the number of subjects, and \(\nu\) is the degrees of freedom from the ANOVA error term, generally N-2. For multiplicative parameters analyzed on the log scale, the limits of the CI are exponentiated to form the confidence interval for \(\theta\), the ratio of the test and reference means. The null hypothesis is not rejected if the exponentiated CI contains the value 1.

The exponentiated difference of means of logs is actually equivalent to a ratio of medians, or geometric means (Schuirmann 87). For an ANOVA of log(x), a normal distribution is assumed on the log scale. The mean and median of a normal distribution are identical. However, when the distribution is transformed back to the standard scale, the location of the median is maintained but the mean is shifted towards the tail of the skewed distribution. The exponentiated confidence interval is therefore interpretable as an interval for the ratio of medians, not means. It turns out that the exponentiated mean of logs is mechanically equivalent to a geometric mean, defined \(\prod_{i=1}^{n} x_n^{1/n}\) (Remington and Schork 1985). The terms median and geometric mean can be used interchangeably in this setting, although this work generally refers to the geometric mean.

If the study is not balanced (i.e., there are \(n_A\) subjects receiving sequence A and \(n_B\) subjects receiving sequence B) then the best unbiased estimator of the treatment difference is actually

\[
\left( \frac{\bar{x}_{tA} + \bar{x}_{tB}}{2} \right) - \left( \frac{\bar{x}_{rA} + \bar{x}_{rB}}{2} \right)
\]

(3.3)

where \(\bar{x}_{tA}\) is the mean response for test formulation of sequence group A, etc. The standard error of the mean difference becomes \(s \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}\), where s is the root mean square error from the ANOVA model (Schuirmann 87). A
balanced study is preferable because it is generally less sensitive to moderate violations of assumptions (Schuirmann 1990).

The best unbiased estimator of the difference of treatment means and the appropriate standard error are obtained from the $\beta$ estimate for the treatment effect in the ANOVA model and its standard error $\text{SE}(\hat{\beta})$, regardless of whether the study is balanced or unbalanced, provided test is coded 1 and reference is coded zero. Equation (3.2) can thus be expressed more generally for additive parameters as:

$$[\hat{\beta} \pm \text{SE}(\hat{\beta}) \times t_{0.975(\nu)}],$$

or for multiplicative parameters analyzed on the log scale as:

$$\exp[\hat{\beta} \pm \text{SE}(\hat{\beta}) \times t_{0.975(\nu)}].$$

Although still in occasional use, it is now generally accepted that a test of the traditional null hypothesis is not appropriate for bioequivalence, since the hypothesis does not specifically test equivalence. Several more appropriate methods have been suggested in recent years.

Following is an overview of these methods which are based on the ANOVA model. They all require the assumptions that the subjects are an independent random sample and the cross-over ANOVA residuals are normally distributed with constant variance, unless the sample sizes are sufficiently large for the relaxing of the normality assumption. While they were originally proposed using the standard 2-period cross-over design, they may be applied to any general linear model appropriate for the study design, including parallel designs and analysis of covariance. The model estimate of the treatment difference, its standard error, and the error degrees of freedom are all that is needed to apply the techniques. For parallel studies, these methods could also be extended to designs with more than two treatments.

3.2.2 The Confidence Interval Approach
Westlake (1972) and Metzler (1974) suggested a new strategy basing bioequivalence directly on confidence intervals. The method is as follows: Construct a $100(1-\alpha)\%$ confidence interval where $\alpha=0.05$ for $\mu_t-\mu_r$ (additive parameters) or for the geometric mean treatment ratio (multiplicative parameters), as described in equations (3.4) and (3.5). Conclude bioequivalence if the CI is completely contained within a pre-specified range defining clinical equivalence. As described earlier, the FDA initially accepted equivalence for multiplicative parameters if the CI for the ratio of treatment geometric means was within $\pm 20\%$, i.e., $0.80 < \mu_t/\mu_r < 1.20$, but has since changed its requirements to $0.80 - 1.25$, which is symmetrical on the log scale.

For parameters following an additive model this method requires choosing a value of the test-reference difference that is therapeutically unimportant, which may have to be made on a study by study basis. It is no longer based, however, on the inappropriate test of the null hypothesis that the difference is equal to zero. The null hypothesis corresponding to this strategy is discussed in the next section.

**Symmetrical Confidence Intervals**

Westlake (1972, 1976) developed a modified confidence interval which is symmetric about equivalence. The standard $100(1-\alpha)$ CI formula for $(\mu_t - \mu_r)$ is $[\bar{\beta} \pm t_{1-\alpha/2}SE(\bar{\beta})]$. Since $t_{\alpha/2} = -(t_{1-\alpha/2})$, the interval is symmetric about the best unbiased estimate of the difference ($\bar{x}_t - \bar{x}_r$). Westlake proposed a modification to the choice of $t_{\alpha/2}$ and $t_{1-\alpha/2}$ which would lead to CIs which are symmetric about zero, or symmetric about 1.0 after exponentiating log CIs. The modified $t$ distribution limits, $k_1$ and $k_2$, are chosen such that

(1) $k_1 + k_2 = 2(\bar{\beta})/SE(\bar{\beta})$, and

(2) 95% of the $t$ distribution lies between $k_1$ and $k_2$. 
The modified limits can be found by trial and error from a detailed tabulation of
the t distribution. Westlake demonstrated that with this method, resulting CIs
are symmetric about zero (exponentiated log CIs are symmetrical about 1.0), and
the confidence is always greater than the nominal 95%. However, the confidence
limits are longer than the standard CI limits. As an example, a standard 95% CI
with limits (.727, 1.15) has limits (.752, 1.248) under his method. Note that if the
equivalence criteria were ±25%, then equivalence would be concluded using the
symmetric method, but not using the standard method. Spriet and Beiler (1978)
provide tables of $k_1$ and $k_2$ for $N-2$ degrees of freedom.

Mandallaz and Mau (81) derived a Bayesian interpretation to Westlake's
symmetrical CIs. A decision in favor of bioequivalence is equivalent to a Bayes
posterior probability of $> 0.95$ that $\mu_t/\mu_r$ is within the symmetrical bioequivalence
range (generally 0.8-1.20), assuming an uninformative vague prior distribution for
the treatment means and variation (Steinijans and Hauschke 90).

Kirkwood (1981) and many others (listed in Steinijans and Diletti 83),
criticized Westlake's symmetrical method on several grounds. Kirkwood argued
that the CI should remain centered around the sample means, so that actual
differences between treatments in the study are plainly visible. Mantel (77)
pointed out that if the designation of test and reference is arbitrary, then
switching the two by taking the reciprocal of the symmetrical confidence limits
yields limits which are no longer symmetrical. Symmetrical CIs are apparently
not appropriate for such cases. Mantel (77) also mentioned that if the test and
reference formulation are very different from one another, it can be impossible to
obtain an interval which is symmetric about zero (or 1.0). Westlake (81) replied
that for most pharmaceutical bioequivalence studies, the formulations are usually
within at most 30% of each other, so this problem does not arise. However, this
may not be the case in other clinical trial applications.
A common complaint of Westlake's interval is that it becomes easier to show equivalence. Since the confidence area is moved around, when the treatment difference or the error variance is large, all 5% can be at one tail of the traditional confidence interval, essentially making a one sided CI.

90% Confidence Intervals

In Westlake's 1981 rebuttal to Kirkwood, he suggested the standard confidence interval could actually be broadened to 100(1-2α)% or 90% with a standard α of 0.05. His justification is as follows: for a traditional test of efficacy of a new drug versus placebo, the FDA considers a test significant if p < 0.05. Since a drug less efficacious than placebo would never be submitted, Westlake argues this corresponds to a one-sided test of size α=0.05. For bioequivalence, the test drug could be better or worse than the reference, so it is truly a two-sided situation. To maintain consistency with FDA guidelines for efficacy testing, each side of the CI should have α=0.05, corresponding to an overall CI of 90%.

Westlake's symmetrical 95% CI is a step between the conventional 95% and 90% CIs. The standard 90% CI is more powerful than the symmetric CI because it has a greater chance of concluding equivalence (Hauck and Anderson 84). Westlake (88) showed for a cross-over study of a test and reference treatment that the sample size to detect an equivalence of Δ when no difference applies using a 100(1 - 2α)% CI is

\[ n = \frac{2(z_{1-\alpha} + z_{1-\beta/2})^2 s^2}{\Delta^2} \]  

and for a traditional study to detect a difference of Δ with power (1 - β) is

\[ n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 s^2}{\Delta^2} \]  

This clearly shows both the reversed roles of α and β, and the appropriateness of a 100(1-2α)% CI instead of 100(1-α)% CI. Westlake uses a t distribution instead of the z, which is appropriate for small studies. Sample size and power are discussed in
more detail in Section 3.9. Blackwelder (82) also discusses the role reversal of alpha and beta.

3.2.3 The Null Hypothesis of Non-equivalence

Two One-sided Tests Procedure

It has been pointed out that the null hypothesis of no treatment difference is not appropriate for a test of equivalence. It is intuitive that a strategy for assessing equivalence is to formulate an appropriate null hypothesis.

Schuirmann (87) suggests what he labels the two one-sided tests procedure. Two null and alternate hypotheses are posed as follows:

\[ H_{01}: \mu_t - \mu_r \leq \Delta_1 \quad H_{A1}: \mu_t - \mu_r > \Delta_1 \]
\[ H_{02}: \mu_t - \mu_r \geq \Delta_2 \quad H_{A2}: \mu_t - \mu_r < \Delta_2 \]

where \( \Delta_1 < 0 < \Delta_2 \). Equivalence is concluded if both \( H_{01} \) and \( H_{02} \) are rejected at significance level \( \alpha = 0.05 \). Each of these tests is conducted as a one-sided t-test with \( df = N-2 \), which for a balanced study are:

\[
t_1 = \frac{(\bar{x}_t - \bar{x}_r) - \Delta_1}{s/\sqrt{2/N}} \geq t_{1-\alpha(\nu)} \quad t_2 = \frac{\Delta_2 - (\bar{x}_t - \bar{x}_r)}{s/\sqrt{2/N}} \geq t_{1-\alpha(\nu)},
\]

or more generally, based on the ANOVA estimate of the treatment difference and its standard error,

\[
t_1 = (\hat{\beta} - \Delta_1)/SE(\hat{\beta}) \geq t_{1-\alpha(\nu)} \quad t_2 = (\Delta_2 - \hat{\beta})/SE(\hat{\beta}) \geq t_{1-\alpha(\nu)}.
\]

The p-value is the maximum of the p-values from the two tests. Generally for multiplicative pharmacokinetic parameters analyzed on the log scale, \( \Delta_1 = \log(0.80) \), and \( \Delta_2 = \log(1.25) \). For additive parameters, appropriate values of \( \Delta_1 \) and \( \Delta_2 \) must be specified. This hypothesis test is exactly equivalent to the 90% CI proposed by Westlake. The 90% CI includes all values of \( \Delta_1 \) and \( \Delta_2 \) (or \( \theta_1 = e^{\Delta_1} \) and \( \theta_2 = e^{\Delta_2} \) for ratios) which if tested would not be rejected by one of the
one-sided tests \((p > 0.05)\). If the CI is completely contained within \((\Delta_1, \Delta_2)\), then Schuirmann's p-value is \(< 0.05\).

The two one-sided tests procedure has become the standard assessment of bioequivalence accepted by the FDA (K. Phillips 90 A). Schuirmann (87) showed his method leads to a more appropriately shaped rejection region than the power approach.

The Interval Hypothesis

Anderson and Hauck (83, Hauck and Anderson 84) suggested formulating a single null hypothesis, so that rejection of the null hypothesis leads directly to the acceptance of bioequivalence. Schuirmann (87) labels this the interval hypothesis. It is stated as follows:

\[
H_0: \mu_t - \mu_r \leq \Delta_1 \text{ OR } \mu_t - \mu_r \geq \Delta_2 \quad H_A: \Delta_1 < \mu_t - \mu_r < \Delta_2
\]

where \(\Delta_1 < \Delta_2\) are symmetrical about the null value, generally log(0.8) and log(1.25). The test statistic for a balanced design is:

\[
T = \frac{\bar{x}_t - \bar{x}_r - \frac{1}{2}(\Delta_1 + \Delta_2)}{s\sqrt{2/N}}
\]

which is more generally (for an unbalanced design) expressed as:

\[
T = \frac{\hat{\beta} - \frac{1}{2}(\Delta_1 + \Delta_2)}{SE(\hat{\beta})}
\]

\(T\) is a measure of the distance of the mean difference from the center of the equivalence interval. \(T\) has a noncentral t distribution with noncentrality parameter \(\delta = \frac{(\mu_t - \mu_r) - \frac{1}{2}(\Delta_1 + \Delta_2)}{\sigma\sqrt{2/N}}\) and df=\(\nu\), the error degrees of freedom from the ANOVA model. Generally the noncentrality parameter is unknown. Anderson and Hauck (83) showed the best approximation is an approximate t distribution. The p-value is obtained by:

\[
p = \left[ t_{\nu}[|T| - \hat{\delta}] - t_{\nu}[|T| - \hat{\delta}] \right], \quad \text{where } \hat{\delta} = \frac{\frac{1}{2}(\Delta_2 - \Delta_1)}{SE(\hat{\beta})}
\]

and \(t_{\nu}\) is the probability corresponding to the cumulative t distribution with \(\nu\)
degrees of freedom. It can be shown that this p-value is the absolute value of the
difference of the two one-sided p-values from Schuirmann’s procedure, where
Schuirmann takes the maximum p-value as the overall p-value. This test is less
conservative than the two 1-sided tests procedure, but suffers an anomaly. For
very large variation, the test can conclude equivalence even when the mean
difference (or ratio) is outside the acceptance interval (Schuirmann 87). It can
also be anti-conservative (Steinijans and Hauschke 90). Rocke (84, 85) also
proposed a special case of the Hauck and Anderson procedure concurrently.

3.2.4 Confidence Intervals for Ratios without Log-Transforming

Sometimes it is of interest to evaluate ratios without log-transforming. For
instance, if the data contains values less than one, the log transform may behave
awkwardly. Other times the data may be normally distributed, but taking the log
leads to worrisome non-normal residuals. For these situations, methods have been
developed which allow for the calculation of confidence intervals for ratios of
treatment means either without an ANOVA or via the ANOVA analysis of a 2
period cross-over study on the original scale.

Locke (84) describes an approximate method for forming confidence intervals
of ratios without log transforming, which consists of calculating the CI of the
difference of means and then dividing the endpoints by the estimate of the
reference mean and adding 1.0. This method is undesirable, because it does not
account for the error associated with the reference mean in the denominator.

Fieller (1954) developed a formula for the CI of the ratio of two means when
the estimators of the means come from a bivariate normal distribution and there
is an independent estimate of the covariance matrix. Fieller shows that for a
balanced study:

$$T^2 = \frac{(\bar{x}_t - \theta \bar{x}_r)^2}{(\hat{\sigma}_t^2 - 2\theta \hat{\sigma}_t \theta \hat{\sigma}_r + \theta^2 \hat{\sigma}_r^2)}.$$  (3.13)
T has a t distribution with \( \nu \) df, usually \( \text{N-2} \) for a two period cross-over study. This formula can be used to test \( H_0: \theta = \theta_0 \) vs \( H_A: \theta \neq \theta_0 \). Alternatively, a 100\((1 - \alpha)\)% CI for \( \theta \) is obtained by finding all values of \( \theta \) such that \( |T| \leq t_{1-\alpha/2, \nu} \). This is done by solving the roots of the equation \( T^2 - (t_{1-\alpha/2, \nu})^2 = 0 \) as a function of \( \theta \) using the quadratic formula. The limits of the confidence interval are
\[
\left[ \frac{b \pm \sqrt{b^2 - 4ac}}{2a} \right],
\]
where
\[
a = \bar{x}_r^2 - \bar{\sigma}^2_r (t_{1-\alpha/2, \nu})^2,
\]
\[
b = 2\bar{\sigma}_r(t_{1-\alpha/2, \nu})^2 - 2\bar{\sigma}_r \bar{x}_r
\]
\[
c = \bar{x}_r^2 - \bar{\sigma}^2_r(t_{1-\alpha/2, \nu})^2.
\]
(3.14)

There is a possibility that the roots are imaginary, or the confidence set is not internal to the roots, but when the assumptions are not badly violated or the number of subjects is reasonably large, the confidence set will almost certainly be an interval containing all positive values with the endpoints of the interval being the roots of the quadratic equation (Locke 84). If the assumptions are met, then the CI is exact, otherwise it is an approximation.

Hauck and Anderson (92) point out that there are two sets of standard errors that can be used in Fieller's formula when applying it to a two period cross-over study. The ANOVA approach (Mandallaz and Mau 81, Fleuhler et al 81) assumes the within-subject variances are equal for the two treatments, and that the correlation between a subject's two observations in a cross-over design is fully explained by the fixed subject effect in the model. This implies \( \sigma_{tr} = 0 \), and \( \bar{\sigma}^2_r = \sigma^2 \) is estimated by \( s^2/N \), where \( s^2 \) is the mean squared error from the ANOVA model with \( \nu \) degrees of freedom. Fieller's formula is thus simplified to:
\[
T^2 = \frac{(\bar{x}_A - \theta \bar{x}_r)^2}{(s^2/N)(1 + \theta^2)}.
\]
(3.15)

Note that with an unbalanced design, unbiased estimates of the treatment means must be obtained via
\[
\left[ \frac{1}{2}(\bar{x}_{tA} + \bar{x}_{tB}) - \theta \bar{x}_{rA} + \bar{x}_{rB} \right]^2,
\]
(3.16)
and \( s^2/N \) becomes \( \frac{s^2}{4}(1/n_A + 1/n_B) \), which equals \( \text{SE}(\bar{\beta})^2/2 \). Mandallaz and Mau (81) compared this technique to Westlake's symmetrical CI, and found Fieller's
formula resulted in a somewhat tighter confidence interval.

Locke (84) suggests a different choice of standard errors which makes neither of the above variance assumptions, and can thus be applied in a more general situation. A bivariate normal distribution is still assumed. The period effect is averaged over, by calculating means and standard deviations for each treatment sequence separately, and then averaging over sequence for each treatment.

Locke proposes using the sample estimates of the variances and covariance. For the unbalanced case:

\[
\hat{\sigma}_i^2 = \frac{1}{n_A + n_B - 2} \sum_{i=1}^{\bar{n}_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{ii})^2
\]

\[
\hat{\sigma}_r^2 = \frac{1}{n_A + n_B - 2} \sum_{i=1}^{\bar{n}_i} \sum_{j=1}^{n_i} (x_{rij} - \bar{x}_{rr})^2
\]

\[
\hat{\sigma}_{tr} = \frac{1}{n_A + n_B - 2} \sum_{i=1}^{\bar{n}_i} \sum_{j=1}^{n_i} (x_{rij} - \bar{x}_{ri})(x_{rij} - \bar{x}_{ri})
\]

and

\[
T^2 = \frac{[\frac{1}{2}(|\bar{x}_{iA} + \bar{x}_{iB}) \cdot \theta_{1/2}^1(|\bar{x}_{rA} + \bar{x}_{rB})]^2}{(\hat{\sigma}_i^2 - 2\hat{\sigma}_{tr} + \theta^2\hat{\sigma}_r^2) \times (1/n_A + 1/n_B)/4}
\] (3.17)

Fieller's method can also be used to evaluate treatment equivalence from parallel designs. It is frequently used in calculating confidence intervals for LD50 and relative potency estimates from bioassay (dose-response) experiments (Yeh 1981).

3.3 Non-parametric Confidence Intervals

The vast majority of bioequivalence trials follow cross-over designs (Steinijans and Diletti 83), but sometimes regardless of whether a parameter should be analyzed on the original or the log scale, assumptions required for ANOVA are not met. The assumptions of normally distributed residuals with constant variance and additivity of the effects of subjects, treatments, and periods may be sufficiently violated so that analysis of variance is inappropriate. For
bioequivalence, this is true for $T_{max}$, which is neither normal nor log-normal. It becomes of interest to apply non-parametric methods with less stringent assumptions. One advantage of these methods is that they are based on individual differences or ratios instead of ratios of group means. They tend to be conservative, however, due to the use of discrete distributions (Steinijans and Hauschke 90).

3.3.1 Based on Wilcoxon Signed-Rank Test

Steinijans and Diletti (83) described a method for obtaining a CI for the median difference (or ratio of medians) based on the Wilcoxon signed-rank test. Assumptions are that there is no period effect, the subjects are an independent random sample, and the errors are from a continuous distribution symmetric about zero but not necessarily identically distributed (Daniel 1990). The signed-rank test is conducted on the test - reference difference for each subject, and tests $H_0$: median $\nu_t - \nu_r = 0$ vs $H_A$: $\nu_t - \nu_r \neq 0$. The CI procedure, described in Hollander and Wolfe (73), and first developed by Tukey, forms a test-based interval of all differences of medians that would not be rejected by the test (generally $p > 0.05$). Since the Wilcoxon distribution is discrete, the coverage probability is always greater than $100(1-\alpha)\%$, but can be refined by interpolation. And since Wilcoxon is a rank-based test, confidence intervals for ratios of medians can be obtained either by exponentiating the confidence limits for the difference of logs, or equivalently by forming a CI directly on the individual subject ratios.

Meineke (87) published a BASIC program that produces CIs using this method for a ratio of medians from a balanced two-period cross-over study.

3.3.2 Based on Pitman’s Permutation Test

Steinijans and Diletti (83) suggested another method with similar
assumptions as Wilcoxon signed-rank but the errors can be from discrete
distributions (i.e., measurements can be discrete). The procedure is based on
Pitman's permutation test using all possible permutations of averages of observed
treatment differences. The coverage probability can get closer to the appropriate
100(1-α)% than the Wilcoxon CI can, but becomes too computer-intensive for
sample sizes above 12 subjects. Steinijans and Diletti (1983) give details for
making the CIs using this method.

3.3.3 Based on Wilcoxon-Mann-Whitney Test

A major flaw of the signed-rank and Pitman methods is the assumption of no
period effect. Hauschke, Steinijans, and Diletti (90) refined the method to
account for period effects. Unequal period effects are rare according to the
authors, but if they occur, the signed rank method may lead to a biased point
estimate and unnecessarily wide confidence intervals. They suggest using their
new method, which they call the distribution-free counterpart to Westlake's 90%
Confidence Interval, in place of the signed-rank test.

Assumptions are that the subjects are an independent random sample, the
measurement is continuous, and the variation of the period 1 - period 2 difference
is the same for sequence groups A and B. The method does not assume normal or
log-normal distributions (Daniel 1990).

The method is conducted as follows: for sequence A, calculate the test-
reference difference, and for sequence B, calculate the reference-test difference.
Then perform the Wilcoxon-Mann-Whitney test (Hollander and Wolfe 1973) on
these two groups. The exact test is conducted by forming the \( n_A \times n_B \) differences
of pairs of observations, one from each sequence group, and comparing them to
the Wilcoxon distribution. In this case, divide each pairwise difference by two, so
the Wilcoxon test will estimate \( \frac{1}{2}[(\nu_t - \nu_r) - (\nu_r - \nu_t)] = \nu_t - \nu_r \). A 100(1-α)% CI for \( \nu_t \).
\( \nu_r \) is obtained from the two pairwise differences which correspond to the largest tail probabilities less than or equal to \( \alpha/2 \) in the discrete Wilcoxon distribution. Equivalence is concluded if the CI is contained entirely within the pre-specified equivalence criteria. For ratios, this procedure can either be calculated using the log transformation and then exponentiating the resulting confidence limits, or by calculating ratios instead of differences in the computation.

Hauschke et al (90) provide a table for confidence limits for sample sizes of 8 to 24 subjects, and Daniel (90) gives limits for sample sizes up to 20 subjects per sequence. For larger sample sizes, parametric ANOVA methods may be reasonable. The standard Wilcoxon-Mann-Whitney test of \( H_0: \nu = \nu_r \) vs \( H_A: \nu \neq \nu_r \) can be approximated by applying parametric methods to the ranks. This has limited use for equivalence, since confidence intervals formed on the ranks lose all information of the original scale, and can not be compared to an equivalence criterion.

3.3.4 Robust Procedures

Yuh (1990) suggests two non-parametric estimates for the test vs. reference ratio, based on individual subject ratios instead of ratios of means. They both do not take period effects into account. The first method is based on order statistics: Calculate the test/reference ratio for each subject, and order them from smallest to largest. The probability that the median is between \( \theta_k \) and \( \theta_{n-k+1} \) is calculated as \( 1 - \left( \frac{1}{2} \right)^{n-1} \sum_{i=0}^{k-1} (\frac{n}{2})_i \). Find the symmetrical set of order statistics which has probability \( \geq 0.90 \). One problem with this approach is that it is very discrete. Confidence limits are limited to the observed values.

Yuh's second suggestion is the generalized maximum likelihood M-estimate with a t-distribution weight. The solution to \( \theta \) requires numerical iterative methods, but can be conducted with a SAS Macro written by the authors (Yuh et
Previous work by the author found the M-estimate to be more efficient than other traditional and robust estimates. An advantage of this estimation is that it down-weights extreme observations, which are likely in small samples and skewed distributions of ratios of bioequivalence measures. However, the authors warn that outliers or bad data points can have a substantial influence on the final estimate whenever individual ratios are analyzed. The authors suggest applying the traditional analyses based on ANOVA, as well as their M-estimate approach. If they agree, report the ANOVA method, if they disagree, examine the data for the cause, and if the data can not be invalidated, report their robust estimate.

3.4 Example

To demonstrate the techniques described in the previous sections, their results are examined for an example dataset. The data is from study #1 of the meta analysis collection of bioequivalence trials of a generic and reference drug formulation for the treatment of a psychiatric disorder, described in Section 1.4.1. Blood levels of each formulation were measured on 21 people in a cross-over design, and AUC, $C_{\text{max}}$, and $T_{\text{max}}$ were calculated. Analysis of variance for AUC and $C_{\text{max}}$ found carry-over and period effects to be non-significant. Confidence Intervals and/or p-values are presented in Tables 3.1 and 3.2 for each method discussed, except Yuh's robust M-estimate, which was not examined due to computational inconvenience. Except for the power approach, equivalence is accepted if the CI is completely contained within the range 0.80 - 1.25 or the p-value is less than 0.05.
### Table 3.1
Equivalence Results for Area Under the Curve (AUC)
Comparing Generic and Reference Formulations of a Psychotropic Drug, Study #1

<table>
<thead>
<tr>
<th>Method</th>
<th>CI for Ratio of Medians</th>
<th>P-Value</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Approach 95% CI, power†</td>
<td>(0.808, 1.118)</td>
<td>power=0.65</td>
<td>Reject</td>
</tr>
<tr>
<td>Westlake Symmetrical 95% CI</td>
<td>(0.829, 1.206)</td>
<td>-</td>
<td>Accept</td>
</tr>
<tr>
<td>90% CI, Schuirmann 2 1-sided tests</td>
<td>(0.831, 1.087)</td>
<td>0.0194</td>
<td>Accept</td>
</tr>
<tr>
<td>Hauck and Anderson</td>
<td>-</td>
<td>0.0183</td>
<td>Accept</td>
</tr>
<tr>
<td>ANOVA Fieller’s Formula 90% CI</td>
<td>(0.921, 1.126)</td>
<td>-</td>
<td>Accept</td>
</tr>
<tr>
<td>Locke Fieller’s Formula 90% CI</td>
<td>(0.913, 1.126)</td>
<td>-</td>
<td>Accept</td>
</tr>
<tr>
<td>Wilcoxon-Mann-Whitney 90% CI</td>
<td>(0.849, 1.130)</td>
<td>-</td>
<td>Accept</td>
</tr>
</tbody>
</table>

†Accept if CI includes 1 and power > 0.80

### Table 3.2
Equivalence Results for Maximum Concentration (C\text{max})
Comparing Generic and Reference Formulations of a Psychotropic Drug, Study #1

<table>
<thead>
<tr>
<th>Method</th>
<th>CI for Ratio of Medians</th>
<th>P-Value</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Approach 95% CI, power†</td>
<td>(0.944, 1.248)</td>
<td>power=0.78</td>
<td>Reject</td>
</tr>
<tr>
<td>Westlake Symmetrical 95% CI</td>
<td>(0.821, 1.218)</td>
<td>-</td>
<td>Accept</td>
</tr>
<tr>
<td>90% CI, Schuirmann 2 1-sided tests</td>
<td>(0.967, 1.218)</td>
<td>0.0236</td>
<td>Accept</td>
</tr>
<tr>
<td>Hauck and Anderson</td>
<td>-</td>
<td>0.0235</td>
<td>Accept</td>
</tr>
<tr>
<td>ANOVA Fieller’s Formula 90% CI</td>
<td>(1.003, 1.328)</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td>Locke Fieller’s Formula 90% CI</td>
<td>(1.004, 1.285)</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td>Wilcoxon-Mann-Whitney 90% CI</td>
<td>(0.984, 1.220)</td>
<td>-</td>
<td>Accept</td>
</tr>
</tbody>
</table>

†Accept if CI includes 1 and power > 0.80
From examining Tables 3.1 and 3.2, it can be seen that for both AUC and C_{max}, the power approach does not have enough power to conclude equivalence. Hauck and Anderson's test agrees closely with Schuirmann's method for these data, because variation is not extreme. Westlake's symmetrical CI is approximate due to the iterative program for finding k_1 and k_2. It behaves similarly to the 90% confidence interval, but has wider intervals. The Wilcoxon non-parametric CI is also fairly similar to the parametric CI. The two Fieller's methods agree with each other fairly well, but tend to have higher limits than other methods, causing rejection of equivalence for C_{max}. Based on the majority of tests concluding acceptance, the bioequivalence of AUC and C_{max} can be concluded for these formulations.

Equivalence for the time to maximum concentration, T_{max}, was examined only for the exact Wilcoxon-Mann-Whitney method. Since evaluating time in terms of ratios is not deemed appropriate by the literature, a 90% CI for the T_r - T_s difference was found to be (-30 min, 0 min). If a 30 minute difference was deemed reasonable by the investigators, then equivalence could be accepted.

3.5 Categorical Data Methods

3.5.1 Parallel Designs to Test Δ from a Dichotomous Response

Often showing equivalence is of interest for applications besides drug bioavailability. Clinical trials with a parallel design and dichotomous response are among the applicable designs to the broader question of equivalence. Blackwelder (82) suggests that for clinical trials, often equivalence can be considered a one-sided test that the percent responding favorably in the test group (π_t) is no worse than Δ% from the percent responding favorably in the reference group (π_r). Depending on the definition of a favorable response, Δ could be positive or
negative. Dunnett and Gent (77) describe an example. They discuss a health care trial to compare the quality of patient care under two practices: physician care, or nurse-practitioner care. The data is in the form of a 2 x 2 table of treatment group vs. adequate or not adequate care.

**Based on Pearson Chi-square**

Dunnett and Gent (77) developed a method to test $H_0$: $\pi_t - \pi_r = \Delta$ from a parallel groups design when the outcome is dichotomous. The method assumes subjects are a random sample from a population, and so subjects in the treatment groups form two binomial distributions. Notation for the response is displayed in Figure 3.1.

**Figure 3.1 Response Notation for Dunnett and Gent’s Method**

care

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Adequate</th>
<th>Not Adequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>Ref.</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

The hypothesis that $\delta = \pi_t - \pi_r$ is no less than $\Delta$, stated $H_0$: $\delta \leq \Delta$ vs. $H_A$: $\delta > \Delta$, is tested with a Pearson chi-square statistic using a modified expected count for cell $n_{ij}$ under the null hypothesis. For a traditional chi-square test, $H_0$ is $\pi_t = \pi_r$, and $E(n_{ij}|H_0)$ is $m_{ij} = (n_i)(n_j)/n$. Dunnett and Gent's modification is:

$$\hat{\pi}_t = \frac{n_1 + n_2\Delta}{n}, \quad \hat{\pi}_r = \frac{n_1 - n_1\Delta}{n},$$

(3.18)

and modified expected counts $m_{ij}^*$ equal (clockwise from top left in Figure 3.1) $n_1\hat{\pi}_t$, $n_1(1-\hat{\pi}_t)$, $n_2(1-\hat{\pi}_r)$, and $n_2\hat{\pi}_r$ (see Figure 3.2). Instead of expected counts based on equal distributions under $H_0$, this test assigns larger expected counts to the cells supportive of the null hypothesis. The example described by the authors
uses $\Delta = -0.1$ as the equivalence criterion. Rodary et al (1989) published an example of an upper-tail test. Blackwelder suggests using a $\Delta$ of $-0.2$. As mentioned in Chapter 1, the choice of $\Delta$ should depend on the specific situation.

**Figure 3.2 Expected Cell Counts for Dunnett and Gent’s Method**

<table>
<thead>
<tr>
<th>Adequate</th>
<th>Not Adequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{n_1(n_1+n_2\Delta)}{n}$</td>
<td>$\frac{n_1(n_2-n_2\Delta)}{n}$</td>
</tr>
<tr>
<td>$\frac{n_2(n_1-n_1\Delta)}{n}$</td>
<td>$\frac{n_2(n_2+n_1\Delta)}{n}$</td>
</tr>
</tbody>
</table>

$n_1$ $n_2$ $n$

If all expected cell counts are $\geq 5$, significance is evaluated via $Q_P$ (equation 2.2) using the modified expected counts. The 1-sided $p$-value is obtained by dividing the probability from a $\chi^2$ table with 1 df by two. The authors suggest using Yate's continuity correction in the chi-square calculation.

Dunnett and Gent (77) explain that $p$-values for a two-tail test (i.e., for the alternate hypothesis $-0.1 < \pi_t - \pi_r < 0.1$), are not equal to twice the 1-tail $p$-value. They suggest creating a table of hypothetical observed frequencies having the same marginal totals but deviating in the opposite direction from expectation, calculating the 1-sided $p$-value as above, and adding the two 1-sided $p$-values together. The hypothetical observed frequencies are obtained by $m_{ij}^* + (m_{ij}^* - n_{ij})$ and rounding to the nearest integer. Occasionally this method leads to hypothetical observed counts off by one count. Dunnett and Gent (77) describe a likelihood ratio method to determine if the hypothetical cell counts are correct.

**Based on the Normal Approximation to the Binomial**

Blackwelder (82) points out that clinical trials usually have large enough
samples for the normal approximation to the binomial. If all expected cell counts are \( \geq 5 \) (Rodary et al 89), the one-sided test of equivalence can be conducted using the test statistic

\[
z = \frac{p_l - p_r - \Delta}{SE}, \text{ where } SE = \sqrt{p_l(1 - p_l)/n_l + p_r(1 - p_r)/n_r}.
\]

(3.19)

The corresponding lower one-sided 100(1-\(\alpha\))% CI for \(p_l, p_r\) is:

\[
[(p_l\cdot p_r) - z_{1-\alpha}SE, 1].
\]

(3.20)

Dunnett and Gent (77) suggest using their modified \(\hat{p}_l\) and \(\hat{p}_r\) in the standard error calculation. They also recommend applying a continuity correction of \((1/n_l + 1/n_r)/2\) to equation (3.19). An upper 1-sided CI (Rodary et al 89) is obtained by

\[
[-1, (p_l\cdot p_r) + z_{1-\alpha}SE].
\]

(3.21)

Generally, it may often be of interest to examine a two-sided 100(1-2\(\alpha\))% CI, whose limits correspond to the upper and lower one-sided 100(1-\(\alpha\))% intervals.

**Gart Method**

Dunnett and Gent (77) discuss another method for obtaining confidence intervals on a difference of proportions first derived by Gart. Due to computational difficulties, Rodary et al (89) suggest this method only be used if there is not sufficient sample size for the chi-square approximation. There is no sufficient statistic for \(\delta\), a difference in percents, but an exact CI can be obtained for the odds ratio, \(\psi\). To test \(H_0: \delta = \Delta\) vs \(H_A: \delta < \Delta\), estimate the expected odds ratio based on the expected counts in Figure 3.2 as:

\[
\hat{\psi} = \frac{(n_{11} + n_2 \Delta)(n_{21} + n_1 \Delta)}{(n_{11} - n_1 \Delta)(n_{21} - n_2 \Delta)}.
\]

(3.22)

Compute the exact 1-tail p-value of the conditional distribution derived by Fisher:

\[
g(n_{11}|n_1, \psi) = \frac{(C_{n_1}^n)(C_{n_{12}}^n) \psi^{n_{11}}}{\min(n_{11}, n_{12}) \sum_{i = \max(0, n_{11} - n_2)} (C_{i}^{n_1})(C_{n_{12} - i}^{n_2}) \psi^i}
\]

(3.23)
This reduces to the hypergeometric (Fisher's exact test) if $\Delta=0$, and hence $\hat{\psi} = 1$. The exact p-value for $\psi$ is interpretable as an approximate p-value for $\delta$. Alternatively, use (3.23) to find the exact CI for $\psi$, as described by Gart and others, perhaps using the computer software package StatXact (Cytel Software 1991). Due to the discrete nature of the distribution, the coverage probability for a $100(1-\alpha)\%$ CI is generally $> 1-\alpha$. An approximate CI for $\psi$ can be obtained via equation (2.5). The CI for $\psi$ is transformed to an approximate CI for $\delta$ by applying the following formulas to the upper and lower bounds of $\psi$, and solving for $x$ and then $\Delta$:

$$\psi = \frac{x(x + n_2 - n_1)}{(n_1 - x)(n_1 - x)}, \quad \Delta = \frac{x}{n_1} - \frac{(n_1 - x)}{n_2}. \quad (3.24)$$

Often it may be of more interest to formulate the hypothesis of equivalence in terms of an odds ratio instead of a difference. In such cases, the confidence interval for $\psi$ is of interest in its own right.

### 3.5.2 Ordered Categorical Data

Often clinical trials have a response that take the form of ordered categories, for instance a study of treatments for duodenal ulcers may report no relief, some relief, or substantial relief. In such cases, it may be of interest to evaluate equivalence in terms of a set of odds ratios, which compare two categories or subsets of categories at a time. Conclude equivalence, for instance, if each of the odds ratio confidence intervals is within 0.8 -1.25.

Mehta et al (1984) extend Dunnett and Gent's method to ordered categorical data from parallel designs using the Wilcoxon midranks test. They provide an algorithm for computing the exact p-value of the Wilcoxon midrank test under arbitrarily specified treatment differences. Notation for the test is as follows: ordered outcomes are numbered 1, 2, .. k. Sample size for the test treatment is $n_t$, for reference is $n_r$. The total number of subjects with outcome $j$ is $t_j$, and the
number of subjects from the test group with outcome \( j \) is \( x_j \). The null hypothesis is based on the odds ratios of each of the \( j = 1 \ldots k-1 \) outcomes vs. the \( k \)th outcome. One hypothesis of interest is that each of the \( j \) odds ratios are no larger than some value \( \psi^* \). \( H_0: \psi_j \geq \psi^*_j \quad j = 1, 2, \ldots k-1 \) vs. \( H_A: \psi_j < \psi^*_j \). The Wilcoxon statistic is \( W = \sum_{j=1}^{k} r_j x_j \), where \( r_j = \left( \sum_{m=0}^{j-1} n_m \right) + 1/2(n_j + 1) \), (and \( n_0 = 0 \)) are the midranks of the \( k \) ordered categories. The exact \( p \)-value is based on networking, and is not easily computed. An asymptotic test is provided, which can be used unless both the sample size is small and the treatment difference is large. The normal approximation is:

\[
T = \frac{W - n_t(n_t + n_r + 1)/2}{\sqrt{\frac{n_t n_r (n_t + n_r + 1)}{12} \sum_{j=1}^{k} (t_j^3 - t_j)}}
\]

\[ (3.25) \]

\( T \) has an approximate normal distribution with mean \( \theta \) and variance 1, where \( \theta \) is equal to the value of the normalized Wilcoxon statistic applied to the expected counts of a \( 2 \times k \) table under \( H_0 \). The expected counts are calculated similar to Dunnett and Gent's, so they conserve the marginal totals, and are compatible with the odds ratios of \( H_0 \). The approximate \( p \)-value is obtained by calculating the \( \alpha \) tail probability of \( z = (T - \theta) \).

This method can be generalized to a 2-sided test if desired, and can be expressed in terms of a set of confidence intervals for \( \psi \), the odds ratio of each outcome vs. the \( k \)th outcome. The set of CIs may be more informative than a single test, and is easily obtained via the StatXact software.

Hauck and Anderson (92) point out that \( T_{\text{max}} \) cannot be analyzed by either parametric methods or non-parametric methods which rely on the absence of ties, due to its discrete nature. They say nothing in the literature is appropriate for analyzing \( T_{\text{max}} \), and suggest this as an area for future research. They did not mention Mehta's method, but Mehta does not discuss how to modify his method
for a cross-over study with period effects.

3.5.3 Categorical Data Example

The above methods are demonstrated for the categorical and ordinal outcomes measured in the clinical trial of delayed onset muscle soreness, described in Section 1.4.3. Treatment A is defined as the test therapy, and B is the reference. For the categorical variables of no pain by 48 hours and 50% pain reduction by 48 hours, the lower 1 tailed test of \( H_0: \pi_t - \pi_r \leq \Delta \) vs. \( H_A: \pi_t - \pi_r > \Delta \) is tested with \( \Delta = -0.10 \). This yields a test of whether the test treatment response is no lower than reference - 0.10. Response proportions are:

<table>
<thead>
<tr>
<th></th>
<th>No pain by 48 hrs</th>
<th>50% pain reduction by 48 hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( P_t )</td>
<td>( P_r )</td>
</tr>
<tr>
<td>Global</td>
<td>0.172</td>
<td>0.169</td>
</tr>
<tr>
<td>Movement</td>
<td>0.125</td>
<td>0.108</td>
</tr>
<tr>
<td>Palpation</td>
<td>0.125</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Presented in Table 3.3 are the p-values from Dunnett and Gent’s \( \chi^2 \) statistic (without Yate’s correction), and the normal approximation from equation (3.19) for each of the three soreness evaluations. Nearly all expected cell counts are \( \geq 5 \). Also presented are approximate two-sided 90% confidence intervals for \( \delta \) based on the normal and Gart approximations. Table 3.4 presents the exact 90% intervals for the odds ratio used in calculating the Gart approximation.

The analysis concludes that the treatment A-B difference in percent response is significantly greater than -0.10 for absence of pain, but not for pain reduction. By examining the confidence intervals, one-sided equivalence could generally be concluded for \( \Delta = -0.20 \). The chi-square and normal approximation yield similar results, and the two types of CIs are nearly identical. Confidence intervals for the odds ratio are much wider than would be needed to claim equivalence using the
0.8-1.25 rule described for bioequivalence.

Table 3.3
Test of \( H_0 \) \( \delta \leq -0.10 \) for the Study of Delayed Onset Muscle Soreness
Using the Chi-squared method and Normal Approximation
and 90% CIs for \( \delta \) based on the Normal and Gart Approximations

<table>
<thead>
<tr>
<th>Measure</th>
<th>( \chi^2 )</th>
<th>Normal Approximation</th>
<th>Gart approx.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-value</td>
<td>P-value</td>
<td>90% CI for ( \Delta )</td>
</tr>
<tr>
<td>No Pain by 48 hrs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>0.054</td>
<td>0.061</td>
<td>(-0.106, 0.112)</td>
</tr>
<tr>
<td>Movement</td>
<td>0.011</td>
<td>0.019</td>
<td>(-0.076, 0.110)</td>
</tr>
<tr>
<td>Palpation</td>
<td>0.001</td>
<td>0.003</td>
<td>(-0.039, 0.135)</td>
</tr>
<tr>
<td>50% Pain Reduction by 48 hrs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>0.174</td>
<td>0.176</td>
<td>(-0.161, 0.118)</td>
</tr>
<tr>
<td>Movement</td>
<td>0.180</td>
<td>0.181</td>
<td>(-0.163, 0.120)</td>
</tr>
<tr>
<td>Palpation</td>
<td>0.107</td>
<td>0.108</td>
<td>(-0.136, 0.153)</td>
</tr>
</tbody>
</table>

Table 3.4
90% Confidence Intervals for the Odds Ratio of Improvement
for Drug A vs Drug B in the study of Delayed Onset Muscle Soreness

<table>
<thead>
<tr>
<th>Measure</th>
<th>90% CI for ( \psi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pain by 48 hrs</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>(0.465, 2.233)</td>
</tr>
<tr>
<td>Movement</td>
<td>(0.468, 3.018)</td>
</tr>
<tr>
<td>Palpation</td>
<td>(0.629, 4.859)</td>
</tr>
<tr>
<td>50% Pain Reduction by 48 hrs</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>(0.497, 1.675)</td>
</tr>
<tr>
<td>Movement</td>
<td>(0.503, 1.660)</td>
</tr>
<tr>
<td>Palpation</td>
<td>(0.576, 1.858)</td>
</tr>
</tbody>
</table>
Mehta's method is applied to the ordered categorical change in pain from baseline to 48 hours. Exact odds ratios for each of some, moderate, and marked pain decrease vs. no change or worsening were calculated from StatXact, based on the Wilcoxon midranks statistic. Table 3.5 presents exact 90% confidence intervals for each ratio. Intervals are generally wide and skewed to the right, suggesting treatment A may have slightly outperformed treatment B. Assessment of equivalence depends on a pre-specified range that the confidence intervals must not exceed. Another way of examining equivalence for ordered categories would be to impose a proportional odds or equal adjacent odds structure and obtain one approximate odds ratio confidence interval.

Table 3.5

90% Confidence Intervals for Odds of Improvement vs. None for Drug A vs B for Three Ordered Categories of Pain Improvement by 48 hours in the Study of Delayed Onset Muscle Soreness

<table>
<thead>
<tr>
<th>Measure</th>
<th>Some vs. None</th>
<th>Moderate vs. None</th>
<th>Marked vs. None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>(0.959, 1.905)</td>
<td>(0.917, 3.803)</td>
<td>(0.895, 5.502)</td>
</tr>
<tr>
<td>Movement</td>
<td>(0.919, 1.901)</td>
<td>(0.838, 3.864)</td>
<td>(0.803, 5.327)</td>
</tr>
<tr>
<td>Palpation</td>
<td>(0.853, 1.524)</td>
<td>(0.706, 3.510)</td>
<td>(0.620, 3.542)</td>
</tr>
</tbody>
</table>

3.6 Bayesian Methods for Bioequivalence

Several authors have examined Bayesian methods for bioequivalence. These references include Selwyn et al (81, 84, 85), Racine-Poon et al 87, Fleuhler et al
(81, 83), Mandallaz and Mau (81), Rodda and Davis (80). The Bayesian methods require the same assumptions as ANOVA methods, and are discussed mostly in the context of 2-period cross-over studies, where the measure of interest is the ratio of treatment means.

The Bayesian equivalence criterion is a high posterior probability (generally 90% or 95%) that the difference in formulation means is less than some fraction of the reference formulation mean (one-sided), or within some fraction of the reference (two-sided), usually 20%. Selwyn et al (81) describe how to calculate the posterior density of the treatment means, and apply numerical integration over the region that corresponds to acceptable equivalence to obtain the posterior probability.

They suggest using a prior probability if one is available from a previous study, but otherwise recommend the use of a non-informative prior. The Bayesian method can be modified to incorporate carry-over effects.

One criticism of the method is that the use of different priors may affect the posterior conclusion, which could lead to a dispute between a pharmaceutical company and a regulatory agency. However, with large sample sizes, inferences are insensitive to the prior chosen. Sewyn et al (81) examine four noninformative priors on the variance components and conclude there is little difference provided the sample size is at least six subjects. Another problem is the computational difficulty of implementing numerical integration.

Sewyn et al (81) suggest that when several parameters are being examined, numerical integration could be performed over regions of higher dimension, or Bonferroni inequalities could be applied.

Mandallaz and Mau (81), and Fluehler (81) suggest it is more informative to present the complete posterior probability distribution of the ratio of formulation means, instead of simply the values of a confidence interval. Steinijans and
Diletti (85) suggest their method of obtaining confidence intervals based on Wilcoxon-Mann-Whitney statistics can be presented in a histogram of the probability distribution, by plotting a histogram of the exact Wilcoxon probabilities for each pairwise test/reference ratio. This is a non-parametric probability density for $\theta$, and shows whether the distribution is skewed or bimodal. When a Bayesian posterior probability distribution (which is a smooth normal curve) is plotted over the histogram, the inadequacies of the normal assumption can be visualized by poor similarities.

Racine-Poon et al (87) suggest a 2-stage Bayesian strategy where the posterior density of treatment differences and variances for the first stage is used as the prior for the second stage. Their strategy allows equivalency acceptance, rejection, or continuation at stage 1, and uses stage 1 to determine the sample size required in stage 2. The authors compared their 2-stage method with standard Bayesian 1-stage methods. The 2-stage method found true differences much more often with only slightly larger sample sizes. Their method is best applied to studies with small samples (generally no more than 8 subjects per stage). A method for sequential Bayesian analysis is also discussed by Sewyl n et al (81).

3.7 Individual Equivalence

Hauck and Anderson (1992) discuss different definitions of bioequivalence. The usual average bioequivalence is based on the mean or median of the population distribution of drug differences (or ratios). A more general population equivalence would include all parameters of the population distribution, or at least a measure of variability. They ask how two formulations can be considered bioequivalent if their variances differ substantially. Under average bioequivalence strategies, a test treatment that is more variable but with equal mean as the reference can be accepted as equivalent (although more sample size may be
needed to prove equivalence for variable formulations).

The FDA approves bioequivalence based on average equivalence, but they have the full expectation that the patient will receive the same clinical effect from the generic drug. This is an emphasis on the individual - yet average bioavailability does not insure this. Individual bioavailability insures that the drugs will be sufficiently similar in most of the people.

Hauck and Anderson (92) suggest individual bioequivalence is necessary only when a patient switches from one drug to another equivalent formulation. The impact of switching on the patient has been a concern for some time. When switching occurs between average bioequivalent formulations, individuals may experience problems, especially for drugs with titrated doses or narrow therapeutic windows such as anticonvulsant and psychoactive drugs. If individual equivalence is established, each person is expected to have fewer problems. When starting on a new drug, individual equivalence is of less importance.

If the within-individual variability is sufficiently large, individual bioequivalence could be hard to show between a reference and itself. Therefore a requirement of individual bioequivalence is not appropriate for all drugs. But Hauck and Anderson ask if a treatment is so variable, is it appropriate to try to show its equivalence? It has been suggested that the average equivalence criteria for some highly variable drugs be wider than the usual [0.80 - 1.25] because it would otherwise be impossible to show average equivalence (Pabst and Jaeger 90).

3.7.1 FDA Tolerance Interval

In the early 1980's the FDA considered a bioequivalence criterion motivated by individual equivalence, called the 75/75 rule. It was applied to crossover trials by requiring that at least 75% of the individual test/reference ratios must be between 0.75 and 1.25 in order to demonstrate equivalence. The rule was
developed to insure the therapeutic effect will be within certain limits for a large percent of people, but received much criticism. It ignores statistical inference and assumes no period or carryover effects.

Haynes (81, 83) examined the 75/75 rule via simulation and concluded it was unfair. The test accepts equivalence more readily for smaller treatment variation, and performs worse when treatment variations are equal but large than when one is smaller than another. Obviously, more variable products have a harder time meeting this criterion, suggesting this should not apply to all formulations, or different criteria should be applied to different drugs. The FDA (Cabana 83) defended the proposed rule, saying it would be applied mostly to drugs with small variation (coefficient of variation < 40%), and would also require the (although inappropriate) traditional ANOVA approach.

Dobbins and Thiagarajan (92) developed the sampling distribution which incorporates a binomial distribution implied by the 75/75 rule, and thus placed the rule in the framework of a statistical hypothesis test. They showed why the test acts in the way it has been criticized by Haynes: the significance level of the test depends directly upon the variability. Increasing sample size when formulations are variable will not reduce the significance level of the test.

Metzler (87) points out the 75/75 rule is affected by variability, but is not a test of variance; rather it was proposed as an alternate for testing average bioavailability. The FDA Bioequivalence Task Force hearing in September 1986 recommended the use of 90% confidence intervals instead of the 75/75 rule (Clin Pharm 1988 7:334-335).

3.7.2 The TIER Method

Anderson and Hauck (90) suggest an individual bioequivalence strategy partially motivated by Dobbins and Thiagarajan (92), and ideas mentioned in
It can be viewed as a statistically proper way of implementing the intent of the 75/75 rule, and is applicable to continuous responses where the test/reference ratio is of interest. It does not account for carry-over or period effects. The TIER (Test of Individual Equivalence Ratios) procedure is as follows: calculate the percent of people in the sample \( p = y/n \) with ratio \( x_i/x_r \) within specified equivalence limits (for instance 0.80, 1.25). Choose \( p_E \), the percent of the population required to be within the equivalent range in order to conclude individual equivalence (i.e., 75%). The p-value, interpreted as the probability that the population percent within the equivalence range is \( \geq p_E \), is calculated as the probability of observing \( y \) or more subjects within the equivalence limits, according to the binomial distribution with parameters \( n \) and \( p_E \):

\[
p\text{-value} = \sum_{i = y}^{n} C^n_i p_E^i (1 - p_E)^{n-i}.
\]  

(3.26)

Conclude individual equivalence if p-value \( \leq 0.05 \). An upper 95\% CI for \( p = y/n \) can be calculated by finding all values of \( p < p_E \) such that \( \sum_{i = y}^{n} C^n_i p^i (1 - p)^{n-i} \leq \alpha \). Conclude with 95\% certainty that at least \( p_{lowerbound} \) of the population will respond within the equivalence criteria.

This method is more stringent than average equivalence. The required sample size is larger, even larger than may be practical for this type of study. If between-subject or within-subject variability is high, either a large sample size is needed to conclude equivalence, or it may be impossible to conclude individual equivalence regardless of sample size or equality of the treatment means.

Drawbacks of TIER are that it reduces information on equivalence to a dichotomous response (ratio in range or not), and within-subject variability reduces the chance of showing equivalence, making the test conservative (actual p-value is \( < \alpha \)). The conservativeness could be improved by conducting extended-period crossover designs, where a subject receives each treatment more than once.
3.7.3 Individual Equivalence Example

To demonstrate results of individual equivalence techniques, the 75/75 rule and TIER method are examined for the same data used to explore average equivalence methods in section 3.4. Whereas average equivalence was accepted (based on the range 0.80-1.25), Table 3.6 indicates individual equivalence is clearly rejected, even with a less stringent equivalence range (0.7-1.43).

Table 3.6
Individual Equivalence Results for
Area under the Curve (AUC) and Maximum Concentration (Cmax)
Comparing Generic and Reference Formulations of a Psychotropic Drug, Study #1

<table>
<thead>
<tr>
<th>Method</th>
<th>Percent of patients within equivalence range*</th>
<th>Lower 1-sided 95% CI for percent</th>
<th>P-Value</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIER method†</td>
<td>0.67</td>
<td>(0.512, 1)</td>
<td>0.744</td>
<td>Reject</td>
</tr>
<tr>
<td>FDA 75/75 Rule</td>
<td>0.48</td>
<td>-</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td>Cmax</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIER method†</td>
<td>0.76</td>
<td>(0.615, 1)</td>
<td>0.367</td>
<td>Reject</td>
</tr>
<tr>
<td>FDA 75/75 Rule</td>
<td>0.57</td>
<td>-</td>
<td>-</td>
<td>Reject</td>
</tr>
</tbody>
</table>

*Equivalence range is test/reference ratio within 0.7-1.43 for TIER and within 0.75-1.25 for FDA Rule.
†Accept if CI lower bound is above 0.75 (p ≤ 0.05).

3.7.4 Further Individual Equivalence Strategies

Hauck and Anderson (91) do not view TIER as 'the final word' for individual bioequivalence, but a start in the process. Another method is to test the subject by formulation interaction in the ANOVA model (Ekbohm and Melander 89).
Designs with replication of formulations in subjects is required to test the interaction. Schuirmann (90) discusses a variety of replicated treatment designs, which also have the ability to control for carry-over effects while estimating main effects. For two formulations Schuirmann recommends three replicated designs. One is a 3 period, 2 sequence trial: sequence A = TRT, sequence B=RTR.

Design Two has 4 periods and sequences: A: TTRR, B: TRRT, C: RRTT, D: RTTR. Each formulation precedes the other and itself an equal number of times.

Design three has 4 sequences and two periods: A: TR, B: RT, C: TT, D: RR. This is inefficient if there are no carry-over effects, because then sequences C and D have no helpful information. A repeated design improves the power because the within-subject error term in the model is halved. Whether this is better financially and logistically would have to be examined.

Gaffney (92) supports replicated treatment designs because they allow examination of within-subject variation. The within-subject variance component for each formulation can be compared as a measure of equal variability. The FDA Bioequivalence Task Force hearing in 1986 recommended that further study should be made regarding intrasubject variability (Clin Pharm 1988 7:334-335).

Haynes (81) proposed the use of an adjusted Pitman-Morgan F test for comparing the variation of the two treatment groups in a 2 period cross-over design. $F = (s_i/s_r)^2$, where $s$ is the standard deviation. He discusses the appropriate F distribution for the cross-over design. Critics (Hauck and Anderson 92, Pabst and Jaeger 90) have pointed out the null hypothesis is in the wrong direction for equivalence. Perhaps future work could convert the test to a confidence interval for the ratio of standard deviations. An ideal test would combine the joint equivalence assessment of geometric means and variation in one over-all test (Hauck and Anderson 92). Much larger sample sizes would be needed to assess variance equality (Metzler and Huang 83).
3.8 Additional Bioequivalence Topics

Equivalence Methods to Interpret Negative Studies

Hauck and Anderson (86) suggest that equivalence methods can be applied to studies attempting to prove a difference, but where no significant differences were found. Instead of examining the power for detecting certain differences, confidence intervals of the parameters show what values the study had reasonable power to detect. For instance a 90% confidence interval on a ratio of (0.9, 1.4) can be interpreted as concluding (with \( \alpha = 0.05 \)) that the ratio is not more than 1.4. A confidence interval of ratios may show that equivalence has been proven, and CIs of differences may show that the treatment difference is not clinically significant, and hence the treatments may be clinically equivalent.

Confidence intervals in studies to show a difference are not new, but the bioequivalence literature emphasizes their usefulness, and contributes the methodology to obtain p-values for tests of equivalence.

Reducing Variability with an Independent Reference

Phillips (90 B) described a bioequivalence method for extremely variable formulations. In the standard two-period cross-over study, give the subject another compound containing a stable isotope at the same time as each formulation under study. Calculate the parameters of interest for the study formulations as well as the isotopes. For each person, form the log difference of the study with the same-time isotope measure. Then perform equivalence on the ratios of these ratios. If the test vs isotope ratio is within 0.8-1.25 of the reference vs isotope ratio, then accept equivalence. Advantages and concerns are discussed by the author.
Assessing Bioequivalence with Observational Data

Kaniwa et al (90) have suggested that a method for estimating population pharmacokinetics that uses extended nonlinear least squares (called the NONMEM method), could evaluate bioequivalence for observational, unbalanced, and noisy data. NONMEM could use observational routine blood samples collected from a larger group of subjects to assess equivalence, instead of conducting a formal equivalence study. Jawien (92) proposed a simple arithmetic estimate of AUC when concentrations are available from only one time point per person, but several different time points overall. This could also be used for observational data, without the use of sophisticated computer programs.

3.9 Power and Sample Size Calculations

The probability of a type I error $\alpha=\text{pr}(\text{reject } H_0 \text{ of non-equivalence when should not})$, is interpreted as consumer risk, and is the concern of regulatory agencies. The probability of a type II error, $\beta=\text{pr}(\text{do not reject } H_0 \text{ when should})$, is traditionally also a concern of regulatory agencies. However, for bioequivalence, if the null hypothesis of non-equivalence is not rejected when it should be, this is more a concern of the producer, since they are hoping to show equivalence. For clinical trials, both $\alpha$ and $\beta$ are important to the investigator.

Westlake (88), Schuirmann (87, 90) and Hauck and Anderson (84) each discuss power and sample size for their proposed procedures. Metzler (91) found little difference in probability of rejection between Schuirmann's 2 1-sided tests procedure, Westlake symmetrical CIs, Anderson and Hauck's method, and Bayesian methods. Therefore, when planning power and sample size for a two period cross-over equivalence study, results for one method should reasonably hold for the others. Sample size formulas are provided for Schuirmann's 2 1-sided test procedure. Most bioequivalence studies have a coefficient of variation less than
33%, with roughly a third having CV less than 10% (Metzler 91). When the CV is 30%, sample size required to show equivalence get fairly large, but most studies usually have 12-18 subjects (Steinijans, Hauschke 1990).

For Schuirmann 2 1-Sided Tests and 90% CI

Although not discussed much in this chapter, some measures of bioavailability are appropriate for parametric analysis, but on the original (additive) scale rather than the log-transform (multiplicative) scale. Sample size formulas are presented here for both situations.

A. Additive model

Phillips (90 A) presents exact power curves for varying coefficients of variation and sample size, but a simple formula for sample size calculation is not possible. Liu and Chow (92) provide an approximate formula that allows calculation with a hand calculator or SAS program. The following must be specified or estimated from previous studies:

- Maximum % difference expected \( \gamma = \left( \frac{\mu_t - \mu_r}{\mu_r} \right) \times 100 \)
  or wanted to detect as equivalent \( \gamma < \Gamma \)
- Variability (in % of ref mean) \( CV = \left( \frac{\sqrt{MSE}}{\mu_r} \right) \times 100 \)
- Upper limit for equivalence \( \Gamma \)
- Alpha and power usually .05 and 0.80 or 0.90

The sample size (number of subjects) for a two-period cross-over equivalence study is: if \( \gamma = 0 \): \[ N = 2[t_{1-\alpha} + t_{1-\beta/2}]^2 \frac{CV^2}{\Gamma^2} \]
if \( \gamma > 0 \): \[ N = 2[t_{1-\alpha} + t_{1-\beta}]^2 \frac{CV^2}{(\Gamma - \gamma)^2} \] (3.27)

where \( t \) is the t distribution with N-2 df. Since N is unknown, it may be easier to calculate sample size with a z. These formulas can be manipulated to solve for power. Liu and Chow (92) provide approximate sample size tables for varying CV.
and power, and an example of how to calculate N using the t distribution.

B. Multiplicative model

Diletti et al (91, 92) presented exact sample sizes and power curves using Phillip's (90 A) method for the multiplicative model. Hauschke et al (92) modified Liu and Chow's approximate sample size formula as follows:

- Maximum ratio expected or \( \theta = \mu_1/\mu_r \), \( \Theta_1 < \theta < \Theta_2 \)
  wanted to detect as equivalent
- Variability (in fraction relative to reference mean) \( CV = \sqrt{\exp(MSE) - 1} \)
- Equivalence limits \( \Theta_1 \) and \( \Theta_2 \) (usually 0.80 and 1.25)

Sample size formulas are:

- If \( \theta = 1 \):
  \[ N = 2[t_{1-\alpha} + t_{1-\beta/2}]^2 CV^2 / (\ln\Theta_2)^2 \]
- If 1 < \( \theta < \Theta_2 \):
  \[ N = 2[t_{1-\alpha} + t_{1-\beta}]^2 CV^2 / (\ln\Theta_2 - \ln\theta)^2 \]
- If \( \Theta_1 < \theta < 1 \):
  \[ N = 2[t_{1-\alpha} + t_{1-\beta}]^2 CV^2 / (\ln\Theta_1 - \ln\theta)^2 \]  

Again, t has N-2 degrees of freedom. When variability is small, the coefficient of variation is approximately equal to the root mean squared error, making this equation comparable to equation 3.6 described by Westlake (where Westlake's \( \Delta \) is equal to \( \ln\Theta_2 \) here).

For Non-parametric Techniques

Sample size and power are not explored by the literature for the Wilcoxon-Mann-Whitney 90\% confidence interval, but Diletti et al (91) claim there should only be a minor loss in power from the parametric calculations when switching to the non-parametric approach.

For Categorical Data from Parallel Studies
Blackwelder (82) and Makuch (78) present sample size formulas for Dunnett and Gent's one sided confidence interval of treatment differences from a parallel study with dichotomous outcomes, based on the normal approximation. Needed is the following information, plus \( \alpha \) and \( \beta \):

- Estimated treatment responses (%) \( \pi_t, \pi_r \)
- Maximum difference expected \( \delta = \pi_t - \pi_r \), upper CI: \( \Delta - \delta > 0 \)
  or wanted to detect as equivalent lower CI: \( \Delta - \delta < 0 \)
- Upper or lower limit for equivalence \( \Delta \) (perhaps .10 or -.10)

For a study with equal sample size per group, the formula is:

\[
\text{n per group} = \frac{[z_{1-\alpha} + z_{1-\beta}]^2[\pi_t(1 - \pi_t) + \pi_r(1 - \pi_r)]}{(\Delta - \delta)^2} \tag{3.29}
\]

Makuch (78) suggests for planning to just use a common estimate of the treatment success rates \( \pi \), so \( \delta = 0 \). If unequal treatment group sizes are desired, define \( k \) such that \( n_t = k \times n_r \). Then,

\[
n_r = \frac{[z_{1-\alpha} + z_{1-\beta}]^2[\pi_t(1 - \pi_t) + k\pi_r(1 - \pi_r)]}{(\Delta - \delta)^2}, \text{ and } n_t = k \times n_r. \tag{3.30}
\]

Rodary et al (89) suggest a slightly modified formula which takes advantages of Dunnett and Gent's estimated variance under \( H_0 \) instead of the observed variance. Their formula for equal group sizes is \( n \) per group =

\[
\frac{[z_{1-\alpha}\sqrt{\pi_r(1 - \pi_r) + (\pi_r + \Delta)(1 - \pi_r - \Delta)} + z_{1-\beta}\sqrt{\pi_r(1 - \pi_r) + \pi_t(1 - \pi_t)}]^2}{(\Delta - \delta)^2}.
\]

\( \tag{3.31} \)

3.10 Conclusion

This chapter has examined many of the statistical techniques and considerations for bioequivalence and clinical equivalence studies. While the majority of the methods were discussed in terms of a two period cross-over
bioavailability study (in the context they were developed), their principles can also be applied to parallel designs, and clinical equivalence trials.

A major concern with bioequivalence studies is that they nearly always measure multiple outcome responses (AUC, C\text{max}, T\text{max}, etc), but the literature rarely recognizes the multiple testing issue, and practically never cites appropriate techniques for handling this problem. In the next chapters, several methods for equivalence testing with multiple endpoints are examined.
Chapter 4
Multiple Endpoint Equivalence Techniques:
Two Univariate Confidence Intervals

4.1 Introduction to Multiple Endpoint Analysis

4.1.1 Current Strategy

Many bioequivalence and clinical equivalence trials evaluate more than one endpoint. Analysis strategies for multiple endpoints, however, are nearly nonexistent in the equivalence literature. Sauter et al (92) suggest appropriate presentation of results from bioequivalence studies in detail, but do not mention multiple comparisons. In particular, one of their examples concluded bioequivalence for extent of absorption but not rate of absorption, and there was no discussion of how this affected the overall conclusion concerning bioequivalence.

Very few publications have discussed multiple endpoint strategies. Westlake (88) mentions the possibility of using a Bonferroni correction to adjust for multiple measures such as AUC, $C_{max}$, and $T_{max}$, but if significance of all three tests is required in order to conclude equivalence then a Bonferroni correction is not necessary. A recent article written in Japanese may address some issues (Tsubaki, Fujita 87). Concheiro et al (86) suggest applying a MANOVA to three equivalence measures. MANOVA, however, tests the traditional hypothesis of no difference, not equivalence. Also, as mentioned in Section 2.4, MANOVA statistics do not differentiate whether treatment differences are in the same or opposite directions across the responses, making their use questionable for equivalence. MANOVA is a global test: it could reject the null hypothesis based
on a linear combination of variables, when each may not be rejected when evaluated individually.

At present, when equivalence studies measure multiple endpoints, a confidence interval is generated for each measurement separately. Bioequivalence is generally concluded if 90% CIs for both area under the curve and maximum concentration are completely contained within pre-specified equivalence limits. In this work, this is referred to as the two (or multiple) univariate confidence interval strategy. For some phenomena, it is customary to select one primary outcome which will define clinical equivalence. Other fields consider multiple measures equally important, but may require only a set number (i.e., any two out of three measurements) to statistically support equivalence before treatment equivalence is accepted. Under this scenario, the \( \alpha \) level for each test should be adjusted to \( 0.05 \times \frac{2}{3} \), using the Hailperin-Rüger method described in Section 1.3.1.

When all outcomes are required to be equivalent in order to conclude treatment equivalence, there is no problem with an increased type-I error rate, since each test must result in a p-value \( \leq 0.05 \). Concern actually arises for power. As the number of outcomes increases, the probability that two treatments will be found significantly equivalent for all outcomes decreases, even when equivalence actually holds. If only a subset of outcomes are required to show equivalence, then the type-I error rate may need to be maintained with a Bonferroni or Hailperin-Rüger strategy, in addition to concerns about decreased power. These methods have not been formally specified as strategies for the equivalence situation.

The bioequivalence literature does not seem to recognize the loss of power when evaluating equivalence for multiple outcomes. In order to avoid under powered studies, sample size calculations should consider multivariate power, instead of simpler univariate power estimates. Sufficient multivariate power may
be more important for clinical equivalence trials, due to costs and ethical considerations when studying large samples of sometimes critically ill participants (e.g., anti-infective studies).

Chapters 4 and 5 evaluate power of the two-univariate confidence interval strategy under a variety of scenarios. This chapter examines continuous responses, while the next focuses on dichotomies. The chapters show that, while the univariate technique ignores the correlation between outcomes, its power is generally better when the endpoints are highly correlated. Chapter 7 suggests multivariate analysis strategies that may gain power, by taking the correlation of the endpoints into account.

4.1.2 Power of the Current Strategy

The power of the two univariate CI strategy depends on the correlation between the outcomes. As an example, consider the situation where the power of detecting equivalence for variable X is 0.90, and the power for variable Y is also 0.90. If variances, population test/reference geometric mean ratios, and equivalence criteria are the same for each response, and if the correlation between the responses is 1.0, then the bivariate power is also 0.90. (If geometric means, variances, or criteria are unequal, bivariate power is more complicated; see Section 4.2.3.) If the variables are completely independent, the bivariate power is $0.90 \times 0.90 = 0.81$. This loss of power becomes more severe as univariate powers decrease, for instance $0.70 \times 0.70 = 0.49$.

This quick rule of thumb provides a range of multivariate power for any number of responses. If responses are positively correlated, multivariate power can not be greater than the minimum univariate power, and is no worse than the product of all univariate powers. This range can be wide, however. If an estimate of the correlation is available, it would be advantageous to get a better
approximation of the power. The rule of thumb does not apply when outcomes are negatively correlated, although this is fairly uncommon. Bivariate power for negatively correlated responses is discussed in Section 4.2.7.

This chapter examines power of the multiple univariate confidence interval strategy for continuous responses when a log-normal distribution can be assumed. The power of this multivariate strategy is assessed for both cross-over and parallel designs under a variety of scenarios, primarily for two endpoints. Univariate power for individual outcomes is also provided for comparison. Results assuming an additive-normal scale would be similar, but are of less interest since equivalence is generally defined in terms of ratios of medians.

In the following sections, numerical integration methods are derived to obtain exact bivariate power for studies with large sample sizes, and simulations are conducted to approximate power for small sample sizes. Cross-over and parallel designs are discussed separately. For each design, a statistical model is specified and the appropriate test statistic and its distribution are derived from the model. Bivariate power of the statistic is evaluated under a variety of scenarios. Features of the influence from correlation are discussed, and an example is provided for illustration.

4.2 Cross-Over Bivariate Normal Distribution

In this section, bivariate power for concluding equivalence from two univariate tests is evaluated for a cross-over design with two responses that are assumed to follow a bivariate log-normal distribution.

4.2.1 Statistical Model

Consider a cross-over design with two continuous response variables, \(x_1\) and \(x_2\) (for example AUC and \(C_{max}\)), and assume that \(\ln(x_1)\) and \(\ln(x_2)\) follow a
bivariate normal distribution. The response can be modeled as follows:

$$\ln(x)_{hijk} = \mu_k + \pi_{jk} + \tau_{\xi(h,j)k} + s_{ik} + e_{ijk}$$  \hspace{1cm} (4.1)$$

where $h =$ sequence (A=TR, B=RT)

$i =$ subject $1...N$ or $1...n_A$ and $1...n_B$ for each sequence group

$j =$ period I, II

$k =$ variable 1, 2

$\xi(h,j) =$ treatment: $\xi(A,I) = t$ $\xi(A,II) = r$ $\xi(B,I) = r$ $\xi(B,II) = t$

$\mu_k =$ fixed response mean for the $k$th variable

$\pi_{jk} =$ fixed period effect for $k$th variable

$\tau_{\xi(h,j)k} =$ fixed treatment effect for the $k$th variable

$s_{ik} =$ random subject effect for the $k$th variable of the $i$th subject

$e_{ijk} =$ random error for the $j$th period and $k$th variable of the $i$th subject.

The following assumptions are imposed on this model:

no carry-over: period effects are the same for each sequence

subject effects are the same for each period

errors from the 2 periods of a subject are independent and identically distributed (IID)

errors and subject effects from different subjects are each IID

subject effects and errors are independent

for each subject, $s_i \sim N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{s1}^2 & \rho_s \sigma_{s1} \sigma_{s2} \\ \rho_s \sigma_{s1} \sigma_{s2} & \sigma_{s2}^2 \end{bmatrix} = \Sigma_s$  \hspace{1cm} (4.2)$$

for each subject and period, $e_{ij} \sim N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e1}^2 & \rho_e \sigma_{e1} \sigma_{e2} \\ \rho_e \sigma_{e1} \sigma_{e2} & \sigma_{e2}^2 \end{bmatrix} = \Sigma_e$  \hspace{1cm} (4.2)$$
4.2.2 Obtaining the Test Statistic

In order to evaluate bivariate power, the distribution of the test statistic must be derived, assuming the above model. Steps involved in obtaining the test statistic and its distribution are described below.

1. Generate period I - period II difference for each subject:

\[
\ln(x)_{hilk} - \ln(x)_{hilIk} = (\mu_k - \mu_k) + (\pi_k - \pi_{lIk}) + (\tau_{\xi(h,l)k} - \tau_{\xi(h,II)k}) + (s_{ik} - s_{ik}) + (e_{ilk} - e_{iIIk})
\]

\[
= (\gamma_k) + (\delta_k \cdot \zeta_h) + (e_{ilk} - e_{iIIk})
\]

where \( \delta_k \) = test-reference difference of treatment effects for the kth variable

\( \zeta_h \) = treatment order indicator : \( \zeta_A = 1, \zeta_B = -1 \)

\( \gamma_k \) = difference in period effects for the kth variable.

For each subject and sequence,

\[
\ln\begin{bmatrix} x_{hII} \\ x_{hil2} \\ x_{hil1} \end{bmatrix} - \ln\begin{bmatrix} x_{hII} \\ x_{hil2} \\ x_{hil1} \end{bmatrix} \sim N\left(\begin{bmatrix} \gamma_1 + \delta_1 \zeta_h \\ \gamma_2 + \delta_2 \zeta_h \end{bmatrix}, 2 \times \Sigma_e \right)
\]

2. Average subjects in each sequence:

\[
\frac{1}{n_h} \sum_{i=1}^{n_h} [\ln(x)_{hilk} - \ln(x)_{hIIk}] = (\gamma_k) + (\delta_k \zeta_h) + \frac{1}{n_h} \sum_{i=1}^{n_h} (e_{ilk} - e_{iIIk}).
\]

For each sequence,

\[
\frac{1}{n_h} \sum_{i=1}^{n_h} \left[ \ln\begin{bmatrix} x_{hI} \\ x_{hil2} \\ x_{hil1} \end{bmatrix} - \ln\begin{bmatrix} x_{hI} \\ x_{hil2} \\ x_{hil1} \end{bmatrix} \right] \sim N\left(\begin{bmatrix} \gamma_1 + \delta_1 \zeta_h \\ \gamma_2 + \delta_2 \zeta_h \end{bmatrix}, 2 \times \Sigma_e \right)
\]

3. Generate (Seq A - Seq B)/2:

\[
d_k = \frac{1}{2} \left[ \frac{1}{n_A} \sum_{i=1}^{n_A} [\ln(x)_{Ailk} - \ln(x)_{AIIk}] - \frac{1}{n_B} \sum_{i=1}^{n_B} [\ln(x)_{Bilk} - \ln(x)_{BIIIk}] \right]
\]

\[
= \frac{1}{2} \left[ (\gamma_k) + (\delta_k \zeta_A) + \frac{1}{n_A} \sum_{i=1}^{n_A} (e_{ilk} - e_{iIIk}) - (\gamma_k) + (\delta_k \zeta_B) - \frac{1}{n_B} \sum_{i=1}^{n_B} (e_{ilk} - e_{iIIk}) \right]
\]
\[
(\delta_k) + \frac{1}{2} \left[ \frac{1}{n_A} \sum_{i=1}^{n_A} (e_{ii} - e_{ii}) - \frac{1}{n_B} \sum_{i=1}^{n_B} (e_{ii} - e_{ii}) \right] \\
\left[ \begin{array}{c}
d_1 \\
d_2 \\
\end{array} \right] \sim N \left( \left[ \begin{array}{c}
\delta_1 \\
\delta_2 \\
\end{array} \right], \frac{1}{2} \frac{1}{n_A} + \frac{1}{n_B} \times \Sigma \right) 
\] (4.4)

It is easily shown by algebra that \( d_k \) can also be calculated by mean[\( \ln(x)_{ik} \)] - mean[\( \ln(x)_{ik} \)] when \( n_A = n_B \). The univariate distribution of \( d_k \) corresponds to the distribution of Schuirmann's two 1-sided test statistic (equation 3.8), assuming \( \sigma_{ek} \) is known (although Schuirmann's statistic is presented in standardized form). Equation (4.4) thus provides the distribution of a bivariate generalization of Schuirmann's test statistics.

Now that the distribution of the test statistic is derived, its power can be evaluated. There are two approaches to calculating power - either direct calculation via integration or numerical integration, or estimation via simulation. If the sample size is large (i.e., \( N \geq 30 \)), then the variance can be assumed known, and equation (4.4) can be used to obtain power directly. Alternatively, simulation could be applied assuming a known variance.

Estimates of variance from any given sample are actually random variables, with a distribution often proportional to the chi-square distribution. As the sample size increases, the estimated variance approaches the population parameter, but is only an estimate of it in small samples. If the variance is assumed unknown and is estimated from the sample, then the univariate test statistic for comparing \( d_k \) to \( \delta_k \) follows a \( t \) distribution with \( N-2 \) degrees of freedom. As the sample size increases, the \( t \) distribution becomes equivalent to a normal \( z \). Usually when \( N \) is 30 or larger, applicability of the \( z \) is reasonable.

However, when the sample size is small (\( N<30 \)), the distribution in equation (4.4) does not apply to the bivariate test statistic, and must be modified to account for the uncertainty in estimating the variances. Exact numerical
integration of power requires knowing the joint multivariate distribution of the
test statistic and the variances; this can lead to a multivariate non-central t, or a
function of Wishart and normal distributions. Laska and Meisner (89) derived the
exact distribution for a similar situation but with only one variance parameter as
a function of a bivariate normal and chi-square. Although derivation of the
appropriate distribution for test statistics in this setting is possible in principle, it
is beyond the purposes of this work, and so power in small samples is
approximated by simulation.

The next sections describe steps to calculate bivariate power via numerical
integration and simulation assuming known and unknown variance.

4.2.3 Power Calculation from Numerical Integration

Numerical integration based on a bivariate normal distribution is applicable
to power calculations if the sample size is large enough that the applicable
variances can be assumed known. Bivariate power is derived as follows:

The univariate power of a test statistic, $d_k$, is defined as the probability of
rejecting the null hypothesis $H_0$. For equivalence, this corresponds to:

$$\Pr(c_L < d_k < c_U). \quad (4.5)$$

The cutpoints $c_L$ and $c_U$ are obtained by applying the fixed type I error, $\alpha \geq
\alpha(\delta_k)=\Pr[\text{reject } H_0 \mid E(d_k)=\delta_k \text{ where } \delta_k<\Delta_1 \text{ or } \delta_k>\Delta_2]$, corresponding to each of
the two 1-sided tests as described in Section 3.2.3. For $c_L$,

$$\alpha \geq \Pr( d_k > c_L \mid \delta_k < \Delta_1).$$

If the variance is assumed known, a standard normal variable is obtained by
subtracting $\delta_k$ and dividing by $\text{SE}(d_k)$ on both sides of the equation. The type I
error will be at a maximum when $\delta_k = \Delta_1$, so this value is used to obtain:

$$\alpha = \Pr\left[ Z > \frac{c_L - \Delta_1}{\text{SE}(d_k)} \right]$$

By applying the $z$-value corresponding to the $\alpha$ upper tail of the normal
distribution, and solving for \( c_L \), one obtains
\[
c_L = \Delta_1 + Z_{1-\alpha} \times \text{SE}(d_k).
\]
Similarly, the upper cutpoint is calculated using:
\[
\alpha \geq \Pr( d_k < c_U | \delta_k > \Delta_2)
\]
and is found to be:
\[
c_U = \Delta_2 + Z_\alpha \times \text{SE}(d_k).
\]
The power of the test is obtained by substituting \( c_L \) and \( c_U \) into equation (4.5):
\[
\text{Power} = \Pr[\Delta_1 + Z_{1-\alpha} \times \text{SE}(d_k) < d_k < \Delta_2 + Z_\alpha \times \text{SE}(d_k)]
\]
Standardizing by subtracting \( \text{E}(d_k) = \delta_k \), dividing by \( \text{SE}(d_k) \) on each side, and using \( Z_\alpha = -Z_{1-\alpha} \) leads to:
\[
\text{Power} = \Pr\left[ \frac{\Delta_1 - \delta_k}{\text{SE}(d_k)} + Z_{1-\alpha} < Z < \frac{\Delta_2 - \delta_k}{\text{SE}(d_k)} - Z_{1-\alpha} \right].
\]
(4.6)
where \( \text{SE}(d_k) = \sqrt{\frac{\sigma^2_{d_k}}{2 \left( \frac{1}{n_A} + \frac{1}{n_B} \right)}} \). This formula displays reasonable properties. As \( N=(n_A+n_B) \to \infty \), and \( \Delta_1 < \delta_k < \Delta_2 \), then power \( \to 1 \). Power \( \to 1 \) fastest when \( \delta_k = 0 \). If \( \delta_k > \Delta_2 \) or \( \delta_k < \Delta_1 \), then as \( N \to \infty \), power \( \to 0 \). Power \( \to 0 \) faster as \( |\delta_k| \) increases. When \( \delta_k = \Delta_1 \) or \( \delta_k = \Delta_2 \), then as \( N \to \infty \), power \( \to \alpha \). Also, for fixed \( N \), power is zero when
\[
\left[ \frac{\Delta_1 - \delta_k}{\text{SE}(d_k)} + Z_{1-\alpha} \geq \frac{\Delta_2 - \delta_k}{\text{SE}(d_k)} - Z_{1-\alpha} \right]
\]
or
\[
\left[ \frac{\Delta_2 - \Delta_1}{\text{SE}(d_k)} \leq 2Z_{1-\alpha} \right]
\]
or
\[
\left[ N \leq \frac{4Z^2_{1-\alpha} \sigma^2_{d_k}}{(\Delta_2 - \Delta_1)^2} \right. \text{ when } n_A = n_B.
\]
If the lower and upper limits are referred to as \( L \) and \( U \), and if \( \delta \) and \( \sigma^2_e \) are assumed known, then the power is calculated by referring to tables of the standard normal distribution, or by numerical integration of the function:
\[
\int_L^U \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz.
\]
(4.7)
To obtain bivariate power of statistics \( d_1 \) and \( d_2 \), the correlation between
variables within a person, $\rho_\epsilon$, must also be known. By standardizing the bivariate
distribution in equation (4.4) and applying the same integration limits, power is
found by numerical integration of a standardized bivariate normal distribution:

$$\int_{L_1}^{U_1} \int_{L_2}^{U_2} \frac{1}{2\pi\sqrt{1-\rho_\epsilon^2}} e^{\frac{-1}{2(1-\rho_\epsilon^2)}[z_1^2 - 2\rho_\epsilon z_1 z_2 + z_2^2]} \, dz_1 \, dz_2. \quad (4.8)$$

If the correlation is zero, the variables are independent, and this formula
simplifies to the product of two univariate standard normals (equation 4.7).
Bivariate power is formally undefined if the correlation is 1.0 since the quantity to
be integrated is null. However, another way of expressing a bivariate distribution
is in terms of conditional probability, $f(z_1, z_2) = f(z_2|z_1) \times f(z_1)$. A standardized
bivariate normal can be expressed as $N(\rho_\epsilon z_1, 1-\rho_\epsilon^2) \times N(0, 1)$, which if the variables
are perfectly correlated reduces to $N(z_1,0) \times N(0,1)$. Bivariate power is expressed
as $\Pr[L_1<z_1<U_1 \text{ and } L_2<z_2<U_2]$. When $\rho_\epsilon=1$, then $z_2=z_1$ with probability 1, so
the bivariate power becomes $\Pr[L_1<z_1<U_1 \text{ and } L_2<z_1<U_2]$, which simplifies to:

$$\Pr[\max(L_1, L_2) < z < \min(U_1, U_2)].$$

If $L_1=L_2$ and $U_1=U_2$, then this is equal to the univariate power in equation (4.6).
However, when the correlation is 1.0 but there are differences for the population
geometric mean ratios, equivalence criteria, or standard errors for the two
responses, then the limits may not be equal. The two univariate powers may be
different and the bivariate power must be evaluated with the above equation.

A constraint imposed by equation (4.6) is that

$$\Delta_1 + Z_{1-\alpha} \times \text{SE}(d_k) < \Delta_2 + Z_\alpha \times \text{SE}(d_k)$$

Plugging in the standards of $\Delta_1=\ln(0.80)$, $\Delta_2=\ln(1.25)$, and $\alpha=0.05$, then SE($d_k$)
must be less than 0.1356 for the univariate or bivariate power to be properly
defined. As an example, for $n_A=n_B=10$, $\sigma_{ek}$ must be less than 0.428 in order for
power to be greater than zero, regardless of $\delta_k$. If these imposed constraints are
not met, numerical integration leads to a negative power estimate. In such cases, power is defined to be zero.

4.2.4 Power Estimation from Simulation Assuming Known Variance

If numerical integration techniques are not straightforward, then power under the scenario of known variance can be approximated by simulation. Simulation involves the repetitive generation of random samples from a distribution with known parameters and evaluation of test statistics from each sample. Power is estimated as the percent of the samples which reject the null hypothesis and support equivalence. This section describes how simulations are conducted to approximate the power derived in Section 4.2.3. It is intended as an introduction to simulation and for comparison against simulations assuming unknown variance.

All parameters of the test statistic distribution ($\delta_1, \delta_2, n_A, n_B, \sigma_{e1}^2, \sigma_{e2}^2,$ and $\rho_e$) must first be specified, and then random observations from the normal distribution based on them are repetitively generated. Since the variance is assumed known, individual observations do not need to be generated to estimate it. For each repetition, one random observation from each of 4 standardized normal deviates $= z_{A1}, z_{A2}, z_{B1}, z_{B2}$ is generated, to represent the average of the observed errors for each sequence-variable combination. For each sequence,

$$
\begin{bmatrix}
  z_{h1} \\
  z_{h2}
\end{bmatrix}
\sim N
\begin{bmatrix}
  0 \\
  0
\end{bmatrix},
\begin{bmatrix}
  1 & 0 \\
  0 & 1
\end{bmatrix}
$$

These observations are transformed so they have the specified variance-covariance structure of the average period I - period II difference for each sequence group by applying a Cholesky decomposition to $C_{hh} = \Sigma_e$. The Cholesky decomposition of a symmetric positive definite matrix $A$ is obtained by finding a lower triangular matrix $L$ such that $A = LL'$ (Strang). For a $2 \times 2$ matrix, the
result is:
\[
\begin{bmatrix}
  \sqrt{a} & 0 \\
  b & \sqrt{\frac{-b^2}{a} + c}
\end{bmatrix}
\begin{bmatrix}
  \sqrt{a} & \frac{b}{\sqrt{a}} \\
  \frac{b}{\sqrt{a}} & \sqrt{\frac{-b^2}{a} + c}
\end{bmatrix}
\]

For each sequence group, the transformed vector is defined
\[
\begin{bmatrix}
  f_{h1} \\
  f_{h2}
\end{bmatrix}
= L \times
\begin{bmatrix}
  z_{h1} \\
  z_{h2}
\end{bmatrix}
\]
so that
\[
f_{h1} = (\sqrt{a}) \times z_{h1}
\]
and
\[
f_{h2} = (\frac{b}{\sqrt{a}}) \times z_{h1} + \left[ \sqrt{\frac{-b^2}{a} + c} \right] \times z_{h2}
\]
The transformed vector has the following distribution:
\[
\begin{bmatrix}
  f_{h1} \\
  f_{h2}
\end{bmatrix}
\sim N
\begin{bmatrix}
  0 \\
  0
\end{bmatrix},
\begin{bmatrix}
  a & b \\
  b & c
\end{bmatrix}
\]

To obtain the specified covariance matrix of the average period I - period II
difference for each sequence, the matrix \( A_h \) is set equal to \( (\frac{2}{n_h} \times \Sigma_e) \). The elements
of \( A_h \) are thus:
\[
\begin{align*}
a &= 2\sigma_{e1}^2/n_h \\
b &= 2\rho_e\sigma_{e1}\sigma_{e2}/n_h \\
c &= 2\sigma_{e2}^2/n_h
\end{align*}
\]
The observed test statistic, the average sequence difference of average period
differences, is found by \( \hat{d}_k = \delta_k + (f_{Ak} - f_{Bk})/2 \). The \( \hat{d}_k \) have a bivariate normal
distribution as described in equation (4.4).

The p-value for each of the statistics \( \hat{d}_k \) is calculated following the two 1-
sided tests procedure (equation 3.8), where Schuirmann’s \( t_{k1} = \frac{\hat{d}_k - \Delta_1}{SE(d_k)} \), \( t_{k2} = \frac{\Delta_2 - \hat{d}_k}{SE(d_k)} \), and \( SE(d_k) = \sqrt{\frac{\sigma_{ek}^2}{2(n_A + n_B)}} \). Since the variance \( \sigma_{ek}^2 \) is assumed known,
the tests are applied to a z distribution instead of a t. The bivariate p-value for
\( \hat{d}_1 \) and \( \hat{d}_2 \) is the maximum of the four univariate 1-sided p-values.

Power is simulated by repeating the above steps many times and calculating
the percent of repetitions in which \( H_0 \) is rejected (p-value \( \leq \alpha \)). Since the
bivariate p-value is based on the univariate p-values, an estimate of the two
univariate powers is also obtained.

The accuracy of the simulation is controlled by the number of repetitions, \( r \). If rejection of the null hypothesis follows a binomial distribution with \( \hat{\Omega} \), the estimated power, is \( SE = \sqrt{\frac{\hat{\Omega}(1-\hat{\Omega})}{r}} \). For most situations, 1,000 repetitions lead to a reasonable standard error. For instance, at \( \hat{\Omega}=.90 \), \( SE = 0.01 \). A smaller \( SE \) is desirable for extreme values of \( \hat{\Omega} \), especially for values less than \( \alpha \). With \( r=10,000 \) and \( \hat{\Omega}=0.05 \), \( SE=0.002 \). Simulation repetitions in this work are always 1,000 for any \( \hat{\Omega}>0.05 \), and 10,000 for \( \hat{\Omega} \leq 0.05 \).

4.2.5 Power Estimation from Simulation Assuming Unknown Variance

The estimated sample variance is in fact a random variable, and when the sample size is small this variation should be taken into account. As discussed above, this work addresses small-sample power via simulation.

Steps for simulation are similar to those described in 4.2.4. The main difference is that instead of simulating the sequence group averages, each subject’s observations are simulated with random variates, so that the sample variance can be estimated. The process for each repetition begins by generating \( 2n_A + 2n_B \) random standardized normal deviates which represent subjects’ observed errors for each sequence-variable combination. Arrange them in vectors \( z_{A1}, z_{A2}, z_{B1}, z_{B2} \). Apply a Cholesky decomposition as described above to the error vectors for each sequence to obtain vectors \( f_{A1}, f_{A2}, f_{B1}, f_{B2} \) of variates with the correct variance-covariance structure of the period I - period II difference for each subject. In the Cholesky decomposition,

\[
\begin{align*}
    a &= 2\sigma_{e2}^2 \\
    b &= 2\rho_0\sigma_{e1}\sigma_{e2} \\
    c &= 2\sigma_{e2}^2.
\end{align*}
\]

Once the errors have the correct covariance matrix, the mean \( \bar{f}_{hh} \) of the transformed errors for each sequence-variable combination is calculated. The
pooled-sequence sample variance for each outcome is calculated as:

$$\widehat{\sigma}_{ek}^2 = \frac{\sum_{h=i}^{g} \sum_{i=1}^{n_h} (f_{hik} - \bar{f}_{hh})^2}{2(n_A + n_B - 2)}$$

The division by 2 is necessary to obtain an estimate of $\sigma_{ek}^2$ instead of $2\sigma_{ek}^2$. The test statistics are obtained by calculating $\widehat{d}_k = \delta_k + (\bar{T}_{Ak} - \bar{T}_{Bk})/2$. The standard error of the test statistic for variable $k$ is equal to $SE(\widehat{d}_k) = \sqrt{\frac{1}{2(n_A + 1)} \widehat{\sigma}_{ek}^2}$. The $\widehat{d}_k$ have a bivariate normal distribution as described in equation (4.4).

The remainder of the simulation steps are identical, except that the two 1-sided tests are evaluated relative to the $t$ distribution with $(n_A + n_B - 2)$ degrees of freedom. Simulation assuming unknown variance is used in this work whenever the total sample size $N$ is less than 30.

4.2.6 Application and Results

Bivariate power as described previously is provided for a variety of situations in Table 4.1 to Table 4.4. Power is calculated for $\alpha = .05$ and .025 (corresponding to 90% and 95% confidence intervals), for balanced designs with sample size per sequence group of 10 and 20. The within-subject correlation $\rho_e$ of $\ln(x_1)$ and $\ln(x_2)$ is fixed at 0.85, 0.75, 0.50, and 0.25, and power is evaluated for a variety of within-subject standard errors of log $x$ ($\sigma_{e1}$ and $\sigma_{e2}$) and treatment geometric mean ratios ($\theta_k = e^{\delta_k}$) for two responses.

Power is defined as the probability of rejecting the null hypothesis; if $H_0$ is actually true, then the power is an estimate of the type I error. When $\delta_k$ is equal to $\Delta_1$ or $\Delta_2$, then the corresponding univariate power is expected to equal $\alpha$. Univariate power is symmetric with respect to $\delta = \ln(\theta)$, so for the most part only $\theta$s of 1.0 and greater are examined. However, bivariate power is not symmetric if one $\theta$ is below 1.0, and one is above. Tables 4.1-4.4 are designed to examine characteristics of power and may have limited use as a tool for bivariate power.
calculation, since many possibilities are not listed.

For comparison, the corresponding univariate power is presented in Table 4.5. Univariate power was calculated from numerical integration for known variance, and simulation otherwise. It can also be evaluated exactly, as by Phillips (90 A) using numerical integration of a bivariate noncentral t distribution for the two Schuirmann test statistics, or by using the approximation by Hauschke et al (92), derived from the sample size formula (equation 3.28), which assumes a central t distribution. All methods yield similar results.

The approximation by Hauschke is a bit conservative, sometimes leading to an estimated power a few percentage points less than other methods (Hauschke 92). Occasionally, however, the approximation is anti-conservative. Discrepancies between the approximation and simulation/numerical integration methods appear for a fixed sample size as the standard deviation increases and the treatment geometric mean ratio is midway between full acceptance ($\theta=1$) and rejection ($\theta=1.25$). The worst discrepancy encountered in Table 4.5 is for $\alpha=0.025$, $n$ per sequence group=10, $\sigma_e=0.30$, and $\theta=1.1$. Simulation reported a power of 0.14, and Hauschke's formula calculated a power of 0.23. The power at the anti-conservative discrepancies is in a range around 0.20; such situations are generally of little interest when planning studies.

It is of academic interest to compare the power of an equivalence test versus a standard test of treatment efficacy. A univariate test of $H_0: \theta=1$ vs. $H_A: \theta \neq 1$ has increasing power as $\theta$ gets farther away from 1.0 for a fixed sample size and variance. Equivalence power increases as $\theta$ gets closer to 1.0. As an example, when $\sigma_e=0.15$, $n$ per sequence group=10, and $\alpha=0.05$, the power for equivalence at $\theta=1.1$ is 0.83, and for $\theta=1.2$ is 0.21. Power for an efficacy test under the same situation is 0.52 and 0.97. Whether the test for a difference or equivalence will be more powerful depends on the true treatment geometric mean ratios.
Table 4.1
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 90% Confidence Intervals
Cross-over Design, N per Sequence Group = 10, Equivalence Range = 0.8-1.25
Obtained via Simulation

<table>
<thead>
<tr>
<th>$\theta_1$, $\theta_2 \rightarrow$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.0</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{e1}$</td>
<td>$\sigma_{e2}$</td>
<td>$\rho_e$</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>0.85</td>
<td>0.99</td>
<td>0.83</td>
<td>0.77</td>
<td>0.59</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.99</td>
<td>0.82</td>
<td>0.75</td>
<td>0.58</td>
<td>0.21</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.99</td>
<td>0.83</td>
<td>0.72</td>
<td>0.59</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.99</td>
<td>0.83</td>
<td>0.70</td>
<td>0.61</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>0.15</td>
<td>0.20</td>
<td>0.85</td>
<td>0.91</td>
<td>0.62</td>
<td>0.61</td>
<td>0.38</td>
<td>0.15</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.91</td>
<td>0.62</td>
<td>0.59</td>
<td>0.39</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.91</td>
<td>0.62</td>
<td>0.56</td>
<td>0.41</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.91</td>
<td>0.63</td>
<td>0.54</td>
<td>0.45</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td>0.15</td>
<td>0.25</td>
<td>0.85</td>
<td>0.70</td>
<td>0.45</td>
<td>0.44</td>
<td>0.24</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.69</td>
<td>0.44</td>
<td>0.43</td>
<td>0.25</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.71</td>
<td>0.44</td>
<td>0.40</td>
<td>0.29</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.71</td>
<td>0.46</td>
<td>0.39</td>
<td>0.32</td>
<td>0.14</td>
<td>0.05</td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>0.85</td>
<td>0.87</td>
<td>0.58</td>
<td>0.54</td>
<td>0.22</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.86</td>
<td>0.58</td>
<td>0.51</td>
<td>0.24</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.84</td>
<td>0.58</td>
<td>0.45</td>
<td>0.27</td>
<td>0.14</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.83</td>
<td>0.58</td>
<td>0.43</td>
<td>0.32</td>
<td>0.14</td>
<td>0.044</td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>0.85</td>
<td>0.68</td>
<td>0.42</td>
<td>0.42</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.67</td>
<td>0.41</td>
<td>0.39</td>
<td>0.14</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.66</td>
<td>0.41</td>
<td>0.34</td>
<td>0.18</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.65</td>
<td>0.42</td>
<td>0.31</td>
<td>0.22</td>
<td>0.12</td>
<td>0.038</td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>0.85</td>
<td>0.46</td>
<td>0.28</td>
<td>0.28</td>
<td>0.08</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.46</td>
<td>0.27</td>
<td>0.26</td>
<td>0.09</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.43</td>
<td>0.27</td>
<td>0.23</td>
<td>0.12</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.42</td>
<td>0.27</td>
<td>0.20</td>
<td>0.14</td>
<td>0.10</td>
<td>0.031</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>0.85</td>
<td>0.58</td>
<td>0.32</td>
<td>0.37</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.56</td>
<td>0.32</td>
<td>0.33</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.53</td>
<td>0.32</td>
<td>0.27</td>
<td>0.11</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.51</td>
<td>0.32</td>
<td>0.24</td>
<td>0.15</td>
<td>0.09</td>
<td>0.030</td>
</tr>
</tbody>
</table>

$\sigma_{ek} = SD(\ln(x_k))$

$\rho_e = corr(\ln(x_{11}) - \ln(x_{11}), \ln(x_{12}) - \ln(x_{12}))$
Table 4.2
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 90% Confidence Intervals
Cross-over Design, N per Sequence Group = 20, Equivalence Range = 0.8-1.25
Obtained via Numerical Integration

<table>
<thead>
<tr>
<th>$\theta_1, \theta_2 \rightarrow$</th>
<th>$\sigma_{e_1}$</th>
<th>$\sigma_{e_2}$</th>
<th>$\rho_e$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.2</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.20</td>
<td>0.20</td>
<td>.85</td>
<td>.99</td>
<td>.89</td>
<td>.85</td>
<td>.73</td>
<td>.23</td>
<td>.16</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.99</td>
<td>.89</td>
<td>.83</td>
<td>.73</td>
<td>.23</td>
<td>.14</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.99</td>
<td>.89</td>
<td>.81</td>
<td>.73</td>
<td>.23</td>
<td>.11</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.99</td>
<td>.89</td>
<td>.80</td>
<td>.73</td>
<td>.23</td>
<td>.08</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.25</td>
<td>.85</td>
<td>.98</td>
<td>.74</td>
<td>.73</td>
<td>.58</td>
<td>.18</td>
<td>.14</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.98</td>
<td>.74</td>
<td>.72</td>
<td>.58</td>
<td>.18</td>
<td>.12</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.98</td>
<td>.74</td>
<td>.69</td>
<td>.59</td>
<td>.18</td>
<td>.09</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.98</td>
<td>.74</td>
<td>.67</td>
<td>.60</td>
<td>.18</td>
<td>.06</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.30</td>
<td>.85</td>
<td>.91</td>
<td>.60</td>
<td>.60</td>
<td>.44</td>
<td>.15</td>
<td>.12</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.91</td>
<td>.60</td>
<td>.59</td>
<td>.45</td>
<td>.15</td>
<td>.11</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.91</td>
<td>.60</td>
<td>.57</td>
<td>.46</td>
<td>.15</td>
<td>.08</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.91</td>
<td>.60</td>
<td>.55</td>
<td>.48</td>
<td>.15</td>
<td>.05</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.25</td>
<td>.85</td>
<td>.97</td>
<td>.73</td>
<td>.67</td>
<td>.42</td>
<td>.17</td>
<td>.12</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.97</td>
<td>.73</td>
<td>.65</td>
<td>.42</td>
<td>.17</td>
<td>.11</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.96</td>
<td>.73</td>
<td>.61</td>
<td>.45</td>
<td>.17</td>
<td>.07</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.96</td>
<td>.73</td>
<td>.57</td>
<td>.47</td>
<td>.18</td>
<td>.05</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.30</td>
<td>.85</td>
<td>.90</td>
<td>.59</td>
<td>.57</td>
<td>.29</td>
<td>.14</td>
<td>.11</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.90</td>
<td>.59</td>
<td>.55</td>
<td>.30</td>
<td>.14</td>
<td>.09</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.90</td>
<td>.59</td>
<td>.51</td>
<td>.34</td>
<td>.14</td>
<td>.06</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.89</td>
<td>.59</td>
<td>.48</td>
<td>.37</td>
<td>.15</td>
<td>.04</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>0.30</td>
<td>.85</td>
<td>.87</td>
<td>.56</td>
<td>.52</td>
<td>.18</td>
<td>.11</td>
<td>.10</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.85</td>
<td>.56</td>
<td>.49</td>
<td>.20</td>
<td>.11</td>
<td>.08</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.84</td>
<td>.55</td>
<td>.44</td>
<td>.25</td>
<td>.13</td>
<td>.06</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.83</td>
<td>.55</td>
<td>.40</td>
<td>.29</td>
<td>.13</td>
<td>.04</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>0.40</td>
<td>.85</td>
<td>.60</td>
<td>.36</td>
<td>.36</td>
<td>.07</td>
<td>.08</td>
<td>.08</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.75</td>
<td>.59</td>
<td>.36</td>
<td>.34</td>
<td>.09</td>
<td>.08</td>
<td>.07</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>.57</td>
<td>.35</td>
<td>.30</td>
<td>.14</td>
<td>.10</td>
<td>.05</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>.55</td>
<td>.35</td>
<td>.26</td>
<td>.18</td>
<td>.10</td>
<td>.03</td>
<td>.04</td>
<td></td>
</tr>
</tbody>
</table>

$\sigma_{ek} = \text{SD}(\ln(x_k))$

$\rho_e = \text{corr}(\ln(x_{I1}) - \ln(x_{II1}), \ln(x_{I2}) - \ln(x_{II2}))$
Table 4.3
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 95% Confidence Intervals
Cross-over Design, N per Sequence Group = 10, Equivalence Range = 0.8-1.25
Obtained via Simulation

<table>
<thead>
<tr>
<th>$\theta_1, \theta_2 \rightarrow$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{e1}$</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
<td>1.25</td>
</tr>
<tr>
<td>$\sigma_{e2}$</td>
<td>.75</td>
<td>.73</td>
<td>.63</td>
<td>.41</td>
<td>.12</td>
<td>.07</td>
<td>.026</td>
</tr>
<tr>
<td>$\rho_e$</td>
<td>.5</td>
<td>.72</td>
<td>.59</td>
<td>.42</td>
<td>.13</td>
<td>.05</td>
<td>.024</td>
</tr>
<tr>
<td>.25</td>
<td>.98</td>
<td>.73</td>
<td>.56</td>
<td>.44</td>
<td>.14</td>
<td>.03</td>
<td>.025</td>
</tr>
<tr>
<td>0.15</td>
<td>.85</td>
<td>.98</td>
<td>.72</td>
<td>.66</td>
<td>.39</td>
<td>.12</td>
<td>.09</td>
</tr>
<tr>
<td>.75</td>
<td>.72</td>
<td>.63</td>
<td>.41</td>
<td>.12</td>
<td>.07</td>
<td>.026</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.59</td>
<td>.42</td>
<td>.13</td>
<td>.05</td>
<td>.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.66</td>
<td>.39</td>
<td>.28</td>
<td>.10</td>
<td>.022</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>.85</td>
<td>.81</td>
<td>.47</td>
<td>.46</td>
<td>.18</td>
<td>.09</td>
<td>.07</td>
</tr>
<tr>
<td>.75</td>
<td>.47</td>
<td>.45</td>
<td>.20</td>
<td>.09</td>
<td>.06</td>
<td>.026</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.46</td>
<td>.24</td>
<td>.09</td>
<td>.04</td>
<td>.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.48</td>
<td>.38</td>
<td>.28</td>
<td>.10</td>
<td>.022</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>.85</td>
<td>.53</td>
<td>.28</td>
<td>.28</td>
<td>.10</td>
<td>.07</td>
<td>.06</td>
</tr>
<tr>
<td>.75</td>
<td>.28</td>
<td>.28</td>
<td>.11</td>
<td>.07</td>
<td>.05</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.28</td>
<td>.14</td>
<td>.07</td>
<td>.03</td>
<td>.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.30</td>
<td>.23</td>
<td>.17</td>
<td>.08</td>
<td>.018</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>.85</td>
<td>.74</td>
<td>.40</td>
<td>.39</td>
<td>.07</td>
<td>.05</td>
<td>.06</td>
</tr>
<tr>
<td>.75</td>
<td>.40</td>
<td>.36</td>
<td>.09</td>
<td>.05</td>
<td>.05</td>
<td>.010</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.39</td>
<td>.30</td>
<td>.14</td>
<td>.07</td>
<td>.03</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.40</td>
<td>.26</td>
<td>.17</td>
<td>.08</td>
<td>.016</td>
<td>.019</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>.85</td>
<td>.51</td>
<td>.24</td>
<td>.25</td>
<td>.04</td>
<td>.04</td>
<td>.05</td>
</tr>
<tr>
<td>.75</td>
<td>.24</td>
<td>.24</td>
<td>.05</td>
<td>.05</td>
<td>.04</td>
<td>.013</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.25</td>
<td>.20</td>
<td>.07</td>
<td>.05</td>
<td>.022</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.25</td>
<td>.17</td>
<td>.10</td>
<td>.06</td>
<td>.013</td>
<td>.018</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>.85</td>
<td>.24</td>
<td>.12</td>
<td>.13</td>
<td>.019</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>.75</td>
<td>.12</td>
<td>.11</td>
<td>.03</td>
<td>.03</td>
<td>.03</td>
<td>.012</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.11</td>
<td>.09</td>
<td>.04</td>
<td>.04</td>
<td>.016</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.13</td>
<td>.08</td>
<td>.05</td>
<td>.04</td>
<td>.008</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>.85</td>
<td>.38</td>
<td>.14</td>
<td>.20</td>
<td>.016</td>
<td>.017</td>
<td>.04</td>
</tr>
<tr>
<td>.75</td>
<td>.14</td>
<td>.18</td>
<td>.024</td>
<td>.022</td>
<td>.03</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.14</td>
<td>.14</td>
<td>.04</td>
<td>.03</td>
<td>.018</td>
<td>.009</td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.14</td>
<td>.11</td>
<td>.06</td>
<td>.04</td>
<td>.011</td>
<td>.012</td>
<td></td>
</tr>
</tbody>
</table>

$\sigma_{ek} = \text{SD}(\ln(x_k))$

$\rho_e = \text{corr}(\ln(x_{I1}) - \ln(x_{I2}), \ln(x_{I1}) - \ln(x_{I2}))$
Table 4.4
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 95% Confidence Intervals
Cross-over Design, N per Sequence Group= 20, Equivalence Range= 0.8-1.25
Obtained via Numerical Integration

<table>
<thead>
<tr>
<th>$\sigma_{e1}$</th>
<th>$\sigma_{e2}$</th>
<th>$\rho_e$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.2</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.85</td>
<td>.99</td>
<td>.81</td>
<td>.76</td>
<td>.56</td>
<td>.15</td>
<td>.10</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.99</td>
<td>.81</td>
<td>.74</td>
<td>.57</td>
<td>.15</td>
<td>.08</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.99</td>
<td>.81</td>
<td>.71</td>
<td>.58</td>
<td>.15</td>
<td>.06</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.99</td>
<td>.81</td>
<td>.68</td>
<td>.59</td>
<td>.15</td>
<td>.03</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.85</td>
<td>.96</td>
<td>.63</td>
<td>.61</td>
<td>.38</td>
<td>.11</td>
<td>.08</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.96</td>
<td>.63</td>
<td>.60</td>
<td>.39</td>
<td>.11</td>
<td>.07</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.96</td>
<td>.63</td>
<td>.57</td>
<td>.41</td>
<td>.11</td>
<td>.04</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.96</td>
<td>.63</td>
<td>.54</td>
<td>.44</td>
<td>.11</td>
<td>.03</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.85</td>
<td>.83</td>
<td>.48</td>
<td>.47</td>
<td>.24</td>
<td>.09</td>
<td>.07</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.83</td>
<td>.47</td>
<td>.46</td>
<td>.26</td>
<td>.09</td>
<td>.06</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.83</td>
<td>.47</td>
<td>.44</td>
<td>.29</td>
<td>.09</td>
<td>.04</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.83</td>
<td>.47</td>
<td>.41</td>
<td>.33</td>
<td>.09</td>
<td>.024</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.85</td>
<td>.94</td>
<td>.61</td>
<td>.54</td>
<td>.21</td>
<td>.09</td>
<td>.07</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.93</td>
<td>.61</td>
<td>.52</td>
<td>.23</td>
<td>.09</td>
<td>.06</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.92</td>
<td>.60</td>
<td>.47</td>
<td>.27</td>
<td>.10</td>
<td>.04</td>
<td>.021</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.92</td>
<td>.60</td>
<td>.43</td>
<td>.31</td>
<td>.10</td>
<td>.023</td>
<td>.023</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.30</td>
<td>.85</td>
<td>.83</td>
<td>.46</td>
<td>.44</td>
<td>.11</td>
<td>.07</td>
<td>.06</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.82</td>
<td>.46</td>
<td>.42</td>
<td>.13</td>
<td>.07</td>
<td>.05</td>
<td>.016</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.80</td>
<td>.46</td>
<td>.38</td>
<td>.18</td>
<td>.08</td>
<td>.03</td>
<td>.021</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.80</td>
<td>.46</td>
<td>.34</td>
<td>.23</td>
<td>.08</td>
<td>.019</td>
<td>.023</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.30</td>
<td>.85</td>
<td>.76</td>
<td>.39</td>
<td>.39</td>
<td>.05</td>
<td>.04</td>
<td>.05</td>
<td>.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.74</td>
<td>.39</td>
<td>.36</td>
<td>.07</td>
<td>.05</td>
<td>.04</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.71</td>
<td>.39</td>
<td>.31</td>
<td>.12</td>
<td>.06</td>
<td>.03</td>
<td>.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.69</td>
<td>.39</td>
<td>.26</td>
<td>.16</td>
<td>.07</td>
<td>.016</td>
<td>.019</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.40</td>
<td>.85</td>
<td>.40</td>
<td>.20</td>
<td>.22</td>
<td>.01</td>
<td>.022</td>
<td>.04</td>
<td>.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td></td>
<td>.39</td>
<td>.20</td>
<td>.20</td>
<td>.02</td>
<td>.03</td>
<td>.03</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td></td>
<td>.36</td>
<td>.20</td>
<td>.17</td>
<td>.05</td>
<td>.04</td>
<td>.019</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.34</td>
<td>.20</td>
<td>.14</td>
<td>.08</td>
<td>.05</td>
<td>.011</td>
<td>.018</td>
<td></td>
</tr>
</tbody>
</table>

$\sigma_{ek} = SD(\ln(x_k))$
$\rho_e = corr(\ln(x_{I_1}) - \ln(x_{I_2}), \ln(x_{I_2}) - \ln(x_{I_2}))$
Table 4.5
Power for Testing Equivalence of a Log-Normal Response
For Method of 100(1-2α)% Confidence Intervals
Cross-over Design, Equivalence Range= 0.8-1.25
Obtained via simulation (n per sequence group=10)
and numerical integration (n per sequence group =20)

<table>
<thead>
<tr>
<th>α</th>
<th>n per seq. group</th>
<th>σ_e</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.2</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.15</td>
<td></td>
<td>.99</td>
<td>.83</td>
<td>.76</td>
<td>.21</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td></td>
<td>.91</td>
<td>.62</td>
<td>.56</td>
<td>.16</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.70</td>
<td>.45</td>
<td>.42</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td></td>
<td>.46</td>
<td>.29</td>
<td>.28</td>
<td>.11</td>
<td>.05</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td></td>
<td>.99</td>
<td>.89</td>
<td>.84</td>
<td>.23</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.98</td>
<td>.74</td>
<td>.68</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td></td>
<td>.91</td>
<td>.60</td>
<td>.54</td>
<td>.15</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td></td>
<td>.60</td>
<td>.39</td>
<td>.35</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>0.025</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.15</td>
<td></td>
<td>.99</td>
<td>.73</td>
<td>.65</td>
<td>.13</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td></td>
<td>.81</td>
<td>.48</td>
<td>.43</td>
<td>.09</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.52</td>
<td>.29</td>
<td>.27</td>
<td>.08</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td></td>
<td>.25</td>
<td>.14</td>
<td>.13</td>
<td>.05</td>
<td>.022</td>
</tr>
<tr>
<td>0.025</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td></td>
<td>.99</td>
<td>.82</td>
<td>.75</td>
<td>.15</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td></td>
<td>.96</td>
<td>.63</td>
<td>.56</td>
<td>.11</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td></td>
<td>.83</td>
<td>.48</td>
<td>.42</td>
<td>.09</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td></td>
<td>.41</td>
<td>.24</td>
<td>.22</td>
<td>.06</td>
<td>.024</td>
</tr>
</tbody>
</table>

σ_e = SD(ln(x))

When planning a study, estimates of ρ_e and σ_e are sometimes available for x_1 and x_2 on the standard scale, but not the log scale. If estimates are obtained from previous studies for which subject-wise data is available, the log-scale estimates can be obtained by simply log-transforming the data, and calculating ̂ρ_e and ̂σ_e. The correlation between variables for a cross-over study is calculated as the within sequence groups correlation between variables for the period I-period II difference, and the within-subject standard deviation is estimated as the within sequence groups standard deviation of the period I-period II difference divided by √2, or by
the mean square error of the appropriate cross-over ANOVA model. If estimates
from previous studies are reported on the observed scale rather than the log scale
and subject-wise data is not available, then a conversion formula of $\sigma_e$ and $\rho_e$ from
the standard scale to the log scale is needed in order to use Tables 4.1-4.5.

A transformation of $\sigma_e$ can be applied by appealing to the moment generating
function (mgf) of the normal distribution. If $f = \ln(g) \sim N(\mu, \sigma^2)$, then $E(e^{\ln(g)}) =
e^{\mu + \frac{1}{2}\sigma^2}$. Since $g = e^{\ln(g)}$ then $E(g) = e^{\mu + \frac{1}{2}\sigma^2}$, and $\text{var}(g)$ found by $E(g^2) - E(g)^2$ is
$(e^{2\mu + \sigma^2})(e^{\sigma^2} - 1)$. The median of $g$ is $e^\mu$. For cross-over studies, $\sigma_e$ is best
evaluated with $f = [\ln(x)_{II} - \ln(x)_{III}]$, which on the standard scale is the period ratio
g = $x_{II}/x_{III}$. For any particular sequence group or outcome:

$$f = \ln(x)_{III} - \ln(x)_{II} \sim N(\gamma_k + \delta_k\zeta_h, 2\sigma^2_e)$$

$$E(g_{hk}) = e^{\gamma_k + \delta_k\zeta_h + \sigma^2_e}$$

$$SD(g_{hk}) = \sqrt{e^{2(\gamma_k + \delta_k\zeta_h)^2 + 2\sigma^2_e}}(e^{2\sigma^2_e} - 1).$$

A good approximation for $\sigma_e$ is the coefficient of variation $CV$ of the period ratios,
$\frac{SD(g)}{E(g)} = \sqrt{e^{2\sigma^2_e} - 1}$. As $\sigma_e \to 0$, the $CV$ approximately equals $\sqrt{2}\sigma_e$. Unfortunately,
most studies do not evaluate the $CV$ of period ratios. An alternative more
practical strategy is to consider one observation per subject, $x$. If subject effects
are assumed fixed, then $CV(x) = \sqrt{e^{2\sigma^2_e}} - 1$. In Tables 4.1-4.4, $CV(x)$ approximates
$\sigma_e$ correctly to within two decimal places for $\sigma_e$ of 0.15, 0.20, and 0.25.
Approximations for $\sigma_e = 0.30$ and 0.40 are 0.31 and 0.42 respectively. Use of the
CV to approximate $\sigma_e$ may yield slightly conservative power estimates for large
variation, but is practical for most applications.

For bivariate power, an estimate of $\rho_e$, the correlation between period I -
period II differences on the log scale is hard to obtain from estimates on the
observed scale when subject-wise data is unavailable. For known population
parameters, the corresponding correlation between the period1/period2 ratio for $x_1$
and $x_2$ can be found by applying the moment generating function of the
multivariate normal distribution $E(e^{t\ln(x)}) = (e^{t\mu + \frac{1}{2} t^2 \Sigma})$. Correlation for $g_1 = \frac{x_{I1}}{x_{II1}}$ and $g_2 = \frac{x_{II2}}{x_{II1}}$ is $\frac{\text{cov}(g_1, g_2)}{\sqrt{\text{var}(g_1)\text{var}(g_2)}}$, where $\text{cov}(g_1, g_2) = E(g_1g_2) - E(g_1)E(g_2)$. $E(g_1g_2)$ is found using the bivariate mgf, and the correlation on the standard scale is found by substituting appropriate functions into the equation and simplifying:

$$\text{corr}(x_{I1}, x_{II2}) = \frac{(e^{2\sigma_{e1}^2 - 1})}{\sqrt{(e^{2\sigma_{e1}^2 - 1})(e^{2\sigma_{e2}^2 - 1})}}. \tag{4.9}$$

In a population, this formula can be used to convert the $\rho_e$ in tables 4.1-4.4 from the period I-II difference of log scale to that for the period ratio on the standard scale. However, the transformation is a non-linear function of parameters, and produces a biased estimate even in large samples. If log scale information is unavailable, but the correlation of period ratios and $\text{CV}(x_k)$ are available, then very rough estimates of $\rho_e$ may be obtained by finding a range of $\rho_e$ corresponding to a confidence interval around the estimated period-ratio correlation. It is unlikely that correlation of period ratios is available from a published study, so this discussion is primarily of academic interest.

### 4.2.7 Conclusions

By examining Tables 4.1-4.5, several patterns in the bivariate equivalence power become apparent. Evaluating the test at or near $H_0$ shows that the multiple univariate test procedure becomes conservative (leading to type I errors of less than $\alpha$) as the variance increases relative to the sample size and the correlation increases. This is not a surprising result, since this multivariate test does not take into account the association between outcomes. Conservativeness arises for larger variances because of the increased role correlation plays on power in this situation (see below).

For fixed variances, as the correlation becomes larger, the power increases for responses with treatment geometric mean ratios on the same side of one, and
decreases for responses with opposite treatment effects. The amount of change is greater for geometric mean ratios midway between full acceptance (1.0) and rejection (1.25). When the treatment geometric mean ratios are on opposite sides of 1.0 and the correlation is positive, the rule of thumb of assuming the correlation is zero and multiplying univariate powers actually identifies the maximum bivariate power, rather than the minimum. The same situation arises if the geometric mean ratios are on the same side of 1.0, but the correlation is negative. If the median ratios are on opposite sides of 1.0, and the correlation is negative, Tables 4.1-4.4 can be used to estimate power by first taking the reciprocal of the smaller ratio, and changing the sign of the correlation. In general, the effect of correlation on bivariate power is stronger as the power decreases.

Insight into bivariate power can be obtained through graphical examination. For a fixed sample size and covariance, the bivariate power follows a 3-dimensional oblong-bell shaped surface as $\delta_1$ and $\delta_2$ vary (the surface on the exponentiated $\theta$ scale is additionally warped due to the log transform). As viewed from above with topographic lines representing power, the surface is represented by concentric ellipses (see Figure 4.1). When the correlation is positive, the power is greater for sets of $\delta$s which have the same sign than those with opposite sign. The power curve is flattened in opposite-sign quadrants, and elongated in same-sign quadrants. This effect is reversed if the correlation is negative. The amount of elongation depends on the correlation. If variance and equivalence criteria for the two responses are equal power is completely bell shaped when the correlation is zero (concentric circles in the topographic view), and pinched to a single plane when the correlation is $\pm 1.0$. When the variances for the two outcomes are equal, the axis of the concentric ellipses is in the plane $\delta_1=\delta_2$. As the variances become more dissimilar, the axis of the ellipse shifts towards the Cartesian axis with the smaller variance.
Figure 4.1
Bivariate Power Density

Zero Correlation

Positive Strong Correlation, Equal Variances

Negative Weak Correlation, $\text{Var}(\delta) < \text{Var}(\delta_1)$
If the power surface is sliced along the plane where $\delta_1=0$, the resulting curve is symmetrical bell-shaped with maximum power at zero. For similar slices not along zero, the resulting curve is still symmetrical, but is shifted towards the direction of the slice. For instance, when variances are equal, a slice in the plane where $\delta_1=0.095$ may yield a bell curve with maximum power at $\delta_2=0.03$.

For a fixed sample size, as the variances increase, the bivariate power decreases, and the oblong-bell surface has less height. In such situations, the elongation due to the correlation is more pronounced, and changes in the shape between different correlation values are larger.

The effect of correlation can also be evaluated by a 2-dimensional plot of the bivariate power against the correlation for fixed sample size, variance, and $\delta$s. Consider the situation where both $\delta$s are positive. If the correlation is positive, the maximum bivariate power is equal to or less than the minimum univariate power, as described in Section 4.2.3. The minimum is the product of the univariate powers, and the power at correlation of 0.25, 0.50, 0.75, and 0.85 are obtained from Tables 4.1-4.4. When the correlation has little effect on power, the plot is nearly a horizontal line, and the product of the univariate powers is not much smaller than the minimum univariate power. When a positive correlation is very influential, the plot is an increasing slightly concave curve. If the power at correlation of 0.5 was estimated as halfway between the minimum univariate power (an estimate for corr=1) and the product of the two univariate powers (assuming corr is zero), the estimate would be a bit anti-conservative. For negative correlations, the maximum power is estimated by the product of the univariates, but the minimum power can not be approximated with a rule of thumb.

It is concluded that correlation plays an important role in bivariate power when each univariate power is moderate to low. When planning a study,
approximating the bivariate power as the product of the univariate powers is a reasonable strategy, but can be conservative. If the correlation is positive and $\delta$s are expected to have the same sign, the estimate can be somewhat improved by interpolating between the estimated maximum and minimum powers.

This evaluation also demonstrated that multiple univariate tests can be quite conservative when the correlation is high and variances are large, with a true type-I error less than $\alpha$.

4.2.8 Example

Bivariate equivalence is illustrated through the comparisons of test and reference formulations of a psychotropic drug described in Section 3.4. Area under the curve (AUC) and maximum concentration ($C_{\text{max}}$) are well described by a log-normal distribution, and have no significant period effects. The 90% confidence interval for the ratio of test/reference geometric means for AUC and $C_{\text{max}}$ are (0.83, 1.09) and (0.97, 1.22) respectively. The p-values from the two 1-sided test procedure are $p=0.019$ and $p=0.024$ (see Tables 3.1 and 3.2). From the cross-over ANOVA model, the mean square errors were found to be 0.251 and 0.216. The geometric mean treatment ratios are 0.950 and 1.085. Sample size is 11 subjects in sequence A and 10 subjects in sequence B. The correlation of period I- period II differences on the log scale is 0.496.

From simulation using the post-hoc estimates, the univariate powers are 0.65 and 0.66 respectively, and bivariate power is 0.39. If the sample represents the true distribution, the study had a 39% chance of concluding equivalence for both measures. Note that 0.39 is less than $0.65 \times 0.66 = 0.43$ because the treatment geometric mean ratios are on opposite sides of 1.0, and the correlation is positive. The maximum power estimate of 0.43 is a reasonable approximation to the true power in this example.
4.3 Parallel Bivariate Normal Distribution

Although most bioequivalence trials are conducted with a cross-over design, those involving measures with small between-subject variation could be conducted with parallel designs. Parallel designs are also more common in clinical equivalence trials, for instance the study described in Section 1.4.2 comparing doses of ramipril and HCTZ for treatment of high blood pressure.

Some clinical equivalence trials are concerned with demonstrating only the one-sided alternative hypothesis that the test treatment is no worse than the reference. This is not appropriate, however, for many biologic outcomes for which too strong a response would be harmful (i.e., blood pressure, glucose levels). Many continuous-response clinical equivalence trials fall within the two-sided testing scenario, and so this approach is evaluated in this section.

As with the cross-over design, a statistical model assuming a log-normal distribution is developed. The distribution of the appropriate test statistic is derived, and the power is evaluated via numerical integration and simulation.

4.3.1 Statistical Model

Assume \( \ln(x) \) follows a normal distribution, and \( \ln(x_1) \) and \( \ln(x_2) \) follow a bivariate normal, where \( x_1 \) and \( x_2 \) are two response variables (for example difference from baseline for sitting and standing systolic blood pressure). The response can be modeled as follows:

\[
\ln(x)_{gik} = \mu_k + \tau_{gk} + s_{ik} + e_{ik}
\]  

(4.10)

where \( g = \) treatment (\( t, r \))

\( i = \) subject 1...\( N \) or 1...\( n_t \) and 1...\( n_r \)

\( k = \) variable 1, 2

\( \mu_k = \) fixed response mean for the \( k \)th variable
\[ \tau_{gk} = \text{fixed treatment effect for the kth variable} \]
\[ s_{ik} = \text{random subject effect for the kth variable of the ith subject} \]
\[ e_{ik} = \text{random error for the kth variable of the ith subject} \]

The following assumptions are imposed on this model:

- subject effects in treatment groups are independent and identically distributed (IID)
- errors in treatment groups are IID
- subjects are independent of each other
- for each subject, subject effects and errors are independent

for each subject:
\[
\begin{bmatrix}
    s_{i1} \\
    s_{i2}
\end{bmatrix} \sim N\left(\begin{bmatrix}
    0 \\
    0
\end{bmatrix}, \begin{bmatrix}
    \sigma_{s1}^2 & \rho_{s1s2} \sigma_{s1} \sigma_{s2} \\
    \rho_{s1s2} \sigma_{s1} \sigma_{s2} & \sigma_{s2}^2
\end{bmatrix}\right) = \Sigma_s
\]

for each subject:
\[
\begin{bmatrix}
    e_{i1} \\
    e_{i2}
\end{bmatrix} \sim N\left(\begin{bmatrix}
    0 \\
    0
\end{bmatrix}, \begin{bmatrix}
    \sigma_{e1}^2 & \rho_{e1e2} \sigma_{e1} \sigma_{e2} \\
    \rho_{e1e2} \sigma_{e1} \sigma_{e2} & \sigma_{e2}^2
\end{bmatrix}\right) = \Sigma_e
\]

4.3.2 Obtaining the Test Statistic

Steps involved in obtaining the bivariate equivalence test statistic and its distribution are described below.

1. Average subjects in each treatment group:
\[
\frac{1}{n_g} \sum_{i=1}^{n_g} [\ln(x)_{gik}] = (\mu_k) + (\tau_{gk}) + \frac{1}{n_g} \sum_{i=1}^{n_g} (s_{ik} + e_{ik}).
\]

For each treatment group,
\[
\frac{1}{n_g} \sum_{i=1}^{n_g} \begin{bmatrix}
    x_{g1i} \\
    x_{g2i}
\end{bmatrix} \sim N\left\{ \begin{bmatrix}
    \mu_1 + \tau_{g1} \\
    \mu_2 + \tau_{g2}
\end{bmatrix}, \frac{1}{n_g} \times (\Sigma_e + \Sigma_s) \right\}
\]
2. Generate Treatment t - Treatment r:

\[ d_k = \frac{1}{n_t} \sum_{i=1}^{n_t} [\ln(x)_{tik}] - \frac{1}{n_r} \sum_{i=1}^{n_r} [\ln(x)_{rik}] \]

\[ = (\mu_k) + (\tau_{tk}) + \frac{1}{n_t} \sum_{i=1}^{n_t} (s_{ik} + e_{ik}) - (\mu_k) - (\tau_{rk}) - \frac{1}{n_r} \sum_{i=1}^{n_r} (s_{ik} + e_{ik}) \]

\[ = (\delta_k) + \left[ \frac{1}{n_t} \sum_{i=1}^{n_t} (s_{ik} + e_{ik}) - \frac{1}{n_r} \sum_{i=1}^{n_r} (s_{ik} + e_{ik}) \right] \]

where \( \delta_k = (\tau_{tk} - \tau_{rk}) \). The test statistic is thus distributed as follows:

\[
\begin{bmatrix}
  d_1 \\
  d_2
\end{bmatrix}
\sim N\left( \begin{bmatrix}
  \delta_1 \\
  \delta_2
\end{bmatrix}, \left[ \frac{1}{d_1} + \frac{1}{d_2} \right] \times (\Sigma_c + \Sigma_r) \right)
\]  

(4.11)

The main difference between parallel and cross-over designs is that subject effects do not cancel out of treatment comparisons. The variance of the test statistic contains components for true (within subject) error, and between-subject error, and is therefore larger than the cross-over test statistic variance. Parallel designs usually require larger sample size for equivalent power to cross-over studies, and are avoided when between-subject variability is extremely large.

For power calculations, it is necessary only to obtain an estimate of the overall combined within-subject and between-subject variance \( \sigma^2_k = \sigma^2_{e_k} + \sigma^2_{x_k} \), as well as the overall correlation \( \rho \) (which is a function of \( \rho_c \) and \( \rho_x \) - see below).

Statistics \( d_1 \) and \( d_2 \) are the continuous-variable versions of Blackwelder's z-score statistic designed for parallel dichotomous trials (equation 3.19), and are the parallel design counterpart to Schuirmann's cross-over design test statistics.

Bivariate power for \( d_1 \) and \( d_2 \) is derived below for large samples in which the variance is assumed known. Evaluation of power using numerical integration and simulation assuming both known and unknown variance is described in the following sections.
4.3.3 Power Calculation from Numerical Integration

The univariate power of the test statistic $d_k$ for the parallel design is derived in exactly the same manner as the cross-over design, applying the principle of the two 1-sided tests procedure. The power is found to be:

$$\text{Power} = \Pr\left[ \frac{\Delta_1 - \delta_k}{\text{SE}(d_k)} + Z_{1-\alpha} < Z < \frac{\Delta_2 - \delta_k}{\text{SE}(d_k)} - Z_{1-\alpha} \right],$$

(4.12)

where $\text{SE}(d_k) = \sqrt{\left(\sigma_{ek}^2 + \sigma_{sk}^2\right) \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$. If the lower and upper limits are referred to as $L$ and $U$, and if $\delta_k$ and $\text{SE}(d_k)$ are assumed known, then the univariate power is calculated by referring to tables of the standard normal distribution, or by numerical integration of the function:

$$\int_{L}^{U} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz.$$

For bivariate power of statistics $d_1$ and $d_2$, if it is also assumed that the correlation between variables, $\rho = \frac{\rho_1 \sigma_{e1} \sigma_{e2} + \rho_2 \sigma_{e1} \sigma_{e2}}{\sqrt{(\sigma_{e1}^2 + \sigma_{e1}^2)(\sigma_{e2}^2 + \sigma_{e2}^2)}}$, is known, then bivariate power is obtained by numerical integration of a standardized bivariate normal distribution:

$$\int_{L_1}^{U_1} \int_{L_2}^{U_2} \frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{1}{2(1 - \rho^2)} \left[z_1^2 - 2\rho z_1 z_2 + z_2^2\right]} \, dz_1 \, dz_2.$$

As with the cross-over design, the constraint on $\text{SE}(d_k)$ imposed by equation (4.12) holds. Power is $> 0$ only if $\text{SE}(d_k) < 0.1356$ for $\Delta_1 = \ln(0.80)$, $\Delta_2 = \ln(1.25)$, and $\alpha = 0.05$. For example, if $n_i = n_r = 50$, then $\sqrt{\sigma_{ek}^2 + \sigma_{sk}^2}$ must be $< 0.678$ in order for power to be greater than zero. More generally, $(\Delta_2 - \Delta_1)^2 \geq [2Z_{1-\alpha} \text{SE}(d_k)]^2$ must hold.
4.3.4 Power Estimation from Simulation Assuming Known Variance

If numerical integration software is unavailable, univariate and bivariate power assuming known variance can be approximated by simulation. All parameters must first be specified, including \( \delta_k, n_t, n_r, \alpha \), and either \( \rho \) and \( \sigma_k^2 \) \((=\sigma_{ek}^2 + \sigma_{sk}^2)\), or \( \rho_e, \rho_s, \sigma_{ek}^2 \), and \( \sigma_{sk}^2 \). Once the distribution under study is fully specified, one random observation is generated to form each of 2 standardized random normal deviates \( z_1, z_2 \) which represent the test-reference difference of the average response for each variable.

The observations are transformed so they have the correct covariance structure of the test statistics by applying a Cholesky decomposition of 
\[
\left( \frac{1}{n_t} + \frac{1}{n_r} \right) \times (\Sigma_e + \Sigma_n),
\]
as described in Section 4.2.4; this yields \( f_1 \) and \( f_2 \). Test statistics \( \hat{d}_1 \) and \( \hat{d}_2 \) are obtained by \( (\delta_1 + f_1) \) and \( (\delta_2 + f_2) \).

The p-value for each \( \hat{d}_k \) is calculated following the principle of the two 1-sided test procedure (equation 3.8), where \( t_{k1} = \frac{\hat{d}_k - \Delta_1}{SE(d_k)} \), \( t_{k2} = \frac{\Delta_2 \hat{d}_k}{SE(d_k)} \), and 
\[
SE(d_k) = \sqrt{(\sigma_{ek}^2 + \sigma_{sk}^2) \left[ \frac{1}{n_t} + \frac{1}{n_r} \right]}.
\]
Since the variance is assumed known, the tests are applied to a z distribution. The bivariate p-value for \( \hat{d}_1 \) and \( \hat{d}_2 \) is defined as the maximum of the four univariate 1-sided p-values.

Power is simulated by repeating the above steps many times, and calculating the percent of repetitions in which \( H_0 \) is rejected (p-value \( \leq \alpha \)). This yields an estimated percent of samples which will reject \( H_0 \) given the specified parameters. Since the bivariate p-value is based on the univariate p-values, an estimate of the univariate powers is also obtained. Simulations are based on repetitions of 1,000 for power estimates \( > 0.05 \), and 10,000 for estimates \( \leq 0.05 \).

4.3.5 Power Estimation from Simulation Assuming Unknown Variance

If sample sizes are small \( (N<30) \) and \( \sigma_{ek} \) and \( \sigma_{sk} \) are assumed unknown, then the power for test statistics \( d_1 \) and \( d_2 \) involves a non-central bivariate t-
distribution. Power under this situation could be obtained via numerical integration of this function, but can easily be approximated via simulation. In this work, only approximations via simulation are addressed.

Steps for simulation are similar to those described in Section 4.3.4. The main difference is that each subject's observations are simulated with random variates, to obtain sample variance estimates. Steps are as follows: generate \(2n_r + 2n_\tau\) random standardized normal deviates which represent subjects' observed errors for each treatment group and outcome. Arrange them in vectors \(z_{t1}, z_{t2}, z_{r1}, z_{r2}\). Apply a Cholesky decomposition of \((\Sigma_e + \Sigma_r)\) to the error vectors for each treatment group to obtain \(f_{t1}, f_{t2}, f_{r1}, f_{r2}\): vectors of variates with the correct variance-covariance structure of \(\log(x_k)\). Calculate the mean \(\overline{f}_{gk}\) of each treatment-variable combination, and the pooled treatment group sample variance for each outcome:

\[
\hat{\sigma}_k^2 = \frac{1}{n+1} \sum_{j=1}^{n} (f_{gik} - \overline{f}_{gk})^2
\]

The test statistics are obtained by \(\hat{d}_k = \delta_k + (\overline{f}_{rk} - \overline{f}_{r})\). The standard error of the test statistic for response \(k\) is estimated by \(SE(\hat{d}_k) = \sqrt{\frac{1}{n_t} + \frac{1}{n_r}} \hat{\sigma}_k^2\). P-values are obtained from the t distribution with \((n_t + n_r - 2)\) degrees of freedom.

The remaining simulation steps are identical to the known variance case. The unknown case is used when the total sample size \(N\) is less than 30. Parallel trials often have more than 30 subjects, since larger sample sizes are needed over a cross-over design to compensate for the additional variability of \(\Sigma_r\). For applications in this chapter, sample sizes are larger than 30.

### 4.3.6 Application and Results

Bivariate power as described above is provided for a variety of situations in Tables 4.6 to 4.9. Power is calculated for \(\alpha = .05\) and \(.025\) (corresponding to 90\% and 95\% confidence intervals), for balanced designs with sample size per
treatment group = 25 and 50. The correlation of \( \ln(x_1) \) and \( \ln(x_2) \), \( \rho \), is fixed at 0.85, 0.75, 0.50, and 0.25, and treatment geometric mean ratios (\( \theta_k = e^{\delta_k} \)) for the two measures vary over several values. Since there is no need to separate variance into within and between subject components, one variance denoted \( \sigma_k^2 \) represents the sum of the two variances in the power tables. All estimates are based on numerical integration.

For comparison, the corresponding univariate power is presented in Table 4.10. Univariate power is calculated by numerical integration assuming known variance. It can also be approximated using a modification of Hauschke's cross-over formula (92). As with the cross-over design, the z-score approximation is occasionally anti-conservative when the standard error is large relative to sample size, and the treatment geometric mean ratio is midway between full acceptance and rejection. This is no cause for concern when planning studies because power in this range is generally low.

In order for Tables 4.6-4.9 to be informative when estimates for correlation and variance from previous studies are on the standard rather than the log scale, a conversion of \( \rho \) and \( \sigma \) from the log to the standard scale is desirable. As with the cross-over design, \( \sigma_k = \text{SD}(\ln(x)) \) is closely approximated by the coefficient of variation, \( CV = \frac{\text{SD}(x)}{\text{E}(x)} \). In a population, the correlation from the log to the standard scale is converted by

\[
\text{corr}(x_1, x_2) = \frac{(e^{\rho \sigma_1 \sigma_2} - 1)}{\sqrt{(e^{\sigma_1^2} - 1)(e^{\sigma_2^2} - 1)}}.
\]

(4.13)

This non-linear transformation yields biased estimates even for large samples, but can be applied to the limits of a confidence interval for the observed-scale correlation to obtain a range in which the log-correlation might lie. Information needed for the transformation is more likely to be published for parallel than cross-over trials, so this transformation may be more useful for parallel trials.
Table 4.6
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 90% Confidence Intervals
Parallel Design, N per Treatment Group = 25, Equivalence Range = 0.8-1.25
Obtained via Numeric Integration

<table>
<thead>
<tr>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$\rho$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.2</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>.85</td>
<td>.99</td>
<td>.91</td>
<td>.88</td>
<td>.79</td>
<td>.25</td>
<td>.18</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>.75</td>
<td>.99</td>
<td>.91</td>
<td>.87</td>
<td>.79</td>
<td>.25</td>
<td>.16</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>.5</td>
<td>.99</td>
<td>.91</td>
<td>.86</td>
<td>.79</td>
<td>.25</td>
<td>.12</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>.25</td>
<td>.99</td>
<td>.91</td>
<td>.84</td>
<td>.79</td>
<td>.25</td>
<td>.09</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.20</td>
<td>.85</td>
<td>.98</td>
<td>.73</td>
<td>.73</td>
<td>.60</td>
<td>.18</td>
<td>.14</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.20</td>
<td>.75</td>
<td>.98</td>
<td>.73</td>
<td>.72</td>
<td>.60</td>
<td>.18</td>
<td>.13</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.20</td>
<td>.5</td>
<td>.98</td>
<td>.73</td>
<td>.70</td>
<td>.61</td>
<td>.18</td>
<td>.09</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.20</td>
<td>.25</td>
<td>.98</td>
<td>.73</td>
<td>.68</td>
<td>.62</td>
<td>.18</td>
<td>.07</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.25</td>
<td>.85</td>
<td>.87</td>
<td>.56</td>
<td>.56</td>
<td>.44</td>
<td>.14</td>
<td>.12</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.25</td>
<td>.75</td>
<td>.87</td>
<td>.56</td>
<td>.56</td>
<td>.44</td>
<td>.14</td>
<td>.11</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.25</td>
<td>.5</td>
<td>.87</td>
<td>.56</td>
<td>.55</td>
<td>.45</td>
<td>.14</td>
<td>.08</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.25</td>
<td>.25</td>
<td>.87</td>
<td>.56</td>
<td>.53</td>
<td>.47</td>
<td>.14</td>
<td>.05</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.85</td>
<td>.97</td>
<td>.72</td>
<td>.66</td>
<td>.40</td>
<td>.17</td>
<td>.12</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.75</td>
<td>.96</td>
<td>.72</td>
<td>.64</td>
<td>.41</td>
<td>.17</td>
<td>.10</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.5</td>
<td>.96</td>
<td>.72</td>
<td>.60</td>
<td>.43</td>
<td>.17</td>
<td>.07</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.25</td>
<td>.96</td>
<td>.72</td>
<td>.56</td>
<td>.46</td>
<td>.17</td>
<td>.05</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.85</td>
<td>.87</td>
<td>.55</td>
<td>.54</td>
<td>.25</td>
<td>.13</td>
<td>.11</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.75</td>
<td>.86</td>
<td>.55</td>
<td>.52</td>
<td>.27</td>
<td>.13</td>
<td>.09</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.5</td>
<td>.86</td>
<td>.55</td>
<td>.48</td>
<td>.31</td>
<td>.14</td>
<td>.06</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.25</td>
<td>.85</td>
<td>.55</td>
<td>.44</td>
<td>.34</td>
<td>.14</td>
<td>.042</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.85</td>
<td>.68</td>
<td>.42</td>
<td>.42</td>
<td>.15</td>
<td>.11</td>
<td>.09</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.75</td>
<td>.67</td>
<td>.42</td>
<td>.41</td>
<td>.18</td>
<td>.11</td>
<td>.08</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.5</td>
<td>.67</td>
<td>.42</td>
<td>.37</td>
<td>.22</td>
<td>.12</td>
<td>.06</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.25</td>
<td>.66</td>
<td>.42</td>
<td>.34</td>
<td>.25</td>
<td>.12</td>
<td>.036</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.85</td>
<td>.81</td>
<td>.50</td>
<td>.47</td>
<td>.13</td>
<td>.09</td>
<td>.09</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.75</td>
<td>.80</td>
<td>.50</td>
<td>.45</td>
<td>.15</td>
<td>.10</td>
<td>.08</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.5</td>
<td>.77</td>
<td>.49</td>
<td>.40</td>
<td>.21</td>
<td>.11</td>
<td>.05</td>
<td>.035</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.25</td>
<td>.76</td>
<td>.49</td>
<td>.35</td>
<td>.25</td>
<td>.12</td>
<td>.035</td>
<td>.05</td>
<td></td>
</tr>
</tbody>
</table>

$\sigma_k = \text{SD}(\ln(x_k))$

$\rho = \text{corr}(\ln(x_1), \ln(x_2))$
Table 4.7
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 90% Confidence Intervals
Parallel Design, N per Treatment Group= 50, Equivalence Range= 0.8-1.25
Obtained via Numerical Integration

<table>
<thead>
<tr>
<th>$\theta_1, \theta_2 \rightarrow$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1$ $\sigma_2$ $\rho$</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
<td>1.25</td>
</tr>
<tr>
<td>0.20 0.20 .85</td>
<td>.99</td>
<td>.94</td>
<td>.91</td>
<td>.84</td>
<td>.27</td>
<td>.19</td>
<td>.05</td>
</tr>
<tr>
<td>.75 0.20 .85</td>
<td>.99</td>
<td>.94</td>
<td>.91</td>
<td>.84</td>
<td>.27</td>
<td>.17</td>
<td>.05</td>
</tr>
<tr>
<td>.5 0.20 .85</td>
<td>.99</td>
<td>.94</td>
<td>.90</td>
<td>.84</td>
<td>.27</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td>.25 0.20 .85</td>
<td>.99</td>
<td>.94</td>
<td>.89</td>
<td>.84</td>
<td>.27</td>
<td>.10</td>
<td>.05</td>
</tr>
<tr>
<td>0.20 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.81</td>
<td>.72</td>
<td>.20</td>
<td>.16</td>
<td>.05</td>
</tr>
<tr>
<td>.75 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.81</td>
<td>.72</td>
<td>.20</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td>.5 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.79</td>
<td>.72</td>
<td>.20</td>
<td>.11</td>
<td>.05</td>
</tr>
<tr>
<td>.25 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.78</td>
<td>.73</td>
<td>.20</td>
<td>.08</td>
<td>.05</td>
</tr>
<tr>
<td>0.20 0.30 .85</td>
<td>.96</td>
<td>.69</td>
<td>.68</td>
<td>.59</td>
<td>.17</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td>.75 0.30 .85</td>
<td>.96</td>
<td>.69</td>
<td>.68</td>
<td>.59</td>
<td>.17</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>.5 0.30 .85</td>
<td>.96</td>
<td>.69</td>
<td>.67</td>
<td>.60</td>
<td>.17</td>
<td>.09</td>
<td>.05</td>
</tr>
<tr>
<td>.25 0.30 .85</td>
<td>.96</td>
<td>.69</td>
<td>.66</td>
<td>.61</td>
<td>.17</td>
<td>.07</td>
<td>.05</td>
</tr>
<tr>
<td>0.25 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.76</td>
<td>.58</td>
<td>.20</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td>.75 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.74</td>
<td>.58</td>
<td>.20</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>.5 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.71</td>
<td>.59</td>
<td>.20</td>
<td>.09</td>
<td>.05</td>
</tr>
<tr>
<td>.25 0.25 .85</td>
<td>.99</td>
<td>.82</td>
<td>.69</td>
<td>.60</td>
<td>.20</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>0.25 0.30 .85</td>
<td>.96</td>
<td>.68</td>
<td>.66</td>
<td>.45</td>
<td>.16</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td>.75 0.30 .85</td>
<td>.96</td>
<td>.68</td>
<td>.65</td>
<td>.45</td>
<td>.16</td>
<td>.11</td>
<td>.05</td>
</tr>
<tr>
<td>.5 0.30 .85</td>
<td>.96</td>
<td>.68</td>
<td>.61</td>
<td>.47</td>
<td>.17</td>
<td>.08</td>
<td>.05</td>
</tr>
<tr>
<td>.25 0.30 .85</td>
<td>.96</td>
<td>.68</td>
<td>.59</td>
<td>.50</td>
<td>.17</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>0.30 0.30 .85</td>
<td>.94</td>
<td>.67</td>
<td>.61</td>
<td>.32</td>
<td>.15</td>
<td>.11</td>
<td>.035</td>
</tr>
<tr>
<td>.75 0.30 .85</td>
<td>.94</td>
<td>.67</td>
<td>.58</td>
<td>.33</td>
<td>.15</td>
<td>.10</td>
<td>.039</td>
</tr>
<tr>
<td>.5 0.30 .85</td>
<td>.93</td>
<td>.66</td>
<td>.54</td>
<td>.36</td>
<td>.16</td>
<td>.07</td>
<td>.044</td>
</tr>
<tr>
<td>.25 0.30 .85</td>
<td>.93</td>
<td>.66</td>
<td>.50</td>
<td>.40</td>
<td>.16</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>0.30 0.40 .85</td>
<td>.75</td>
<td>.46</td>
<td>.45</td>
<td>.15</td>
<td>.11</td>
<td>.09</td>
<td>.035</td>
</tr>
<tr>
<td>.75 0.40 .85</td>
<td>.74</td>
<td>.46</td>
<td>.43</td>
<td>.17</td>
<td>.11</td>
<td>.08</td>
<td>.039</td>
</tr>
<tr>
<td>.5 0.40 .85</td>
<td>.73</td>
<td>.45</td>
<td>.39</td>
<td>.22</td>
<td>.12</td>
<td>.06</td>
<td>.044</td>
</tr>
<tr>
<td>.25 0.40 .85</td>
<td>.72</td>
<td>.45</td>
<td>.36</td>
<td>.26</td>
<td>.12</td>
<td>.036</td>
<td>.05</td>
</tr>
</tbody>
</table>

$\sigma_k = SD(\ln(x_k))$
$\rho = corr(\ln(x_1), \ln(x_2))$
Table 4.8
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 95% Confidence Intervals
Parallel Design, N per Treatment Group= 25, Equivalence Range= 0.8-1.25
Obtained via Numerical Integration

<table>
<thead>
<tr>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$\rho$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.15</td>
<td>.85</td>
<td>.99</td>
<td>.85</td>
<td>.80</td>
<td>.65</td>
<td>.16</td>
<td>.11</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.99</td>
<td>.85</td>
<td>.79</td>
<td>.65</td>
<td>.16</td>
<td>.09</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.99</td>
<td>.85</td>
<td>.76</td>
<td>.65</td>
<td>.16</td>
<td>.06</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.99</td>
<td>.85</td>
<td>.74</td>
<td>.66</td>
<td>.16</td>
<td>.04</td>
<td>.025</td>
</tr>
<tr>
<td>0.15</td>
<td>0.20</td>
<td>.85</td>
<td>.95</td>
<td>.62</td>
<td>.61</td>
<td>.41</td>
<td>.11</td>
<td>.08</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.95</td>
<td>.62</td>
<td>.60</td>
<td>.42</td>
<td>.11</td>
<td>.07</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.95</td>
<td>.62</td>
<td>.57</td>
<td>.44</td>
<td>.11</td>
<td>.05</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.95</td>
<td>.62</td>
<td>.55</td>
<td>.46</td>
<td>.11</td>
<td>.03</td>
<td>.025</td>
</tr>
<tr>
<td>0.15</td>
<td>0.25</td>
<td>.85</td>
<td>.77</td>
<td>.43</td>
<td>.43</td>
<td>.24</td>
<td>.08</td>
<td>.07</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.77</td>
<td>.43</td>
<td>.43</td>
<td>.26</td>
<td>.08</td>
<td>.06</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.77</td>
<td>.43</td>
<td>.41</td>
<td>.29</td>
<td>.08</td>
<td>.04</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.77</td>
<td>.43</td>
<td>.39</td>
<td>.32</td>
<td>.08</td>
<td>.02</td>
<td>.025</td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.85</td>
<td>.93</td>
<td>.59</td>
<td>.53</td>
<td>.19</td>
<td>.09</td>
<td>.07</td>
<td>.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.92</td>
<td>.59</td>
<td>.51</td>
<td>.22</td>
<td>.09</td>
<td>.06</td>
<td>.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.91</td>
<td>.59</td>
<td>.46</td>
<td>.26</td>
<td>.10</td>
<td>.04</td>
<td>.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.91</td>
<td>.59</td>
<td>.42</td>
<td>.30</td>
<td>.10</td>
<td>.02</td>
<td>.023</td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.85</td>
<td>.76</td>
<td>.41</td>
<td>.41</td>
<td>.08</td>
<td>.06</td>
<td>.06</td>
<td>.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.76</td>
<td>.41</td>
<td>.39</td>
<td>.11</td>
<td>.07</td>
<td>.05</td>
<td>.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.74</td>
<td>.41</td>
<td>.34</td>
<td>.16</td>
<td>.07</td>
<td>.03</td>
<td>.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.74</td>
<td>.41</td>
<td>.30</td>
<td>.20</td>
<td>.08</td>
<td>.018</td>
<td>.023</td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.85</td>
<td>.50</td>
<td>.28</td>
<td>.28</td>
<td>.04</td>
<td>.05</td>
<td>.05</td>
<td>.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.49</td>
<td>.28</td>
<td>.27</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
<td>.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.48</td>
<td>.27</td>
<td>.23</td>
<td>.10</td>
<td>.06</td>
<td>.025</td>
<td>.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.48</td>
<td>.27</td>
<td>.21</td>
<td>.13</td>
<td>.06</td>
<td>.014</td>
<td>.023</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.85</td>
<td>.68</td>
<td>.33</td>
<td>.34</td>
<td>.026</td>
<td>.023</td>
<td>.05</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td>.66</td>
<td>.33</td>
<td>.32</td>
<td>.05</td>
<td>.03</td>
<td>.04</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td>.62</td>
<td>.33</td>
<td>.27</td>
<td>.09</td>
<td>.05</td>
<td>.025</td>
<td>.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td>.60</td>
<td>.33</td>
<td>.22</td>
<td>.13</td>
<td>.06</td>
<td>.014</td>
<td>.018</td>
</tr>
</tbody>
</table>

$\sigma_k = SD(\ln(x_k))$
$\rho = \text{corr}(\ln(x_1), \ln(x_2))$
### Table 4.9
Power for Testing Equivalence of Bivariate Log-Normal Responses
For Method of Two Univariate 95% Confidence Intervals
Parallel Design, N per Treatment Group= 50, Equivalence Range= 0.8-1.25

Obtained via Numerical Integration

<table>
<thead>
<tr>
<th>$\theta_1, \theta_2 \rightarrow$</th>
<th>$\sigma_1$</th>
<th>$\sigma_2$</th>
<th>$\rho$</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.20  0.20  .85]</td>
<td>0.20</td>
<td>0.20</td>
<td>0.85</td>
<td>.99</td>
<td>.89</td>
<td>.85</td>
<td>.73</td>
<td>.17</td>
<td>.12</td>
</tr>
<tr>
<td>[0.75  0.20  .85]</td>
<td>0.75</td>
<td>0.20</td>
<td>0.85</td>
<td>.99</td>
<td>.89</td>
<td>.84</td>
<td>.73</td>
<td>.17</td>
<td>.10</td>
</tr>
<tr>
<td>[0.25  0.20  .85]</td>
<td>0.25</td>
<td>0.20</td>
<td>0.85</td>
<td>.99</td>
<td>.89</td>
<td>.82</td>
<td>.73</td>
<td>.17</td>
<td>.07</td>
</tr>
<tr>
<td>[0.20  0.25  .85]</td>
<td>0.20</td>
<td>0.25</td>
<td>0.85</td>
<td>.99</td>
<td>.72</td>
<td>.72</td>
<td>.56</td>
<td>.13</td>
<td>.10</td>
</tr>
<tr>
<td>[0.75  0.25  .85]</td>
<td>0.75</td>
<td>0.25</td>
<td>0.85</td>
<td>.99</td>
<td>.72</td>
<td>.71</td>
<td>.56</td>
<td>.13</td>
<td>.08</td>
</tr>
<tr>
<td>[0.25  0.25  .85]</td>
<td>0.25</td>
<td>0.25</td>
<td>0.85</td>
<td>.99</td>
<td>.72</td>
<td>.68</td>
<td>.57</td>
<td>.13</td>
<td>.06</td>
</tr>
<tr>
<td>[0.20  0.30  .85]</td>
<td>0.20</td>
<td>0.30</td>
<td>0.85</td>
<td>.92</td>
<td>.57</td>
<td>.57</td>
<td>.41</td>
<td>.10</td>
<td>.08</td>
</tr>
<tr>
<td>[0.75  0.30  .85]</td>
<td>0.75</td>
<td>0.30</td>
<td>0.85</td>
<td>.92</td>
<td>.57</td>
<td>.56</td>
<td>.41</td>
<td>.10</td>
<td>.07</td>
</tr>
<tr>
<td>[0.25  0.30  .85]</td>
<td>0.25</td>
<td>0.30</td>
<td>0.85</td>
<td>.92</td>
<td>.57</td>
<td>.54</td>
<td>.43</td>
<td>.10</td>
<td>.05</td>
</tr>
<tr>
<td>[0.20  0.25  .85]</td>
<td>0.20</td>
<td>0.25</td>
<td>0.85</td>
<td>.98</td>
<td>.72</td>
<td>.65</td>
<td>.38</td>
<td>.12</td>
<td>.08</td>
</tr>
<tr>
<td>[0.75  0.25  .85]</td>
<td>0.75</td>
<td>0.25</td>
<td>0.85</td>
<td>.98</td>
<td>.72</td>
<td>.63</td>
<td>.39</td>
<td>.12</td>
<td>.07</td>
</tr>
<tr>
<td>[0.25  0.25  .85]</td>
<td>0.25</td>
<td>0.25</td>
<td>0.85</td>
<td>.98</td>
<td>.72</td>
<td>.59</td>
<td>.42</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>[0.20  0.30  .85]</td>
<td>0.20</td>
<td>0.30</td>
<td>0.85</td>
<td>.92</td>
<td>.56</td>
<td>.54</td>
<td>.24</td>
<td>.09</td>
<td>.07</td>
</tr>
<tr>
<td>[0.75  0.30  .85]</td>
<td>0.75</td>
<td>0.30</td>
<td>0.85</td>
<td>.92</td>
<td>.56</td>
<td>.52</td>
<td>.26</td>
<td>.09</td>
<td>.06</td>
</tr>
<tr>
<td>[0.25  0.30  .85]</td>
<td>0.25</td>
<td>0.30</td>
<td>0.85</td>
<td>.91</td>
<td>.56</td>
<td>.48</td>
<td>.30</td>
<td>.10</td>
<td>.04</td>
</tr>
<tr>
<td>[0.30  0.30  .85]</td>
<td>0.30</td>
<td>0.30</td>
<td>0.85</td>
<td>.89</td>
<td>.53</td>
<td>.48</td>
<td>.13</td>
<td>.07</td>
<td>.06</td>
</tr>
<tr>
<td>[0.75  0.30  .85]</td>
<td>0.75</td>
<td>0.30</td>
<td>0.85</td>
<td>.88</td>
<td>.53</td>
<td>.45</td>
<td>.15</td>
<td>.07</td>
<td>.05</td>
</tr>
<tr>
<td>[0.25  0.30  .85]</td>
<td>0.25</td>
<td>0.30</td>
<td>0.85</td>
<td>.85</td>
<td>.52</td>
<td>.36</td>
<td>.24</td>
<td>.09</td>
<td>.019</td>
</tr>
<tr>
<td>[0.30  0.40  .85]</td>
<td>0.30</td>
<td>0.40</td>
<td>0.85</td>
<td>.59</td>
<td>.31</td>
<td>.32</td>
<td>.03</td>
<td>.04</td>
<td>.05</td>
</tr>
<tr>
<td>[0.75  0.40  .85]</td>
<td>0.75</td>
<td>0.40</td>
<td>0.85</td>
<td>.58</td>
<td>.31</td>
<td>.30</td>
<td>.06</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>[0.25  0.40  .85]</td>
<td>0.25</td>
<td>0.40</td>
<td>0.85</td>
<td>.55</td>
<td>.31</td>
<td>.22</td>
<td>.14</td>
<td>.07</td>
<td>.015</td>
</tr>
</tbody>
</table>

$\sigma_k = SD(\ln(x_k))$

$\rho = corr(\ln(x_1), \ln(x_2))$
Table 4.10
Power for Testing Equivalence of a Log-Normal Response
For Method of 100(1-2α)% Confidence Intervals
Parallel Design, Equivalence Range= 0.8-1.25
Obtained via Numerical Integration

<table>
<thead>
<tr>
<th>α</th>
<th>n per group</th>
<th>σ</th>
<th>θ →</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.2</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>25</td>
<td>.15</td>
<td></td>
<td>.99</td>
<td>.91</td>
<td>.87</td>
<td>.25</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.20</td>
<td></td>
<td>.98</td>
<td>.73</td>
<td>.67</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td>.87</td>
<td>.56</td>
<td>.51</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.30</td>
<td></td>
<td>.68</td>
<td>.43</td>
<td>.39</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>0.05</td>
<td>50</td>
<td>.20</td>
<td></td>
<td>.99</td>
<td>.94</td>
<td>.90</td>
<td>.27</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td>.99</td>
<td>.82</td>
<td>.76</td>
<td>.20</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.30</td>
<td></td>
<td>.96</td>
<td>.69</td>
<td>.62</td>
<td>.17</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.40</td>
<td></td>
<td>.75</td>
<td>.47</td>
<td>.42</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td>0.025</td>
<td>25</td>
<td>.15</td>
<td></td>
<td>.99</td>
<td>.85</td>
<td>.79</td>
<td>.16</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.20</td>
<td></td>
<td>.95</td>
<td>.62</td>
<td>.55</td>
<td>.11</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td>.77</td>
<td>.43</td>
<td>.38</td>
<td>.08</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.30</td>
<td></td>
<td>.50</td>
<td>.29</td>
<td>.26</td>
<td>.07</td>
<td>.025</td>
</tr>
<tr>
<td>0.025</td>
<td>50</td>
<td>.20</td>
<td></td>
<td>.99</td>
<td>.89</td>
<td>.84</td>
<td>.17</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td>.99</td>
<td>.72</td>
<td>.65</td>
<td>.13</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.30</td>
<td></td>
<td>.92</td>
<td>.57</td>
<td>.50</td>
<td>.10</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.40</td>
<td></td>
<td>.59</td>
<td>.34</td>
<td>.30</td>
<td>.07</td>
<td>.025</td>
</tr>
</tbody>
</table>

σ = SD(ln(x))

The shape of the bivariate power surface for parallel and cross-over designs is the same, so the description in Section 4.2.7 is applicable here. When the study hypothesis is one-sided rather than two-sided (i.e., drug t is no worse than drug r) power tables in this chapter do not apply. Unlike the traditional test of a difference, the 2-sided equivalence power cannot be tricked or manipulated to get the 1-sided power. For 2-sided, the power is defined as the area of a normal curve sandwiched between two lines, but for one-sided, the power extends from one line to the end of the curve. Graphical evaluation suggests that power for a 1-sided test can be converted into 2-sided power for the same α under certain conditions. If the 1-sided power is greater than 0.5, and Δ₁ and Δ₂ are symmetrical about
zero, then 2-sided power = 2 \times (1\text{-sided power - 0.5}). If the 1\text{-sided power is less than 0.5, then 2-sided power is zero.}

4.3.7 Example

A clinical equivalence trial of two drugs for the treatment of hypertension was described in Section 1.4.2. Bivariate power is examined here for the comparison of treatments 2 and 5 for the visit 1 - visit 9 difference in supine systolic and diastolic blood pressure. Since the log transformation is made first, and then the visits are subtracted, evaluation is for the treatment ratio of the (visit 1)/(visit 9) ratio blood pressure change.

The sample size is 44 subjects for treatment 2 and 45 for treatment 5. The geometric mean treatment 2 vs 5 ratio is 1.0001 for diastolic, and 0.988 for systolic, with standard deviations on the log scale of 0.083 and 0.099, respectively. The 90\% CI of the treatment geometric mean ratio for supine diastolic is (0.97, 1.03), and for supine systolic is (0.95, 1.02). The correlation between the responses (on the scale of difference of logs) is 0.69.

The variation of the test statistic is small relative to the sample size: within-subject variability is reduced by averaging three blood pressure measurements, and between-subject variability is reduced by evaluating the visit 1 - visit 9 change. The confidence intervals are clearly well within the standard range of 0.80-1.25. Since variability is low, evaluating a tighter criterion for equivalence is of interest. In this example we consider an equivalence range of 0.95-1.053.

The univariate post-hoc power for equivalence based on 0.95-1.053 was estimated by numerical integration as 0.80 and 0.51 for diastolic and systolic. The bivariate power is estimated as 0.44, slightly larger than the estimate of 0.80 \times 0.51 = 0.41. Even though the \theta s are on opposite sides of 1.0, they are each so close to 1.0 that a negative effect of the correlation on the power is not realized. If this
study reflects the true population, then it had a 44% chance of finding treatments 2 and 5 equivalent for both systolic and diastolic blood pressure. If the equivalence range is changed to 0.9-1.11, then bivariate power jumps to 0.99.

Although a test of the log-normal assumption was not rejected, blood pressure measurements are generally analyzed on the additive rather than multiplicative scale. An alternative strategy which does not imply a log-normal distribution is Fieller’s formula. A confidence interval for the ratio of visit 1 - visit 9 blood pressure differences based on Fieller’s formula is obtained via equation (3.14), setting $\tilde{\sigma}_r=0$, estimating $\tilde{\sigma}^2_f$ and $\tilde{\sigma}^2_r$ as $s^2_p/n_f$ and $s^2_p/n_r$, respectively, where $s^2_p$ is the pooled treatment variance, and applying a $t$ with $n_f+n_r-2$ degrees of freedom. The 90% CI based on Fieller’s formula is (0.61, 1.57) for diastolic and (0.48, 1.29) for systolic. Confidence intervals are much wider because they are based on visit1-visit9 differences rather than ratios. In this chapter, univariate and bivariate power is evaluated only for the log-normal ANOVA approach. It is of interest in future research to evaluate the power of Fieller’s method for cross-over and parallel trials through simulation.

### 4.3.8 Extension to More Than Two Outcomes

When more than two outcomes are evaluated for equivalence, power of the multiple univariate CI approach decreases quickly. Multivariate power is generally undesirably small for a clinical trial unless both the univariate powers and correlations between the outcomes are high. As an example, consider six outcomes, each with a standard deviation of $\sigma=0.15$, with correlation of 0.85 between each outcome pair. When $\alpha=0.05$, $n$ per treatment group=25, and treatment geometric mean ratios are all 1.1, the univariate power is 0.91, bivariate is 0.88, and multivariate power for six outcomes is 0.81 (obtained via an extension of simulation described in 4.3.4). If the correlation between outcome pairs is zero,
the multivariate power is \((0.91)^6 = 0.57\).

The rough approximation of multiplying the univariate powers does not provide an accurate result for multiple outcomes with moderate correlations; as the number of responses increases, the effect of the correlation structure on the power also increases. When planning trials with several outcomes, it is therefore advisable to evaluate multivariate power through simulation.

Since the multivariate CI procedure does not make use of the correlation, it is a conservative test, and becomes more conservative for multiple outcomes. The type I error decreases from 0.05 as the number of outcomes increases. Intuitively it seems that analysis methods which account for the correlation should be less conservative, and result in higher power. Chapters 7 and 8 recommend several analysis strategies which may have more power in settings with many outcomes.

4.4 Conclusions

This chapter has evaluated bivariate power for the current strategy of concluding treatment equivalence if both univariate responses support equivalence, assuming a log-normal distribution. Chapter 5 evaluates power for dichotomous responses. Bivariate power can be much less than each univariate power, as shown by the rule of thumb of multiplying the univariate powers. Although the testing method ignores the correlation between outcomes, power is affected by it. When the treatment geometric mean ratios are on the same side of 1.0, positive correlation increases power, and negative correlation decreases power. The effect of correlation is stronger when variance is large relative to sample size and treatment geometric mean ratios are midway between full acceptance and rejection (i.e., both univariate powers are moderate). When one univariate power is relatively high or low, the multivariate power, regardless of correlation, is fairly well estimated by the rule of thumb of multiplying univariate powers.
Chapter 5
Bivariate Dichotomous Responses from Parallel Designs

5.1 Introduction

This chapter evaluates power of bivariate equivalence from dichotomous responses. Bivariate equivalence is defined as in Chapter 4: the p-value is the maximum of the univariate p-values, and equivalence is concluded if lack of equivalence is contradicted for both of the univariate outcomes. Dichotomous responses are generally studied with parallel designs, so cross-over designs are not considered.

Unlike Chapter 4, only one-sided tests are evaluated. For example, in the study of delayed onset muscle soreness described in Section 1.4.3, the proportion of subjects with $\geq 50\%$ pain reduction by 48 hours for active range of motion is 59\% for the test treatment, and 62\% for the reference. It would be a concern if the test treatment had a significantly smaller response rate, but would be beneficial if the test treatment was better than reference. If the test was expected to be better than the reference, then a standard test of a difference is appropriate.

There are several ways to compare test and reference response rates. The clinical equivalence literature most often uses the difference of proportions. Equivalence is concluded if the test response proportion minus the reference $\pi_t - \pi_r$ is no worse than $\Delta$, where $\Delta$ is often -0.10 or -0.20. Equivalence can also be concluded if $\pi_t / \pi_r$ is significantly greater than a ratio $\Theta$. Reasonable values of $\Theta$ are 0.9 or 0.8. A third measure is the odds ratio $\psi$. The test treatment is equivalent to reference if $\psi$ is not less than $\Psi$, perhaps 0.5.

This chapter assumes the response rates are for favorable outcomes rather
than unfavorable. An equivalence trial of two drugs may wish to show that the rate of an adverse event for the test treatment is no higher than reference by a specified amount, but this can be reworded to a hypothesis that the event-free rate for test is not significantly less than reference. Response rates evaluated in this chapter are generally above 0.5, but could potentially be less. Ulcer studies are an example of equivalence trials where two dichotomous responses are of interest. Ulcer healing is evaluated by both the absence of ulcer pain, and the absence of ulcers through endoscopic evaluation.

Univariate equivalence tests for differences of proportions are described in Section 3.5.1. Approximate power is described in Section 3.9, and also by Gould (93). This chapter evaluates bivariate power of test statistics based on differences of proportions, ratios of proportions, and odds ratios. Following the organization of Chapter 4, the large sample bivariate distribution is derived for the test statistic and power is evaluated by numerical integration. Approximate methods for small sample situations are discussed, and small-sample power is evaluated via simulation.

5.1.1 Response Notation

Response notation for the bivariate response is described in Figure 5.1:

**Figure 5.1**
Notation for Bivariate Dichotomous Response

<table>
<thead>
<tr>
<th>Observed Counts</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
<th>n_t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>n_{11}</td>
<td>n_{12}</td>
<td>n_{21}</td>
<td>n_{22}</td>
<td>n_t</td>
</tr>
<tr>
<td>Reference</td>
<td>n_{r11}</td>
<td>n_{r12}</td>
<td>n_{r21}</td>
<td>n_{r22}</td>
<td>n_r</td>
</tr>
<tr>
<td></td>
<td>n_{11}</td>
<td>n_{12}</td>
<td>n_{21}</td>
<td>n_{22}</td>
<td>N</td>
</tr>
</tbody>
</table>
Observed Proportions

<table>
<thead>
<tr>
<th>Outcome 1,2</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>$P_{t11}$</td>
<td>$P_{t12}$</td>
<td>$P_{t21}$</td>
<td>$P_{t22}$</td>
</tr>
<tr>
<td>Reference</td>
<td>$P_{r11}$</td>
<td>$P_{r12}$</td>
<td>$P_{r21}$</td>
<td>$P_{r22}$</td>
</tr>
</tbody>
</table>

1

Estimated Expected Counts

<table>
<thead>
<tr>
<th>Outcome 1,2</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>$m_{t11}$</td>
<td>$m_{t12}$</td>
<td>$m_{t21}$</td>
<td>$m_{t22}$</td>
</tr>
<tr>
<td>Reference</td>
<td>$m_{r11}$</td>
<td>$m_{r12}$</td>
<td>$m_{r21}$</td>
<td>$m_{r22}$</td>
</tr>
</tbody>
</table>

| $n_{11}$ | $n_{12}$ | $n_{21}$ | $n_{22}$ |

$N$

In a population, the ps are replaced by $\pi$s. Proportions are obtained by dividing the number observed for each category by the total number of subjects in a treatment group. The estimated expected cell count for treatment $g$ and response profile $i,j$ under the null hypothesis of no association between treatment and response is defined $m_{gij} = \frac{n_{ij}n_g}{N}$.

For each outcome, the response proportions are labeled as shown in Figure 5.2, so that the marginal responses for each outcome are defined as:

$P_{t1} = P_{t11} + P_{t12}$

$P_{t2} = P_{t11} + P_{t21}$

$P_{r1} = P_{r11} + P_{r12}$

$P_{r2} = P_{r11} + P_{r21}$

The estimated expected cell counts are defined $m_{tk} = \frac{n_{k1}n_t}{N}$ and $m_{rk} = \frac{n_{k1}n_r}{N}$.

For treatment $g$, the responses for outcomes 1 and 2 can be presented as in Figure 5.3. The population odds of positive response on outcome 2 for positive vs negative outcome 1 for each treatment group is $\psi_{g12} = \frac{\pi_{g11} \times \pi_{g22}}{\pi_{g12} \times \pi_{g21}}$. If marginal proportions $\pi_{g1}$, $\pi_{g2}$, and the odds ratio $\psi_{g12}$ are specified, then the bivariate cell probabilities $\pi_{g11}$, $\pi_{g12}$, $\pi_{g21}$, and $\pi_{g22}$ are obtained by rewriting $\psi_{g12} = \frac{\pi_{g11}(1-\pi_{g1}-\pi_{g2}+\pi_{g11})}{(\pi_{g1}-\pi_{g11})(\pi_{g2}-\pi_{g11})}$, and solving for $\pi_{g11}$ as a quadratic equation. The other cell probabilities are obtained by subtraction.
Figure 5.2
Notation for Univariate Dichotomous Response

**Observed Counts**

<table>
<thead>
<tr>
<th>Outcome K</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>$n_{tk}$</td>
<td>$n_{r \cdot n_{tk}}$</td>
</tr>
<tr>
<td>Reference</td>
<td>$n_{rk}$</td>
<td>$n_{r \cdot n_{rk}}$</td>
</tr>
<tr>
<td></td>
<td>$n_{k,1}$</td>
<td>$n_{k,2}$</td>
</tr>
</tbody>
</table>

**Observed Proportions**

<table>
<thead>
<tr>
<th>Outcome K</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>$p_{tk}$</td>
<td>$1 - p_{tk}$</td>
</tr>
<tr>
<td>Reference</td>
<td>$p_{rk}$</td>
<td>$1 - p_{rk}$</td>
</tr>
</tbody>
</table>

**Estimated Expected Counts**

<table>
<thead>
<tr>
<th>Outcome K</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>$m_{tk}$</td>
<td>$n_t \cdot m_{tk}$</td>
</tr>
<tr>
<td>Reference</td>
<td>$m_{rk}$</td>
<td>$n_r \cdot m_{rk}$</td>
</tr>
<tr>
<td></td>
<td>$n_{k,1}$</td>
<td>$n_{k,2}$</td>
</tr>
</tbody>
</table>

Figure 5.3
Notation for Outcome 1 vs Outcome 2 for Treatment g

**Outcome 2**

<table>
<thead>
<tr>
<th>Outcome 1</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>$\pi_{g11}$</td>
<td>$\pi_{g12}$</td>
</tr>
<tr>
<td>N</td>
<td>$\pi_{g21}$</td>
<td>$\pi_{g22}$</td>
</tr>
<tr>
<td></td>
<td>$\pi_{g2}$</td>
<td>$1 - \pi_{g2}$</td>
</tr>
</tbody>
</table>
If the outcome 1 vs 2 odds ratios for the treatment groups are assumed identical, then all population parameters above can be completely specified by the sample size \((n_1, n_2)\), the marginal positive response probabilities \((\pi_{t1}, \pi_{t2}, \pi_{r1}, \pi_{r2})\), and the odds ratio \((\psi_{12})\). This work will always assume that the outcome odds ratio \(\psi_{12}\) is the same for each treatment group. If this were not assumed, a broader definition of equivalence must be addressed, in which not only the relationship between \(\pi_t\) and \(\pi_r\) is evaluated, but a two-sided equivalence criterion is also placed on \(\psi_{t12}\) and \(\psi_{r12}\). For purposes of this work, it is sufficient to estimate \(\psi_{12}\) with the Mantel-Haenszel estimate of the common odds ratio if homogeneity of the odds ratios is not rejected.

5.2 Difference of Proportions

This section develops a bivariate model for the difference between the test and reference response rates for two outcomes, and derives the bivariate power of the test statistic for both large and small samples. The univariate hypothesis is stated: \(H_0: \delta_k = (\pi_{tk} - \pi_{rk}) < \Delta\), where \(\Delta\) is negative (test worse than reference), vs \(H_A: \delta_k = (\pi_{tk} - \pi_{rk}) \geq \Delta\) (test within equivalence range of reference).

The data in Figure 5.1 follow a multinomial distribution, and marginal response rates for each treatment group-outcome combination follow a binomial distribution. When proportions are not extremely near one or zero, and sample size is large enough so that expected cell counts (Figure 5.2) for each outcome are all \(\geq 10\), then by the central limit theorem, the normal approximation to the binomial distribution can be applied with the assurance that results derived from the normal distribution are similar to exact binomial results.

When expected cell counts are smaller (i.e., one or more expected counts per outcome between 3-10), then the normal approximation can still be applied, but its usefulness becomes increasingly crude, leading to p-values smaller and power
larger than actual. Modifications to the approximation which might improve performance are discussed and evaluated.

If sample sizes are so small or proportions are so extreme that some expected counts are less than 3, then normal approximation methods could be applied, but their accuracy is questionable. One alternate strategy is exact methods for odds ratios discussed in Section 3.5.1.

5.2.1 Model Assuming Normal Distribution

If we assume the normal approximation to the binomial holds, and that the treatment groups are independent but not necessarily identically distributed, then for each treatment group the positive response proportions for the two outcomes can be assumed to have the following distribution in large samples (Fleiss):

$$
P \left[ \frac{g_1}{g_2} \right] \sim N \left[ \begin{array}{c} \pi_{g_1} \\ \pi_{g_2} \end{array} \right] \begin{bmatrix} \frac{\pi_{g_1}(1 - \pi_{g_1})}{n_g} & \frac{\pi_{g_1} \pi_{g_22} - \pi_{g_12} \pi_{g_21}}{n_g} \\ \frac{\pi_{g_1} \pi_{g_22} - \pi_{g_12} \pi_{g_21}}{n_g} & \frac{\pi_{g_2}(1 - \pi_{g_2})}{n_g} \end{bmatrix} = \Sigma_g \tag{5.1}$$

For each treatment, the correlation between outcome 1 and 2, also called the $\phi$ coefficient (Fleiss), is

$$
\rho_g = \frac{\pi_{g_1} \pi_{g_22} - \pi_{g_12} \pi_{g_21}}{\sqrt{\pi_{g_1}(1 - \pi_{g_1}) \pi_{g_2}(1 - \pi_{g_2})}} \tag{5.2}
$$

A further assumption imposed on the bivariate model is that the odds ratio of outcome 2 for positive vs negative outcome 1 is the same for each treatment ($\psi_{12} = \psi_{r12}$). In Chapter 4, it was assumed that treatment groups have equal variance and equal correlation. Such a requirement for dichotomies also places restrictions on response proportions. Since the odds ratios for different treatment groups are equal, but response proportions are different, then the variances and correlations from different treatment groups can be unequal.

For each treatment group, the correlation is zero when the odds ratio is 1.0. The odds ratio and correlation increase as ($\pi_{g11}$ to $\pi_{g1}$ and $\pi_{g12}$ to 0), and/or ($\pi_{g22}$ to $\pi_{g2}$
and \( \pi_{g21} \rightarrow 0 \). If \( \pi_{g12} = 0 \) and \( \pi_{g21} = 0 \), then \( \pi_{g11} = \pi_{g1} = \pi_{g2} \), and \( \rho_g = 1.0 \).

While the \( \phi \) coefficient is the applicable correlation for the bivariate normal distribution, a correlation more suited to categorical data is preferred for descriptive purposes. Since the relationship is specified by an odds ratio, the Goodman-Kruskal Gamma, called Yule’s Q for a \( 2 \times 2 \) table, is particularly attractive (Daniel):

\[
\Gamma_g = \frac{\pi_{g11}\pi_{g22} - \pi_{g12}\pi_{g21}}{\pi_{g11}\pi_{g22} + \pi_{g12}\pi_{g21}} = \frac{\psi_{g12} - 1}{\psi_{g12} + 1}
\]  

(5.3)

When \( \psi_{g12} = 12 \), \( \Gamma_g = 0.85 \), and when \( \psi_{g12} = 1.5 \), \( \Gamma_g = 0.2 \). The \( \phi \) coefficient is generally smaller, and depends on the marginal response proportions as well as the odds ratio. For instance, in the examples discussed in Section 5.2.5, when \( \psi_{12} = 12 \), \( \phi \) ranges from 0.30 to 0.50, and when \( \psi_{12} = 1.5 \), \( \phi = 0.03-0.09 \). The \( \phi \) coefficient appears to get larger for the same odds ratio as the proportions get closer to 0.5.

Dunnett and Gent (77) suggest calculating estimated expected cell counts under the equivalence null hypothesis \( \delta = \Delta \) instead of \( \delta = 0 \), and basing the variance in equation (5.1) on these modified expected cell counts, as described in Section 3.5.1. For large samples with marginal response proportions \( \leq 0.95 \), one might find equation (5.1) and their method to be similar in terms of power to support equivalence. Differences in results between the direct (5.1) method and Dunnett and Gent methods may be more noticeable in small sample settings. Both methods will be evaluated under large and small sample scenarios (described in Section 5.2.3, below) to evaluate their bivariate power.

The statistic used to test \( \delta_k = (\hat{\pi}_{tk} - \pi_{rk}) < \Delta \) is obtained by subtracting test-reference response proportions (\( \hat{\delta}_k = p_{tk} - p_{rk} \)):

\[
\begin{bmatrix}
\hat{\delta}_1 \\
\hat{\delta}_2
\end{bmatrix}
\sim N
\begin{bmatrix}
\delta_1 \\
\delta_2
\end{bmatrix}, \Sigma = \Sigma_t + \Sigma_r
\]

(5.4)
\[ \Sigma = \begin{bmatrix}
\frac{\pi_{t1}(1 - \pi_{t1})}{n_t} + \frac{\pi_{r1}(1 - \pi_{r1})}{n_r} & \frac{\pi_{t11}\pi_{t22} - \pi_{t12}\pi_{t21}}{n_t} + \frac{\pi_{r11}\pi_{r22} - \pi_{r12}\pi_{r21}}{n_r} \\
\frac{\pi_{t11}\pi_{t22} - \pi_{t12}\pi_{t21}}{n_t} + \frac{\pi_{r11}\pi_{r22} - \pi_{r12}\pi_{r21}}{n_r} & \frac{\pi_{t21}(1 - \pi_{t21})}{n_t} + \frac{\pi_{r21}(1 - \pi_{r21})}{n_r}
\end{bmatrix} \]

The correlation of outcomes 1 and 2 for the test-reference difference, \( \rho_3 \), is \( \frac{\text{cov}(\hat{\delta}_1, \hat{\delta}_2)}{\sqrt{\text{var}(\hat{\delta}_1)\text{var}(\hat{\delta}_2)}} \), and is essentially a weighted average of the test and reference correlations. Since the variance is assumed known in large sample situations, the sample proportions \( \pi_s \) can be substituted for the population \( \pi_s \)s to calculate \( \Sigma \).

In the univariate case, equivalence is concluded if \( \frac{\hat{\delta}_k - \Delta}{\text{SE}(\hat{\delta}_k)} \) is found to be significantly greater than zero. This corresponds to Blackweilder's normal approximation described in equation (3.19).

### 5.2.2 Bivariate Power Assuming Normality

As in Chapter 4, if the sample size is large, then power of the univariate and bivariate tests can be evaluated directly with numerical integration. In this section, large-sample power of the test statistic is derived. This is similar to power for parallel studies described in Chapter 4, except that the test is one-sided.

For a one-sided test of \( H_0: \delta_k < \Delta \) vs \( H_A: \delta_k \geq \Delta \), the univariate power of the test statistic, \( \hat{\delta}_k \), is defined as the probability of rejecting the null hypothesis:

\[ \Pr(\hat{\delta}_k > c). \quad (5.5) \]

The cutpoint is obtained by applying the fixed type I error, \( \alpha = \Pr(\text{reject } H_0 | H_0) \):

\[ \alpha \geq \Pr(\hat{\delta}_k > c | \delta_k < \Delta) \]

A standard normal variable is obtained by subtracting \( \delta_k \) and dividing by \( \text{SE}(\hat{\delta}_k) \).

The type I error is maximized when \( \delta_k = \Delta \), so this is substituted for \( \delta_k \) to obtain:

\[ \alpha = \Pr\left[Z > \frac{c - \Delta}{\text{SE}(\hat{\delta}_k)}\right] \]

By applying the z-value corresponding to the \( \alpha \) upper tail of the normal distribution, and solving for \( c \), one obtains
\[ c = \Delta + Z_{1-\alpha} \times \text{SE}(\hat{\delta}_k) \]

The power of the test is obtained by substituting \( c \) into equation (5.5):

\[
\text{Power} = \Pr[\Delta + Z_{1-\alpha} \times \text{SE}(\hat{\delta}_k) < \hat{\delta}_k]
\]

Subtracting \( \text{E}(\hat{\delta}_k) = \delta_k \) and dividing by \( \text{SE}(\hat{\delta}_k) \) on each side leads to:

\[
\text{Power} = \Pr\left[ \frac{\Delta - \delta_k}{\text{SE}(\hat{\delta}_k)} + Z_{1-\alpha} < Z \right], \quad (5.6)
\]

where \( \text{SE}(\hat{\delta}_k) = \sqrt{\frac{\pi_{tk}(1-\pi_{tk})}{n_t} + \frac{\pi_{rk}(1-\pi_{rk})}{n_r}} \). With reversal of direction from multiplication by -1, the power is also stated:

\[
\text{Power} = \Pr\left[ Z < \frac{\delta_k - \Delta}{\text{SE}(\hat{\delta}_k)} + Z_{\alpha} \right], \quad (5.7)
\]

Equation (5.6) displays reasonable properties. As \( n_t \) and \( n_r \to \infty \), and \( \Delta < \delta_k \), then power\( \to 1 \). Power\( \to 1 \) faster as \( \delta_k \) increases. If \( \delta_k < \Delta \), then as \( n_t \) and \( n_r \to \infty \), power\( \to 0 \). Power\( \to 0 \) faster as \( \delta_k \) decreases. When \( \delta_k = \Delta \), then as \( n_t \) and \( n_r \to \infty \), power\( \to \alpha \).

If the lower limit is referred to as \( L \), and if \( \pi_{tk} \) and \( \pi_{rk} \) are assumed known, then the power is calculated by numerical integration of the standard normal distribution from \( L \) to \( \infty \). Alternately, for the opposite direction, the distribution is integrated from \( -\infty \) to \( U \), where \( U \) is the limit in (5.7).

For bivariate power of statistics \( \hat{\delta}_1 \) and \( \hat{\delta}_2 \), if it is also assumed that the \( \pi_{g11} \), \( \pi_{g12} \), \( \pi_{g21} \), and \( \pi_{g22} \) are known, then bivariate power is obtained by numerical integration of a standardized bivariate normal distribution:

\[
\int_{L_1}^{\infty} \int_{L_2}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_\delta^2}} e^{-\frac{1}{2(1-\rho_\delta^2)}[z_1^2 - 2\rho_\delta z_1 z_2 + z_2^2]} \, dz_1 \, dz_2 \quad (5.8)
\]

where \( \rho_\delta \) is described in equation (5.4). Unlike Chapter 4, a one-sided test has no constraint on the standard error in order for power to be properly defined. In this work, bivariate power for Dunnnett and Gent's modified test is estimated through simulation rather than numerical integration.
5.2.3 Modifications for Small Samples

When sample sizes are large, the variance is assumed known, and the test statistics and power can be evaluated using $\Sigma$. But when sample sizes are small (some expected cell counts between 3-10), the univariate test must be modified to counteract anti-conservativeness of the normal approximation. Bivariate power of the modified tests is evaluated by simulation, described in Section 5.2.4.

Two sets of modifications are considered. Method A employs a variance estimate based on Dunnett and Gent's modified expected cell counts under the null hypothesis that $\pi_{tk} = (\pi_{rk} + \Delta)$. Their standard error is:

$$SE'(\delta_k) = \sqrt{\frac{\hat{\pi}_{tk} (1-\hat{\pi}_{tk})}{n_t} + \frac{\hat{\pi}_{rk} (1-\hat{\pi}_{rk})}{n_r}}$$

(5.9)

where $\hat{\pi}_{tk} = \frac{n_{tk} + n_{rk} + \Delta n_r}{N}$, and $\hat{\pi}_{rk} = \frac{n_{tk} + n_{rk} - \Delta n_t}{N}$ as described in equation (3.18).

The expected proportions are derived by setting $\hat{\pi}_{tk} = (\hat{\pi}_{rk} + \Delta)$, $\hat{m}_{tk} = n_t (\hat{\pi}_{rk} + \Delta)$, $\hat{m}_{rk} = n_r (\hat{\pi}_{rk})$, then setting $\hat{m}_{tk} + \hat{m}_{rk} = n_k$, and solving for $\hat{\pi}_{rk}$. This method is not easily converted to a confidence interval for $\delta_k$, since under the null hypothesis $\delta_k = \Delta$, and the standard error contains the parameter $\Delta$.

Dunnett and Gent originally derived this test as a $\chi^2$ statistic based on their modified expected cell counts. They showed that although the results from the $\chi^2$ test and normal approximation are not identical in this non-null case, they agree very closely. In order to make the normal approximation more conservative, the authors suggest applying a continuity correction of

$$z = \frac{\delta_k - \Delta - 0.5[(1/n_t + 1/n_r)]}{SE'(\delta_k)}$$

However, univariate power simulations using this correction demonstrated that in the range of response proportions examined the test was too conservative, yielding type I errors well below 0.05. An alternate correction of

$$z = \frac{\delta_k - \Delta - 1/[2(n_t + n_r)]}{SE(\delta_k)}$$

was found to be more powerful while maintaining a significance level near 0.05.
Simulation results are presented using the alternate correction.

Method B employs the standard error:

$$\text{SE}^*(\hat{\delta}_k) = \sqrt{\frac{\hat{\pi}^*_{tk}(1-\hat{\pi}^*_{tk})}{n_t-1} + \frac{\hat{\pi}^*_{rk}(1-\hat{\pi}^*_{rk})}{n_r-1}}$$

where

$$\hat{\pi}^*_{gk} = \frac{(n_{gk}+0.5)}{n_{g}+1} \quad (5.10)$$

This variance has two conservative components: division by $n_g-1$ in the denominator which makes the variance larger; and modified observed response proportions which are necessary because in small samples it is possible that no (or all) events may occur, leading to a standard error of zero and an undefined p-value. The addition of 0.5 in the numerator and 1 in the denominator insures that $\hat{\pi}^*_{gk}$ will be slightly larger than zero if $n_{gk}=0$, and slightly less than 1 if $n_{gk}=n_g$. When not under these scenarios, the modification leads to a slightly larger variance than the standard formula.

A third conservative component is provided by the continuity correction, which makes the absolute value of the z-score smaller. Since two levels of conservatism are already in place, it is appropriate to apply a correction which is smaller than the traditional value. Therefore, the alternate correction of $-1/[2(n_t+n_r)]$ is applied to this method as well. Power simulations found that Method B with no continuity correction had slightly better power than when the alternate correction was applied, while still maintaining type I error near 0.05. This suggests that perhaps no continuity correction is required for Method B.

For Dunnett and Gent's Method A, when both observed response proportions are either 0 or 1, then the standard error may be undefined (square root of a negative number). A possible adjustment is $\hat{\pi}_{tk} = \frac{(n_{tk}+n_{rk}+\Delta n_r+0.5)}{(N+1)}$, and $\hat{\pi}_{rk} = \frac{(n_{tk}+n_{rk}-\Delta n_t+0.5)}{(N+1)}$. This adjustment is assured to work only when $|\Delta n_t|$ and $|\Delta n_r| < 0.5$. In Method A power simulations described below, undefined standard errors occur rarely (between 0-4 times in 1000 or 10,000 repetitions). Due to programming, the presented bivariate power is slightly overestimated by this problem (at worst by 0.004). Simulation from Method A
using the adjusted standard error produced bivariate power estimates that are slightly smaller (for sample sizes of 40, within 0.01, for n=15, within 0.03).

Although for continuous variables, it is customary to apply a t-score to normal approximations when sample sizes are small, this is not done for dichotomies due to uncertainty about the appropriate degrees of freedom. With the continuity correction and other corrective modifications, the test statistic can be reasonably compared to a z-score.

5.2.4 Bivariate Power Assuming Normal Approximation

In small sample situations, adjustments to the standard error estimate are not sufficient to make the bivariate normal distribution of the test statistic applicable. Since the exact distribution is unknown, bivariate power is evaluated through simulation. Simulations are conducted similar to Chapter 4: a sample with known properties is repetitively drawn with random numbers, test statistics are evaluated, and univariate power is estimated as the percent of test statistics that reject the null hypothesis. Bivariate power is defined as the percent of times that both univariate test statistics reject the null hypothesis. As in Chapter 4, power greater than 0.05 is estimated with 1,000 repetitions, and power ≤ 0.05 is estimated with 10,000 repetitions.

Data for each treatment group is generated by simulating a random uniform variate on the interval from zero to one for each subject. The number is converted to a bivariate response using the specified \( \pi_{gij} \) (i,j=1,2 see Figure 5.1). For each treatment group, the \( \pi_{gij} \) add up to 1.0. The region 0-1 is partitioned into four intervals, each with length corresponding to a bivariate outcome probability. The subject is assigned the bivariate response which corresponds to the interval in which their random number falls. Once all subjects are simulated, the observed \( p_{gij} \) are calculated as \( n_{gij}/n_g \), and test statistics are evaluated.
5.2.5 Applications and Results

Tables 5.1.1 and 5.1.2 present bivariate power for testing that $\pi_r$ is no worse than $\pi_r - 0.10$, for reference response proportions of 0.9 and 0.85. Tables 5.2.1 and 5.2.2 provide power for testing $\Delta = -0.20$ for more moderate response proportions. For large sample sizes, bivariate power of the direct method is obtained by numerical integration. Simulation is used to estimate power for Method A (equation 5.9) in large and small sample settings, and for Method B (equation 5.10) in small samples. The alternate continuity correction is applied to Methods A and B. As expected, for large samples little difference in power was found between Method A and the direct method. In small samples, Method B has slightly less power than A, demonstrating that B is a more conservative test.

For the large sample size scenarios, the marginal expected cell counts for each response (Figure 5.2) are all greater than 10, supporting the use of the normal approximation to the binomial. The bivariate expected cell counts (Figure 5.1) on the whole have at least 75% of counts >10, and 25% from 5-10. Expected cell counts less than 5 occur rarely. This supports the application of the bivariate normal distribution used in the numerical integration for power. In small samples, all situations have 50% of marginal expected cell counts <10. In most cases, 25% are <5, occasionally 50% are less than 5, but no marginal expected cell counts are less than 3. The methods applied are designed for situations when one or more expected counts per outcome are between 3-10. Bivariate expected cell counts are for the most part quite small, but they are not required to be large since power is assessed through simulation.

Power is evaluated when the odds ratio between outcomes is 1.5, 3, 6, and 12. The $\phi$-coefficient in the bivariate normal distribution ranges from <0.10 for $\psi_{12}=1.5$ to as much as 0.5 for $\psi_{12}=12$. Correlations are larger when response
proportions are less extreme. Correlation has little effect on bivariate power when power is large and variance is small, but as response proportions are closer to 0.5 and as sample size decreases the effect of correlation is more pronounced.

In these examples, power to detect 1-sided bivariate equivalence is reasonable when all expected cell counts are large, but is quite small when expected counts are small. In small sample situations, power to show that test is at least as good as reference is reasonably high only when test is actually better than reference. The bivariate test is conservative, with many power estimates below 0.05.

For reference, Table 5.3 provides the corresponding univariate power, obtained from numerical integration and simulation. Based on the last column of Table 5.3 it can be seen that both small sample methods result in Type-I errors close to 0.05: the modifications appear to have corrected the anti-conservativeness of the normal approximation. Univariate power (1-\(\beta\)) can also be approximated by several formulas. For large samples, the standard formula is \(\Pr[z \leq z_{1-\beta}]\) determined from the standard normal distribution where:

\[
z_{1-\beta} = \frac{\delta - \Delta}{\text{SE}(\delta)} + z_\alpha
\]  
(5.11)

This corresponds to the definition of power derived in equation (5.7). For small samples, the continuity correction is incorporated as:

\[
z_{1-\beta} = \frac{\delta - \Delta - 0.5[(1/n_1 + 1/n_r)]}{\text{SE}(\delta)} + z_\alpha
\]  
(5.12)

however, the alternate continuity correction may be more appropriate. The other two modifications of Method B are not employed in power calculations. Since the postulated response proportions are specified and are not one or zero, the modified response proportions in equation (5.10) are not needed, although they could be applied if desired. Division by \(n_{g} - 1\) in the denominator of the standard error is not necessary because in power calculations it is assumed the population response proportions are known. Approximations using equation (5.12) with the alternate continuity correction are similar to row B of Table 5.3, but can be off by \(\pm 0.02\).
For Method A, \( SE'(\hat{\delta}) \) can be substituted into the approximate power equations. This leads to estimates a few percent smaller than simulated power in row A of Table 5.3. Applying the Dunnett and Gent modifications only under \( H_0 \) leads to the equation discussed by Rodary et al (89) for balanced designs with \( n_t = n_r = N/2 \):

\[
z_{1-\beta} = \frac{\sqrt{\frac{N}{2}}(\delta - \Delta) + z_\alpha \sqrt{\pi_r(1-\pi_r) + (\pi_r+\Delta)(1-\pi_r-\Delta)}}{\sqrt{\pi_r(1-\pi_r) + (\pi_t)(1-\pi_t)}}
\]

Gould (93) recommends another version:

\[
z_{1-\beta} = \frac{\sqrt{\frac{N}{2}}(\delta - \Delta) + z_\alpha \sqrt{2(\pi_r(1-\pi_r) + \Delta(1-\delta/2 - 2\pi_r)}}}{\sqrt{\pi_r(1 - \pi_r) + (\pi_t)(1 - \pi_t)}}
\]

(5.14)

A continuity correction could also be added to either formula. Approximate power from these methods generally leads to smaller power than equation (5.12) using \( SE'(\hat{\delta}) \) and the alternate correction. Most results agree within 0.02, but when power is large and sample sizes are small, the Rodary and Gould approximations are too small (i.e., 0.80 vs. 0.63, 0.62 vs. 0.52). The Gould approximation yields slightly smaller results than equation (5.13).
Table 5.1.1
Power for Testing Test-Reference Percent Response is no Less than $\Delta = -0.10$
for Two Dichotomous Outcomes. Parallel Design, $\alpha = 0.05$, n per group = 200,
Obtained via Numerical Integration (Direct Method) and Simulation (Method A)

<table>
<thead>
<tr>
<th>$\delta_1, \delta_2$ $\rightarrow$</th>
<th>$\pi_{r1}$</th>
<th>$\pi_{r2}$</th>
<th>$\psi_{12}$</th>
<th>0.0</th>
<th>-0.02</th>
<th>0.0</th>
<th>0.0</th>
<th>-0.05</th>
<th>-0.075</th>
<th>0.0</th>
<th>-0.02</th>
<th>-0.05</th>
<th>-0.075</th>
<th>-0.05</th>
<th>-0.075</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90 0.90 12</td>
<td>.92</td>
<td>.70</td>
<td>.44</td>
<td>.18</td>
<td>.27</td>
<td>.07</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90 0.85 12</td>
<td>.85</td>
<td>.62</td>
<td>.37</td>
<td>.16</td>
<td>.23</td>
<td>.06</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90 0.90 12</td>
<td>.91</td>
<td>.71</td>
<td>.46</td>
<td>.17</td>
<td>.26</td>
<td>.07</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90 0.85 12</td>
<td>.85</td>
<td>.62</td>
<td>.37</td>
<td>.15</td>
<td>.23</td>
<td>.06</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\delta_k = \pi_{tk} - \pi_{rk}$.
$\pi_{rk}$ = proportion with positive response for treatment r, variable k
$\psi_{12}$ = odds ratio between variable 1 and 2 (same for each treatment)
Table 5.1.2
Power for Testing Test-Reference Percent Response is no Less than Δ = 0.10
for Two Dichotomous Outcomes. Parallel Design, α = 0.05, n per group = 40
Obtained via Simulation

<table>
<thead>
<tr>
<th>δ₁, δ₂ →</th>
<th>π₀₁</th>
<th>π₀₂</th>
<th>ψ₁₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.0</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

Method A

|          | 0.90| 0.90| 12  | .69 | .26 | .10 | .07 | .15 | .026| .041 |
|          | 0.90| 0.85| 12  | .56 | .21 | .09 | .06 | .12 | .022| .034 |
|          | 0.90| 0.90| 6   | .68 | .23 | .09 | .06 | .13 | .019| .036 |
|          | 0.90| 0.85| 6   | .54 | .19 | .09 | .05 | .12 | .018| .030 |
|          | 0.90| 0.90| 3   | .67 | .21 | .08 | .05 | .13 | .015| .028 |
|          | 0.90| 0.85| 3   | .54 | .16 | .08 | .05 | .12 | .012| .025 |
|          | 0.90| 0.90| 1.5 | .66 | .19 | .07 | .043| .13 | .011| .022 |
|          | 0.90| 0.85| 1.5 | .52 | .16 | .06 | .038| .12 | .010| .021 |

Method B

|          | 0.90| 0.90| 12  | .64 | .21 | .08 | .05 | .13 | .020| .031 |
|          | 0.90| 0.85| 12  | .51 | .18 | .08 | .05 | .12 | .018| .031 |
|          | 0.90| 0.90| 6   | .63 | .18 | .07 | .05 | .11 | .013| .028 |
|          | 0.90| 0.85| 6   | .49 | .16 | .07 | .05 | .11 | .014| .026 |
|          | 0.90| 0.90| 3   | .62 | .16 | .06 | .041| .11 | .010| .023 |
|          | 0.90| 0.85| 3   | .49 | .14 | .07 | .04 | .11 | .009| .022 |
|          | 0.90| 0.90| 1.5 | .61 | .15 | .05 | .033| .11 | .007| .019 |
|          | 0.90| 0.85| 1.5 | .48 | .13 | .06 | .034| .11 | .007| .019 |

δₖ = πₖ - πᵣₖ
πᵣₖ = proportion with positive response for treatment r, variable k
ψ₁₂ = odds ratio between variable 1 and 2 (same for each treatment)
Table 5.2.1
Power for Testing Test-Reference Percent Response is no Less than $\Delta = -0.20$
for Two Dichotomous Outcomes. Parallel Design, $\alpha = 0.05$, $n$ per group = 100
Obtained via Numerical Integration (Direct Method) and Simulation (Method A)

<table>
<thead>
<tr>
<th>$\delta_1$, $\delta_2 \rightarrow$</th>
<th>$\pi_{r1}$</th>
<th>$\pi_{r2}$</th>
<th>$\psi_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
<td>-0.05</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Direct Method                     |           |           |             |            |            |             |
|-----------------------------------|------------|------------|-------------|
|                                   | 0.70       | 0.70       | 12          |           |           |             |
| 0.70                              | 6          | 3          | 1.5         |           |           |             |
| 0.70                              | 6          | 3          | 1.5         |           |           |             |
|                                   | 0.70       | 0.60       | 12          |           |           |             |
| 0.70                              | 6          | 3          | 1.5         |           |           |             |
|                                   | 0.70       | 0.60       | 12          |           |           |             |
| 0.70                              | 6          | 3          | 1.5         |           |           |             |

| Method A                          |           |           |             |            |            |             |
|-----------------------------------|------------|------------|-------------|
|                                   | 0.70       | 0.70       | 12          |           |           |             |
| 0.70                              | 6          | 3          | 1.5         |           |           |             |
|                                   | 0.70       | 0.60       | 12          |           |           |             |
| 0.70                              | 6          | 3          | 1.5         |           |           |             |

$\delta_k = \pi_{tk} - \pi_{rk}$
$\pi_{rk}$ = proportion with positive response for treatment $r$, variable $k$
$\psi_{12}$ = odds ratio between variable 1 and 2 (same for each treatment)
Table 5.2.2
Power for Testing Test-Reference Percent Response is no Less than $\Delta = -0.20$
for Two Dichotomous Outcomes. Parallel Design, $\alpha = 0.05$, n per group = 15

Obtained via Simulation

<table>
<thead>
<tr>
<th>$\delta_1$, $\delta_2 \rightarrow$</th>
<th>$\pi_{r1}$</th>
<th>$\pi_{r2}$</th>
<th>$\psi_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$\pi_{r1}$</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

Method A

<table>
<thead>
<tr>
<th></th>
<th>0.70</th>
<th>0.70</th>
<th>12</th>
<th>.46</th>
<th>.18</th>
<th>.11</th>
<th>.07</th>
<th>.12</th>
<th>.031</th>
<th>.038</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td>.42</td>
<td>.17</td>
<td>.09</td>
<td>.05</td>
<td>.12</td>
<td>.023</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>.40</td>
<td>.14</td>
<td>.07</td>
<td>.044</td>
<td>.11</td>
<td>.016</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td></td>
<td>.37</td>
<td>.11</td>
<td>.06</td>
<td>.034</td>
<td>.09</td>
<td>.011</td>
<td>.020</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.70</th>
<th>0.60</th>
<th>12</th>
<th>.39</th>
<th>.17</th>
<th>.11</th>
<th>.06</th>
<th>.12</th>
<th>.028</th>
<th>.036</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td>.38</td>
<td>.15</td>
<td>.08</td>
<td>.05</td>
<td>.11</td>
<td>.024</td>
<td>.034</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>.35</td>
<td>.13</td>
<td>.07</td>
<td>.044</td>
<td>.11</td>
<td>.018</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td></td>
<td>.33</td>
<td>.11</td>
<td>.05</td>
<td>.034</td>
<td>.09</td>
<td>.011</td>
<td>.020</td>
</tr>
</tbody>
</table>

Method B

<table>
<thead>
<tr>
<th></th>
<th>0.70</th>
<th>0.70</th>
<th>12</th>
<th>.40</th>
<th>.15</th>
<th>.09</th>
<th>.06</th>
<th>.10</th>
<th>.028</th>
<th>.035</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td>.36</td>
<td>.14</td>
<td>.08</td>
<td>.05</td>
<td>.10</td>
<td>.021</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>.34</td>
<td>.12</td>
<td>.06</td>
<td>.038</td>
<td>.09</td>
<td>.015</td>
<td>.023</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td></td>
<td>.32</td>
<td>.09</td>
<td>.044</td>
<td>.028</td>
<td>.07</td>
<td>.010</td>
<td>.016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.70</th>
<th>0.60</th>
<th>12</th>
<th>.35</th>
<th>.15</th>
<th>.10</th>
<th>.06</th>
<th>.12</th>
<th>.027</th>
<th>.034</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td>.34</td>
<td>.13</td>
<td>.08</td>
<td>.05</td>
<td>.10</td>
<td>.022</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>.31</td>
<td>.12</td>
<td>.06</td>
<td>.041</td>
<td>.10</td>
<td>.017</td>
<td>.027</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td></td>
<td>.29</td>
<td>.10</td>
<td>.05</td>
<td>.030</td>
<td>.08</td>
<td>.010</td>
<td>.018</td>
</tr>
</tbody>
</table>

$\delta_k = \pi_{rk} - \pi_{r_k}$

$\pi_{r_k}$ = proportion with positive response for treatment r, variable k

$\psi_{12}$ = odds ratio between variable 1 and 2 (same for each treatment)
Table 5.3
Power for Testing (Test - Reference) Percent Response is no less than \( \Delta \)
Parallel Design, \( \alpha = 0.05 \)
Obtained via Numerical Integration (Direct Method) and Simulation (Methods A, B)

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>( n ) per group</th>
<th>( \pi_r )</th>
<th>( \delta \rightarrow )</th>
<th>0.0</th>
<th>-0.02</th>
<th>-0.05</th>
<th>-0.075</th>
<th>-0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.10</td>
<td>200 D</td>
<td>0.90</td>
<td>0.85</td>
<td>.95</td>
<td>.82</td>
<td>.45</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td></td>
<td>.88</td>
<td>.70</td>
<td>.37</td>
<td>.16</td>
<td>.05</td>
</tr>
<tr>
<td>-0.10</td>
<td>200 A</td>
<td>0.90</td>
<td>0.85</td>
<td>.95</td>
<td>.82</td>
<td>.45</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.85</td>
<td></td>
<td>.88</td>
<td>.70</td>
<td>.37</td>
<td>.15</td>
<td>.05</td>
</tr>
<tr>
<td>( \delta \rightarrow )</td>
<td>0.05</td>
<td>0.0</td>
<td>-0.05</td>
<td>-0.075</td>
<td>-0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.10</td>
<td>40 A</td>
<td>0.90</td>
<td>0.85</td>
<td>.81</td>
<td>.43</td>
<td>.16</td>
<td>.10</td>
<td>.051</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.63</td>
<td>.33</td>
<td>.15</td>
<td>.08</td>
<td>.045</td>
</tr>
<tr>
<td>-0.10</td>
<td>40 B</td>
<td>0.90</td>
<td>0.85</td>
<td>.78</td>
<td>.38</td>
<td>.14</td>
<td>.08</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.60</td>
<td>.31</td>
<td>.13</td>
<td>.08</td>
<td>.043</td>
</tr>
<tr>
<td>( \delta \rightarrow )</td>
<td>0.0</td>
<td>-0.05</td>
<td>-0.1</td>
<td>-0.15</td>
<td>-0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.20</td>
<td>100 D</td>
<td>0.70</td>
<td>0.60</td>
<td>.93</td>
<td>.73</td>
<td>.44</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.89</td>
<td>.69</td>
<td>.41</td>
<td>.18</td>
<td>.05</td>
</tr>
<tr>
<td>-0.20</td>
<td>100 A</td>
<td>0.70</td>
<td>0.60</td>
<td>.93</td>
<td>.73</td>
<td>.43</td>
<td>.17</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.89</td>
<td>.70</td>
<td>.42</td>
<td>.17</td>
<td>.05</td>
</tr>
<tr>
<td>( \delta \rightarrow )</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.1</td>
<td>-0.15</td>
<td>-0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.20</td>
<td>15 A</td>
<td>0.70</td>
<td>0.60</td>
<td>.61</td>
<td>.33</td>
<td>.15</td>
<td>.09</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.52</td>
<td>.30</td>
<td>.14</td>
<td>.09</td>
<td>.052</td>
</tr>
<tr>
<td>-0.20</td>
<td>15 B</td>
<td>0.70</td>
<td>0.60</td>
<td>.55</td>
<td>.29</td>
<td>.14</td>
<td>.08</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.50</td>
<td>.30</td>
<td>.15</td>
<td>.09</td>
<td>.052</td>
</tr>
</tbody>
</table>

\( \delta = \pi_t - \pi_r \)
\( \pi_r \) = proportion with positive response for treatment \( r \)
D/A/B = standard error method (Direct/A/B)

### 5.3 Ratios of Proportions

This section considers ratios of proportions, defined \( \theta = \pi_t / \pi_r \). The one-sided hypothesis that the test treatment is no worse than \( \Theta \) percent of the reference is
stated $H_0: \pi_t < \Theta \pi_r$, or $\pi_t - \Theta \pi_r < 0$ vs $H_1: \pi_t - \Theta \pi_r \geq 0$, where $\Theta$ is less than 1.0, generally 0.9 or 0.8. If a high response rate is bad rather than good, then the inequalities in the hypothesis are switched, and $\Theta$ is generally 1.1 or 1.2 (Gould 93). As with continuous responses, symmetry on the ratio scale suggests $\frac{1}{0.9}=1.11$ and $\frac{1}{0.8}=1.25$ are perhaps more appropriate upper-sided values for $\Theta$.

This section evaluates power of univariate and bivariate test statistics for ratios of proportions. Derivations and application steps are very similar to those for differences of proportions. The distribution of the univariate test statistic is derived assuming large samples, modifications to test statistics for small samples are described, and power is calculated for large and small sample settings.

5.3.1 Model Assuming Normal Distribution

The distribution for this statistic assumes the same model as described in 5.2.1, with the modification that $p_{t1}$ and $p_{t2}$ are multiplied by $\Theta$, where $\Theta$ is the equivalence criterion. Using the normal approximation to the binomial, and the knowledge that $\text{var}(ax) = a^2 \text{var}(x)$, the test and reference components of the test statistic are distributed as:

\[
\begin{bmatrix}
    p_{t1} \\
    p_{t2}
\end{bmatrix}
\sim N
\begin{bmatrix}
    \pi_{t1} \\
    \pi_{t2}
\end{bmatrix},
\begin{bmatrix}
    \frac{\pi_{t1}(1 - \pi_{t1})}{n_t} & \frac{\pi_{t11}\pi_{t22} - \pi_{t12}\pi_{t21}}{n_t} \\
    \frac{\pi_{t11}\pi_{t22} - \pi_{t12}\pi_{t21}}{n_t} & \frac{\pi_{t2}(1 - \pi_{t2})}{n_t}
\end{bmatrix} = \Sigma_t
\]

\[
\begin{bmatrix}
    \Theta p_{r1} \\
    \Theta p_{r2}
\end{bmatrix}
\sim N
\begin{bmatrix}
    \Theta \pi_{r1} \\
    \Theta \pi_{r2}
\end{bmatrix},
\begin{bmatrix}
    \frac{\Theta \pi_{r1}(1 - \pi_{r1})}{n_r} & \frac{\Theta^2(\pi_{r11}\pi_{r22} - \pi_{r12}\pi_{r21})}{n_r} \\
    \frac{\Theta^2(\pi_{r11}\pi_{r22} - \pi_{r12}\pi_{r21})}{n_r} & \frac{\Theta^2 \pi_{r2}(1 - \pi_{r2})}{n_r}
\end{bmatrix} = \Sigma_{\Theta r}
\]

The odds ratio of outcome 2 for positive vs negative outcome 1 is again assumed to be equal for the two treatment groups, and the correlation between outcome 1 and 2 in treatment group $g$ is $\rho_g$ as described in equation (5.2). The test statistic is obtained by subtracting test - reference treatment groups, to
obtain $\hat{d}_1 = p_{t1} - \Theta p_{r1}$ and $\hat{d}_2 = p_{t2} - \Theta p_{r2}$, with the distribution:

$$
\begin{bmatrix}
\hat{d}_1 \\
\hat{d}_2
\end{bmatrix} \sim N
\begin{bmatrix}
d_1 = \pi_{t1} - \Theta \pi_{r1} \\
d_2 = \pi_{t2} - \Theta \pi_{r2}
\end{bmatrix}, \quad \Sigma_\theta = \Sigma_t + \Sigma_\theta
$$

(5.15)

The correlation of outcomes 1 and 2 for the statistic $d$, $\rho_{\hat{d}_1, \hat{d}_2}$, is

$$
\rho_{\hat{d}_1, \hat{d}_2} = \frac{\text{cov}(\hat{d}_1, \hat{d}_2)}{\sqrt{\text{var}(\hat{d}_1)\text{var}(\hat{d}_2)}}
$$

where

$$
\text{var}(\hat{d}_k) = \left[ \frac{\pi_{tk}(1 - \pi_{tk})}{n_t} + \frac{\Theta^2 \pi_{rk}(1 - \pi_{rk})}{n_r} \right]
$$

and

$$
\text{cov}(\hat{d}_1, \hat{d}_2) = \left[ \frac{\pi_{t11}\pi_{t22} - \pi_{t12}\pi_{t21}}{n_t} + \frac{\Theta^2 (\pi_{r11}\pi_{r22} - \pi_{r12}\pi_{r21})}{n_r} \right]
$$

(5.16)

In the univariate case, one-sided equivalence is concluded if $\frac{\hat{d}_k}{\text{SE}(\hat{d}_k)}$ is significantly greater than zero (Gould 93). Bivariate equivalence requires equivalence to be concluded for both univariate outcomes.

The statistic $\hat{d}_k$ has some undesirable properties. For tests where a high response rate is good, $\Theta$ is less than 1.0 and the variance of $\hat{d}_k$ is smaller than the variance of the difference of proportions $\delta_k$. However, when a high response rate is bad, $\Theta$ is greater than 1.0 and the variance is larger: the variance for the test of $\Theta$ and its reciprocal are not the same. In addition, the test statistic itself is not complementary: if response proportions are switched from positive to negative outcomes, then $\theta_k^* = \frac{1 - \pi_{tk}}{1 - \pi_{rk}} \neq \frac{1}{\theta_k}$. Results are not compatible when response proportions are switched from positive to negative and if the labeling of test and reference is arbitrary, results are also not compatible when test and reference are switched.

Perhaps a more appropriate strategy for $\theta$ would involve modeling the bivariate distribution of the natural logarithm of the proportion ratios, or assessing equivalence with Fieller's method. Although these methods again do not provide complementary test statistics, the effect of $\Theta^2$ on the variance is removed.
It can be shown by algebra that when the lopsided effect of \( \Theta^2 \) on the variance is removed, then the complementary test statistics when \( \pi_{rk} \) is a positive response proportion are \( \frac{\pi_{tk}}{\pi_{rk}} \leq (c+1) \) and \( \frac{(1-\pi_{tk} + c)}{1-\pi_{rk}} \geq (c+1) \).

Since the proposed method is asymmetric with respect to \( \Theta \), the most conservative approach is to define the test such that \( \Theta > 1.0 \), as done by Gould (93). However, in order to maintain consistency with the rest of the chapter, power is evaluated in this section assuming \( \Theta < 1.0 \). All methods are directly applicable to the complementary case. The difference in power between the two approaches is evaluated for an example in Section 5.5.

5.3.2 Bivariate Power Assuming Normality

In this section, large-sample power of the test statistic is derived. The one-sided hypothesis that test is no worse than a percent of reference is stated \( H_0 \): \( d_k = (\pi_t - \Theta \pi_r) < 0 \) vs \( H_A \): \( d_k \geq 0 \), where \( \Theta < 1.0 \). The univariate power of the test statistic \( \hat{d}_k \) is defined as the probability of rejecting the null hypothesis. Similar steps are followed as in section 5.2.2 to obtain

\[
\text{Power} = \Pr \left[ \frac{-d_k}{\text{SE}(\hat{d}_k)} + Z_{1-\alpha} < Z \right],
\]

(5.17)

where \( \text{SE}(\hat{d}_k) = \sqrt{\frac{\pi_{tk}(1-\pi_{tk})}{n_t} + \frac{\Theta_k^2 \pi_{rk}(1-\pi_{rk})}{n_r}} \). With reversal of direction, then:

\[
\text{Power} = \Pr \left[ Z < \frac{d_k}{\text{SE}(\hat{d}_k)} + Z_{\alpha} \right],
\]

(5.18)

If the lower limit is referred to as \( L \), and if \( \pi_{tk} \) and \( \pi_{rk} \) are assumed known, then the power is calculated by numerical integration of the standard normal distribution from \( L \) to \( \infty \). Alternately, the distribution is integrated from \( -\infty \) to \( U \), where \( U \) is the limit in (5.18).

If it is also assumed that the \( \pi_{g11}, \pi_{g12}, \pi_{g21}, \) and \( \pi_{g22} \) are known, then
bivariate power of statistics \( \hat{d}_1 \) and \( \hat{d}_2 \) is obtained by numerical integration from \( L_1 \) to \( \infty \) and \( L_2 \) to \( \infty \) of a standardized bivariate normal distribution with correlation \( \rho_2 \), as described in equation (5.16).

5.3.3 Modifications for Small Samples

When sample sizes are smaller so that some marginal expected counts are between 3-10, then the test statistics are modified to account for the anticonservativeness of the normal approximation. Two modifications are presented, corresponding to ratio versions of the methods presented in section 5.2.3.

The ratio version of Method A involves deriving the expected cell counts under the null hypothesis that \( \pi_{tk} = \Theta \pi_{rk} \). Modified expected counts are obtained by setting \( \hat{\pi}_{tk} = \Theta \hat{\pi}_{rk} \), \( \hat{m}_{tk} = n_t(\Theta \hat{\pi}_{rk}) \), \( \hat{m}_{rk} = n_r(\hat{\pi}_{rk}) \), then using \( \hat{m}_{tk} + \hat{m}_{rk} = n_{k,1} \) to solve for \( \hat{\pi}_{rk} \) and \( \hat{\pi}_{tk} \) to obtain:

\[
\hat{\pi}_{rk} = \frac{n_{tk} + n_{rk}}{\Theta n_t + n_r} \quad \hat{\pi}_{tk} = \frac{\Theta (n_{tk} + n_{rk})}{\Theta n_t + n_r}
\] (5.19)

The modified univariate standard error is \( SE'(\hat{d}_k) = \sqrt{\frac{\hat{\pi}_{tk}(1-\hat{\pi}_{tk})}{n_t} + \frac{\Theta^2 \hat{\pi}_{rk}(1-\hat{\pi}_{rk})}{n_r}}. \)

Unlike Section 5.2.3, the standard error for Method A is never undefined when the response proportions are all one or zero, due to the formulas for the modified expected cell counts. The modified expected counts are applicable to a \( \chi^2 \) test of \( d_k \), but as in Section 5.2.3, results based on a normal approximation are expected to be very similar.

In small samples, a continuity correction should be applied. The correction that would apply when testing \( \Theta \) with Method A is \( z = \frac{\hat{d}_k - 0.5[1/n_t + \Theta/n_r]}{SE(\hat{d}_k)}. \) This is derived by manipulating the equation for the square root of the \( \chi^2 \) corrected formula for one cell (= observed count - expected count - 1/2) to the form \( \frac{\hat{d}_k - \text{correction}}{SE} \). The SE achieved in this manner is not exactly equal to \( SE(\hat{d}_k) \), since as Dunnett and Gent explained, the \( z \) and \( \chi^2 \) are not exactly identical under
the non-null case. As in Section 5.2.3, power of the test with this continuity correction was found to be too conservative for the response proportions evaluated. The alternate correction of \(-0.5[1/(n_r+n_r)]\) was again found to have better power while maintaining a significance level of 0.05.

Method B employs the same modifications as in Section 5.2.3, so the standard error is 
\[
SE^*(\hat{\delta}_k) = \sqrt{\frac{\hat{\pi}_{tk}(1-\hat{\pi}_{tk})}{n_r-1} + \frac{\Theta^2 \hat{\pi}_{rk}(1-\hat{\pi}_{rk})}{n_r-1}}
\]
where 
\[
\hat{\pi}_{pk} = \frac{(n_{sk}+0.5)}{n_g+1}.
\]
The alternate continuity correction is also applied to Method B, although the method has slightly better power and still maintains significance level near 0.05 without a continuity correction.

5.3.4 Bivariate Power Assuming Normal Approximation

Simulation of univariate and bivariate power follows the same steps as described in Section 5.2.4, but is applied to the statistics in Section 5.3.3.

5.3.5 Applications and Results

Table 5.4 presents bivariate power for testing whether test is no worse than 0.9 \times\text{reference for } \pi_r=0.90 \text{ and } 0.85. Table 5.5 provides power for testing } \Theta = 0.8, \text{ with more moderate response proportions. Each table has a large-sample section obtained from numerical integration, and a small-sample section obtained by simulation using Method A with the alternate continuity correction described in Section 5.2.3. Method B with the same correction resulted in power estimates slightly less than Method A, by no more than 0.02, indicating that Method B is a more conservative approach. Method A was not evaluated for large samples since the direct method and Method A resulted in similar bivariate power for } \delta \text{ in Section 5.2. These power estimates may be considered anti-conservative, since the } \Theta \text{ being tested is } < 1.0, \text{ although the significance level is maintained at } 0.05.

For large sample sizes, marginal expected cell counts for each response are
>10. The bivariate expected cell counts have at least 75% of counts >10, and
25% in the range 5-10. Expected cell counts <5 occur only occasionally,
supporting the application of the bivariate normal distribution in bivariate power
calculation. In small samples, all situations have 50% of marginal expected cell
counts <10, but no expected cell counts are less than 3. Bivariate expected cell
counts are quite small, but are not required to be large since power is assessed
through simulation.

Table 5.4
Power for Testing Test/Reference Percent Response is no less than Θ = 0.9
for Two Dichotomous Outcomes. Parallel Design, α = 0.05,
Obtained via Numerical Integration (n=200) and Simulation (n=40)

|               | 1.0  | 1.0  | 1.0  | 1.053 | 0.95 | 0.925 | 1.0  \\
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0</td>
<td>0.95</td>
<td>0.925</td>
<td>0.95</td>
<td>0.95</td>
<td>0.925</td>
<td>0.9</td>
</tr>
<tr>
<td>n per group=200</td>
<td>Direct Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>12</td>
<td>.88</td>
<td>.41</td>
<td>.17</td>
<td>.42</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td>.88</td>
<td>.40</td>
<td>.17</td>
<td>.42</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.88</td>
<td>.40</td>
<td>.16</td>
<td>.42</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td>.87</td>
<td>.39</td>
<td>.16</td>
<td>.42</td>
<td>.18</td>
</tr>
<tr>
<td>0.90</td>
<td>0.85</td>
<td>12</td>
<td>.77</td>
<td>.32</td>
<td>.14</td>
<td>.32</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td>.76</td>
<td>.31</td>
<td>.14</td>
<td>.32</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.76</td>
<td>.31</td>
<td>.14</td>
<td>.32</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td>.75</td>
<td>.30</td>
<td>.13</td>
<td>.32</td>
<td>.14</td>
</tr>
<tr>
<td>n per group=40</td>
<td>Method A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>12</td>
<td>.20</td>
<td>.09</td>
<td>.06</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td>.18</td>
<td>.07</td>
<td>.05</td>
<td>.11</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.16</td>
<td>.06</td>
<td>.045</td>
<td>.11</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td>.15</td>
<td>.06</td>
<td>.038</td>
<td>.11</td>
<td>.026</td>
</tr>
<tr>
<td>0.90</td>
<td>0.85</td>
<td>12</td>
<td>.17</td>
<td>.08</td>
<td>.06</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td>.15</td>
<td>.07</td>
<td>.05</td>
<td>.11</td>
<td>.035</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.14</td>
<td>.06</td>
<td>.043</td>
<td>.10</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td></td>
<td>.12</td>
<td>.06</td>
<td>.034</td>
<td>.10</td>
<td>.024</td>
</tr>
</tbody>
</table>

θ_k = π_{1k}/π_{rk}
π_{rk} = proportion with positive response for treatment r, variable k
ψ_{12} = odds ratio between variable 1 and 2 (same for each treatment)
Table 5.5
Power for Testing Test/Reference Percent Response is no less than $\Theta = 0.8$
for Two Dichotomous Outcomes. Parallel Design, $\alpha = 0.05$,
Obtained via Numerical Integration ($n=100$) and Simulation ($n=15$)

<table>
<thead>
<tr>
<th>$\pi_{r1}$</th>
<th>$\pi_{r2}$</th>
<th>$\psi_{12}$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.0</th>
<th>1.11</th>
<th>0.9</th>
<th>0.9</th>
<th>0.85</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n per group=100</td>
<td>Direct Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.70</td>
<td>12</td>
<td>.65</td>
<td>.29</td>
<td>.14</td>
<td>.31</td>
<td>.17</td>
<td>.06</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>.63</td>
<td>.28</td>
<td>.13</td>
<td>.31</td>
<td>.15</td>
<td>.045</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.62</td>
<td>.27</td>
<td>.12</td>
<td>.31</td>
<td>.13</td>
<td>.035</td>
<td>.045</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>.60</td>
<td>.25</td>
<td>.11</td>
<td>.31</td>
<td>.11</td>
<td>.025</td>
<td>.041</td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.60</td>
<td>12</td>
<td>.53</td>
<td>.23</td>
<td>.12</td>
<td>.24</td>
<td>.14</td>
<td>.05</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>.51</td>
<td>.22</td>
<td>.11</td>
<td>.24</td>
<td>.12</td>
<td>.040</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.50</td>
<td>.21</td>
<td>.11</td>
<td>.24</td>
<td>.11</td>
<td>.031</td>
<td>.045</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>.48</td>
<td>.20</td>
<td>.10</td>
<td>.24</td>
<td>.09</td>
<td>.022</td>
<td>.041</td>
<td></td>
</tr>
<tr>
<td>n per group=15</td>
<td>Method A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.70</td>
<td>12</td>
<td>.12</td>
<td>.07</td>
<td>.05</td>
<td>.09</td>
<td>.044</td>
<td>.025</td>
<td>.034</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>.11</td>
<td>.05</td>
<td>.040</td>
<td>.08</td>
<td>.033</td>
<td>.018</td>
<td>.027</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.08</td>
<td>.05</td>
<td>.032</td>
<td>.06</td>
<td>.026</td>
<td>.012</td>
<td>.021</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>.06</td>
<td>.036</td>
<td>.022</td>
<td>.05</td>
<td>.017</td>
<td>.006</td>
<td>.014</td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.60</td>
<td>12</td>
<td>.09</td>
<td>.07</td>
<td>.042</td>
<td>.08</td>
<td>.037</td>
<td>.022</td>
<td>.029</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>.08</td>
<td>.05</td>
<td>.034</td>
<td>.06</td>
<td>.030</td>
<td>.016</td>
<td>.024</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.07</td>
<td>.040</td>
<td>.025</td>
<td>.05</td>
<td>.020</td>
<td>.011</td>
<td>.018</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>.044</td>
<td>.025</td>
<td>.017</td>
<td>.037</td>
<td>.014</td>
<td>.008</td>
<td>.012</td>
<td></td>
</tr>
</tbody>
</table>

$\theta_k = \pi_{nk}/\pi_{rk}$
$
\pi_{rk} =$ proportion with positive response for treatment r, variable k
$\psi_{12} =$ odds ratio between variable 1 and 2 (same for each treatment)

Each table evaluates power when the odds ratio between outcomes is 1.5, 3, 6, and 12. The $\phi$-coefficient in the bivariate normal distribution is essentially the same as that for differences of proportions (from 0.03 to 0.5). Correlation has a small effect on bivariate power when univariate powers are large and variance is small, but as response proportions are closer to 0.5 and as sample size decreases, the effect of correlation is more pronounced. Power to detect 1-sided bivariate
equivalence is reasonable when all expected cell counts are large, but is small when expected counts are small. For small samples, power to show that test is no worse than reference is large only when test is better than reference. The bivariate test is conservative, with many power estimates below 0.05, even for large sample sizes.

For reference, Table 5.6 provides the corresponding univariate power, obtained from numerical integration and simulation using Method A with the alternate continuity correction. Continuity corrected Method B power estimates are slightly smaller than Method A. Based on the last column of Table 5.6, it appears that Method A results in Type-I errors of the desired 0.05.

Table 5.6
Power for Testing Test/Reference Percent Response is no less than Θ
Parallel Design, α = 0.05
Obtained via Numerical Integration (n large) and Simulation (n small)

<table>
<thead>
<tr>
<th>Θ</th>
<th>n per group</th>
<th>π_r</th>
<th>θ →</th>
<th>1.053</th>
<th>1.0</th>
<th>0.95</th>
<th>0.925</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>200 D .90</td>
<td></td>
<td></td>
<td>.99</td>
<td>.93</td>
<td>.42</td>
<td>.17</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.85</td>
<td></td>
<td></td>
<td>.99</td>
<td>.80</td>
<td>.32</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td>40</td>
<td>A .90</td>
<td></td>
<td></td>
<td>.76</td>
<td>.38</td>
<td>.15</td>
<td>.09</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.85</td>
<td></td>
<td></td>
<td>.56</td>
<td>.30</td>
<td>.13</td>
<td>.09</td>
<td>.05</td>
</tr>
<tr>
<td>0.80</td>
<td>100 D .70</td>
<td></td>
<td></td>
<td>.99</td>
<td>.77</td>
<td>.31</td>
<td>.14</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td></td>
<td></td>
<td>.92</td>
<td>.61</td>
<td>.24</td>
<td>.12</td>
<td>.05</td>
</tr>
<tr>
<td>15</td>
<td>A .70</td>
<td></td>
<td></td>
<td>.45</td>
<td>.24</td>
<td>.13</td>
<td>.08</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td></td>
<td></td>
<td>.30</td>
<td>.19</td>
<td>.10</td>
<td>.07</td>
<td>.05</td>
</tr>
</tbody>
</table>

θ = π_l/π_r,
π_r = proportion with positive response for treatment r
D/A=method (Direct/A)
Univariate power can be approximated by several formulas. In large samples, the standard is:

$$z_{1-\beta} = \frac{d}{\text{SE}(d)} + z_{\alpha}$$  \hspace{1cm} (5.20)

which corresponds to large-sample power derived in equation (5.18). For small samples, the alternate continuity correction is added:

$$z_{1-\beta} = \frac{d - 1/2[1(n_{1r} + n_{2r})]}{\text{SE}(d)} + z_{\alpha}$$  \hspace{1cm} (5.21)

Method A can be approximated by substituting $\text{SE}'(\hat{d})$ into these formulas, or by using the modified expected cell counts under $H_0$. The formula for a balanced design is:

$$z_{1-\beta} = \frac{\sqrt{N(d) + z_{\alpha}}\sqrt{\Theta^2(\pi_r)(1 - \pi_r) + (\Theta\pi_r)(1 - \Theta\pi_r)}}{\sqrt{\Theta^2(\pi_r)(1 - \pi_r) + (\pi_t)(1 - \pi_t)}}$$  \hspace{1cm} (5.22)

Gould (93) suggests a different version. A continuity correction can also be added. Equation (5.21) with Method A standard error and equation (5.22) both produce power estimates that are less than those obtained by simulation. They were generally within a few percent, but occasionally were worse (ie 0.45 vs. 0.39).

5.4 Odds Ratios

A third way of assessing the equivalence of test and reference response proportions is with odds ratios. For each outcome, the odds of positive response for test vs. reference is $\psi_k = \frac{\pi_{tk}(1 - \pi_{rk})}{\pi_{rk}(1 - \pi_{tk})}$ (this should not be confused with $\psi_{g12}$, the odds ratio for outcome 1 vs 2 for treatment group g). One-sided equivalence is accepted if $\psi_k$ is either greater than or less than some pre-specified value, $\Psi$. For instance, if high response rates are good, the one-sided hypothesis is $H_0$: $\psi_k < \Psi$ vs $H_A$: $\psi_k \geq \Psi$, where $\Psi < 1.0$, generally 0.5. If a high response rate is bad, then the inequalities in the hypothesis are switched, and $\Psi > 1.0$, perhaps 2.0.

Partial motivation of the values 0.5 and 2.0 for $\Psi$ stems from evaluating
\[ \frac{\pi_{ik}(1-\pi_{rk})}{\pi_{rk}(1-\pi_{tk})} = \frac{c_1}{c_2}. \] For positive response rates, evaluating equivalence for the ratios separately requires \( \frac{\pi_{ik}}{\pi_{rk}} > c_1 \) and \( \frac{(1-\pi_{ik})}{(1-\pi_{rk})} < c_2 \), or \( \frac{(1-\pi_{rk})}{(1-\pi_{tk})} > c_2 \). If \( c_1 = 0.80 \) and \( c_2 = 1.25 \), then \( c_1/c_2 = 0.64 \), so 0.64 is a reasonable value for \( \Psi \). Similarly, for negative response rates a reasonable \( \Psi \) is 1.56. Variance for odds ratios tends to be larger, however, so less conservative values such as 0.5 and 2.0 may be more appropriate.

5.4.1 Model Assuming Normal Distribution

Under large samples, we can assume that the natural logarithm of the sample odds ratio \( \tilde{\lambda}_k = \ln(\tilde{\psi}_k) \) follows a normal distribution with mean \( \lambda_k = \ln(\psi_k) \), the population log-odds ratio, and variance estimated by the sum of the reciprocals of the observed cell counts (Fleiss). For each outcome,

\[
\tilde{\lambda}_k \sim N \left[ \lambda_k, \frac{1}{n_{ik}(1-\pi_{ik})} + \frac{1}{n_{rk}(1-\pi_{rk})} \right]
\]

(5.23)

The variance is estimated in the sample by \( \left[ \frac{1}{n_{ik} + n_{rk}} + \frac{1}{n_{nt} + n_{nr}} \right] \).

The bivariate distribution for outcomes 1 and 2 under large samples is:

\[
\begin{bmatrix}
\tilde{\lambda}_1 \\
\tilde{\lambda}_2
\end{bmatrix}
\sim N
\begin{bmatrix}
\lambda_1 \\
\lambda_2
\end{bmatrix},
\begin{bmatrix}
v_1 & \text{cov} \\
\text{cov} & v_2
\end{bmatrix} = \Sigma_{\lambda}
\]

where \( \text{cov}(\tilde{\lambda}_1, \tilde{\lambda}_2) = \frac{\pi_{t11}\pi_{r22} - \pi_{t12}\pi_{r21}}{n_t(\pi_{t1})(1-\pi_{t1})(1-\pi_{t2})} + \frac{\pi_{r11}\pi_{t22} - \pi_{r12}\pi_{t21}}{n_r(\pi_{r1})(1-\pi_{r1})(1-\pi_{r2})} \)

(5.24)

The variance-covariance matrix is derived from the formula

\[
\Sigma_{\lambda} = A_2D^{-1}a_1V_pA'_pD^{-1}a'_1A_2'
\]

(5.25)

(Koch and Imrey 85). The vector of log-odds ratios is \( \lambda = A_2\ln A_1p \), where \( p \) is an \((8 \times 1)\) vector of bivariate responses = \((p_{t11}...p_{t22}, p_{r11}...p_{r22})'\). \( V_p \), the covariance matrix of the responses assuming a multinomial distribution, is an \((8 \times 8)\) block diagonal matrix with each block equal to \([D_{x} - \pi_{p}\pi_{p}']/n_g\), where \( \pi_{p} \) are the vectors
of responses for each treatment group \(=(\pi_{g11}...\pi_{g22})'\) (Koch and Imrey 85). \(A_1\) forms the two univariate responses for each treatment:

\[
A_1 = \begin{bmatrix}
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1
\end{bmatrix} \otimes I_2 , \text{ where } \otimes \text{ is the left Kronecker product, } a_1 = A_1 p, \text{ and } D_{a_1}
\]

is the diagonal matrix with elements \(a_1\). \(A_2\) forms the log-odds ratios: \(A_2 = \begin{bmatrix}
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & 1 & 1
\end{bmatrix}\). This is the framework assumed for logistic regression, except that for logistic regression, the \(A_2\) matrix forms log-odds, or logits, instead of their ratios. The correlation between outcome 1 and 2 is \(\rho_\lambda = \frac{\text{cov}(\lambda_1, \lambda_2)}{\sqrt{\text{var}(\lambda_1) \text{var}(\lambda_2)}}\).

In the univariate case, one-sided equivalence is concluded if \(\frac{\hat{\lambda}_k - \Lambda}{\text{SE}(\hat{\lambda}_k)}\) is significantly greater than zero, where \(\Lambda = \ln(\Psi)\). Bivariate equivalence applies if equivalence is concluded for both univariate outcomes.

5.4.2 Bivariate Power Assuming Normality

The univariate 1-sided hypothesis is stated \(H_0: \ln \left[ \frac{\pi_{tk}(1-\pi_{rk})}{\pi_{rk}(1-\pi_{tk})} \right] < \ln(\Psi)\), or \(\lambda_k < \Lambda\) where \(\Psi < 1\) (test worse than reference), vs \(H_A: \lambda_k \geq \Lambda\) (test within equivalence range of reference). The univariate power of the test statistic \(\hat{\lambda}_k\) is defined as the probability of rejecting the null hypothesis. By following steps similar to Section 5.2.2, power is found to be:

\[
\text{Power} = \Pr \left[ \frac{\Lambda - \lambda_k}{\text{SE}(\hat{\lambda}_k)} + Z_{1-\alpha} < Z \right], \quad (5.26)
\]

where \(\text{SE}(\hat{\lambda}_k)\) is estimated as \(\sqrt{\frac{1}{n_{tk}} + \frac{1}{n_{tr}n_{tk}} + \frac{1}{n_{rk}} + \frac{1}{n_{rt}n_{rk}}}\). If the direction is reversed, then power is defined as

\[
\text{Power} = \Pr \left[ Z < \frac{\lambda_k - \Lambda}{\text{SE}(\hat{\lambda}_k)} + Z_\alpha \right], \quad (5.27)
\]

If the lower limit is referred to as \(L\), and if \(\pi_{tk}\) and \(\pi_{rk}\) are assumed known, then the power is calculated by numerical integration of the distribution from \(L\) to \(\infty\).
Alternately, for the opposite direction, the distribution is integrated from $-\infty$ to $U$, where $U$ is the upper limit in (5.27).

If it is also assumed that $\pi_{g11}$, $\pi_{g12}$, $\pi_{g21}$, and $\pi_{g22}$ are known, then bivariate power of statistics $\hat{\lambda}_1$ and $\hat{\lambda}_2$ is obtained by numerical integration from $L_1$ to $\infty$ and $L_2$ to $\infty$ of a standardized bivariate normal distribution with correlation $\rho_h$.

5.4.3 Modifications for Small Samples

When sample sizes are smaller so that some expected counts are between 3-10, then the test statistics are modified to correct for crudeness of the normal approximation.

The odds ratio version of Method A involves deriving the expected cell counts under the null hypothesis that $\psi_k = \Psi$. For each outcome the formula for the odds ratio is used to solve $\hat{\pi}_{rk} = \frac{\hat{\pi}_{tk}}{\hat{\pi}_{tk} + \Psi(1-\hat{\pi}_{tk})}$. This is plugged into the equations $\hat{m}_{tk} = n_t \hat{\pi}_{tk}$, $\hat{m}_{rk} = n_r \hat{\pi}_{rk}$, and then $\hat{m}_{tk} + \hat{m}_{rk} = n_{k,1}$ is used to solve for $\hat{\pi}_{rk}$ and $\hat{\pi}_{tk}$ in terms of the observed counts and $\Psi$. The modified expected counts and corresponding standard error are (Fleiss):

$$m_{tk} = \frac{\Psi(n_{tk} + n_{k,1}) + (n_r - n_{k,1}) - \sqrt{[\Psi(n_{tk} + n_{k,1}) + (n_r - n_{k,1})]^2 - 4n_{tk}n_{k,1}\Psi(1-\Psi)}}{2(\Psi-1)}$$

$$m_{rk} = n_{k,1} - m_{tk}$$

$$SE'(^{\hat{\lambda}_k}) = \sqrt{\frac{1}{m_{tk}} + \frac{1}{n_r m_{tk}} + \frac{1}{m_{rk}} + \frac{1}{n_r m_{rk}}}$$ (5.28)

Fleiss presents Method A as a $\chi^2$ test of $\psi_k = \Psi$, but like Dunnett and Gent's method, it is expected that the normal approximation with modified expected cell counts will produce very similar results to the $\chi^2$ test.

Method B avoids undefined odds ratios and SEs by adding 0.5 to each observed count. The modified log-odds ratio and its standard error (Fleiss) are:

$$\hat{\lambda}_k^* = \ln \left[ \frac{(n_{tk} + 0.5)(n_r - n_{rk} + 0.5)}{(n_{rk} + 0.5)(n_r - n_{tk} + 0.5)} \right]$$
\[
\text{SE}^* (\lambda_k) = \sqrt{\frac{1}{n_{tk} + 0.5} + \frac{1}{(n_r - n_{tk}) + 0.5} + \frac{1}{n_{rk} + 0.5} + \frac{1}{(n_r - n_{rk}) + 0.5}}. \tag{5.29}
\]

Neither method incorporates a continuity correction because the z score for log-odds is actually a conservative approximation to the \(\chi^2\) test. However, power simulations of these methods presented in Section 5.4.5 suggest that the normal approximation is anti-conservative in small samples when the response proportions are near 0.5. A continuity correction of the form used for \(\delta\) and \(\theta\) yields smaller power estimates in this range and better controls the type I error at 0.05.

Fleiss suggests that Method A is preferable to Method B in small samples. However, when all observed proportions for an outcome are either 0 or 1, then the Method A test statistic and standard error are undefined (division by zero). A proposed adjustment involves using \(\hat{\lambda}_k^*\) as the test statistic, and adding 0.5 to the modified expected cell counts in \(\text{SE}' (\lambda_k)\). In simulations of unmodified Method A, undefined standard errors occur rarely, but could lead to overestimates of power. Simulation using the adjusted estimate and \(\text{SE}\) produce power estimates that are slightly larger, generally by 0.01 for the examples. The tables below provide estimates from the adjusted method.

### 5.4.4 Bivariate Power Assuming Normal Approximation

Simulation of univariate and bivariate power follows the same steps as described in Section 5.2.4, but is applied to the statistics in Section 5.4.3.

### 5.4.5 Applications and Results

Tables 5.7 and 5.8 present bivariate power for testing whether the test vs reference odds ratio is no worse than 0.5, for large and moderate response proportions. The test response proportion can be solved from \(\pi_{rk}\) and \(\psi_k\) by
\[
\pi_{tk} = \frac{\psi_k \pi_{rk}}{1 + \pi_{rk} (\psi_k - 1)}.
\]

Each table has a large-sample section obtained by numerical
integration, and a small-sample section obtained by simulation using the modified Method A formulas. Method B power estimates are larger than Method A (by about 0.01) for Tables 5.7-5.8, indicating that Method A is slightly more conservative.

Each table evaluates power when the odds ratio between outcomes is 1.5, 3, 6, and 12. The correlation corresponding to the odds ratios are basically the same as those for tests of $\Delta$ and $\Theta$. The effect of correlation is seen consistently in each column, generally changing power by a difference of 0.03 to 0.07 from odds ratios of 1.5 to 12. Overall, the bivariate power is barely desirable for the large sample size presented. For small sample sizes, power to show that test is no worse than reference is reasonable only when test is better than reference. The numerous p-values below 0.05 show the conservativeness of the bivariate test.

For large sample sizes, marginal expected cell counts for each response are $>10$. At least 75% of bivariate expected counts are $>10$, and 25% are in the range 5-10. Expected cell counts $<5$ occur only occasionally, supporting the application of the bivariate normal distribution in bivariate power calculations. In small samples, all situations have 50% of marginal expected cell counts $<10$, but no expected cell counts are less than 3. Bivariate expected cell counts are quite small, but are not required to be large since power is assessed through simulation.

For reference, Table 5.9 provides the corresponding univariate power obtained from numerical integration and simulation using modified Method A. Method B power estimates are larger than Method A by a few percent. From the last column of Table 5.9, it appears the modified Method A is reaching the correct Type-I error of 0.05, but is perhaps a bit anti-conservative in small samples.
Table 5.7

Power for Testing Odds Ratio for Test vs Reference is no less than $\Psi = 0.5$

for Two Dichotomous Outcomes. Parallel Design, $\alpha = 0.05$

Obtained via Numerical Integration ($n=200$) and Simulation ($n=40$)

<table>
<thead>
<tr>
<th>$\pi_{r1}$</th>
<th>$\pi_{r2}$</th>
<th>$\psi_{12}$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.0</th>
<th>1.25</th>
<th>0.8</th>
<th>0.8</th>
<th>0.667</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>n per group=200</td>
<td>Direct Method</td>
<td></td>
<td>.50</td>
<td>.34</td>
<td>.20</td>
<td>.39</td>
<td>.25</td>
<td>.10</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>12</td>
<td>.48</td>
<td>.32</td>
<td>.19</td>
<td>.38</td>
<td>.23</td>
<td>.08</td>
<td>.042</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>.56</td>
<td>.40</td>
<td>.23</td>
<td>.48</td>
<td>.28</td>
<td>.10</td>
<td>.043</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.60</td>
<td>.45</td>
<td>.27</td>
<td>.52</td>
<td>.30</td>
<td>.12</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>.64</td>
<td>.50</td>
<td>.30</td>
<td>.61</td>
<td>.35</td>
<td>.17</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>n per group=40</td>
<td>Method A</td>
<td></td>
<td>.08</td>
<td>.06</td>
<td>.05</td>
<td>.08</td>
<td>.05</td>
<td>.027</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>12</td>
<td>.07</td>
<td>.05</td>
<td>.04</td>
<td>.07</td>
<td>.05</td>
<td>.040</td>
<td>.021</td>
<td>.020</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>.09</td>
<td>.07</td>
<td>.06</td>
<td>.09</td>
<td>.06</td>
<td>.041</td>
<td>.023</td>
<td>.018</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>.11</td>
<td>.08</td>
<td>.07</td>
<td>.11</td>
<td>.07</td>
<td>.042</td>
<td>.017</td>
<td>.015</td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td>.13</td>
<td>.10</td>
<td>.09</td>
<td>.13</td>
<td>.09</td>
<td>.044</td>
<td>.014</td>
<td>.012</td>
</tr>
</tbody>
</table>

$\psi_k = \pi_{tk}(1-\pi_{tk})/\pi_{r_k}(1-\pi_{tk})$

$\pi_{r_k}$ = proportion with positive response for treatment $r$, variable $k$

$\psi_{12}$ = odds ratio between variable 1 and 2 (same for each treatment)
Table 5.8
Power for Testing Odds Ratio for Test vs Reference is no less than $\Psi = 0.5$
for Two Dichotomous Outcomes. Parallel Design, $\alpha = 0.05$
Obtained via Numerical Integration ($n=100$) and Simulation ($n=15$)

<table>
<thead>
<tr>
<th>$\psi_1$, $\psi_2 \rightarrow$</th>
<th>$\pi_{r1}$</th>
<th>$\pi_{r2}$</th>
<th>$\psi_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.8</td>
<td>0.667</td>
</tr>
<tr>
<td>n per group=100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.70</td>
<td>12</td>
<td>.59</td>
</tr>
<tr>
<td>6</td>
<td>.57</td>
<td>.39</td>
<td>.22</td>
</tr>
<tr>
<td>3</td>
<td>.55</td>
<td>.37</td>
<td>.20</td>
</tr>
<tr>
<td>1.5</td>
<td>.54</td>
<td>.35</td>
<td>.19</td>
</tr>
<tr>
<td>0.70</td>
<td>0.60</td>
<td>12</td>
<td>.62</td>
</tr>
<tr>
<td>6</td>
<td>.60</td>
<td>.41</td>
<td>.23</td>
</tr>
<tr>
<td>3</td>
<td>.59</td>
<td>.40</td>
<td>.22</td>
</tr>
<tr>
<td>1.5</td>
<td>.57</td>
<td>.37</td>
<td>.20</td>
</tr>
<tr>
<td>n per group=15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.70</td>
<td>12</td>
<td>.14</td>
</tr>
<tr>
<td>6</td>
<td>.13</td>
<td>.10</td>
<td>.07</td>
</tr>
<tr>
<td>3</td>
<td>.11</td>
<td>.08</td>
<td>.06</td>
</tr>
<tr>
<td>1.5</td>
<td>.07</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>0.70</td>
<td>0.60</td>
<td>12</td>
<td>.15</td>
</tr>
<tr>
<td>6</td>
<td>.13</td>
<td>.10</td>
<td>.07</td>
</tr>
<tr>
<td>3</td>
<td>.11</td>
<td>.08</td>
<td>.06</td>
</tr>
<tr>
<td>1.5</td>
<td>.09</td>
<td>.06</td>
<td>.044</td>
</tr>
</tbody>
</table>

$\psi_k = \pi_{rk}(1-\pi_{rk})/\pi_{rk}(1-\pi_{rk})$

$\pi_{rk} =$ proportion with positive response for treatment $r$, variable $k$
$\psi_{12} =$ odds ratio between variable 1 and 2 (same for each treatment)
Table 5.9
Power for Testing Odds Ratio for Test vs Reference is no less than $\Psi$
Parallel Design, $\alpha = 0.05$
Obtained via Numerical Integration (n large) and Simulation (n small)

<table>
<thead>
<tr>
<th>$\Psi$</th>
<th>n per group</th>
<th>$\psi \rightarrow$</th>
<th>1.25</th>
<th>1.0</th>
<th>0.8</th>
<th>0.667</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>200 D .90</td>
<td></td>
<td>.84</td>
<td>.67</td>
<td>.43</td>
<td>.24</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.85</td>
<td></td>
<td>.93</td>
<td>.80</td>
<td>.54</td>
<td>.29</td>
<td>.05</td>
</tr>
<tr>
<td>40</td>
<td>A .90</td>
<td></td>
<td>.29</td>
<td>.22</td>
<td>.15</td>
<td>.10</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.85</td>
<td></td>
<td>.41</td>
<td>.29</td>
<td>.20</td>
<td>.12</td>
<td>.045</td>
</tr>
<tr>
<td>100</td>
<td>D .70</td>
<td></td>
<td>.89</td>
<td>.73</td>
<td>.46</td>
<td>.25</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td></td>
<td>.93</td>
<td>.78</td>
<td>.50</td>
<td>.26</td>
<td>.05</td>
</tr>
<tr>
<td>15</td>
<td>A .70</td>
<td></td>
<td>.39</td>
<td>.27</td>
<td>.19</td>
<td>.14</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>.60</td>
<td></td>
<td>.40</td>
<td>.30</td>
<td>.20</td>
<td>.14</td>
<td>.07</td>
</tr>
</tbody>
</table>

$\psi = \pi_t(1-\pi_r)/\pi_r(1-\pi_t)$
$\pi_r$ = proportion with positive response for treatment r
D/A = method (Direct/A)

Approximate univariate power can be obtained by several formulas. The standard formula is $z_{1-\beta} = \frac{\lambda - A}{SE(\lambda)} + z_\alpha$. Estimates for Method A can be obtained by substituting the standard error from equation (5.28). A formula applying Method A only under the null hypothesis is also possible, but is not derived here. In small samples, the approximate power provided estimates sometimes larger, sometimes smaller than the simulations. Although generally accurate, the approximation significantly underestimated power for small samples when response rates are near 0.5. This anomaly would be improved if the alternate continuity correction were applied to Methods A and B in small sample simulations.
5.5 Example

The clinical equivalence trial of topical treatments for delayed-onset muscle soreness, described in Section 1.4.3 and evaluated for univariate equivalence in Section 3.5.3, is examined as an example of the above methods. Two responses of particular interest are 50% pain reduction by 48 hours assessed during palpation (outcome 1) and during range of motion (outcome 2). The data is:

<table>
<thead>
<tr>
<th>Outcome 1,2</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>32</td>
<td>3</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>Reference</td>
<td>30</td>
<td>5</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

The odds ratio of positive movement response for positive vs negative palpation is 12 for the reference treatment, and 40.88 for test. The Mantel-Haenszel estimate of the common odds ratio is 19.7; the Breslow-Day test for homogeneity of the odds ratios found reasonable homogeneity (p=0.208). In bivariate power calculations, the common odds ratio is rounded to 20. Marginal response proportions are as follows: movement: \( p_t = 0.594, p_r = 0.615 \), palpation: \( p_t = 0.547, p_r = 0.538 \).

Marginal expected cell counts are all well above 10, so direct methods are used to evaluate test statistics. Univariate power is evaluated by numerical integration, but 50% of bivariate expected cell counts are between 3-10, so bivariate power is evaluated via simulation. Univariate equivalence results are presented in Table 5.10. One sided equivalence that test is no worse than some amount of reference is concluded if the p-value is < 0.05 or the equivalence parameter is less than the lower bound of the confidence interval.
Table 5.10
Univariate Analysis for Example Data

<table>
<thead>
<tr>
<th>Outcome</th>
<th>$\delta$=0.02</th>
<th>$\Delta$=0.2</th>
<th>P-value</th>
<th>lower 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movement</td>
<td>$\theta$=0.96</td>
<td>$\Theta$=0.8</td>
<td>0.019</td>
<td>(-0.163, 1)</td>
</tr>
<tr>
<td></td>
<td>$\psi$=0.91</td>
<td>$\Psi$=0.5</td>
<td>0.047</td>
<td>(0.505, $\infty$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Palpation</th>
<th>$\delta$=0.01</th>
<th>$\Delta$=0.2</th>
<th>P-value</th>
<th>lower 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta$=1.02</td>
<td>$\Theta$=0.8</td>
<td>0.009</td>
<td>(-0.136, 1)</td>
</tr>
<tr>
<td></td>
<td>$\psi$=1.03</td>
<td>$\Psi$=0.5</td>
<td>0.020</td>
<td>(0.578, $\infty$)</td>
</tr>
</tbody>
</table>

The lower CIs for $\delta$ and $\psi$ match those given in Tables 3.3 and 3.4, except that the CIs for $\psi$ are slightly larger since they are based on the normal approximation to the log-odds instead of an exact method. The CI for $\theta$ is obtained by Fieller's formula. It cannot be properly obtained from the test statistic in Section 5.3.1 because $\Theta$ appears in the standard error formula. Table 5.11 presents the post-hoc power from each method.

Table 5.11
Univariate and Bivariate Power for Example Data

<table>
<thead>
<tr>
<th>Tested</th>
<th>Univariate Power</th>
<th>Bivariate Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Movement</td>
<td>Palpation</td>
</tr>
<tr>
<td>$\Delta$=0.2</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>$\Theta$=0.8</td>
<td>0.37</td>
<td>0.43</td>
</tr>
<tr>
<td>$\Psi$=0.5</td>
<td>0.51</td>
<td>0.66</td>
</tr>
</tbody>
</table>

For each of these methods, the correlation ($\phi$ coefficient) is 0.62. Fixing the odds ratio at 20 for both outcomes, and fixing the marginal percents yielded the following cell counts on which the bivariate power is estimated:
<table>
<thead>
<tr>
<th>Outcome 1,2</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>31</td>
<td>4</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>Reference</td>
<td>31</td>
<td>4</td>
<td>9</td>
<td>21</td>
</tr>
</tbody>
</table>

Power is largest for $\delta$ and smallest for $\theta$. However, if $\pi_r$ is fixed at the observed level, a $\delta$ of -0.2 corresponds to $\pi_r = 0.415$ for movement and 0.338 for palpation, $\theta = 0.8$ corresponds to $\pi_r = 0.492$ and 0.430, and $\psi = 0.5$ corresponds to $\pi_r = 0.444$ and 0.368. To fairly compare power of the three methods, $\Delta$, $\Theta$, and $\Psi$ must be set by finding power for a fixed $\pi_{r1}$ or $\pi_{r2}$. For example, when $\pi_{r1}$ is set at 0.45, then power appears in Table 5.12. Since $\pi_{r2} \neq \pi_{r1}$, the three methods are actually only compatible for movement. It appears that when the parameters are set to be fairly equivalent, then power is best for ratios, and worse for odds ratios.

To demonstrate the non-reciprocal nature of the test based on $\theta$, power was re-calculated for the situation where the outcome is the proportion of negative responses. All methods described earlier are directly applicable, except the sign of test statistics is reversed. Table 5.13 shows power for the situation compatible to Table 5.12: the probability of a negative response for movement is set to $\pi_{r1} = 0.55$. The tested values for $\Delta$ and $\Psi$ are reciprocals, and power is identical. The power for $\Theta$, however, is much lower, and the tested value is not the reciprocal of the value in Table 5.12.

Table 5.12

Univariate and Bivariate Power of $\pi_{r1} = 0.45$ for Example Data

<table>
<thead>
<tr>
<th>Tested</th>
<th>Univariate Power</th>
<th>Bivariate Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Movement</td>
<td>Palpation</td>
</tr>
<tr>
<td>$\Delta = -0.165$</td>
<td>0.51</td>
<td>0.63</td>
</tr>
<tr>
<td>$\Theta = 0.732$</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>$\Psi = 0.512$</td>
<td>0.49</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Table 5.13  
Power for Testing Proportion of Negative Responses  
When $\pi_{11}=0.55$ for Example Data

<table>
<thead>
<tr>
<th>Tested</th>
<th>Univariate Power</th>
<th>Bivariate Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Movement</td>
<td>Palpation</td>
</tr>
<tr>
<td>$\Delta=0.165$</td>
<td>0.51</td>
<td>0.63</td>
</tr>
<tr>
<td>$\Theta=1.429$</td>
<td>0.39</td>
<td>0.61</td>
</tr>
<tr>
<td>$\Psi=1.952$</td>
<td>0.49</td>
<td>0.64</td>
</tr>
</tbody>
</table>

5.6 Conclusions

Hypothesis tests about the univariate or bivariate equivalence of two categorical responses can be specified in terms of differences of proportions, ratios of proportions, or odds ratios. For each case, in large samples an underlying bivariate normal distribution is applicable, and can be used to calculate bivariate power. When expected counts are small, the test statistics are modified to counteract the anti-conservativeness of the normal approximation. Bivariate power is then estimated by simulation. For small sample sizes, power to detect that test is no worse than reference is reasonable only when test performs better than reference.

Compared to bivariate power for continuous responses described in Section 4.2.7, some similar trends exist, but are less apparent since the variance and correlation change with the response proportions. As noted, two separate univariate tests are conservative because they do not take into account the correlation of the responses.

The correlation in the bivariate distribution of the test statistics for each method are basically the same. When the equivalence criteria are set equal for a
fixed $\pi_1$, and $\pi_{r1} = \pi_{r2}$ and $n_1 = n_r$, the correlations are identical. The correlation is larger when the response proportions are closer to 0.5. The effect of correlation on bivariate power is stronger when the power is not too near one or zero, and the variance is large. For $\delta$ and $\theta$, as response proportions are closer to 0.5 and as sample size decreases, variance increases, and the effect of correlation is more pronounced. For $\psi$, the standard error is relatively larger, and decreases as the response proportions get closer to 0.5. The effect of correlation is stronger for $\psi$, and is seen consistently in Tables 5.8 and 5.9.

Unlike Chapter 4, power increases with correlation even for responses on opposite sides of zero (for $\delta$) or 1.0 (for $\theta$ and $\psi$), because the test is 1-sided. The minimum bivariate power when the correlation is zero is equal to the multiplication of the univariate powers, and the maximum bivariate power when the correlation is 1.0 is at most the minimum univariate power.

When $\theta < 1.0$, the variance for the test statistic $\tilde{d}$ is always smaller than the variance for $\tilde{\delta}$. For each statistic, the variance and power get smaller as the response proportions get closer to 1.0. The variance for $\lambda$, although in different units, is larger; variance and power get larger as the response proportions are closer to 1.0.

Power of the different methods can be compared when the three equivalence criteria are set to compare the same $\pi_1$, as done in the example. Univariate power is largest for $\theta$, and smallest for its conservative complement $\theta^*$. When response proportions are large, the power for $\theta$ is just a little better than $\delta$ (variances are similar since $\Theta$ is near 1.0), and $\psi$ is far behind ($SE(\lambda)$ is large). The power for $\theta^*$ should be just a little less than that for $\delta$. When response proportions are close to 0.5, like in the example, the power for $\theta$ is much larger ($SE(\tilde{d})$ is small due to small $\Theta$), and the power for $\psi$ is only slightly less than that for $\delta$. Power for $\theta^*$ is quite low.
Differences of proportions have been suggested by most of the literature (Dunnett and Gent 77, Rodary et al 89, Blackwelder 82), and produce the most powerful test when response proportions are near one or zero. However, different specifications of \( \Delta \) are required depending on the level of the reference response proportions (i.e., as the response proportions get closer to 0 or 1, a tighter definition of equivalence is desirable). It should be noted that similar problems occur for continuous responses: when variation is small, a tighter equivalence criteria is desirable, and vice versa.

Odds ratios, however, appear to be fairly stable as the reference proportions vary. For instance, if \( \pi_r = 0.9 \) and \( \delta = -0.075 \), then \( \psi = 0.52 \), but when \( \pi_r = 0.7 \) and \( \delta = -0.15 \), \( \psi \) is still 0.52. Equivalence criteria for odds ratios can be applied to many proportions without being modified, making the universal use of one definition of equivalence based on odds ratios an incentive for using this parameter. Power for the test of equivalence based on odds ratios is better when response proportions are near 0.5.

If categorical and continuous responses are combined to examine multivariate equivalence, then it may be advantageous for the test statistics to have the same equivalence criteria. In these cases, ratios of proportions may be most appropriate. The definition \( \theta^* \) should be used, so \( \Theta > 1.0 \). The conservativeness of this method is minimal when response proportions are near one or zero.
Chapter 6
Individual Equivalence from Cross-Over Designs

6.1 Introduction

In Chapter 4, equivalence of continuous responses from cross-over studies is not rejected if a 100\((1-2\alpha)\)% confidence interval for the treatment geometric mean ratio lies within 0.80-1.25. Since this definition is based on the geometric mean ratio, it is referred to as average equivalence. As discussed in Section 3.7, it is also of interest to evaluate individual equivalence, which by definition places acceptable bounds on an individual's ratio of treatment response, instead of the average over all people. Individual equivalence is particularly important for patients switching from the reference to the new treatment, because it assures that most individuals will respond similarly to the two treatments.

Two methods suggested for testing individual equivalence are described in Section 3.7: the FDA 75/75 rule, and the TIER method. The 75/75 rule required 75% of the subjects in a study to have treatment response ratios within 0.75-1.25. The TIER method uses a binomial distribution, and concludes individual equivalence if a 95% CI of the percent of the population whose ratios are within the equivalence criteria is greater than the pre-specified percent. Disadvantages of both methods are discussed in Chapter 3.

This chapter introduces an alternate approach to individual equivalence, based on prediction intervals for log-normal responses from a two-period cross-over design. With no carry-over or period effect, individual equivalence applies if a 100\((1-2\alpha)\)% prediction interval for the treatment ratio of a new subject entering
the study is completely contained within two predetermined limits \( \Theta_1 \) and \( \Theta_2 \). The format of the chapter is similar to Chapter 4: the test criterion and its distribution are presented for the univariate and bivariate case, and power is examined for both large and small samples. In addition, reasonable choices for \( \alpha \), \( \Theta_1 \), and \( \Theta_2 \) are discussed.

### 6.2 Statistical Model and Test Statistic

In chapter 4 (equation 4.4), it was shown for a cross-over design that the bivariate average equivalence test statistics are

\[
\begin{bmatrix}
\frac{d_1}{d_2}
\end{bmatrix} \sim N \left( \begin{bmatrix}
\delta_1 \\
\delta_2
\end{bmatrix}, \frac{1}{2} \left[ \frac{1}{n_A} + \frac{1}{n_B} \right] \times \Sigma_e \right)
\]

where \( d_k \) is one half the average period I - period II difference of log-responses of outcome \( k \) for sequence A minus that for sequence B, and \( \Sigma_e \) is the within-subject covariance matrix:

\[
\Sigma_e = \begin{bmatrix}
\sigma_{e1}^2 & \rho_e \sigma_{e1} \sigma_{e2} \\
\rho_e \sigma_{e1} \sigma_{e2} & \sigma_{e2}^2
\end{bmatrix}
\]

Bivariate average equivalence is supported if 100(1-2\( \alpha \))% confidence intervals for \( \delta_1 \) and \( \delta_2 \) are completely within the equivalence limits \( \Delta_1 \) and \( \Delta_2 \).

To examine individual equivalence, consider a new study subject from the same population as the rest of the participants. If this person were to receive the test and reference treatments in some conceptually simultaneous way such that they are both considered period 1 of two different sequences (in other words assuming that no carry-over or period effect applies to this new patient so it does not matter which drug is received first), then the treatment difference for outcome \( k \) for the new person \( i \) can be defined

\[
d_{ik} = \ln(x)_{ikt} - \ln(x)_{irk}.
\]
Since the average statistic \( d_k \) is based on the presence of two sequences and periods, the formula for \( d_{ik} \) is a simplification of this. The joint distribution for two outcomes is calculated following step one in section 4.2.2 to obtain:

\[
\begin{bmatrix}
  d_{i1} \\
  d_{i2}
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  \gamma_1 + \delta_1 \zeta_h \\
  \gamma_2 + \delta_2 \zeta_h
\end{bmatrix},
2\Sigma_e
\right),
\]  

(6.3)

where \( \gamma_k \) is the period effect for outcome \( k \) and \( \zeta_h \) is a treatment order indicator for sequence \( h \). Since it is assumed that this person has no period effect, then \( \gamma_k = 0 \), and since \( d_{ik} \) is defined as the test-reference difference, then \( \zeta_h = 1 \) and the distribution becomes

\[
\begin{bmatrix}
  d_{i1} \\
  d_{i2}
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  \delta_1 \\
  \delta_2
\end{bmatrix},
2\Sigma_e
\right).
\]  

(6.4)

All subjects are assumed independent, so the difference between the new subject’s statistic and that for the study sample is:

\[
\begin{bmatrix}
  d_{i1} - d_1 \\
  d_{i2} - d_2
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  0 \\
  0
\end{bmatrix},
2 + \frac{1}{n_A} + \frac{1}{n_B}\Sigma_e
\right).
\]  

(6.5)

For each response, a standardized normal variable is obtained by dividing by the standard error:

\[
z = \frac{d_{ik} - d_k}{\text{SE}(d_{ik} - d_k)}, \quad \text{where} \quad \text{SE}(d_{ik} - d_k) = \sigma_{yk}\sqrt{2 + \frac{1}{n_A} + \frac{1}{n_B}}.
\]  

(6.6)

This statistic can be used to test hypotheses about the value of \( d_{ik} \) for a new participant. A 100(1-2\( \alpha \))% prediction interval for the new subject’s treatment log-difference for outcome \( k \) is obtained by:

\[
d_k \pm z_{1-\alpha}\text{SE}(d_{ik} - d_k).
\]  

(6.7)

For an 80% prediction interval, \( z_{1-\alpha} = 1.282 \). The prediction interval for an individual’s treatment response ratio is obtained by exponentiating the limits in
equation (6.7). In small samples, the standardized normal $z$ can be replaced by a $t$ with $n_A+n_B-2$ degrees of freedom. Based on this interval, one can argue that among all new subjects from the same population (with no carry-over or period effect), $100(1-2\alpha)\%$ will have a treatment ratio within the limits of the prediction interval. If $\alpha=0.10$, we are 80% certain that a new individual’s treatment ratio will lie within the interval limits.

A proposed definition of individual equivalence concludes equivalence if the entire exponentiated $100(1-2\alpha)\%$ prediction interval lies completely within prespecified limits $\Theta_1$ and $\Theta_2$. Equivalence limits could be set at 80% within 0.80 and 1.25, or they may be more lenient. As explained in Section 6.3, the individual equivalence limits of $\Theta_1$, $\Theta_2$, and $\alpha$ may vary depending on the value of $\sigma_e$. As in previous chapters, bivariate individual equivalence is accepted only if equivalence is supported for both measures univariately.

A $p$-value for testing whether a new individual’s treatment ratio will lie within $\Theta_1-\Theta_2$ is obtained by a modification of the two 1-sided tests procedure. The hypotheses are stated as:

\[ H_{01}: d_{ik} \leq \Delta_1 \quad H_{A1}: d_{ik} > \Delta_1 \]
\[ H_{02}: d_{ik} \geq \Delta_2 \quad H_{A2}: d_{ik} < \Delta_2 \]

where $\Delta=\ln(\Theta)$ and $\Delta_1<0<\Delta_2$. Individual Equivalence for outcome $k$ is concluded if both $H_{01}$ and $H_{02}$ are rejected at significance level $\alpha$. Each test is conducted as a one-sided $t$-test with degrees of freedom $\nu=n_A+n_B-2$:

\[ t_1 = \frac{d_{ik} - \Delta_1}{\text{SE}(d_{ik} - \Delta_1)} \geq t(\nu) \]
\[ t_2 = \frac{\Delta_2 - d_{ik}}{\text{SE}(d_{ik} - \Delta_2)} \geq t(\nu) \]

(6.8)

If individual equivalence must be assured for 80% of the population, then the significance level is set at $\alpha=0.10$. 

6.3 Choosing Individual Equivalence Limits

Since the standard error for the individual equivalence test is much larger than that for average equivalence, power to show individual equivalence is smaller for the same sample size. This is partially counter-acted by allowing the interval to be less than the standard 90% CI for average equivalence, and by relaxing the equivalence limits.

There is no single set of preferred parameters \((\alpha, \Theta_1, \Theta_2)\) to define individual equivalence, because as within-subject variability increases, individual equivalence becomes harder to prove, and a more lenient definition of individual equivalence may be needed so that the definition is at least met in the population. We have no chance of concluding individual equivalence from a trial if the equivalence criterion is not met by the population. For instance, with a reasonable sample size, we can never show that a 100(1-2\(\alpha\))% prediction interval is within \(\Theta_1 - \Theta_2\) if at least 100(1-2\(\alpha\))% of the population does not lie within these limits.

Assuming a log-normal distribution, no period effect, and infinite sample size, the population test-reference log-difference of outcome \(k\) for either sequence is \(\sim N [\delta_k, 2\sigma_{\epsilon k}^2]\). We can therefore say that 100(1-2\(\alpha\))% of the population's individual test-reference log-differences lie within:

\[
\delta_k \pm z_{1-\alpha} \sqrt{2\sigma_{\epsilon k}^2}.
\] (6.9)

In order for a study to conclude individual equivalence for a particular criterion, the population probability interval for the same \(\alpha\) must lie within the equivalence limits \(\Delta_1 - \Delta_2\). Otherwise, a study will not be able to demonstrate individual equivalence even if the sample size is infinite. Note that the sample prediction interval in equation (6.7) tends to the population interval in equation (6.9) as \(n_A\) and \(n_B \to \infty\).

The parameters of a 'provable' individual equivalence definition (1-2\(\alpha\), \(\Delta_1\), and \(\Delta_2\)) are thus limited by the population values \(\delta_k\) and \(\sigma_{\epsilon k}^2\). We must have:
\[ \Delta_1 < \delta_k - z_{1-\alpha}\sqrt{2\sigma_{ek}^2}, \text{ and } \delta_k + z_{1-\alpha}\sqrt{2\sigma_{ek}^2} < \Delta_2. \]  

(6.10)

For a stringent definition of individual equivalence to apply, \( \delta_k \) must be near zero and \( \sigma_{ek}^2 \) must be small. When \( \delta_k \) is farther away from zero, then \( \sigma_{ek}^2 \) must be smaller in order to accept the same definition of individual equivalence.

One way of refining the choice of \( \Delta_1 \) and \( \Delta_2 \) is to consider limits that support equivalence between reference and itself in the population. It does not make sense to attempt to show that a 100(1-2\( \alpha \))% prediction interval for the ratio of treatment responses is within \( \Theta_1-\Theta_2 \) if at least 100(1-2\( \alpha \))% of the population's individual ratios between two repeats of the reference treatment does not also lie within these limits.

Up to this point, it has been assumed that test and reference have equal variance, but they may have different variances \( \sigma_{ekr}^2 \) and \( \sigma_{ekt}^2 \). In a population with responses following a log-normal distribution and having no carry-over or period effect, the difference of the log-response between two separate measurements of the reference for both outcomes is

\[
\begin{bmatrix}
    d_1 \\
    d_2
\end{bmatrix} \sim N \left( \begin{bmatrix}
    0 \\
    0
\end{bmatrix}, 2\Sigma_{er} \right). 
\]

(6.11)

where

\[
\Sigma_{er} = \begin{bmatrix}
    \sigma_{e1r}^2 & \rho_{er}\sigma_{e1r}\sigma_{e2r} \\
    \rho_{er}\sigma_{e1r}\sigma_{e2r} & \sigma_{e2r}^2
\end{bmatrix}.
\]

If each person in the population repeated the reference treatment twice, then 100(1-2\( \alpha \))% of the population's individual reference-reference differences of logs for response \( k \) lie within

\[ 0 \pm z_{1-\alpha}\sqrt{2\sigma_{ekr}^2}. \]  

(6.12)

When sample size is infinite, the test of individual equivalence should conclude that reference is equivalent to itself. This is guaranteed for a specified \( \alpha \) only if \( \Delta_1 \) and \( \Delta_2 \) are set to \( \pm z_{1-\alpha}\sqrt{2\sigma_{ekr}^2} \) as a minimum.

In keeping with definitions of average equivalence, it is also desirable to
conclude individual equivalence in the population if the geometric mean treatment ratio is within average equivalence limits, i.e., $0.80 \leq \theta_k \leq 1.25$. However, individual equivalence should not be accepted in a population if $\theta_k > 1.25$ or $\theta_k < 0.8$. Therefore, to assure that at most $100(1-2\alpha)\%$ of the population individual treatment differences of logs are within reasonable limits, the maximum limits should be $\Delta_1 = \ln(0.80) - z_{1-\alpha}\sqrt{2}\sigma_{ek}$ and $\Delta_2 = \ln(1.25) + z_{1-\alpha}\sqrt{2}\sigma_{ek}$. This formula assumes test and reference have the same variance. In the more general case, $2\sigma_{ek}^2$ is replaced by $(\sigma_{ekr}^2 + \sigma_{ekt}^2)$ or $2\sigma_{ekr}^2$.

Combining the maximum requirements for $\Delta_1$ and $\Delta_2$ based on $\delta_k$ and the minimum requirements based on $\text{SE}(d_k)$ for a difference of reference and itself, then a reasonable strategy is to define:

$$\Delta_1 = \ln(0.80) - z_{1-\alpha}\sqrt{2}\sigma_{ekr}$$

$$\Delta_2 = \ln(1.25) + z_{1-\alpha}\sqrt{2}\sigma_{ekr}$$

(6.13)

This criterion may be quite stringent if the test treatment is much more variable than the reference. If one is willing to allow acceptance of individual equivalence if $\sigma_{ekt}/\sigma_{erk} \leq \epsilon$, where $\epsilon > 1.0$, perhaps 1.25, then $\sqrt{2}\sigma_{erk}$ could be replaced by $\sqrt{1+\epsilon^2}\sigma_{erk}$ in equation (6.13).

Two methods have been presented for choosing $\Delta_1$ and $\Delta_2$: the formula presented in equation (6.13), and the bounds imposed by the population parameters as described in equations (6.9) and (6.10). When test and reference have different variances, $\Sigma_{et}$ is defined similarly to $\Sigma_{err}$, and equation (6.9) is generalized such that $100(1-2\alpha)\%$ of the population test-reference differences of logs lie within

$$\delta_k \pm z_{1-\alpha}\sqrt{\sigma_{ekr}^2 + \sigma_{ekt}^2}.$$  

(6.14)

If values of $\Delta_1$ and $\Delta_2$ determined by equation (6.13) lead to limits that are not within the population interval for the same $\alpha$, then equivalence by that definition can not be proven, and either the study should not be conducted or more lenient limits should be devised.
The equivalence limits obtained from equation (6.13) impose reasonable properties. For large $|\delta_k|$, the variance must be small to accept equivalence, and for large variance, $\delta_k$ must be closer to 0 for acceptance. One potential disadvantage is that if the test treatment is less variable than the reference, individual equivalence could be accepted by a study even if $\theta_k > 1.25$ or $< 0.80$. Both methods of obtaining equivalence limits require estimates of $\sigma_{ekr}$, so this strategy works best when there are reliable estimates of these values from previous studies.

Some reasonable individual equivalence definitions that hold in the population based on equation (6.10) are displayed in Table 6.1 for several choices of $\alpha$, $\theta$, and $\sigma_e$, assuming test and reference have equal variance. When $\theta$ and $\sigma_e$ are as specified, then at least $100(1-2\alpha)\%$ of the population individual treatment ratios lie within $\Theta_1$ and $\Theta_2$. The study of a response with these parameters which applies the corresponding equivalence criterion will have power $> 0$ for accepting individual equivalence. However, for each $\theta$, these are the tightest limits that could lead to acceptance of equivalence.

In Table 6.1, the variance for test and reference are assumed equal. Since this is the case, the column for $\theta = 1.25$ yields $\Theta_1$ and $\Theta_2$ that would be derived from equation (6.13). Therefore, the equivalence limits in this column are the preferable limits for the corresponding $\sigma_e$ regardless of the value of $\theta$.

The limits presented in Table 6.1 were rounded up to the nearest standard fraction, so that equivalence limits are chosen from the set of 0.8-1.25, 0.75-1.33, 0.7-1.43, 0.67-1.5, 0.6-1.67, 0.5-2.0, 0.4-2.5, or 0.33-3.0. This rounding introduces extra leniency in the limits, so for an actual study it may be preferable to base limits on the actual calculated values resulting from equation (6.13). Choices of $100(1-2\alpha)$ tested for consideration were 90%, 85%, 80%, 75%, 70%, 67%, 60%, or 50%. Alpha was chosen so that rounding of intervals was minimal.
Table 6.1
Percent of Population Individual Treatment Ratios Within Specified Intervals
For $\theta=1.0$ and 1.25 and Equal Test and Reference Variance

<table>
<thead>
<tr>
<th>$\sigma_e$</th>
<th>$\theta=1.0$</th>
<th>$\theta=1.25$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent Interval</td>
<td>Percent Interval</td>
</tr>
<tr>
<td>0.15</td>
<td>80% (0.75, 1.33)</td>
<td>80% (0.6, 1.67)</td>
</tr>
<tr>
<td></td>
<td>70% (0.8, 1.25)</td>
<td>60% (0.67, 1.5)</td>
</tr>
<tr>
<td>0.20</td>
<td>80% (0.7, 1.43)</td>
<td>90% (0.5, 2.0)</td>
</tr>
<tr>
<td></td>
<td>67% (0.75, 1.33)</td>
<td>67% (0.6, 1.67)</td>
</tr>
<tr>
<td>0.25</td>
<td>85% (0.6, 1.67)</td>
<td>80% (0.5, 2.0)</td>
</tr>
<tr>
<td></td>
<td>67% (0.7, 1.43)</td>
<td>50% (0.6, 1.67)</td>
</tr>
<tr>
<td>0.30</td>
<td>75% (0.6, 1.67)</td>
<td>85% (0.4, 2.5)</td>
</tr>
<tr>
<td></td>
<td>50% (0.75, 1.33)</td>
<td>70% (0.5, 2.0)</td>
</tr>
<tr>
<td>0.40</td>
<td>75% (0.5, 2.0)</td>
<td>75% (0.4, 2.5)</td>
</tr>
<tr>
<td></td>
<td>50% (0.67, 1.5)</td>
<td>60% (0.5, 2.0)</td>
</tr>
</tbody>
</table>

When $\theta \neq 1.0$, the actual distribution of treatment log-differences is not centered around zero, although the equivalence limits are centered ($\Theta_1$ and $\Theta_2$ are reciprocals). Using the probability interval defined in equation (6.9) yields conservative estimates. For instance, based on (6.9), when $\theta=1.25$ and $\sigma_e=0.15$, 80% of the population's individual ratios lie within 0.95-1.64. We are thus assured that at least 80% lie within the centered limits 0.6-1.67, but the actual percent could be quite a bit larger, to account for the interval between 0.6 and 0.95.

The exact percent of population individual log-differences between $\Delta_1$-$\Delta_2$ is
found in a manner similar to a two one-sided tests procedure. The probability of the following z scores is calculated and added to form a value \( \alpha' = (\alpha'_1 + \alpha'_2) \). The actual percent within \( \Delta_1 \) and \( \Delta_2 \) is 100(1-\( \alpha' \)):

\[
\begin{align*}
  z_1 &= \frac{\delta - \Delta_1}{\sqrt{2\sigma_e^2}} \geq z_{1 - \alpha'_1} \\
  z_2 &= \frac{\Delta_2 - \delta}{\sqrt{2\sigma_e^2}} \geq z_{1 - \alpha'_2}.
\end{align*}
\] (6.15)

For the example, 91% of individual ratios are within 0.6-1.67. The actual limits for which 80% of the population ratios lie are therefore narrower than 0.6-1.67.

The test is therefore conservative for \( \delta \neq 0 \), and becomes increasingly more conservative as \( |\delta| \) increases. This conservativeness, although perhaps undesirable, also exists for the test of average equivalence, and is what Westlake was attempting to correct with his symmetric confidence intervals described in Section 3.2.2. This effect may not be completely undesirable, since it should become harder to show equivalence as \( |\delta| \) increases, and this may somewhat counteract the anti-conservativeness of the test at the point where the population interval is equal to the equivalence limits, as described in upcoming Sections 6.4 and 6.6.

The use of equation (6.13) makes a decision rule that is equally fair for every \( \sigma_e \) for a fixed \( \alpha \). While equally stringent for all treatments, this leads to a different equivalence criterion for each value of \( \sigma_e \). This could sometimes lead to confusion about results or controversy over the appropriate estimate of \( \sigma_e \) for each treatment and response. Such obstacles appear to be inevitable when individual equivalence is evaluated.

The remainder of this chapter is devoted to evaluating the magnitude of univariate and bivariate power for the prediction interval method of evaluating individual equivalence. For simplicity, it is assumed that test and reference have equal variance \( \sigma_{ek}^2 \) for response \( k \).
6.4 Power Assuming Known Variance

In this section it is assumed the sample size is large enough so that estimates of the within-subject variance for response \( k \) are assumed to equal \( \sigma^2_{ek} \). If the individual equivalence definition for a given \( \delta_k \) and \( \sigma_{ek} \) is such that the inequality of equation (6.10) holds, then power will be greater than zero for any sample size.

The univariate power of the test statistic, \( d_k \), is defined as the probability of rejecting the null hypothesis \( H_0 \). For individual equivalence, this corresponds to \( H_0: \Delta_1 < d_{ik} < \Delta_2 \) vs. \( H_A: \Delta_1 < d_{ik} < \Delta_2 \). The probability that the null hypothesis is rejected is:

\[
\text{Power} = \Pr(c_L < d_k < c_U). \quad (6.16)
\]

The cutpoints \( c_L \) and \( c_U \) are obtained exactly as they were for average equivalence in Section 4.2.3, but are described here in terms of prediction intervals rather than a two 1-sided tests procedure. The null hypothesis is rejected if the prediction interval for the pre-specified \( \alpha \) is within the equivalence limits. Applying equation (6.7), we obtain:

\[
\text{Power} = \Pr[\Delta_1 < d_k - z_{1-\alpha} \times \text{SE}(d_{ik} - d_k) \text{ and } d_k + z_{1-\alpha} \times \text{SE}(d_{ik} - d_k) < \Delta_2]
\]

\[
= \Pr[\Delta_1 + z_{1-\alpha} \times \text{SE}(d_{ik} - d_k) < d_k < \Delta_2 - z_{1-\alpha} \times \text{SE}(d_{ik} - d_k)]
\]

\[
= \Pr(c_L < d_k < c_U). \quad (6.17)
\]

This is the same result obtained in Section 4.2.3 for average equivalence, except that the standard error is for \( (d_{ik} - d_k) \) instead of \( d_k \).

To obtain a standardized variable, subtracting \( E(d_k) = \delta_k \) and dividing by \( \text{SE}(d_k) = \sigma_{ek} \sqrt{\frac{1}{2} \mu_A + \frac{1}{2} \mu_B} \) for each component of equation (6.17) leads to:

\[
\Pr\left[ \frac{\Delta_1 - \delta_k + Z_{1-\alpha} \text{SE}(d_{ik} - d_k)}{\text{SE}(d_k)} < Z < \frac{\Delta_2 - \delta_k - Z_{1-\alpha} \text{SE}(d_{ik} - d_k)}{\text{SE}(d_k)} \right] \quad (6.18)
\]

If the lower and upper limits are referred to as \( L \) and \( U \), and if \( \delta_k \) and \( \sigma^2_{ek} \) are assumed known, then the power is calculated by numerical integration of the standard normal distribution from \( L \) to \( U \).
Equation (6.18) displays reasonable properties. As $N \to \infty$, and $\Delta_1 < \delta_k - z_1 - \alpha \sqrt{2\sigma^2_{ek}} < \delta_k + z_1 - \alpha \sqrt{2\sigma^2_{ek}} < \Delta_2$, then power $\to 1$. Power $\to 1$ fastest when $\delta_k = 0$ and $\sigma^2_{ek}$ is small. If the population interval is not within $\Delta_1 - \Delta_2$, then power $\to 0$ as $N \to \infty$. If the population interval limits are exactly $\Delta_1 - \Delta_2$, then as $N \to \infty$, power $\to 0$.

This formula displays slightly different properties than definitions of power in previous chapters. When the test is right on the border of acceptance or rejection and sample sizes are infinite, the power is zero instead of the usual $\alpha$. This is desirable: in very large samples, there is less less than a 100$\alpha$% chance of rejecting the null hypothesis when it should not be rejected.

However, when sample sizes are not infinite and the population limits for the chosen $\alpha$ are equal to $\Delta_1$ and $\Delta_2$, then the numerators in equation (6.18) are not equal to zero, and power $\to 0$. Even for moderate to large sample sizes, power can be large when the population limits are at or slightly wider than $\Delta_1 - \Delta_2$. For example, when $\sigma_{ek} = 0.20$ and $\theta_k = 1.164$, individual equivalence based on 80% of treatment ratios lying within 0.6-1.67 is just barely missed in the population. The power to accept individual equivalence, however, is over 0.40 for sample sizes from $N = 40$ to $N = 2000$. The magnitude of this type 1 error varies by sample size and depends on the values of the equivalence limits and the population parameters. As sample sizes get very small, the power curve (obtained by simulation methods described in the next section) flattens, and power tends to be small in the region where population limits are close to the equivalence criteria.

One solution to this problem is to make the individual equivalence criterion more stringent. For instance, perhaps instead of basing $\Delta_1$ and $\Delta_2$ on $\ln(0.80)$ and $\ln(1.25)$ in equation (6.13), a more conservative strategy would be to use tighter limits such as $\ln(0.87)$ and $\ln(1.15)$, so that type 1 error for $\theta_k = 0.80$ or 1.25 is properly controlled. As mentioned previously, the conservative nature of
the test when $|\delta_k| > 0$ may also compensate for large type I error.

To obtain bivariate power of two responses, the correlation between variables within a person, $\rho_e$, must be known. By standardizing the bivariate distribution in equation (6.1) and applying the integration limits as previously described, the bivariate power is found by numerical integration from $L_1$ to $U_1$ and $L_2$ to $U_2$ of a standardized bivariate normal distribution with correlation $\rho_e$, where $L_k$ and $U_k$ are defined as in equation (6.18).

6.5 Power Assuming Unknown Variance

Since individual equivalence criteria are stringent, it is unlikely that study sample sizes will be less than $N=30$ generally considered large enough for the variance to be assumed known and the normal distribution to apply.

However, if sample size is small, then the univariate test statistics (equation 6.8) follow a $t$ distribution with $n_A + n_B - 2$ degrees of freedom, and power can not be evaluated with a normal distribution. A feasible approach to assessing univariate and bivariate power in small sample settings is through simulation. Steps for simulation are exactly the same as those described in Section 4.2.5, except that the estimated standard error is for $SE(d_{ik} - d_k)$, not $SE(d_k)$, and the test statistics evaluated are the modified two one-sided tests in equation (6.8). It is also possible to follow similar modifications of Section 4.2.4 to simulate bivariate power with known variance if numerical integration software is not available.

6.6 Application and Results

Univariate and bivariate power for individual equivalence is provided for a variety of situations in Tables 4.2-4.5, for a balanced design with 20 and 30 subjects per sequence group. Equivalence limits are roughly based on equation
(6.13); however, so that power can be evaluated as the variance changes, the same limits are evaluated for a range of variances. Power is examined for criteria of 80% and 70% of individual treatment ratios within either 0.6-1.67 or 0.5-2.0, and for a variety of within-subject standard deviations of \( \ln(x) \) \((\sigma_{e1} \text{ and } \sigma_{e2})\), and treatment geometric mean ratios \((\theta_k = e^{\delta_k})\). The within-subject correlation of \( \ln(x_1) \) and \( \ln(x_2) \), \( \rho_e \) is fixed at 0.85, 0.75, 0.50, and 0.25. The tables presented are based on simulation assuming a known variance with 10,000 sample estimates (repeats) for each value.

Table 6.2

<table>
<thead>
<tr>
<th>Power for Testing Individual Equivalence of a Log-Normal Response</th>
<th>From a Two Period Cross-Over ANOVA Prediction Interval</th>
<th>N per Sequence Group= 20, Obtained via Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Equivalence Criteria</td>
<td>( \theta \rightarrow \sigma_e )</td>
<td>1.0</td>
</tr>
<tr>
<td>80% 0.6-1.67</td>
<td>.15</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.00</td>
</tr>
<tr>
<td>70% 0.6-1.67</td>
<td>.20</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.67</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>.00</td>
</tr>
<tr>
<td>80% 0.5-2.0</td>
<td>.20</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>.00</td>
</tr>
<tr>
<td>70% 0.5-2.0</td>
<td>.20</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>.73</td>
</tr>
</tbody>
</table>

\[ \sigma_e = SD(\ln(x)) \]
Table 6.3
Power for Testing Individual Equivalence of a Log-Normal Response
From a Two Period Cross-Over ANOVA Prediction Interval
N per Sequence Group= 30, Obtained via Simulation

<table>
<thead>
<tr>
<th>Individual Equivalence Criteria</th>
<th>$\theta$ $\rightarrow$ $\sigma_e$</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.2</th>
<th>1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% 0.6-1.67</td>
<td>.15</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.98</td>
<td>.69</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>1.0</td>
<td>.91</td>
<td>.86</td>
<td>.16</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.76</td>
<td>.18</td>
<td>.13</td>
<td>.003</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>70% 0.6-1.67</td>
<td>.20</td>
<td>1.0</td>
<td>.99</td>
<td>.99</td>
<td>.81</td>
<td>.41</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.99</td>
<td>.84</td>
<td>.79</td>
<td>.18</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.78</td>
<td>.30</td>
<td>.25</td>
<td>.02</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>80% 0.5-2.0</td>
<td>.20</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>1.0</td>
<td>.99</td>
<td>.99</td>
<td>.88</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.99</td>
<td>.82</td>
<td>.77</td>
<td>.25</td>
<td>.08</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>70% 0.5-2.0</td>
<td>.20</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>1.0</td>
<td>.99</td>
<td>.99</td>
<td>.89</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td>.40</td>
<td>.84</td>
<td>.53</td>
<td>.48</td>
<td>.13</td>
<td>.05</td>
</tr>
</tbody>
</table>

$\sigma_e = SD(ln(x))$
Table 6.4
Power for Testing Individual Equivalence of Two Bivariate Log-Normal Responses
From a Two Period Cross-Over ANOVA Prediction Interval
N per Sequence Group = 20, Obtained via Simulation

<table>
<thead>
<tr>
<th>$\theta_1$, $\theta_2$</th>
<th>$\sigma_{e1}$</th>
<th>$\sigma_{e2}$</th>
<th>$\rho_e$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.2</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Equivalence Criteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20 0.20 .85</td>
<td>.99</td>
<td>.86</td>
<td>.81</td>
<td>.67</td>
<td>.20</td>
<td>.14</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.86</td>
<td>.80</td>
<td>.67</td>
<td>.20</td>
<td>.12</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.86</td>
<td>.78</td>
<td>.67</td>
<td>.20</td>
<td>.08</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.86</td>
<td>.76</td>
<td>.68</td>
<td>.20</td>
<td>.06</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20 0.25 .85</td>
<td>.65</td>
<td>.22</td>
<td>.22</td>
<td>.07</td>
<td>.01</td>
<td>.01</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.22</td>
<td>.22</td>
<td>.09</td>
<td>.01</td>
<td>.01</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.22</td>
<td>.21</td>
<td>.13</td>
<td>.01</td>
<td>.006</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.22</td>
<td>.20</td>
<td>.15</td>
<td>.01</td>
<td>.004</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20 0.30 .85</td>
<td>.68</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.04</td>
<td>.04</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.04</td>
<td>.04</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.04</td>
<td>.04</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.04</td>
<td>.04</td>
<td>.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25 0.25 .85</td>
<td>.98</td>
<td>.78</td>
<td>.72</td>
<td>.52</td>
<td>.22</td>
<td>.16</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.78</td>
<td>.70</td>
<td>.52</td>
<td>.22</td>
<td>.14</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.78</td>
<td>.67</td>
<td>.54</td>
<td>.22</td>
<td>.10</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.78</td>
<td>.64</td>
<td>.56</td>
<td>.22</td>
<td>.07</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25 0.30 .85</td>
<td>.67</td>
<td>.32</td>
<td>.32</td>
<td>.11</td>
<td>.04</td>
<td>.04</td>
<td>.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.32</td>
<td>.32</td>
<td>.13</td>
<td>.04</td>
<td>.04</td>
<td>.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.32</td>
<td>.30</td>
<td>.17</td>
<td>.04</td>
<td>.03</td>
<td>.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.32</td>
<td>.28</td>
<td>.21</td>
<td>.04</td>
<td>.02</td>
<td>.009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30 0.30 .85</td>
<td>.95</td>
<td>.74</td>
<td>.69</td>
<td>.47</td>
<td>.26</td>
<td>.20</td>
<td>.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.74</td>
<td>.67</td>
<td>.48</td>
<td>.26</td>
<td>.18</td>
<td>.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.74</td>
<td>.64</td>
<td>.50</td>
<td>.27</td>
<td>.14</td>
<td>.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.74</td>
<td>.60</td>
<td>.52</td>
<td>.27</td>
<td>.11</td>
<td>.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30 0.40 .85</td>
<td>.73</td>
<td>.51</td>
<td>.51</td>
<td>.50</td>
<td>.18</td>
<td>.18</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75</td>
<td>.51</td>
<td>.51</td>
<td>.50</td>
<td>.18</td>
<td>.18</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.51</td>
<td>.51</td>
<td>.50</td>
<td>.17</td>
<td>.17</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.51</td>
<td>.50</td>
<td>.50</td>
<td>.17</td>
<td>.16</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\sigma_{ek} = \text{SD}(\ln(x_k)) \quad \rho_e = \text{corr}(\ln(x_{11}), \ln(x_{12}), \ln(x_{11}) - \ln(x_{12}))$
Table 6.5
Power for Testing Individual Equivalence of Two Bivariate Log-Normal Responses
From a Two Period Cross-Over ANOVA Prediction Interval
N per Sequence Group = 30, Obtained via Simulation

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\sigma_{e1}$</th>
<th>$\sigma_{e2}$</th>
<th>$\rho_{e}$</th>
<th>1.0</th>
<th>1.0</th>
<th>1.1</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalence Criteria</td>
<td>80% of treatment ratios within 0.6-1.667</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>.85</td>
<td></td>
<td></td>
<td>.99</td>
<td>.91</td>
<td>.88</td>
<td>.78</td>
<td>.16</td>
<td>.10</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.99</td>
<td>.91</td>
<td>.87</td>
<td>.78</td>
<td>.16</td>
<td>.09</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.99</td>
<td>.91</td>
<td>.85</td>
<td>.78</td>
<td>.15</td>
<td>.06</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.99</td>
<td>.91</td>
<td>.84</td>
<td>.78</td>
<td>.15</td>
<td>.04</td>
<td>.015</td>
</tr>
<tr>
<td>0.20</td>
<td>0.25</td>
<td>.85</td>
<td></td>
<td></td>
<td>.77</td>
<td>.18</td>
<td>.18</td>
<td>.08</td>
<td>.003</td>
<td>.003</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.77</td>
<td>.18</td>
<td>.18</td>
<td>.09</td>
<td>.003</td>
<td>.003</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.77</td>
<td>.18</td>
<td>.18</td>
<td>.12</td>
<td>.003</td>
<td>.002</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.77</td>
<td>.18</td>
<td>.18</td>
<td>.14</td>
<td>.002</td>
<td>.001</td>
<td>.00</td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>.85</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.30</td>
<td>.30</td>
<td>.02</td>
<td>.02</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.30</td>
<td>.30</td>
<td>.02</td>
<td>.02</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.30</td>
<td>.30</td>
<td>.02</td>
<td>.02</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.30</td>
<td>.30</td>
<td>.02</td>
<td>.02</td>
<td>.002</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>.85</td>
<td></td>
<td></td>
<td>.99</td>
<td>.84</td>
<td>.79</td>
<td>.63</td>
<td>.18</td>
<td>.13</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.99</td>
<td>.84</td>
<td>.78</td>
<td>.63</td>
<td>.18</td>
<td>.11</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.99</td>
<td>.84</td>
<td>.75</td>
<td>.63</td>
<td>.18</td>
<td>.08</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.99</td>
<td>.84</td>
<td>.73</td>
<td>.63</td>
<td>.18</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>0.25</td>
<td>0.30</td>
<td>.85</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.30</td>
<td>.13</td>
<td>.02</td>
<td>.02</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.30</td>
<td>.15</td>
<td>.02</td>
<td>.02</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.29</td>
<td>.18</td>
<td>.02</td>
<td>.01</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.78</td>
<td>.30</td>
<td>.28</td>
<td>.21</td>
<td>.02</td>
<td>.01</td>
<td>.002</td>
</tr>
<tr>
<td>0.30</td>
<td>0.30</td>
<td>.85</td>
<td></td>
<td></td>
<td>.98</td>
<td>.81</td>
<td>.76</td>
<td>.58</td>
<td>.24</td>
<td>.18</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.98</td>
<td>.81</td>
<td>.74</td>
<td>.58</td>
<td>.24</td>
<td>.16</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.98</td>
<td>.81</td>
<td>.71</td>
<td>.60</td>
<td>.24</td>
<td>.12</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.98</td>
<td>.81</td>
<td>.70</td>
<td>.61</td>
<td>.24</td>
<td>.09</td>
<td>.07</td>
</tr>
<tr>
<td>80% of treatment ratios within 0.5-2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.40</td>
<td>.85</td>
<td></td>
<td></td>
<td>.84</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.13</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.75</td>
<td></td>
<td></td>
<td>.84</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.13</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td>.84</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.13</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.25</td>
<td></td>
<td></td>
<td>.84</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.13</td>
<td>.13</td>
<td>.05</td>
</tr>
</tbody>
</table>

$\sigma_{ek} = \text{SD(ln}(x_k))$  $\rho_{e} = \text{corr(ln}(x_{I1}) - \text{ln}(x_{I1}), \text{ln}(x_{I2}) - \text{ln}(x_{I2}))$
An approximation to the univariate power can be obtained by a modification of the cross-over average equivalence power formula by Hauschke et al (92), described in equation (3.28). For balanced designs, when \( \Delta_1 \) and \( \Delta_2 \) are symmetric the formula is:

\[
t_{1-\beta} = \frac{\Delta_2 |\delta_k| + t_{\alpha, N-2} \sigma_{ek} \sqrt{2/N}}{\sigma_{ek} \sqrt{2/N}}
\]  

and the approximate power \( = \text{pr}[t_{N-2} \leq t_{1-\beta}] \), where \( t_{1-\beta} \) is the value of a \( t \) random variable with \( N-2 \) degrees of freedom. If \( \delta_k = 0 \), then the formula is

\[
t_{1-\beta/2} = \frac{\Delta_2 + t_{\alpha, N-2} \sigma_{ek} \sqrt{2/N}}{\sigma_{ek} \sqrt{2/N}}
\]  

and the power \( = 1 - 2(1 - \text{pr}[t_{N-2} \leq t_{1-\beta/2}]) \) if \( t_{1-\beta/2} \geq 0 \), and is set to zero otherwise. Equations 6.19 and 6.20 do not easily convert to a formula for sample size. The approximate formula yields reasonable estimates for power in Table 6.2 and 6.3, but often produces estimates a few percent less than those obtained by simulation (The worst approximation encountered is for \( \sigma_e = 0.25 \), \( \delta = 0 \): 0.58 vs 0.65 from simulation). If the power from simulation is zero and \( \delta_k \neq 0 \), power estimates from (6.19) are slightly larger than zero. When \( N > 30 \), then the \( t \) random variable in (6.19) and (6.20) can be replaced by a standard normal.

The individual equivalence criteria used for Tables 6.2-6.5 do not hold in the population for all given values of \( \theta \) and \( \sigma_e \). For instance, in Tables 6.2 and 6.3, 80% of the population treatment ratios are not within 0.6-1.67 when \( \sigma_e = 0.20 \) and \( \delta \geq 1.2 \), when \( \sigma_e = 0.25 \) and \( \delta \geq 1.1 \) or for any value of \( \delta \) when \( \sigma_e = 0.30 \). Except where the power is so low that it is zero, the power estimates for these situations are larger for Table 6.2, where \( N = 40 \) than for Table 6.3 where \( N = 60 \) (i.e., when \( \sigma_e = 0.25 \) and \( \delta = 1.1 \), the power for \( N = 40 \) is 0.22, and the power for \( N = 60 \) is 0.18).

This effect of decreasing power with increasing sample size is due to the type-
1 error anomaly described previously, and is visible in both univariate and bivariate settings. This test has a unique version of type-1 error which depends on sample size: there is some power for finding equivalence for criterion at or more stringent than that met by the population, even when sample size is large. False power is highest when the population limits correspond to the equivalence limits.

By examining Tables 6.4 and 6.5, several familiar patterns for power of bivariate equivalence are apparent. Power increases with correlation when treatment geometric mean ratios for the two responses are on the same side of 1.0, and decreases with increased correlation for responses with opposite treatment effects. The effect of correlation is greater when geometric mean ratios are farther from one so that univariate powers are each moderate to low. Graphical representation of bivariate power follows a similar shape as described in Section 4.2.7.

It should be noted that individual equivalence is a stringent criterion, and power for a particular definition appears to be more determined by \( \sigma_{ek} \) than the sample size. For a fixed \( \delta_k \) and sample size, the window between a power of 1 and zero for changing \( \sigma_{ek} \) is small. Marked differences in power are seen between different definitions of individual equivalence, but for a given \( \sigma_{ek} \) and \( \delta_k \), power changes with sample size only moderately.

If within-subject variances for two responses are very dissimilar, power of an equivalence criterion for the less variable response may be 1.0, while power for the other may be very low. Bivariate power is then equal to the small univariate power, and correlation between the responses has no effect on the bivariate power. The examples provided in Tables 6.4 and 6.5 were picked carefully so that whenever possible univariate powers are not too high or low so the effect of correlation on bivariate power is fairly apparent.
6.7 Example

As an example, consider the bioequivalence study of test and reference formulations of a psychotropic drug described in Section 1.4.1, and evaluated for power of average bioavailability in Section 4.2.8. Area under the curve (AUC) and maximum concentration (C_{max}) each fit a log-normal distribution, and have no significant period effects. From the cross-over ANOVA model, the mean square errors are found to be 0.251 and 0.216. The geometric mean treatment ratios are 0.950 and 1.085. Sample size is 11 subjects in sequence A and 10 subjects in sequence B. The correlation of period I-period II differences on the log scale is 0.496.

The 80% individual treatment ratio prediction intervals for AUC and \(C_{\text{max}}\) are (0.59-1.54) and (0.72-1.64). If individual equivalence is defined as 80% of individual ratios within 0.6-1.67, equivalence would be accepted for \(C_{\text{max}}\), but not AUC. The 70% prediction interval for AUC is (0.65, 1.4), so individual equivalence based on a definition of 70% of ratios within 0.6-1.67 would be accepted for both variables. Assuming that the observed mean square errors are in fact the true within-subject standard deviations, reasonable equivalence limits (\(\Theta_1, \Theta_2\)) for 80% prediction intervals based on equation (6.13) are (0.51, 1.97) for AUC, and (0.54, 1.85) for \(C_{\text{max}}\). Based on these limits, individual equivalence is accepted for both responses. In Section 3.7.3, the TIER method rejected individual equivalence for both AUC and \(C_{\text{max}}\) based on 75% of individual treatment ratios within 0.7-1.43.

From small-sample simulation using the post-hoc estimates, the univariate power of individual equivalence (80% 0.6-1.67) is 0.41 and 0.64 respectively, and bivariate power is 0.26. If the sample represents the true distribution, the study had a 26% chance of finding both measures equivalent. In order to have reasonable power to show bivariate individual equivalence with this definition, a
much larger sample size is needed. If the same study had 40 subjects per sequence group, then the bivariate power would only increase to 40%. However, if the equivalence criterion was changed to 70% of ratios within 0.6–1.667, then the post-hoc bivariate power for the study with the current sample size increases to 70%.

6.8 Discussion and Conclusions

This chapter has presented a method for evaluating individual equivalence based on prediction intervals from cross-over trials with no carry-over. While the period effect can be incorporated in the ANOVA model and $\sigma_e$ is adjusted for this effect, new individuals are assumed to have no period effect, so it seems best to apply this method only in the absence of period effect. The method would require a different extension to parallel trials, since in parallel studies an individual receives only one treatment and individual treatment ratios can not be formed.

An advantage of the proposed individual equivalence method over Anderson and Hauck’s TIER method is that it is formulated similarly to standard methods used for average equivalence in an ANOVA framework, and is implemented easily with minimal modifications to these techniques.

Of some concern is the occurrence of type 1 error $> \alpha$ for values of $\delta_k$ where the population limits are equal to or slightly wider than the equivalence limits for a 100(1-2$\alpha$)% interval, even for moderate to large samples. It is possible that this effect is sufficiently counteracted by the conservativeness for large $|\delta_k|$ that stems from the shift of the center of the distribution from zero to $\delta_k$. In the bivariate case, conservativeness is also added by the fact that the test does not take into account the correlation of the outcomes.

Appropriate equivalence limits applicable to prediction intervals are functions of the within-subject variance for each response, so individual equivalence limits
vary from study to study. The method for choosing limits presented in this chapter assures reasonable properties at the population level: individual equivalence is accepted between the reference treatment and itself, and individual equivalence is not accepted if the geometric mean treatment ratio $\delta_k$ is greater than standard limits for average equivalence.

As is desired, it becomes harder to accept individual equivalence as either $|\delta_k|$ or the within-subject variance get larger. If test and reference variances are different, individual equivalence is harder to accept if the test treatment is more variable than reference, and easier to accept if the test is less variable than reference. For some reference treatments that are quite variable, reasonable limits for evaluating individual equivalence of test treatments may be so large as to be clinically unimportant. In these instances, a test for individual equivalence may not be appropriate.

Individual equivalence can be defined in many ways. Another method is to consider departure from equivalence as a fixed instead of random effect. If certain subgroups, for instance males and females, respond to the treatment differently, a treatment by subgroup interaction could result. An appropriate strategy is to evaluate average equivalence for each subgroup separately, and conclude overall average equivalence only if it is accepted for each subgroup. This subgroup analysis conducts more tests on smaller samples, and hence has less power. Acceptance of equivalence for two or more subgroups is a stronger result than average equivalence, but may not be as stringent as individual equivalence defined in this chapter.
Chapter 7
Multivariate Models for Equivalence

7.1 Introduction

A concern with separate univariate confidence intervals for each outcome, as evaluated in Chapters 4 and 5, is that the intervals do not account for the correlation between the outcomes. Even if separate CIs are the goal, modeling the estimates in a multivariate framework could lead to variance reduction and tighter confidence intervals for each univariate outcome. This chapter examines a variety of strategies for multivariate modeling which lead to univariate CIs that may be based on smaller variances. Multivariate strategies can also lead to global confidence intervals for averages of the outcomes, which always have smaller variance because of the averaging process.

Strategies are examined for log-normal and normal responses from cross-over or parallel designs, and categorical responses from parallel designs. Methods applied include weighted least squares, survey data linear and logistic regression, and mixed linear models. All discussion and examples are described in terms of two outcomes, but they can be generalized to multiple responses. Since this is a fairly novel application to equivalence analysis, this chapter is devoted to description of methods, applications, and examples, but does not evaluate power of the applications.

7.2 Methodology

The inappropriateness of MANOVA, or the general multivariate linear model, for equivalence studies has already been discussed. Multivariate tests such
as the Hotelling $T^2$ are omnibus tests that may be misleading if treatment ratios are heterogeneous across responses. The framework for testing homogeneity is also limited. In addition, MANOVA requires complete data from all observations and requires homogeneous covariance between responses across groups. For multiple outcomes, $\beta$ estimates and standard errors are the same as ANOVA estimates: no reduction in variance is achieved from the correlation of the responses. Because of these disadvantages, MANOVA models are not evaluated in this chapter. Multivariate methods which have improvements over MANOVA are presented in this section, and are applied to equivalence studies for the remainder of the chapter.

7.2.1 Weighted Least Squares

Weighted least squares (WLS) can be applied to evaluate equivalence by fitting multivariate models to functions of response outcomes. Assumptions are that functions of the data follow a normal distribution, and that the sample size in each subgroup is large enough for the sample covariance matrix to be a good approximation of the population covariance. Only categorical predictors are allowed so that sample size is large in each subgroup. WLS must usually have no missing data: all responses must be present for each subject.

The dependent variable in weighted least squares is a vector consisting of a function of the responses. Responses can be categorical, ordinal, or continuous. Functions of interest are subgroup log-odds for categorical data, and subgroup mean of logs (or logs of means) for continuous responses. The following discussion is based on Koch and Imrey (1985).

Let $s$ be the number of subgroups, and $F$ be a vector of subgroup functions. WLS assumes that $F$ is multivariate normal $F \sim N[\mu, V_F]$. Since the population parameters are not known, they are estimated in the sample by $\hat{F}$ and $V_\hat{F}$. 
A weighted regression model is fit to the vector $F$, assuming that $\mu = X\beta$. $X$ is a full rank $(s \times p)$ design matrix of known coefficients, and $\beta$ is a $(p \times 1)$ vector of unknown parameters. The parameter vector is estimated by the weighted least squares estimator $b$. Setting $F = Xb$, and solving for $b$ using weighted least squares produces:

$$b = (X'V_F^{-1}X)^{-1}X'V_F^{-1}F$$
$$\hat{V}_b = (X'V_F^{-1}X)^{-1}$$  \hspace{1cm} (7.1)

$\hat{V}_b$ is the Taylor series approximation of the variance of $b$, and is a consistent estimator of it. The estimated parameter vector has a multivariate normal distribution $b \sim N[\beta, (X'V_{\mu}^{-1}X)^{-1}]$. Since Taylor series is used for the variance estimation and since it is assumed $V_F$ is the correct variance of the subgroup means when calculating $b$, we are assured of reasonable estimates only by the central limit theorem, so sample sizes supporting each component of $\hat{V}_b$ should be 30 or more. If sample sizes are too small, standard error estimates are unstable. For categorical data, the estimate $b$ is also referred to as the minimum modified $\chi^2$ estimate of $\beta$. It minimizes the function $(\text{observed-expected})^2/(\text{expected})^2$, but is modified by the Taylor series linear approximation of the variance. A $100(1-\alpha)\%$ CI for the $p$th element of $\beta$ is obtained by

$$\left[ b_p \pm Z_{1-\alpha/2} \sqrt{\hat{V}_{b_p}} \right]$$  \hspace{1cm} (7.2)

Predicted values of $F$ are obtained by $\hat{F} = Xb$. Goodness of fit of the model is evaluated by $Q = (F-\hat{F})'V_F^{-1}(F-\hat{F})$. For large samples and when the model fits, $Q$ has a $\chi^2$ distribution with $df = s-p$. The model is assumed to be a reasonable fit to the data if the null hypothesis of adequate fit is not rejected. If goodness of fit is rejected, further predictors such as other covariates or interactions may need to be added to the $X$ matrix.

Hypotheses about the parameters can be tested by contrasts of the $b$ vector. The hypothesis is stated $H_0: C\beta = 0$, where $C$ is a full rank $(c \times p)$ matrix. It is tested by

$$Q_C = b'C'(C\hat{V}_bC')^{-1}Cb.$$  \hspace{1cm} (7.3)
When the model fits and the null hypothesis is true, $Q_C$ follows a $\chi^2$ distribution with $df=c$. If $C$ is a contrast which restricts or simplifies the model, $Q_C$ is the goodness of fit statistic $Q$ for the larger model minus that for the simplified model. Linear functions of the parameters are estimated by $Cb$ where $C$ is $(1 \times p)$, and their variance is estimated as $C\hat{V}_bC'$.

When assumptions are met, weighted least squares is an efficient method for fitting a linear model to response functions. Advantages of weighted least squares over ANOVA are that equal variance in subgroups is not required, independence between responses is not assumed, and separate models can be fit to two endpoints simultaneously. It is more efficient than ANOVA because it takes account of the covariance of the outcomes and repeated measures when it calculates $b$ and $\hat{V}_b$ by incorporating $V_F$ in their calculation. Efficiency in estimating $\beta$ is gained by the weighting: for a categorical predictor, the estimate $b$ is a weighted average of subgroup responses, with weight proportional to the inverse of the variance, so that less variable subgroups have a larger weight. The variance of the estimates is expected to be reduced from ANOVA variance estimates because it incorporates the covariance of the responses.

WLS produces the more efficient weighted beta estimate only when the number of parameters estimated is less than the number of subgroups. If the $X$ matrix is square, then $(X'V_F^{-1}X)^{-1}=X^{-1}V_FX^{-1}$, and the estimate simplifies to:

$$b = X^{-1}V_F X^{-1}X'V_F^{-1}F = X^{-1}F$$

$$\hat{V}_b = X^{-1}V_F X^{-1}$$

The improvement in $b$ due to weighting by the inverse variance is eliminated. Also, it can be shown that the standard error of $b_p$ does not take advantage of the covariance between responses. Such a square matrix occurs for a parallel design with two treatments and two responses. There are four subgroups and four parameters: the intercept and treatment effect for each response. If two separate
univariate ANOVAs are fit to the cell means for each response, $X$ is a square $2 \times 2$ matrix. The univariate estimates are:

$$
\hat{\beta}_k = (X'_kX_k)^{-1}X'_kF_k = X_k^{-1}X_k'X_k^{-1}X_k'F_k = X_k^{-1}F_k \\
\hat{\sigma}^2_\beta = (X'_kX_k)^{-1} = X_k^{-1}X_k'X_k^{-1} = \sigma^2_k
$$

(7.4)

The $\beta$ estimates from WLS are identical to those obtained by two univariate ANOVA models, and the variance of the betas in WLS is also similar except for accounting for heterogeneity of variance across groups. Since the most efficient use of weighted least squares stems from models with more subgroups than covariates, only designs which meet this criterion are examined in this chapter.

### 7.2.2 Survey Data Regression

Survey data regression can be applied to continuous data on the actual or log scale when assumptions for weighted least squares are not met. Usually this is when subgroup sample sizes are less than 30 or there are continuous covariates. Survey regression fits a model to the observations, not the subgroup means. While it is specifically designed to analyze data from complex surveys, it can be used in its simplest form as a method for modeling multivariate responses.

In such applications, every observation is given an equal weight, meaning that each subject had an equal chance of being in the study. It is assumed the selection occurred with replacement. A subject could be selected into the study more than once, but this never occurs because the population size is assumed large enough so that the probability of duplicate selection is essentially zero.

For the applications in this chapter, similar models are applied as WLS, and the $X$ matrices described for WLS are the essence matrices for survey regression. $F$, a vector of subgroup functions, is replaced by $Y$, the vector of observed responses. If there are $N$ subjects and $k$ responses, $Y$ is an $(Nk \times 1)$ vector. Often $Y$ is the log of the actual data. The following discussion of survey regression is

If it is assumed the selection probabilities are equal, the method applies standard least squares to estimate \( \beta \), but accounts for the correlation of outcomes in the calculation of \( \text{var}(\hat{\beta}) \). The estimate of beta, called the Horvitz-Thompson estimator, is \( \hat{\beta} = [X'WX]^{-1}X'WY \). For applications in this chapter all weights are equal to 1, so \( W \) is an identity matrix, and the estimate simplifies to the standard least squares estimate. The variance-covariance matrix of \( \hat{\beta} \) is calculated by a Taylor series approximation. The \( (p \times 1) \) estimate vector is first linearized for each stratum \( h \), person \( i \), and response \( j \) by using a Taylor series first order approximation of:

\[
Z_{hij} = [X'WX]^{-1}x_{hij}'[y_{hij}-x_{hij}\hat{\beta}]w_{hij}.
\]

Since the weights are all 1, and there are no design stratification variables, the linearized vector is simplified to:

\[
Z_{ij} = [X'X]^{-1}x_{ij}'[y_{ij}-x_{ij}\hat{\beta}],
\]

where \( x_{ij} \) is the \( (1 \times p) \) row vector of the \( X \) matrix corresponding to response \( y_{ij} \). \( Z_{ij} \) is a \( (p \times 1) \) column vector. For person \( i \), the \( J_i \) observations are added together, so that the linearized column vector for a person is

\[
Z_i = [X'X]^{-1}\sum_{j=1}^{J_i}x_{ij}'[y_{ij}-x_{ij}\hat{\beta}].
\] (7.5)

If observations for a subject are missing, they are simply not summed over. Therefore survey data techniques can use partial information, whereas weighted least squares cannot.

For one element of the \( \hat{\beta} \) vector, \( \hat{\beta}_p \), the variance is calculated as:

\[
\hat{\text{V}}(\hat{\beta}_p) = NS^2, \text{ where } S^2 = \frac{\sum_{i=1}^{N} (Z_{ip} - \bar{Z}_p)^2}{N-1}.
\]

\( Z_{ip} \) is the \( p \)th element of the \( i \)th observation’s linearized vector and \( \bar{Z}_p = \frac{\sum_{i=1}^{N} Z_{ip}}{N} \).

The covariance between two elements of \( \hat{\beta} \) is

\[
\text{Cov}(Z_1, Z_2) = N\sum_{i=1}^{N} \frac{(Z_{i1} - \bar{Z}_1)(Z_{i2} - \bar{Z}_2)}{N-1}.
\]

In general, the variance-covariance matrix of the elements of the vector \( \hat{\beta} \) is
\[ \hat{\mathbf{V}(Z)} = \frac{N}{N-1} \sum_{i=1}^{N} [Z_i - \bar{Z}][Z_i - \bar{Z}]' \]  

(7.6)

where \( \bar{Z} \) is the \( (p \times 1) \) vector of averages over the \( N \) subjects: \( \bar{Z} = \frac{N}{N} \sum_{i=1}^{N} Z_i \).

The variance estimate is a weighted sum of variances from each individual, in which responses from the same individual are summed together. Hence, the variance estimate takes the correlated structure into account. To summarize, survey data regression applies ordinary least squares regression ignoring the correlation structure, and then attempts to correct the variance of the \( \beta \) estimates with the above formula. This corrected estimate is essentially comparable to

\[ \hat{\mathbf{V}_\beta} = (X'X)^{-1}X'\mathbf{V}_\mathbf{F}X(X'X)^{-1}. \]  

(7.7)

Survey regression produces identical results as generalized estimating equations (GEE) when the GEE working correlation is an identity matrix.

Contrasts are specified by \( C \) matrices the same as in weighted least squares, but a variety of test types are available. The most preferable is the Satterthwaite adjusted \( F \), which provides estimated degrees of freedom to improve the \( \chi^2 \) approximation and accounts for the finite number of subjects in the study. Linear functions of parameters are estimated by \( C\hat{\beta} \), and their variance is estimated as \( C\hat{\mathbf{V}}_\beta C' \). Goodness of fit is supported by non-significance of effects not in the model (i.e., other covariates or interactions).

Survey regression has several advantages over weighted least squares. If there is missing data, the available information for a subject can still be used. The method does not require large sample sizes for subgroups, so it can be applied to broader situations with continuous covariates or many categorical predictors. The central limit theorem is invoked to assure that \( \hat{\mathbf{V}}_\beta \) is a reasonable estimate, so the number of subjects should be at least 30, and the behavior of estimates improves as sample size increases.

The estimates, however, are less efficient than weighted least squares. The \( \beta \) estimates do not take advantage of the covariance of responses, and the variance
estimates of the $\beta$s are less efficient than WLS, due to their approximation technique. In small samples, if the improvement from the correlation of responses is not great, the variance of survey data regression parameters may sometimes even be larger than those obtained from ANOVA.

This method can be applied using the survey analysis package SUDAAN (Research Triangle Institute 1991). A survey regression counterpart to logistic regression is also available. Survey logistic regression obtains standard logistic regression maximum likelihood estimates, and corrects the variances for the covariance structure of the data in a similar Taylor series linearization technique.

### 7.2.3 Mixed Linear Model

An alternate strategy to weighted least squares is the mixed linear model, which allows the modeling of random effects and repeated measurements where there is correlation between $Y$ values for certain observations. The covariance structure between outcomes can also be modeled. The mixed model yields an efficient estimator, provided the covariance of the responses is correctly specified. A summary of mixed linear methodology is presented in the SAS Proc Mixed Manual (SAS, 1992).

The mixed model begins with a strung-out vector of responses, in the same format as survey data regression. If there are $N$ subjects, each with $k$ responses, then the vector of responses is $Y_{(Nk \times 1)}$. The mixed model is specified as

$$Y_{(Nk \times 1)} = X_{(Nk \times p)}\beta_{(p \times 1)} + Z_{(Nk \times kq)}d_{(kq \times 1)} + e_{(Nk \times 1)}$$

where $X\beta$ specifies the fixed effects, as in a standard GLUML. $Z$ is a block diagonal matrix with $Z_1...Z_N$ as the diagonal elements, where the $Z_i$ are the design matrices for the random deviation regressions for each subject. $Zd$ specifies random deviations from the population regression lines for each subject $i$. The final term $e$ is the within-subject error for each subject and response. Unlike the
GLUM, the errors from one subject are not required to be independent, but independence is assumed across subjects. The model assumes \( Y \sim N(X\beta, \Sigma) \) where \( \Sigma = \text{Diag}(\Sigma_1, \Sigma_N) \), and \( \Sigma_i = Z_i V_d Z_i' + V_e \). In other words, it is assumed \( d_i \sim \text{NID}[0, V_d] \) independently of \( e_i \sim \text{NID}[0, V_e] \).

The covariance of a subject's observations are split into 2 components: the covariance of the random regression coefficients, \( V_d \), and the covariance of the subject's repeated measurement deviations about the random regression line, \( V_e \).

If \( \Sigma \) is re-written as \( ZGZ' + R \), where \( G \) and \( R \) are \((Nk \times Nk)\) block diagonal matrices with blocks of \( V_d \) and \( V_e \) respectively, then the \( d \) coefficients and the \( \beta \) parameters are obtained by:

\[
\bar{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \\
V_{\bar{\beta}} = (X'\Sigma^{-1}X)^{-1} \\
\bar{d} = GZ'\Sigma^{-1}(Y - X\bar{\beta}).
\] (7.8)

Estimates are solved by an iterative restricted maximum likelihood (REML) procedure. The covariance structure of both \( d_i \) and \( e_i \) can be modeled by a variety of imposed structures. The REML estimates are assumed to be approximately normal, which is generally met if the number of subjects is large (i.e., 30 or more).

Goodness of fit of the model can be assessed by comparison to a model with more parameters using a log likelihood ratio test. When a model is fit, a log likelihood is produced from the REML estimation. \( Q = -2[\text{loglikelihood(model2)} - \text{loglikelihood(model1)}] \), \( \text{df} = \text{number of parameters in 1 - number of parameters in 2} \), including \( \beta \) and covariance estimates. If the test is not significant, then the simpler model is assumed to fit. Goodness of fit can also be tested by non-significance of terms such as other covariates or interactions.

In multivariate equivalence, we are interested in the covariance between measurements for a person, but do not often assume that the subject
measurements are random effects, and we are not interested in coefficients of random regression lines. Therefore, in this discussion, Z and d are not specified, leading to a simplification of the model \( Y = X\beta + e \). \( \text{Var}(Y_i) \) is simplified to \( V_e \), the covariance between the \( k \) responses. The beta estimate becomes \( \hat{\beta} = (X'V_e^{-1}X)^{-1}X'V_e^{-1}Y \), which is a similar form as weighted least squares. When the covariance structure \( V_e \) is left unstructured, then although obtained through different strategies, the mixed model and weighted least squares produce very similar beta estimates and standard errors.

Mixed models can be fit with SAS Proc Mixed (SAS, 1992), using a repeated statement to identify the correlated observations and to impose structure on the covariance of the responses. Apart from a potential problem from lack of convergence of the REML estimates, the mixed linear model has several advantages over WLS. When some responses for a subject are missing, the available responses are included in the analysis. Sometimes pooling of covariances in \( V_F \) is desirable in WLS (see below), but WLS provides no framework to test the appropriateness of such pooling. Mixed models provide a variety of structures to impose pooling, and the log-likelihood ratio test is available to assess goodness of fit of the imposed structure.

7.3 Log-Normal Responses

This section examines multivariate models for assessing equivalence when the responses are assumed to follow log-normal distributions. They are discussed for parallel and cross-over designs with two outcomes, but can easily be generalized to more outcomes. The strategies outlined involve forming two univariate confidence intervals with estimates from a bivariate model, and generating an average or weighted average of responses to form one global confidence interval. These methods are applied with weighted least squares, survey data regression, and
mixed models.

7.3.1 Parallel Design

As explained above, weighted least squares is more efficient than ANOVA if the number of parameters fit is fewer than the number of subgroups. For this reason, consider a parallel design with two centers, two treatments, and two outcomes, so there are eight means based on four groups of subjects defined by (center \times treatment \times response). For valid use of weighted least squares, assume that there are at least 30 subjects per group.

Two Univariate CIs

Two separate parallel ANOVA models can be fit for the log of each response with an intercept, center, and treatment effect. The $\beta$ estimates for treatment and their standard errors can be used to form 90% CIs for the geometric mean treatment ratio, as described in Chapters 3 and 4, except the degrees of freedom are reduced by one to account for estimation of the center effect. As previously discussed, this is a valid method of analysis but ignores the correlation of the responses.

An alternate strategy which takes advantage of the correlation between responses is to fit a multivariate model in weighted least squares, where $\beta$ parameters for intercept, center, and treatment are estimated for the two responses simultaneously. Since there are eight subgroups but only six parameters, the design matrix is not square and WLS produces a more efficient estimate of $\beta$ than two separate ANOVAs. The WLS model is fit as follows:

Define $F$ as an $(8 \times 1)$ vector of subgroup means of log responses such that $y_{cgki} = \ln(x_{cgki})$ for each response in center $c$, treatment group $g$, and outcome $k$. Then

$$F = [\bar{y}_{11}, \bar{y}_{12}, \bar{y}_{1r1}, \bar{y}_{1r2}, \bar{y}_{21}, \bar{y}_{22}, \bar{y}_{2r1}, \bar{y}_{2r2}]'$$
\( V_F \) is an \((8 \times 8)\) covariance matrix with the following structure:

\[
V_F = \begin{bmatrix}
v_{111} & c_{1t} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_{1t} & v_{112} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & v_{1r1} & c_{1r} & 0 & 0 & 0 & 0 \\
0 & 0 & c_{1r} & v_{1r2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & v_{2t1} & c_{2t} & 0 & 0 \\
0 & 0 & 0 & 0 & c_{2t} & v_{2t2} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & v_{2r1} & c_{2r} \\
0 & 0 & 0 & 0 & 0 & 0 & c_{2r} & v_{2r2}
\end{bmatrix}
\] (7.9)

The \((\text{center} \times \text{treatment} \times \text{response})\) subgroup variances are calculated as

\[ v_{cgk} = \frac{n_{cg}}{n_{cg}(n_{cg} - 1)} \sum_{i=1}^{n_{cg}} (y_{cgki} - \bar{y}_{cgk})^2, \]

where \(n_{cg}\) is the number of observations in that subgroup.

The covariance between the two responses within the four patient groups is calculated as \( c_{cg} = \frac{n_{cg}}{n_{cg}(n_{cg} - 1)} \sum_{i=1}^{n_{cg}} (y_{cg1i} - \bar{y}_{cg1})(y_{cg2i} - \bar{y}_{cg2}) \). Division by \(n_{cg}\) yields a variance-covariance matrix of the means of logs, instead of individual observations. The covariance for a subject can not be calculated unless both measurements are present, so subjects with one or more missing responses must be excluded from analysis. Zero covariance between centers and treatments indicates these groups are independent.

The \(X\) matrix used to evaluate the treatment log-difference with reference cell coding has an intercept and effects for center and treatment for each of the responses:

\[
X = \begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\] (7.10)

Weighted least squares produces a \((6 \times 1)\) vector of the \(b\) estimate and its covariance matrix \(V_b\) by equation (7.1). The parameters are interpreted as:

\(b_1\): response 1 intercept: weighted mean for center 2 reference treatment

\(b_2\): weighted increment in response 1 mean due to center 1
b_3: weighted increment in response 1 mean due to test treatment
b_4: response 2 intercept: weighted mean for center 2 reference treatment
b_5: weighted increment in response 2 mean due to center 1
b_6: weighted increment in response 2 mean due to test treatment

If the goodness of fit test is not rejected, then the model can be considered a reasonable fit. The 100(1-\alpha)% confidence intervals for the treatment geometric mean ratio for each response are obtained by \( \exp\left[ b_3 \pm Z_{1-\alpha/2} \sqrt{V_{b_3}} \right] \) and \( \exp\left[ b_6 \pm Z_{1-\alpha/2} \sqrt{V_{b_6}} \right] \). Bivariate equivalence is accepted if 90% confidence intervals are contained within 0.8-1.25 for both responses.

If the center effect is not significant, reducing the model by removing center may be considered. While center is usually left in the model because it is a design factor, often it is desirable to remove other covariates such as baseline status or gender. Center effects for each response can be tested with a contrast such as \( C = [0 1 0 0 0 0 0] \) for response 1, which is a test of \( b_2 = 0 \), or can be tested by fitting a reduced model without the second and/or fifth columns of the X matrix, subtracting the goodness of fit chi-square values from the two models and comparing to the chi-square distribution with 1 or 2 degrees of freedom. If desired, one center effect could be left in, and the other removed. Even if center were removed from the model, the X matrix for WLS would be \( (8 \times 4) \), not \( (4 \times 4) \) so WLS still has an advantage over ANOVA in that it uses the covariance in its estimate of \( \beta \) and \( V_\beta \). More categorical covariates can be included, provided there is enough sample size per group.

**One Average CI**

A further reduction in the model is to test whether the treatment ratios for the two responses are combinable to form one global confidence interval. Homogeneity of the test-reference difference in mean of logs can be assessed by a
contrast \( C = [0 \ 0 \ 1 \ 0 \ 0 \ -1] \). If this test is not rejected, then it can be assumed that
the treatment difference for the two responses are homogeneous, and the ratios
from the two responses can be averaged together. A new model can be fit with
only 5 parameters, where \( b_3 \) becomes the weighted increment in the average
response due to test treatment:

\[
X = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1
\end{bmatrix}
\] (7.11)

In WLS, \( b_3 \) is a weighted average of responses, with weight proportional to the
inverse covariance matrix, so that the less variable response has a larger weight.
The standard error of this estimate is always less than the standard errors of the
individual responses, because of the averaging process. A categorical covariate is
not necessary for this global confidence interval to have reduced variance over two
ANOVA CIs, but the presence of a covariate is beneficial in obtaining an
increased reduction in variance and an improved \( \beta \) estimate.

The impact of the correlation between the responses on the weighted average
is a topic for future research. It is likely that less correlated responses provide a
more noticeable decline in standard error over the two individual SEs, since
\( \text{var}(\frac{a+b}{2}) = \frac{1}{4}\text{var}(a) + \frac{1}{4}\text{var}(b) + \frac{1}{4}\text{cov}(a,b) \). A 100(1-\( \alpha \))% CI for the global
treatment ratio is obtained by \( \exp[b_3 \pm Z_{1-\alpha/2} \sqrt{V_{b_3}}] \). Global bivariate equivalence
is accepted if a 90% confidence interval is contained within 0.8-1.25.

Forming a global confidence interval for a weighted average of treatment
genometric mean ratios has several advantages. There is no reduction in power to
detect equivalence from multiple univariate CIs, as examined in Chapter 4. In
fact, since WLS has improved efficiency at estimating coefficients, and because
the standard error is smaller due to averaging, power of the global test is likely to
be greater than power for each univariate response analyzed by ANOVA.

Disadvantages of this strategy stem from the possibility of accepting that the geometric mean treatment ratios are homogeneous when in fact they are not. If sample sizes are not large, there may not be reasonable power for testing whether the ratios are homogeneous. Homogeneity might be accepted only because of low power, and ratios which should not be combined might be averaged. This could be particularly bad if treatment ratio estimates are on opposite sides of 1.0: the weighted average might suggest equivalence even when each response may lack support individually.

Instead of applying weights from the variance structure, other pre-determined weights may be imposed. For instance, if it is twice as important for response 1 to be within equivalence limits than for response 2, then a weighted average can be obtained by \( C\beta \) applied to the 6 parameter model with \( C = [0 0 2 \frac{2}{3} 0 0 \frac{1}{3}] \). A direct average is obtained by \( C = [0 0 \frac{1}{2} 0 0 \frac{1}{2}] \). The inverse variance is not involved in estimating \( C\beta \), so the weighting component of efficiency is missing from the estimate in this type of average. However, efficiency is still present in the initial \( \beta \) estimates, and variance is still reduced by the averaging process. This application is fair even if the test of homogeneity is rejected, since one is not improving the estimate of the average with weights from the inverse of the variance matrix. However, when homogeneity is rejected, any type of averaging may be debatable.

If there are three or more responses, modeling can be conducted in a similar fashion as described above. If homogeneity of the treatment ratios is rejected, then contrasts can be formed to test if certain pairs or subsets of treatment ratios are homogeneous. Homogeneous sets can be grouped together to form CIs based on weighted averages, while the other responses can be evaluated separately. Equivalence can be accepted if all CIs (individual responses and weighted averages
of response subsets) are contained within 0.8-1.25.

It should be noted that averaging responses is of more interest for clinical equivalence than bioequivalence, especially when the purpose is to show broad equivalence, based on the combined information of many responses. This becomes of more interest as the number of responses increases, and power to show each of them univariately equivalent decreases. Univariate analysis of each response is still of interest for description. An example of a clinical equivalence trial with many outcomes is the blood pressure study described in Section 1.4.2, which measured standing and supine systolic and diastolic blood pressure at many visits.

Mixed Model and Survey Data Regression Methods

All of the previous strategies were explained in terms of weighted least squares, but can also be approached with a mixed linear model or survey data regression. The main difference in model specification is that these methods are designed to model each observation, instead of subgroup means. The dependent variable is the log of the observed value for each response, and the covariance matrix of responses is the same as for WLS, except that since this is the covariance of observations, not means, division by $n_{cg}$ is not applied. The $X$ matrix has the same form as described above, but the $X$ for weighted least squares is the essence matrix of observation-based matrices: each 1 becomes a column of 1s of length $n_{cg}$. The beta parameters and their variance are estimated differently for these two modeling methods, as described in the methodology sections. Adjustment of categorical covariates can have a slightly different interpretation when the study is not balanced, since WLS adjusts to the (weighted) average of subgroup means, whereas observation-based analyses adjust to the overall sample average.

When the assumptions for weighted least squares are met, then it may be
best to apply WLS, since it is more efficient than survey data regression, and avoids convergence problems which may arise in the mixed model. However, when subgroup sample sizes are not large enough for WLS or there are continuous covariates or missing data, then mixed models or survey data regression may be more appropriate. Each of these still requires log-normal distributions, and works best with large overall sample sizes.

Both mixed models and survey data regression can be used to fit two univariate CIs based on the bivariate model using the X matrix in equation (7.10), and one average CI from the model specified by (7.11). However, since survey data regression does not make use of the correlation structure when estimating the \( \beta \) coefficients, model (7.11) is actually a direct average, identical to that obtained by the contrast \( C = [0 \ 0 \ 1 \ 2 \ 0 \ 0 \ 2 2] \).

### 7.3.2 Cross-Over Design

All methods described for parallel designs are directly applicable to cross-over designs, assuming no carry-over effects (which can be tested for each response in an ANOVA as described in Chapter 3). Since cross-over designs have an additional factor for period, other categorical covariates are not necessary to see a gain in efficiency from weighted least squares. Other covariates could be included, but for simplicity consider the minimum subgroups for the bivariate analysis of a cross-over design with two treatments and two responses. Define \( F \) as an \((8 \times 1)\) vector of subgroup means of log responses such that \( y_{p\xi ki} = \ln(x_{p\xi ki}) \) for each response in period \( p \), treatment \( \xi \), and outcome \( k \):

\[
F = [\bar{y}_{111} \ \bar{y}_{112} \ \bar{y}_{1r1} \ \bar{y}_{1r2} \ \bar{y}_{211} \ \bar{y}_{212} \ \bar{y}_{2r1} \ \bar{y}_{2r2}]'.
\] (7.12)

The covariance matrix of the response means has a more complicated structure to account for correlation of the two periods for each subject, and is most easily referenced in terms of \( s \), the number of subgroups:
\[
V_F = \begin{bmatrix}
  v_1 & c_{12} & 0 & 0 & 0 & 0 & c_{17} & c_{18} \\
  c_{12} & v_2 & 0 & 0 & 0 & 0 & c_{27} & c_{28} \\
  0 & 0 & v_3 & c_{34} & c_{35} & c_{36} & 0 & 0 \\
  0 & 0 & c_{34} & v_4 & c_{45} & c_{46} & 0 & 0 \\
  0 & 0 & c_{35} & c_{45} & v_5 & c_{56} & 0 & 0 \\
  0 & 0 & c_{36} & c_{46} & c_{56} & v_6 & 0 & 0 \\
  c_{17} & c_{18} & 0 & 0 & 0 & 0 & v_7 & c_{78} \\
  c_{18} & c_{28} & 0 & 0 & 0 & 0 & c_{78} & v_8 
\end{bmatrix}
\]

(7.13)

Elements of \( V_F \) are calculated in the same manner as the parallel design, including division by the subgroup sample size to obtain covariance of the means.

The models for producing multiple univariate CIs and global average CIs for cross-over designs are exactly the same as described for parallel studies where center is replaced by period. Like center, period is a design factor, and is not removed from the model even if it is non-significant.

7.3.3 Covariate Adjustment

Covariate adjustment can be done in ANOVA, survey regression, and mixed models by adding a factor to the X matrix. In multivariate models, covariates can be applied to the two responses separately by having each covariate take the value 0 for the other outcome, or can be applied to both responses simultaneously by repeating the value of the covariate over both responses in the design matrix.

Weighted Least Squares can apply covariate adjustment by including the mean of the covariate for each subgroup in the outcome vector \( F \). The two response groups may have the same or a different covariate. The covariate is not restricted to the log scale. In the X matrix, all columns described by matrices in (7.10) and (7.11) have a zero for the covariate response, and additional columns are added which average the covariate over all appropriate subgroups. The assumption is that the covariate is equally distributed across the subgroups, which
is generally upheld in randomized studies. The resulting ratio of treatment geometric means is adjusted to the situation where the covariate is similarly distributed in all groups.

7.3.4 Smoothed Covariance Matrix in WLS

Weighted least squares requires large sample sizes in each subgroup for the appropriate estimate of the covariance matrix. If smaller sample sizes are available, the covariance matrix can still be validly estimated by pooling the variance and covariance estimates from some subgroups (Turney et al 1992). Covariance estimates are then based on larger sample sizes, allowing valid application of weighted least squares. Pooling can be based on assumptions of equal variances, and appropriateness can be confirmed with a variance ratio test.

For a parallel trial with two centers, two treatments, and two responses, the \((8 \times 8)\) covariance matrix has 8 variance and 4 covariance estimates, each based on the number of subjects with treatment \(g\) in center \(c\). The minimum sample size needed for valid weighted least squares is \(4 \times 30 = 120\). However, if the assumption of homogeneity of covariance across centers and treatments is imposed, then pooling the variance estimate for a response and the covariance estimate between the two responses is possible regardless of treatment or center. If sample sizes per treatment × center group are all equal, then the pooled variance-covariance matrix of the mean of logs can be represented by:

\[
V_{SF} = \begin{bmatrix}
v_1 & c_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_{12} & v_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & v_1 & c_{12} & 0 & 0 & 0 & 0 \\
0 & 0 & c_{12} & v_2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & v_1 & c_{12} & 0 & 0 \\
0 & 0 & 0 & 0 & c_{12} & v_2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & v_1 & c_{12} \\
0 & 0 & 0 & 0 & 0 & 0 & c_{12} & v_2
\end{bmatrix}
\]
Elements of the pooled covariance matrix are obtained for observations, and are then divided by $n_{sg}$ to obtain variances and covariances of means. If some subgroup sample sizes are different, then pooled estimates in $V_{SF}$ will accordingly vary across subgroups. The minimum sample size required for estimation of this matrix in weighted least squares is reduced from 120 to 30 total subjects.

For two-period cross-over trials, the standard $(8 \times 8)$ matrix for two sequences, two treatments, and two outcomes has 8 variances and 12 covariances. It requires 30 subjects per sequence group, or 60 total subjects for proper estimation. One choice of smoothed covariance is:

$$V_{SF} = \begin{bmatrix}
    v_1 & c_1 & 0 & 0 & 0 & 0 & c_2 & c_4 \\
    c_1 & v_2 & 0 & 0 & 0 & 0 & c_4 & c_3 \\
    0 & 0 & v_1 & c_1 & c_2 & c_4 & 0 & 0 \\
    0 & 0 & c_1 & v_2 & c_4 & c_3 & 0 & 0 \\
    0 & 0 & c_2 & c_4 & v_1 & c_1 & 0 & 0 \\
    0 & 0 & c_4 & c_3 & c_1 & v_2 & 0 & 0 \\
    c_2 & c_4 & 0 & 0 & 0 & 0 & v_1 & c_1 \\
    c_4 & c_3 & 0 & 0 & 0 & 0 & c_1 & v_2
\end{bmatrix}$$

For simplicity, $V_{SF}$ is depicted assuming $n_A = n_B$. Since $n_h$ varies between sequences if the study is unbalanced, the pooled covariances of the means actually differ somewhat as the sample size differs per sequence. The elements of $V_{SF}$ for sequence $h$ are calculated as follows:

$$\text{pooled variance for response } 1 = \frac{1}{n_h(2n_A + 2n_B - 4)} \sum_{p=1}^{2} \sum_{t_1=1}^{n_h} \sum_{t_2=1}^{n_h} (y_{p,t_1} - \bar{y}_{p})^2.$$  

Referring to notation in (7.13),

\[\text{pooled } v_1 = \frac{n_A(n_A-1)(v_1+v_7) + n_B(n_B-1)(v_3+v_5)}{n_h(2n_A + 2n_B - 4)}\]

\[\text{pooled } v_2 = \frac{n_A(n_A-1)(v_2+v_8) + n_B(n_B-1)(v_4+v_6)}{n_h(2n_A + 2n_B - 4)}\]

The pooled $c_1$ is the covariance between the 2 responses when periods and treatments are the same:
\[ c_1 = \frac{1}{n_A(2n_A + 2n_B - 4)} \sum_{p=1}^{2} \sum_{i=1}^{r} \sum_{t=1}^{n_A} (y_{p\xi t 1} - y_{p\xi t 2})(y_{p\xi t 2} - y_{p\xi t 2}), \]

or

\[ c_1 = \frac{n_A(n_A - 1)(c_{12} + c_{38}) + n_B(n_B - 1)(c_{34} + c_{56})}{n_A(2n_A + 2n_B - 4)} \]

C_2 and c_3 are the covariances between the 2 treatments when the sequence and the response are the same; c_2 is for response 1, c_3 is for response 2:

\[ c_2 = \frac{n_A(n_A - 1)(c_{17}) + n_B(n_B - 1)(c_{35})}{n_A(n_A + n_B - 2)} \]

\[ c_3 = \frac{n_A(n_A - 1)(c_{28}) + n_B(n_B - 1)(c_{46})}{n_A(n_A + n_B - 2)} \]

C_4 is the covariance between responses for the same sequence, but for different treatments:

\[ c_4 = \frac{n_A(n_A - 1)(c_{18} + c_{27}) + n_B(n_B - 1)(c_{36} + c_{45})}{n_A(2n_A + 2n_B - 4)} \]

If desired, the covariance matrix may further be reduced by pooling c_1 and c_4, and possibly c_2 and c_3. All estimates except c_2 and c_3 are based on twice the number of subjects in the study; c_2 and c_3 are based on the number of subjects. Therefore, this pooled covariance structure allows the valid application of WLS for a cross-over study with 30 or fewer subjects. Analysis results based on smoothed covariance matrices in WLS are more similar to results from ANOVA because of the more homogeneous covariance matrix.

Since survey data regression does not impose a covariance structure, smoothing is not directly applicable. Survey regression can be applied in small sample sizes, but due to its inefficiencies at estimating the covariance, it may result in larger standard errors than two separate ANOVA models.

In the mixed model, smoothing of the covariance matrix is possible through modeling its structure. Unlike WLS, the goodness of fit of the imposed covariance structure is tested with a log likelihood ratio test. Unfortunately SAS Proc Mixed has a limited repertoire of structures designed primarily for repeated measurement of a single outcome. The only smoothing that can be done in Proc Mixed for
cross-over designs is pooling the estimates from each sequence, so that there are 4 variance and 6 covariance estimates. No appropriate pattern of smoothing is available for parallel designs with multivariate responses.

7.3.5 Example

The previously described methods are applied to the bioequivalence of two formulations of a psychotropic drug as described in Section 1.4.1. In order for sample sizes to be large enough for weighted least squares, the results from three of the seven studies were pooled, for a total of 62 subjects.

Since this is a cross-over design, controlling for covariates other than period is not necessary to achieve benefits from WLS. However, it should be noted that controlling for period adds little efficiency over ANOVA, since the period parameter can be removed by simply modeling the period1-period2 difference. The X matrix in WLS would then be square, and WLS would have few gains over ANOVA. Improvements from multivariate strategies may be more apparent if an additional covariate was added to the ANOVA and multivariate models. An example of covariate adjustment for a parallel design is presented in Section 7.4.

Bioequivalence is evaluated for AUC and $C_{max}$ by 90% CIs for treatment geometric mean ratios from on univariate and multivariate modeling in Table 7.1.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>C_{max}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>SE</td>
</tr>
<tr>
<td>ANOVA</td>
<td>-0.031</td>
<td>0.047</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.044</td>
<td>0.046</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>-0.044</td>
<td>0.046</td>
</tr>
<tr>
<td>Survey Reg.</td>
<td>-0.031</td>
<td>0.101</td>
</tr>
</tbody>
</table>
Based on the 0.8-1.25 criterion, bioequivalence is accepted for all models. Weighted least squares and the mixed linear model produced essentially identical results. The estimates are different from ANOVA, due to weighting by the inverse covariance matrix in the $\beta$ estimate, and the variance is very slightly reduced over ANOVA. Survey regression produced the same estimate as ANOVA, but a larger variance, due to its inefficient variance estimation method.

A limitation of survey regression stems from the fact that within-subject parameters are of most interest, but the method is substantially influenced by between-subject variance through use of the computational strategy which produces estimates as if within-subject responses are independent. For cross-over designs this could be partially resolved by focusing analysis on within-subject differences, or a combination of within-subject differences and sums. If additional covariates are included in the models, further variance reduction from all three multivariate strategies would be expected.

Table 7.2 presents results from model-based (weighted) and directly combined averages of the treatment ratios across responses. Homogeneity of responses was supported by all models (WLS $p=0.226$, Mixed Model $p=0.227$, Survey Regression $p=0.741$). Averaging reduced variances somewhat. Weighted and direct averages provide fairly similar results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model-based Average</th>
<th>Direct Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{\beta}$</td>
<td>SE</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.021</td>
<td>0.041</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>-0.021</td>
<td>0.041</td>
</tr>
<tr>
<td>Survey Reg.</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
In order to demonstrate the use of smooth covariance matrices in WLS, the same bioequivalence example is evaluated, but only for Study #1. Sample size per sequence group is \( n_A = 11 \), \( n_B = 10 \). The order of the outcomes is sorted by period (1, 2), treatment (t, r), and response (1, 2), as described in (7.12). The covariance matrix of the log-responses for the observations is:

\[
V_Y = \begin{bmatrix}
0.270 & 0.170 & 0 & 0 & 0 & 0 & 0.153 & 0.100 \\
0.170 & 0.164 & 0 & 0 & 0 & 0 & 0.080 & 0.076 \\
0 & 0 & 0.514 & 0.425 & 0.548 & 0.512 & 0 & 0 \\
0 & 0 & 0.425 & 0.408 & 0.504 & 0.494 & 0 & 0 \\
0 & 0 & 0.548 & 0.504 & 0.724 & 0.615 & 0 & 0 \\
0 & 0 & 0.512 & 0.494 & 0.615 & 0.605 & 0 & 0 \\
0.153 & 0.080 & 0 & 0 & 0 & 0 & 0.147 & 0.104 \\
0.100 & 0.076 & 0 & 0 & 0 & 0 & 0.104 & 0.141 \\
\end{bmatrix}
\]

After smoothing as described in 7.3.4, the resulting matrix is:

\[
V_{SY} = \begin{bmatrix}
0.403 & 0.318 & 0 & 0 & 0 & 0 & 0.320 & 0.288 \\
0.318 & 0.320 & 0 & 0 & 0 & 0 & 0.288 & 0.274 \\
0 & 0 & 0.403 & 0.318 & 0.320 & 0.288 & 0 & 0 \\
0 & 0 & 0.318 & 0.320 & 0.288 & 0.274 & 0 & 0 \\
0 & 0 & 0.320 & 0.288 & 0.403 & 0.318 & 0 & 0 \\
0 & 0 & 0.288 & 0.274 & 0.318 & 0.320 & 0 & 0 \\
0.320 & 0.288 & 0 & 0 & 0 & 0 & 0.403 & 0.318 \\
0.288 & 0.274 & 0 & 0 & 0 & 0 & 0.318 & 0.320 \\
\end{bmatrix}
\]

The pooled covariance matrix of mean responses \( V_{SF} \) is obtained by dividing \( V_{SY} \) by:

\[
N = \begin{bmatrix}
11 & 11 \\
11 & 11 \\
10 & 10 & 10 & 10 \\
10 & 10 & 10 & 10 \\
10 & 10 & 10 & 10 \\
11 & 11 \\
11 & 11 \\
\end{bmatrix}
\]
All pooled variances and covariances are based on 42 observations, except for $c_2$ and $c_3$ which are based on 21.

The variances and covariances for sequence A and sequence B are actually quite different in this example. A variance ratio test found many significant differences between the variances from sequence B and sequence A which are averaged in the smoothing process. This suggests that smoothing is perhaps inappropriate for these data, and weighted least squares requires modified covariance estimation. When the smoothing process averages unequal quantities, then the usual weighted least squares method does not yield appropriate standard errors. However, a modification of the estimate of $V_\beta$ can at least partially correct for this:

$$\hat{V}_b^* = (X'V_{sf}^{-1}X)^{-1}X'V_{sf}^{-1}V_FV_{sf}^{-1}X(X'V_{sf}^{-1}X)^{-1}$$

(7.13)

If $V_{sf}=V_F$, then $\hat{V}_b^* = \hat{V}_b$. If the smoothed matrix is very unlike the original matrix, then the modified variance estimate $\hat{V}_b^*$ is more appropriate.

Table 7.3 presents 90% confidence intervals for equivalence based on univariate and smoothed-covariance multivariate results. WLS estimates are presented for the usual variance estimate, and the improved estimate from equation (7.13). For AUC, inappropriate smoothing increased variance above ANOVA, and the correction brought it back to ANOVA levels. Smoothing resulted in no variance change for $C_{max}$, but the correction improved it.

Mixed model results are presented both for unsmoothed covariance and partially smoothed, where the sequences are pooled together. The log likelihood test for comparison of the two mixed models was significant ($p=0.005$), again supporting that sequences should not be averaged in this example. For $C_{max}$, the variance from the smoothed model was worse than from the unsmoothed mixed model. In this example, the inappropriate smoothing models still achieved a smaller variance estimate than survey regression.
Table 7.3
Univariate Confidence Intervals for Equivalence of AUC and $C_{max}$ for Study #1 from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\beta}$</th>
<th>SE</th>
<th>90% CI</th>
<th>$\hat{\beta}$</th>
<th>SE</th>
<th>90% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>-0.051</td>
<td>0.078</td>
<td>(0.831-1.087)</td>
<td>0.082</td>
<td>0.067</td>
<td>(0.967-1.218)</td>
</tr>
<tr>
<td>WLS smooth</td>
<td>-0.051</td>
<td>0.089</td>
<td>(0.821-1.100)</td>
<td>0.082</td>
<td>0.067</td>
<td>(0.972-1.211)</td>
</tr>
<tr>
<td>WLS smooth*</td>
<td>-0.051</td>
<td>0.078</td>
<td>(0.836-1.080)</td>
<td>0.082</td>
<td>0.064</td>
<td>(0.976-1.206)</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>-0.048</td>
<td>0.077</td>
<td>(0.838-1.084)</td>
<td>0.091</td>
<td>0.062</td>
<td>(0.989-1.214)</td>
</tr>
<tr>
<td>MM smooth</td>
<td>-0.052</td>
<td>0.077</td>
<td>(0.834-1.080)</td>
<td>0.081</td>
<td>0.067</td>
<td>(0.971-1.212)</td>
</tr>
<tr>
<td>Survey Reg.</td>
<td>-0.051</td>
<td>0.195</td>
<td>(0.689-1.310)</td>
<td>0.082</td>
<td>0.172</td>
<td>(0.818-1.439)</td>
</tr>
</tbody>
</table>

* = corrected variance from equation (7.13)

Homogeneity of treatment ratios between responses is not well supported for this example (WLS-smooth p=0.099, WLS-smooth* p=0.054, Mixed Model p=0.045, Smoothed Mixed Model p=0.053). Survey regression did not reject homogeneity, but the variances were much larger.

The example used in this section stressed the multivariate methods to their limits. The reduction in variance from a cross-over design with no additional covariates is apparently minimal, and actually may lead to a worsening of variance in survey regression. The addition of covariates to both univariate and multivariate models should improve the multivariate performance.

7.4 Responses Following a Normal Distribution

Often it is not appropriate to assume responses have log-normal distributions, for instance time to maximum concentration in bioequivalence studies, or blood pressure measurements in the clinical study of ramipril and HCTZ described in Section 1.4.2. This section describes univariate and multivariate equivalence strategies for data assumed to follow a normal distribution for parallel designs.
When the responses are assumed to be normal, univariate analyses can find a confidence interval for the difference of treatment means. But when ratios are of interest, CIs of treatment mean ratios can be obtained with Fieller's Formula, described in Section 3.2.4. Taylor series approximation can also be used to estimate the ratio of treatment means. Consider one response from a parallel design, with mean response per group $Y = [\bar{y}_r, \bar{y}_s]^T$, and variance of the means $V_Y = \begin{bmatrix} v_r & 0 \\ 0 & v_s \end{bmatrix}$ where $v_r$ is the sample variance divided by the sample size. Taylor series can be used to find the approximate mean and variance of the log of treatment means as follows: $F = \log(Y)$, $V_F = D_Y^{-1} V_Y D_Y^{-1}$, where $D_Y$ is a diagonal matrix with the elements of $Y$ along the diagonal. The difference of logs of means can be obtained by $G = AF$ where $A = [1, -1]$, and the variance is $V_G = AV_F A'$. The steps can be combined by $G = A \log(Y)$, $V_G = AD_Y^{-1} V_Y [AD_Y^{-1}]'$ (Koch and Imrey 85). A $100(1-\alpha)$% confidence interval for the ratio of treatment means is obtained by $\exp(G \pm Z_{1-\alpha/2} \sqrt{V_G})$.

A drawback of Taylor series approximation to obtain a CI of treatment mean ratios is that it can be applied only when all means are positive, so the log of the mean exists. $Y$ can be extended to two or more outcomes with no change in the formulas except that the $A$ matrix has more columns and one row for each response. The assumption that a non-linear function (such as the log) is similar to its Taylor series linearized value is reasonable when sample sizes are large, preferably 30 subjects per subgroup, as is required for weighted least squares.

Treatment group means can be adjusted for a categorical covariate without the use of models. If $Y$ consists of means for each subgroup separately, i.e., $Y = [\bar{y}_{1r}, \bar{y}_{1s}, \bar{y}_{2r}, \bar{y}_{2s}]^T$, then category-adjusted means are obtained by $CY$, where $C = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ with variance $CV_Y C'$. This strategy can be applied to logs of means for Taylor series estimation, or to straight means for Fieller's formula. Adjusting for a continuous covariate requires modeling. Adjusted treatment
means for continuous and/or categorical covariates are obtainable from ANOVA.

Neither Fieller's Method nor Taylor series approximation work well if either of the treatment means is very close to zero; the methods perform best if the means are two standard errors away from zero. This requirement could be a problem for small responses such as change from baseline, or responses with a very large variance.

Just as for log-normal data, if there are two responses, it is possible to increase the efficiency of the estimates and reduce their variability by applying multivariate modeling techniques to the estimation of the treatment means and their variance. In the next sections, such methods are proposed for a parallel design with two normal responses.

7.4.1 WLS Modeling of Logs of Means

As previously explained, improvement of WLS beyond ANOVA can be achieved for parallel designs when a covariate (either categorical or continuous) is adjusted, since the design matrix has more rows than columns.

Categorical Covariates

Consider a parallel design with two responses, two treatments, and one categorical covariate with two groups, such as center, as described in Section 7.3.1. If the data have normal distributions rather than log-normal, and if each of the eight subgroup means is greater than zero, then the vector of mean of logs, defined $F$ in 7.3.1, can be replaced by the log of means $F^* = \log(Y)$. The Taylor series estimate of variance is $V_F^* = D_y^{-1}V_y D_y^{-1}$ (Koch and Imrey 85). All weighted least squares modeling techniques described in 7.3.1 are directly applicable to $F^*$ and $V_F^*$. It is expected that the improved efficiency and variance reduction imposed by WLS should produce a more precise estimate with a smaller variance than straight category-adjusted Taylor series. Center-adjusted treatment
mean ratios obtained from WLS are adjusted to a weighted-average of center logs of means, with more weight from the center with less variance.

Continuous Covariates

Continuous covariate adjustment can be applied in a related manner, by first obtaining a vector $Y^*$ of response means and covariate means for each of the four treatment-response groups. Provided that all response and covariate means are positive, $\mathbf{F}^*$ and $\mathbf{V}_P$ can be calculated. Log-transforming the covariate means is not necessary for analysis but makes calculation of $\mathbf{V}_P$ more straightforward. Models for obtaining two univariate CIs and average CIs can be fit as described in Section 7.3.1, with additional columns in the $X$ matrix for describing variation of the covariate means per response, as described in 7.3.3, so that covariate adjusted results are obtained.

7.4.2 Evaluating Covariate-Adjusted Means

Survey regression and mixed linear models can not be modified to evaluate logs of means instead of means of logs because they operate at the observation level. However, it is possible to obtain the advantages of variance reduction through multivariate application by obtaining covariate adjusted response means that can then be applied to Taylor series and Fieller's Formula methods of treatment mean ratio estimation. Survey regression and mixed models, and also similar techniques in WLS, can adjust for either continuous or categorical covariates, and can obtain improved estimates through variance reduction by accounting for correlation of multiple outcomes.

Categorical Covariates

Adjustment for a categorical covariate such as center can be applied in survey regression, mixed models, and WLS by fitting a model with the $X$ matrix
defined in equation (7.10) to Y, the vector of responses on the standard scale. Center adjusted means for each treatment group and response are obtained by 

\[ Y_A = C\beta, \text{ where } C = \begin{bmatrix} 1 & 1 & \frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \]

\( Y_A \) can be applied to form univariate confidence intervals of ratios for each response using either Taylor series or Fieller’s Formula.

**Continuous Covariates**

Continuous covariate adjusted means for each treatment-response group are obtained in survey regression and mixed models by fitting an X matrix having columns for each covariate and predictors for the four cell means (test and reference for responses 1 and 2). Covariates can be adjusted for each response separately or both simultaneously as described in Section 7.3.3. Covariate adjusted means are then obtained for each group by a contrast which sets the covariate equal to the overall average. For example, adjustment of a separate covariate for each response is obtained with:

\[ C = \begin{bmatrix} \bar{x}_1 & 1 \\ \bar{x}_2 & 1 \\ \bar{x}_1 & 1 \\ \bar{x}_2 & 1 \end{bmatrix} \]

where \( \bar{x}_k \) is the covariate mean for response \( k \). The resulting means can be applied to univariate Fieller’s and Taylor series methods to obtain confidence intervals for treatment mean ratios adjusted to the situation where the covariate is equally distributed between the treatment groups.

A similar strategy of obtaining covariate-adjusted means is available in WLS. A model with the same structure as the covariate-adjusted WLS model described in Section 7.4.1 is fit to the group means (not logs of means). Instead of an
intercept, a cell mean is fit for each treatment-response subgroup, with other X columns averaging the group means for the covariates.

For any multivariate model, a direct average of response treatment mean ratios can be obtained in Taylor series by multiplying $G$, the vector of differences of treatment logs of means, by a matrix $A_2 = [0.5 \ 0.5]$. No such averaging is available in Fieller's method. WLS methods for obtaining model-based weighted averages of treatment mean ratios can also be applied when covariate adjusted means are all positive.

One advantage of covariate adjustment on the original scale combined with Fieller's formula is that neither the responses nor mean responses are required to be positive, since they are never log-transformed. Taylor series requires the adjusted means to be positive, but otherwise adjustment on the original scale is more flexible than WLS modeling of logs of means.

While the standard error of estimates may be expected to be lowest for WLS or mixed models because they are more efficient than survey regression, the strategy with lower estimated variance may differ depending on individual data.

7.4.3 Example

The methods described above are applied to the parallel clinical equivalence study comparing ramipril and HCTZ for change in blood pressure. As in Section 4.3.7, the treatments of 2.5 mg ramipril (test treatment) vs. 12.5 mg HCTZ (reference) are compared for supine systolic and diastolic blood pressure change from baseline. There are 44 subjects in the test treatment group, and 45 in the reference group. The outcome is the (visit 1 - visit 9) difference in blood pressure. Covariates for adjustment are the visit 1 systolic and diastolic blood pressures, each used to adjust the relevant outcome. There is no categorical covariate such as center, but a category (called diacat) based on the baseline diastolic blood
pressure was created for example purposes (<106 vs. ≥ 106). This cutpoint was chosen so that sample sizes in each subgroup are over 20 - less than desirable for category adjusted WLS, but not too small for example purposes.

Blood pressure and change in blood pressure are generally assumed to follow a normal rather than log-normal distribution. While change from baseline can be negative, the mean change for each diacat-treatment-response subgroup is positive, and so weighted least squares methods can be applied to logs of means.

Tables 7.4 and 7.5 present confidence intervals for the ratio of treatment means from methods described in Section 7.4. The Taylor series \( \hat{\beta} \) column in the tables is the estimate of the log of the ratio of treatment means for difference from baseline. All Fieller's formula methods are based on equation (3.14), without pooling the variance between treatment groups. The \( t \) distribution with \( n_r+n_r-2 \) df is used for approximate purposes. More conservative approaches would be to subtract an additional degree of freedom for each covariate, or to set the degrees of freedom to \( \min\{n_r-1, n_r-1\} \).

From this analysis, it is concluded that the treatments are not equivalent, even after adjustment for covariates, based on 90% CIs within 0.8-1.25. For diastolic blood pressure, the standard error is reduced by adjusting for either the categorical or continuous version of baseline diastolic blood pressure. While estimates vary somewhat, all of the multivariate methods result in a slight reduction in variance from the corresponding ANOVA adjustment strategy.
### Table 7.4
Confidence Intervals for Equivalence of Diastolic Blood Pressure from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Adjustment Method</th>
<th>Taylor Series</th>
<th>Fieller's Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>-0.018</td>
<td>0.271</td>
</tr>
<tr>
<td>Diacat Adjusted</td>
<td>0.009</td>
<td>0.260</td>
</tr>
<tr>
<td>ANOVA - Diacat</td>
<td>0.011</td>
<td>0.256</td>
</tr>
<tr>
<td>ANOVA - covar</td>
<td>0.061</td>
<td>0.258</td>
</tr>
<tr>
<td>Diacat WLS log-mean</td>
<td>-0.037</td>
<td>0.256</td>
</tr>
<tr>
<td>WLS</td>
<td>0.041</td>
<td>0.249</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>0.041</td>
<td>0.248</td>
</tr>
<tr>
<td>Survey Reg</td>
<td>0.011</td>
<td>0.254</td>
</tr>
<tr>
<td>Covar WLS log-mean</td>
<td>0.059</td>
<td>0.255</td>
</tr>
<tr>
<td>WLS</td>
<td>0.060</td>
<td>0.257</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>0.071</td>
<td>0.257</td>
</tr>
<tr>
<td>Survey Reg</td>
<td>0.061</td>
<td>0.256</td>
</tr>
</tbody>
</table>

### Table 7.5
Confidence Intervals for Equivalence of Systolic Blood Pressure from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Adjustment Method</th>
<th>Taylor Series</th>
<th>Fieller's Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>-0.213</td>
<td>0.278</td>
</tr>
<tr>
<td>Diacat Adjusted</td>
<td>-0.201</td>
<td>0.280</td>
</tr>
<tr>
<td>ANOVA - Diacat</td>
<td>-0.199</td>
<td>0.279</td>
</tr>
<tr>
<td>ANOVA - covar</td>
<td>-0.144</td>
<td>0.257</td>
</tr>
<tr>
<td>Diacat WLS log-mean</td>
<td>-0.238</td>
<td>0.277</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.187</td>
<td>0.271</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>-0.187</td>
<td>0.269</td>
</tr>
<tr>
<td>Survey Reg</td>
<td>-0.199</td>
<td>0.275</td>
</tr>
<tr>
<td>Covar WLS log-mean</td>
<td>-0.151</td>
<td>0.257</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.151</td>
<td>0.258</td>
</tr>
<tr>
<td>Mixed Model</td>
<td>-0.134</td>
<td>0.258</td>
</tr>
<tr>
<td>Survey Reg</td>
<td>-0.144</td>
<td>0.259</td>
</tr>
</tbody>
</table>
Adjusting for the diastolic baseline category basically resulted in no variance reduction for systolic blood pressure from univariate strategies. Diastolic baseline apparently has little effect on change in systolic pressure. This shows a limitation of multivariate category adjustment: both responses must be adjusted for the same category, whether this is desirable or not. A substantial reduction in variance is seen when systolic baseline is adjusted for, although multivariate models showed no improvement over ANOVA.

Table 7.6 presents results from averaging over responses for WLS modeling of logs of means. Homogeneity of responses was accepted for both continuous and categorical covariate adjustment ($p=0.267$ and $p=0.325$ respectively). Direct and model-based averaging using Taylor series and WLS could also have been applied to adjusted means obtained from any of the other multivariate covariate adjustment techniques. Additional variance reduction is obtained due to the averaging of systolic and diastolic blood pressure treatment ratios.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model-based Average</th>
<th>Direct Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$  SE</td>
<td>90% CI</td>
</tr>
<tr>
<td>WLS log-mn Dia</td>
<td>-0.111  0.245</td>
<td>(0.598-1.340)</td>
</tr>
<tr>
<td>WLS log-mn cov</td>
<td>-0.042  0.238</td>
<td>(0.648-1.418)</td>
</tr>
</tbody>
</table>

The multivariate strategies evaluated in this example each produce relatively comparable variance reduction, with no strategy clearly performing better or worse than others, although WLS and mixed models appear to be preferable for Diacat adjustment in this example. Weighted least squares and mixed model category adjustment produced very similar results, but results differ for covariate adjustment because of the different strategies used to adjust for continuous
covariates. Survey regression produced estimates identical to ANOVA estimates, with variances sometimes higher and sometimes lower than the univariate counterpart.

This section has demonstrated that when treatment ratios of several normally distributed responses are of interest, then narrower confidence intervals for estimates may be obtained by applying multivariate strategies which take advantage of the correlation between the responses. Some methods require that all subgroup means for responses be positive. These methods could also be applied to a cross-over design.

7.5 Dichotomous Response

Similar WLS and survey regression strategies as described in 7.3 can be applied to multiple dichotomous responses. The most common situation is a parallel trial with two treatments and two dichotomous responses, as described in Chapter 5. While Chapter 5 explores three analysis strategies, evaluation of test vs reference odds ratios are best applied to modeling through logistic regression. The direct method of univariate analysis described in Section 5.4.1 is equivalent to a logistic regression model with an intercept and treatment effect. Adjustment for a categorical covariate can easily be implemented by adding the covariate predictor in the model. Proportion differences and ratios can also be modeled by weighted least squares, but will not be evaluated in this chapter.

Multivariate modeling strategies which take account of the correlation between responses may have advantages over univariate analysis by achieving a more efficient estimate with reduced variance. Weighted least squares again requires a covariate to achieve improvement over univariate strategies. If there is a categorical covariate such as center, then WLS is applied exactly as described in 7.3.1, except the modeled response $F$ is the logit for each subgroup. $F$ and $V_F$ are
calculated as described in Section 5.4.1, except the matrices are doubled in size to account for the two centers, and the $A_2$ matrix forms log-odds instead of log-odds ratios: $A_2 = [1 \ -1] \otimes I_5$. Smoothing of covariance matrices may be cruder for categorical data, since variances are determined by the proportions in each subgroup. GEE methods circumvent this problem by structuring the correlation instead of the covariance matrix.

Categorical responses can also be evaluated with survey logistic regression. The $\beta$ estimates are exactly the same as achieved with logistic regression (described in Section 2.1.3), but the variance matrix of the coefficients is adjusted for the correlation structure of the data in a manner similar to that described in Section 7.2.2.

7.5.1 Example

As an example, consider a study of the treatment of duodenal ulcers. The two responses of interest are healing of the ulcer by endoscopy exam (outcome 1) and reported absence of ulcer pain (outcome 2). Figure 7.1 depicts hypothetical data for responses from a study with two treatments stratified by patient gender.

**Figure 7.1**

Ulcer Study Example of Bivariate Dichotomous Response

<table>
<thead>
<tr>
<th>Outcome 1,2</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men Test</td>
<td>130</td>
<td>30</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Men Ref.</td>
<td>130</td>
<td>40</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Women Test</td>
<td>115</td>
<td>25</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>Women Ref.</td>
<td>120</td>
<td>20</td>
<td>10</td>
<td>50</td>
</tr>
</tbody>
</table>
Table 7.7 presents 90% confidence intervals for the odds of positive response for test vs. reference treatment for unadjusted and gender adjusted logistic regression, and gender adjusted WLS and survey logistic regression models. In this example, survey regression resulted in identical results as standard logistic regression, but the WLS weighted log-odds estimate achieved a slightly lower estimated variance.

Table 7.7
Confidence Intervals of Odds Ratios for Equivalence of Ulcer Treatments from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>( \hat{\beta} )</th>
<th>SE</th>
<th>90% CI</th>
<th>( \hat{\beta} )</th>
<th>SE</th>
<th>90% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic un.</td>
<td>-0.138</td>
<td>0.166</td>
<td>(0.662-1.145)</td>
<td>0.059</td>
<td>0.153</td>
<td>(0.824-1.365)</td>
</tr>
<tr>
<td>Logistic Adj.</td>
<td>-0.141</td>
<td>0.168</td>
<td>(0.658-1.145)</td>
<td>0.059</td>
<td>0.154</td>
<td>(0.824-1.367)</td>
</tr>
<tr>
<td>Survey Reg.</td>
<td>-0.141</td>
<td>0.168</td>
<td>(0.658-1.145)</td>
<td>0.059</td>
<td>0.154</td>
<td>(0.823-1.368)</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.107</td>
<td>0.164</td>
<td>(0.686-1.178)</td>
<td>0.026</td>
<td>0.152</td>
<td>(0.799-1.318)</td>
</tr>
</tbody>
</table>

Table 7.8 presents results from model-based and direct averages of the treatment odds ratios across responses. Homogeneity across responses was supported (WLS p=0.408, Survey Regression p=0.236). Unlike survey linear regression, survey logistic regression reports different results for the model-based and direct averages because of the iterative maximum likelihood method used to calculate \( \beta \) estimates. The model-based average for WLS is a weighted average with weights inversely proportional to the variance. Averaging reduces variances and tightens confidence intervals. The different averaging strategies and multivariate methods provide similar results.

We can conclude that the two ulcer treatments are equivalent since the estimated odds ratio for either response is significantly greater than 0.5 for a one-sided test, or internal to (0.5, 2.0) for a two-sided test. Multivariate analysis by
WLS produced somewhat more efficient estimates with reduced variance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Model-based Average</th>
<th>Direct Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>SE</td>
</tr>
<tr>
<td>WLS</td>
<td>-0.030</td>
<td>0.136</td>
</tr>
<tr>
<td>Survey Reg.</td>
<td>-0.032</td>
<td>0.137</td>
</tr>
</tbody>
</table>

7.6 Other Multivariate Strategies

This chapter has evaluated strategies for continuous and categorical responses. In future work, similar multivariate models could also be developed for equivalence trials with several ordered categorical responses. The outcome could be evaluated with either a proportional odds model or a ratio based on a transformation of the Mann-Whitney rank measure of association (Carr, Hafner, Koch 89). It is also of interest to develop strategies for analysis when a combination of responses is measured, for instance one log-normal and one categorical response. Since the modeling strategies for log-normal, categorical, and potentially ordinal responses estimate treatment ratios based on the log scale, it may be possible to combine the methods to form a global equivalence measure for a study which has any combination of outcome types.

A non-parametric method of combining many continuous responses is achieved through an extension of O'Brien's rank-sum test described in Section 1.3.1. O'Brien's method involves ranking the measures, summing the ranks from all outcomes for each person, and performing a t-test on the summed-ranks to compare treatment groups. An extension to equivalence studies involves first multiplying each outcome for the reference therapy by 0.80, then conducting
O'Brien's method to obtain a 1-sided test of whether the new treatment is less than than 80% of the reference. The procedure is repeated, multiplying the reference by 1.25, and testing whether the test treatment is greater than 125% of the reference. If both the 80% and 125% p-values are less than 0.05, then global equivalence of the treatments is accepted. One disadvantage is that this method does not lead directly to a confidence interval. However a CI can be produced by multiplying the standard drug values by a range of numbers until the cut-points are found which just lead to a p-value of 0.05 or greater.

This modification of O'Brien's method may not be applicable to ordered categorical or dichotomous responses, because it is meaningless to multiply such responses by a number like 0.80 or 1.25. Other strategies may need to be developed for these outcomes. The rank-sum test may be less powerful than weighted least squares since the responses are converted to ranks. It also has no way to test for homogeneity of the responses. However, when sample sizes are reasonably large and many continuous responses are measured, this method can be used to achieve an overall test of equivalence without imposing assumptions of normal or log-normal distributions.

7.7 Conclusions

In this chapter, multivariate models for equivalence were evaluated for log-normal, normal, and categorical responses. The models employed weighted least squares, mixed linear models, and survey regression. By taking the correlation between several outcomes into account, a reduction in variance for each univariate estimate can be achieved. The improvement in variance varies with method and application. Only slight variance reduction was seen in examples, but reduction may be more pronounced when the covariates explain a larger amount of the overall variation for each outcome, and when the outcomes are highly correlated.
Methods were evaluated for two responses, but would likely experience further variance reduction as the number of outcomes increases.

Methods for combining results from multiple responses were also presented in order to form confidence intervals for global assessment of equivalence. Global assessment is most helpful for clinical equivalence trials when many responses are measured, and it is appropriate when homogeneity of response treatment ratios is accepted. The assessment of combined outcomes appears to be a more powerful method of evaluating overall equivalence than the separate univariate analysis of each endpoint since the averaging process decreases the variance of the estimate and only one confidence interval is formed.
Chapter 8
Analysis Strategies for
Placebo Controlled Equivalence Trials

8.1 Introduction

As discussed in Section 1.2.2, a flaw of clinical equivalence trials is the lack of a placebo. Statistical analysis may suggest that the test treatment is clinically equivalent to the reference, but when neither the test nor the reference are compared to a placebo, the efficacy of the test drug can only be assumed based on historical placebo controls from previous trials of the reference treatment. In some trials a placebo is not possible, due to ethical reasons or the severity of the disease under investigation. In many situations of less acute illness, however, it is feasible to administer a placebo. The objectives of a study with a test, reference, and placebo treatment might be to show that both active treatments are significantly better than placebo, and that the test and reference treatments are equivalent.

In a placebo-controlled equivalence trial, information from the placebo can be incorporated into the equivalence testing. For instance, instead of evaluating the ratio of test and reference means, one could evaluate the ratio of mean differences from placebo, which tests whether the active treatments are equally different from placebo. Equivalence could be accepted if a 90% confidence interval for the ratio of mean difference from placebo is within a pre-specified range, possibly 0.8-1.25.

Gould (1993) suggested adjusting the ratio of treatment means with a placebo group. For dichotomous responses, he refers to the ratio of mean difference from
placebo as the reduction in risk relative to placebo. He also suggests equivalence might be based on the relative loss due to not using the test treatment, defined as the ratio of difference from placebo divided by the simple ratio of test and reference response rates.

Evaluation of equivalence in this manner is of little interest for bioequivalence, since administration of placebo generally yields no response for bioavailability measures such as AUC and $C_{max}$ or skin blanching in a dermatologic study. The placebo group mean is equal to zero, and the ratio of active treatment-placebo differences simplifies to the standard ratio of active treatment means. Controlling for placebo is applicable, however, to any outcome response that is non-zero when no treatment is applied, such as blood pressure, change in blood pressure, or ulcer healing rates. This method is applicable to parallel trials or cross-over designs where the subject receives all three treatments or one active treatment and placebo.

Differences from placebo are of most interest for responses on an additive rather than multiplicative scale, for instance normally distributed rather than log-normal responses. Active vs placebo ratios may be of more interest for log-normal responses, but for most study designs, the placebo cancels when the test vs reference ratio of ratios is formed.

This chapter evaluates analysis techniques to form confidence intervals for ratios of treatment mean differences from placebo for parallel trials with normally distributed or dichotomous responses. As in Chapter 7, multivariate techniques such as weighted least squares, survey regression, and mixed linear models are evaluated for their ability to improve estimates over standard univariate strategies when multiple response criteria are evaluated.
8.2 Parallel Trials with Normally Distributed Responses

Ratios of treatment mean differences from placebo for continuous responses are limited in this chapter to those following a normal distribution. Non-parametric techniques for skewed distributions or small sample sizes are possible, but are not evaluated here. Methods discussed in this section are described for parallel trials, but are also applicable to cross-over designs.

Consider a parallel trial evaluating two active treatments and a placebo for two responses following a normal distribution. The ratio of test vs reference treatment mean difference from placebo for response \( k \) is formed as:

\[
\theta_{pk} = \frac{(\mu_{tk} - \mu_{pk})}{(\mu_{rk} - \mu_{pk})} \tag{8.1}
\]

Equivalence is accepted if a 100(1-2\( \alpha \))% confidence intervals for \( \theta_{pk} \) is completely contained within pre-specified equivalence limits \( \Theta_1 \) and \( \Theta_2 \). \( \theta_{pk} \) is estimated in the sample by

\[
r_{pk} = \frac{(\bar{y}_{tk} - \bar{y}_{pk})}{(\bar{y}_{rk} - \bar{y}_{pk})}. \tag{8.2}
\]

The equivalence criterion of \( \Theta_1 \leq \theta_{pk} \leq \Theta_2 \) can be rewritten as:

\[
\Theta_1 + (1-\Theta_1)\frac{\mu_p}{\mu_r} \leq \frac{\mu_t}{\mu_r} \leq \Theta_2 + (1-\Theta_2)\frac{\mu_p}{\mu_r} \tag{8.3}
\]

If \( \mu_p \) is zero, this simplifies to the standard definition of equivalence evaluated in Chapter 4. However, as \( \mu_p \) increases with respect to \( \mu_r \), the criterion becomes increasingly more stringent. If \( \mu_p = \mu_r \), then equivalence is essentially impossible to prove. Not only is this criterion more stringent than the standard definition, but the placebo mean introduces additional variability in the ratio estimate. The usual limits of 0.8-1.25 may be too stringent for this equivalence definition.
Perhaps more appropriate equivalence criteria may be 0.5-2.0 or 0.6-1.67.

Some analysis techniques described in the next section require the numerator and denominator to be positive, so it is assumed that $\mu_p < \mu_r$. If outcomes are positive but such that a smaller response is better, for instance time to pain relief, then $\theta_p$ can be formulated as $\frac{(\mu_p - \mu_r)}{\mu_p \mu_r}$. If outcomes are negative (i.e., change from baseline), then their sign can be reversed.

### 8.2.1 Univariate Techniques

There are two methods which can be used to obtain a confidence interval for $\theta_{pk}$. Since the numerator and denominator of the ratio are on the additive scale, $r_{pk}$ can not be formed by log-transforming the data. Confidence intervals for ratios are obtainable by Fieller’s formula or by Taylor series approximation, as described in Section 7.4.

For Fieller’s formula, the numerator and denominator of $r_{pk}$, as well as their variances and covariance are applied to equation (3.14) to form a confidence interval for $\theta_{pk}$. The $t$ distribution in the equation has $(n_r + n_r + n_p - 3)$ degrees of freedom. The differences of means may be positive or negative, but they must be large relative to the variance (i.e. $\geq 2$ standard errors from zero). The differences of means and their covariance matrix are obtained as follows: If the vector of treatment means is $F = [\bar{y}_t \; \bar{y}_r \; \bar{y}_p]'$, and the variance of the means is a matrix $V_F$, then the active-placebo differences are obtained by $D=A_1F$, $V_D=A_1V_FA_1'$, where $A_1 = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$.

An approximate confidence interval for $\theta_{pk}$ based on Taylor series is obtained by starting with $D$ and $V_D$. The ratio estimate is formed by taking the log and then subtracting the reference from the test treatment mean: $R=A_2\log(D)$, $V_R=A_2\text{Diag}(D)^{-1}V_D\text{Diag}(D)^{-1}A_2'$, where $A_2 = [1 \; -1]'$. Since the numerator and denominator in equation (8.2) are log-transformed, both must be positive in order
to apply Taylor series. The $100(1-2\alpha)\%$ CI for $\theta_{pk}$ is obtained by $\exp(R \pm z_{1-\alpha}\sqrt{V_R})$.

If there are multiple responses, then Taylor series methods can be used to obtain confidence intervals for the average of the ratios from each response. If $R$ and $V_R$ are expanded so that they correspond to the log-differences and their covariance for two responses, then a direct average is obtained by $A = A_3 R$

$V_A = A_3 R A_3'$, where $A_3 = [0.5 \ 0.5]$. A weighted average where the weights are determined by the inverse of the variance is obtained by weighted least squares. A simple $X$ matrix $= [1 \ 1]'$ is applied to $R$ to achieve the weighted average $b$. Homogeneity of the ratios is evaluated by a $\chi^2$ test of the contrast $CR$, where $C=[1 \ -1]$. As discussed in Chapter 7, although direct averaging may be applicable when log ratios are heterogeneous, no form of averaging is recommended unless there is acceptable homogeneity. Taylor series, Fieller's formula, and WLS methods work best if each treatment group has a reasonable sample size (i.e. $n_g=30$).

Treatment means in equation (8.2) can be adjusted for categorical or continuous covariates before estimating $\theta_{pk}$. Categorical covariates can be adjusted without a model by generalizing the vector of treatment means $F$ into a vector with a mean for each treatment $\times$ category group. Adjusted means (means that would apply if the category levels were equally distributed in each treatment group) are obtained by $F_A = A_4 F$, where $A_4$ takes the simple average of the category group means for each treatment group. Fieller's formula and Taylor series methods described above are then applied to form confidence intervals based on $F_A$. When a covariate is adjusted for, an additional degree of freedom is removed from the $t$ distribution in Fieller's formula.

Adjustment for a continuous covariate requires a model. For response $k$, either continuous or categorical covariates can be adjusted with an ANOVA
model which has an intercept, the covariate, and indicators for each active treatment. Parameters corresponding to the treatments are estimates of covariate adjusted active-placebo differences of treatment means. Since ANOVA does not account for the correlation between multiple responses, a WLS model-based average or Taylor series direct average of $\theta_p$ across responses is not possible. When there are multiple responses, models which take account of the correlation between outcomes may be more advantageous due to their ability to form global confidence intervals and to produce more efficient univariate estimates of $\theta_{pk}$.

### 8.2.2 Multivariate Techniques

As demonstrated in Chapter 7, ratios and corresponding standard errors from multiple responses can often be estimated more efficiently by a multivariate technique which takes advantage of the correlation between responses. Weighted least squares, survey regression, and mixed linear models were described in Section 7.2. Application of these methods in this setting is similar to applications described in Section 7.4 for normally distributed data. In this section, the three methods are applied to obtain treatment means adjusted for either a categorical or a continuous response. Once active-placebo differences of adjusted means are obtained, each of the techniques to obtain univariate or global confidence intervals for $\theta_p$ described in Section 8.2.1 are applicable.

#### Categorical Covariate

Since weighted least squares is applied to group means instead of individual observations, its application is a little different than survey regression or mixed models. Assume that there are three treatments, two responses, and one dichotomous covariate per response. The $(12 \times 1)$ vector of response means is a concatenation of the vector for each each category $c$ and response $k$ $F_{ck} = [y_{tck}$
\( \bar{y}_{rck} \bar{y}_{pck}' \), sorted first by category and then response. The weighted least squares method for categorical adjustment is to fit an \( X \) matrix with an intercept and indicators for the active treatments and covariate for each response. For one response, the applicable design matrix is:

\[
\begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix}
\]  

(8.4)

For two responses, the complete design matrix is a \((12 \times 8)\) block diagonal with the above matrix as the blocks. The \( b \) parameters for the active treatments are estimates of the difference between active treatment and placebo adjusted for the categorical covariate. These parameters can then be applied to Taylor series and Fieller's formula methods for obtaining confidence intervals for \( \theta_{pk} \), and to Taylor series and WLS methods for obtaining confidence intervals for the average of \( \theta_p \) across the responses.

Survey regression and mixed linear models apply the same model as weighted least squares to obtain estimates of category-adjusted differences of means, but since these methods operate at the observation level, the design matrix for WLS is the essence matrix of that applied to mixed models and survey regression. In order to achieve the most general results, the mixed model covariance matrix for each treatment \( \times \) category subgroup is left unstructured. Smoothing could be achieved by imposing equality of covariances across treatment groups or covariate groups or both.

**Continuous Covariate**

As described in Chapter 7, weighted least squares adjusts for continuous covariates by including the covariate means for each treatment \( \times \) response as additional response means in the vector \( F \). Covariate adjustment is achieved by a
model which fits an intercept and indicators for active treatments to each response, and has columns for each covariate which produce a weighted average of treatment group covariate means. For each response, the applicable design matrix is:

\[
\begin{bmatrix}
1 & 1 \\
1 & 1 \\
1 & 1 \\
\end{bmatrix}
\]

(8.5)

For two responses, the design matrix is a \( (12 \times 8) \) block diagonal with the above matrix as the blocks. The treatment parameters are estimates of the difference between active treatment and placebo adjusted for the continuous covariate.

For survey regression and mixed models, adjustment of a continuous covariate is obtained by fitting an \( X \) matrix exactly like that used for categorical covariates, except the columns for the covariates have values of the covariate for each observation instead of a categorical indicator. As explained in Chapter 7, covariates can apply to only one response by taking the value zero for all observations of the other response, or to both responses at once by having one column in the \( X \) matrix for the covariate. The resulting estimates of covariate adjusted active - placebo means can be applied to confidence interval methods described in Section 8.2.1.

The methods described in this section can be generalized to settings with any combination of multiple categorical or continuous covariates, although WLS methods may be limited due to subgroup sample size requirements.

### 8.2.3 Example

All of the methods described in the above sections are applied to the parallel clinical trial comparing dose levels of ramipril and HCTZ, described in Section 1.4.2, and used as an example in Section 7.4.3. When confidence intervals for \( \theta_p \), comparing treatments 2 and 5 in Figure 1.3 (2.5 mg ramipril vs. 12.5 mg HCTZ)
were formed, the active treatment - placebo differences of means for change in diastolic blood pressure were very small. The differences of means (around 3.0 for each treatment) were less than 2 standard errors from zero (SE=1.9), and Fieller’s formula did not produce meaningful confidence intervals. Therefore, two other treatments are compared in this example: treatment 8 (10 mg ramipril and 12.5 mg HCTZ) is considered the test treatment, and treatment 11 (5 mg ramipril and 25 mg HCTZ) is labeled the reference. The placebo group is treatment 1. Sample size per group is 43, 44, and 42 participants respectively.

Tables 8.1 and 8.2 provide 90% Taylor series and Fieller’s formula confidence intervals for $\theta_p$ for supine diastolic and systolic blood pressure change from baseline. The log of the ratio $\log(r_{ph})$ and its standard error are also presented for Taylor series. Results are presented for unadjusted means, for means adjusted for the categorical covariate of diastolic blood pressure at baseline $< 106$ vs $\geq 106$, and for means adjusted for the continuous covariate of baseline diastolic or systolic blood pressure respectively. Adjustment techniques from ANOVA, WLS, mixed linear models and survey regression are presented.

The treatment groups are not balanced with respect to the categorical covariate based on diastolic blood pressure. The cutpoint was defined in Chapter 7 such that participants in treatment groups 2 and 5 were equally split, but in each of these three treatments there are substantially more subjects in the $\geq 106$ category. Results from category adjustment in weighted least squares should be viewed cautiously in light of the small sample sizes in some subgroups.

In Chapter 7 it was demonstrated that mixed models with unstructured covariance matrices produce very similar results to WLS. This is again apparent for categorical covariate adjustment. Results from the two methods are different for continuous covariate adjustment because of the different manner in which adjustment is applied. Survey regression produces the same estimates as ANOVA
adjustment, but generally produces a somewhat smaller standard error. Fieller's formula confidence intervals are always a bit wider than those from Taylor series.

For categorical covariate adjustment, the direct averaging method results in lower variance than ANOVA adjustment in this example, although as expected all forms of univariate adjustment decrease the standard error from the unadjusted method to some degree. Multivariate methods for categorical covariate adjustment have basically no decrease in variance for diastolic pressure, and only mild improvement for systolic.

For both responses, continuous covariate adjustment from WLS produces a standard error lower than either mixed models or survey regression, although for systolic blood pressure this difference is less pronounced. Multivariate standard errors are much lower than the SE from ANOVA for systolic, while apparently little is gained by multivariate analysis for diastolic pressure.

If equivalence were accepted when the 90% confidence interval is contained within 0.8-1.25, then equivalence of treatments 8 and 11 is rejected for both systolic and diastolic blood pressure. However, if a more lenient criterion such as (0.75-1.33) or (0.7-1.43) were applied, then depending on the method applied, equivalence would be accepted for systolic blood pressure. Diastolic equivalence would be accepted if the criterion was (0.6-1.667).

Table 8.3 presents confidence intervals for averages of systolic and diastolic \( \theta_p \) for both WLS model-based weighted-average and Taylor series direct-average techniques. Standard errors from these methods are improved over standard errors for systolic and diastolic evaluated separately. For weighted least squares, homogeneity of the responses was accepted for all adjustment methods (\( p=0.3 \) or larger). Model-based averaging tended to result in estimates with smaller standard errors than direct averaging. Global equivalence of treatments 8 and 11 can be concluded based on a criterion of (0.7-1.43).
Table 8.1

Confidence Intervals for Placebo-Controlled Equivalence

For Diastolic Blood Pressure

from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>Taylor Series</th>
<th>Fieller's Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log(r_p)</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.135</td>
<td>0.198</td>
</tr>
<tr>
<td>Direct Category Adj.</td>
<td>0.138</td>
<td>0.184</td>
</tr>
<tr>
<td>ANOVA Cat. Adj.</td>
<td>0.150</td>
<td>0.188</td>
</tr>
<tr>
<td>ANOVA Continuous Adj.</td>
<td>0.135</td>
<td>0.194</td>
</tr>
<tr>
<td>WLS Cat. Adj.</td>
<td>0.169</td>
<td>0.192</td>
</tr>
<tr>
<td>Mixed Model Cat. Adj.</td>
<td>0.169</td>
<td>0.190</td>
</tr>
<tr>
<td>Survey Reg. Cat. Adj.</td>
<td>0.150</td>
<td>0.184</td>
</tr>
<tr>
<td>WLS Cont. Adj.</td>
<td>0.136</td>
<td>0.188</td>
</tr>
<tr>
<td>Mixed Model Cont. Adj.</td>
<td>0.135</td>
<td>0.197</td>
</tr>
<tr>
<td>Survey Reg. Cont. Adj.</td>
<td>0.135</td>
<td>0.195</td>
</tr>
</tbody>
</table>

Table 8.2

Confidence Intervals for Placebo-Controlled Equivalence

For Systolic Blood Pressure

from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>Taylor Series</th>
<th>Fieller's Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log(r_p)</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>-0.022</td>
<td>0.183</td>
</tr>
<tr>
<td>Direct Category Adj.</td>
<td>-0.012</td>
<td>0.170</td>
</tr>
<tr>
<td>ANOVA Cat. Adj.</td>
<td>-0.004</td>
<td>0.176</td>
</tr>
<tr>
<td>ANOVA Continuous Adj.</td>
<td>0.047</td>
<td>0.167</td>
</tr>
<tr>
<td>WLS Cat. Adj.</td>
<td>0.004</td>
<td>0.169</td>
</tr>
<tr>
<td>Mixed Model Cat. Adj.</td>
<td>0.004</td>
<td>0.168</td>
</tr>
<tr>
<td>Survey Reg. Cat. Adj.</td>
<td>-0.004</td>
<td>0.172</td>
</tr>
<tr>
<td>WLS Cont. Adj.</td>
<td>0.071</td>
<td>0.148</td>
</tr>
<tr>
<td>Mixed Model Cont. Adj.</td>
<td>0.054</td>
<td>0.151</td>
</tr>
<tr>
<td>Survey Reg. Cont. Adj.</td>
<td>0.047</td>
<td>0.151</td>
</tr>
</tbody>
</table>
Table 8.3
Confidence Intervals for Global Placebo-Controlled Equivalence

<table>
<thead>
<tr>
<th>Method</th>
<th>Model-based Average</th>
<th>Direct Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log($r_p$)</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.039</td>
<td>0.172</td>
</tr>
<tr>
<td>Direct Cat. Adj.</td>
<td>0.047</td>
<td>0.158</td>
</tr>
<tr>
<td>WLS Cat. Adj.</td>
<td>0.060</td>
<td>0.160</td>
</tr>
<tr>
<td>MM Cat. Adj.</td>
<td>0.060</td>
<td>0.159</td>
</tr>
<tr>
<td>Survey Reg. Cat</td>
<td>0.060</td>
<td>0.160</td>
</tr>
<tr>
<td>WLS Cont. Adj.</td>
<td>0.074</td>
<td>0.148</td>
</tr>
<tr>
<td>MM Cont. Adj.</td>
<td>0.056</td>
<td>0.151</td>
</tr>
<tr>
<td>Survey Reg Cont</td>
<td>0.052</td>
<td>0.151</td>
</tr>
</tbody>
</table>

8.3 Parallel Trials with Dichotomous Responses

For dichotomous responses, $\theta_{pk}$ is formed in terms of differences of positive response rates:

$$\theta_{pk} = \frac{(\pi_{tk} - \pi_{pk})}{(\pi_{rk} - \pi_{pk})} \quad (8.6)$$

$\theta_{pk}$ is estimated in a sample by

$$r_{pk} = \frac{(p_{tk} - p_{pk})}{(p_{rk} - p_{pk})} \quad (8.7)$$

If comparisons with placebo were based on ratios or odds ratios between each active treatment and placebo, then placebo would cancel out leaving the equivalence definitions for ratios and odds ratios described in Sections 5.3 and 5.4.

All of the discussion in Section 8.2 is directly applicable to dichotomous responses. Equivalence is concluded if a 100(1-2\alpha)% confidence interval for $\theta_{pk}$ is completely contained within pre-specified limits $\Theta_1$ and $\Theta_2$. Confidence intervals for $\theta_{pk}$ can be formed by Taylor series or Fieller's formula exactly as described in Section 8.2.1. Taylor series and weighted least squares can be used to obtain direct and weighted averages of ratio estimates from multiple responses.
The main difference between continuous and dichotomous univariate strategies is the vector of responses, \( F \). For two responses, \( F = [p_{r1} \ p_{r2} \ p_{p1} \ p_{p2}]' \) is a vector of positive response proportions instead of treatment means. \( F \) and its variance \( V_F \) are obtained as follows: With two dichotomous responses, there are four response categories per treatment group as depicted in Figure 5.1. Let \( P \) be the \((12 \times 1)\) vector of bivariate responses. The variance matrix assuming a multinomial distribution is estimated as \( V_P \), a block diagonal with each block equal to \([D_{Pg} - P_g P_g']/n_g\), where \( P_g \) is the vector of responses for treatment group \( g \). The marginal positive response proportions are obtained by \( F = AP \), \( V_F = AV_P A' \), where \( A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \).

Adjustment for categorical and continuous covariates is limited for dichotomous responses. Direct category adjustment of response proportions is obtainable by a simple average of category subgroups for each treatment, as described in Section 8.2.1. Analysis of variance is not applicable to dichotomies, however, and since the statistic \( \theta_{pk} \) is not based on odds ratios, logistic regression is also not applicable.

### 8.3.1 Multivariate Techniques

As with continuous responses, when more than one outcome is evaluated it is expected that multivariate analyses which take the correlation between responses into account will lead to smaller confidence intervals for each \( \theta_{pk} \). Multivariate techniques in this chapter are limited to weighted least squares, although generalized estimating equations are also applicable. As mentioned previously, increased efficiency from weighted least squares is not achieved when the \( X \) matrix is an identity, so a covariate must be adjusted for. This section considers the situation where there is one categorical covariate with two levels.

Categorical adjustment in WLS follows the same steps as described in Section
8.2.2. The X matrix as displayed in equation (8.4) is applied to form a weighted average of response proportions from the two covariate categories, with weight inversely proportional to the variance. The resulting covariate adjusted proportions are then applied to Taylor series and Fieller’s formula methods to obtain univariate confidence intervals for $\theta_{rk}$, and to Taylor series or an additional WLS model to form confidence interval for an average ratio across responses. Adjustment for a continuous covariate can also be applied.

8.3.2 Example

Consider an extension of the example described in Section 7.5.1, a study comparing two treatments for the healing of duodenal ulcers. An additional treatment group receiving placebo is evaluated for healing by endoscopy exam (outcome 1) and reported absence of ulcer pain (outcome 2). Figure 8.1 displays hypothetical data for responses from all treatment groups stratified by patient gender.

**Figure 8.1**

Ulcet Study Example of Bivariate Dichotomous Response

<table>
<thead>
<tr>
<th>Outcome 1,2</th>
<th>Y,Y</th>
<th>Y,N</th>
<th>N,Y</th>
<th>N,N</th>
<th>N,N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men Test</td>
<td>130</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Men Ref.</td>
<td>130</td>
<td>40</td>
<td>15</td>
<td>15</td>
<td>200</td>
</tr>
<tr>
<td>Men Placebo</td>
<td>70</td>
<td>30</td>
<td>20</td>
<td>80</td>
<td>200</td>
</tr>
<tr>
<td>Women Test</td>
<td>115</td>
<td>25</td>
<td>15</td>
<td>45</td>
<td>200</td>
</tr>
<tr>
<td>Women Ref.</td>
<td>120</td>
<td>20</td>
<td>10</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Women Plcb.</td>
<td>65</td>
<td>15</td>
<td>10</td>
<td>110</td>
<td>200</td>
</tr>
</tbody>
</table>
Tables 8.4 and 8.5 present 90% confidence intervals for $\theta_{pk}$ for endoscopy exam and self-reported absence of pain. The direct gender adjustment method results in identical estimates as the unadjusted because the genders are balanced in all treatment groups. A very slight reduction in standard error is achieved by the adjustment. Weighted least squares produces a slightly smaller variance than direct adjustment, although the improvement is fairly small for both responses. Fieller's formula confidence intervals are very similar to Taylor series, but tend to be slightly wider. Equivalence of treatments for each response is accepted for the criterion of $(0.8 \leq \theta_{pk} \leq 1.25)$.

Table 8.4

Confidence Intervals for Placebo-Controlled Equivalence
of Treatments for Endoscopy Exam
from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>Taylor Series</th>
<th>Fieller's Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\log(r_p)$</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>-0.080</td>
<td>0.097</td>
</tr>
<tr>
<td>Direct Gender Adjusted</td>
<td>-0.080</td>
<td>0.096</td>
</tr>
<tr>
<td>WLS Gender Adjusted</td>
<td>-0.082</td>
<td>0.093</td>
</tr>
</tbody>
</table>

Table 8.5

Confidence Intervals for Placebo-Controlled Equivalence
of Treatments for Reported Absence of Pain
from Several Univariate and Multivariate Strategies

<table>
<thead>
<tr>
<th>Method</th>
<th>Taylor Series</th>
<th>Fieller's Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\log(r_p)$</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.044</td>
<td>0.116</td>
</tr>
<tr>
<td>Direct Gender Adjusted</td>
<td>0.044</td>
<td>0.116</td>
</tr>
<tr>
<td>WLS Gender Adjusted</td>
<td>0.017</td>
<td>0.111</td>
</tr>
</tbody>
</table>
Table 8.6 presents confidence intervals for global equivalence of the test and reference treatments. Standard errors are decreased from Tables 8.4 and 8.5 due to averaging, and confidence intervals are tighter. Homogeneity was accepted in weighted least squares for all methods (p > .25).

<table>
<thead>
<tr>
<th>Method</th>
<th>Model-based Average</th>
<th>Direct Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\log(r_p)$</td>
<td>SE</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>-0.038</td>
<td>0.089</td>
</tr>
<tr>
<td>Direct Gen. Adj.</td>
<td>-0.038</td>
<td>0.088</td>
</tr>
<tr>
<td>WLS Gen. Adj.</td>
<td>-0.048</td>
<td>0.085</td>
</tr>
</tbody>
</table>

8.4 Conclusion

This chapter has evaluated a definition of equivalence of two treatments based on the ratio of mean difference from placebo. This method for evaluating equivalence is best applied to clinical trials when it is ethical and practical to apply a placebo, and when responses from administration of placebo are non-zero.

The primary advantage of $\theta_{ph}$ is that the equivalence of treatments is evaluated in terms of their improvement over placebo. Equivalence becomes increasingly harder to detect as the difference between the reference and placebo gets smaller. The pitfalls of the historical placebo assumption invoked by clinical equivalence studies with no placebo group are avoided. By subtracting the placebo mean response, a study is much more likely to find equivalence between two active treatments when there is a large active treatment effect. If there is little treatment effect, equivalence is difficult to prove. In fact, some methods for evaluating $\theta_{ph}$ should not be applied unless the mean difference between reference and placebo is at least twice the standard error of that difference.
Studies with test, reference, and placebo groups may wish to show equivalence between the active treatments via $\theta_{pk}$, as well as demonstrate a significant difference between the test and placebo treatments. Power required to accept placebo-controlled equivalence via $\theta_{pk}$ is a topic for future research.

Several univariate and multivariate analysis techniques for obtaining confidence intervals for $\theta_{pk}$ were evaluated in this chapter for parallel designs with normally distributed or dichotomous responses. For the bivariate response examples evaluated, slight decreases in variance were obtained by multivariate models which take account of the correlation between responses. The level of improvement from multivariate modeling may differ for each application, and is expected to improve as the number of responses increases.
Chapter 9
Summary and Future Research

9.1 Summary

Often the purpose of clinical trials is to evaluate whether two treatments are equivalent to one another. This arises in two situations: a bioequivalence trial where a new formulation of a drug is tested for equivalence of average drug blood levels to the reference formulation; and clinical equivalence trials where a new treatment is compared for equivalence to the reference for clinical manifestations such as pain reduction or healing rates. For continuous responses, equivalence is generally accepted if a 90% confidence interval for the ratio of the mean response for test and reference treatments is contained within a pre-specified limit, usually 0.8-1.25. When multiple response criteria are evaluated, investigators might conclude overall treatment equivalence if equivalence is accepted for all outcomes, based on a set of univariate confidence intervals. In this situation, there is no concern for increased type-1 error. However, the power for finding equivalence of all endpoints decreases as the number of outcomes increases.

Power of the multiple 90% confidence interval method was evaluated for bivariate log-normal data from parallel and cross-over designs based on numerical integration and simulation. Although the multiple confidence interval strategy ignores the correlation between responses, bivariate power is affected by the correlation. For fixed variance and sample size, the bivariate power increases with correlation for responses with ratios on the same side of 1.0, and decreases for responses with opposite treatment effects. The effect of correlation is more pronounced when variance is large relative to sample size and treatment ratios are
farther from 1.0 (i.e., univariate power is moderately low). The two univariate confidence interval strategy can be conservative, resulting in a bivariate type-1 error less than $\alpha$. Graphical representation of bivariate power was evaluated.

When responses are dichotomous, equivalence is often evaluated with a one-sided confidence interval. Equivalence can be defined in terms of differences of proportions, ratios of proportions, or odds ratios. Univariate and bivariate power were evaluated for each definition for parallel designs with two dichotomous responses. In large samples, test statistics are based on the normal approximation to the binomial, and power is evaluated via numerical integration. For small samples, test statistics are modified to counteract the anti-conservativeness of the normal approximation and power is evaluated by simulation.

Correlation affects bivariate power of dichotomous responses in similar ways as continuous data: the increase or decrease depends on the directions of the treatment effects, and correlation plays a bigger role when each univariate power is moderate to low. For differences and ratios of proportions, the effect of correlation is more pronounced as response proportions are closer to 0.5 (i.e., as variance increases). Odds ratios have larger variation for the same sample size, so the effect of correlation is more pronounced for odds ratios. Odds ratio variation decreases (and hence power increases and effect of correlation decreases) as response proportions get closer to 0.5. As found for log-normal data, the bivariate test can have type-1 error less than $\alpha$.

Advantages and disadvantages of each equivalence definition were evaluated. The test statistic for ratios of proportions is not symmetrical, and should always be applied in a conservative manner such that the 1-sided equivalence limit for the treatment ratio is greater than 1.0. Power for this statistic is fairly low. Of the three definitions, power is higher for methods based on differences of proportions, especially when response proportions are near 1.0. When response
proportions are near 0.5, power for differences and odds ratios is fairly similar. Equivalence based on differences of proportions requires different equivalence limits for varying levels of reference response proportions. Limits for odds ratios can be applied across a wide range of proportions, leading to a more stable definition. Limits for ratios of proportions are similar to those used for log-normal data.

Some authors suggest that equivalence is better evaluated in terms of the ratio between an individual’s response for two treatments, instead of a ratio of average responses. This is a more stringent criterion, and insures that individuals switching from reference to test treatment may expect a similar response. A method of evaluating individual equivalence was presented based on creating ANOVA prediction intervals for an individual’s ratio of responses from cross-over designs with log-normal data. This method assumes no period and no carry-over effect. Appropriate equivalence limits for prediction intervals were suggested based on the reference variance.

Univariate and bivariate power for the test of individual equivalence were evaluated by simulation. Since the test is stringent, power is affected much more by the variance than by changing sample size. For a fixed equivalence limit and sample size, the window between power of zero and 1.0 is fairly small for changing variance. Consequently, the effect of correlation on bivariate power was not strong for most scenarios evaluated, but is apparent when each univariate power is moderate to low. An anomaly found for this test is that power can be substantially larger than \( \alpha \) near the border between the null and alternative hypotheses. No solution for this problem was suggested.

The multiple univariate confidence interval strategy does not account for the correlation between responses. It was proposed that univariate confidence intervals based on multivariate models which take account of the correlation may
have decreased variance, and hence better power. Multivariate modeling was evaluated through weighted least squares, mixed linear models, and survey regression. Confidence intervals for equivalence based on these models were examined for log-normal data from parallel and cross-over designs, and for normally distributed data using these methods combined with Fieller's formula or Taylor series approximation. Survey logistic regression and weighted least squares were applied to odds ratios for dichotomous responses. Efficiency in estimation from weighted least squares and mixed linear models is achieved only if a covariate is adjusted for. Multivariate strategies were evaluated for examples, and in general achieved slight decreases in variance compared to univariate analyses.

One concern with clinical equivalence is the lack of a placebo to assure efficacy of the active treatments. A definition of equivalence was evaluated based on ratios of mean differences from placebo. Univariate and multivariate analysis strategies were evaluated for parallel designs with normally distributed and dichotomous responses. This criterion is more stringent than standard equivalence based on treatment mean ratios. Equivalence becomes harder to prove as the reference becomes more similar to placebo. Multivariate strategies applied to examples again achieved slight decreases in variance compared to univariate analyses.

9.2 Future Research

Bivariate power was evaluated for equivalence based on confidence intervals for ratios of treatment geometric means for log-normal data. Confidence intervals for ratios can also be evaluated using Fieller's formula, which does not assume a log-normal distribution. Univariate and bivariate power of Fieller's formula methods similar to those applied in Chapter 4 are of interest for parallel and cross-over trials.
For dichotomous responses, the ratio of treatment response proportions was evaluated based on the difference $\pi_t - \theta \pi_r$. Undesirable properties were noted because this test is not symmetrical: converting response proportions from positive to negative events results in different conclusions. Alternate methods for obtaining ratios of proportions based on Fieller's formula or by Taylor series approximation are of interest. Univariate and bivariate power of these methods as evaluated by simulation could be compared with results presented in Chapter 5.

Additional definitions of individual equivalence could be examined in more detail. For instance, a less stringent definition could regard departure from equivalence as a fixed effect which can be interpretable as a treatment by subgroup interaction. In the face of an interaction, equivalence would be evaluated for subgroups separately. Evaluation of power under this scenario is of interest. Power for the prediction interval method was examined for situations where the test and reference have equal variance. Properties of power when the treatment group variances are different are of interest. The large type 1 error near the border of the null hypothesis should be corrected. Anderson and Hauck's TIER method for evaluating individual equivalence puts a confidence interval on an estimate of the population proportion that will respond within equivalence limits. Perhaps the proposed method based on prediction intervals requires an additional level of confidence applied to it.

It is of interest to evaluate power of multivariate methods in Chapter 7 through simulation. Applications of multivariate methods could also be extended to combinations of responses, for instance one dichotomous and one normally distributed response. In addition, it is of interest to compare power of the univariate and multivariate methods applied for placebo-controlled equivalence in Chapter 8 with the standard methods evaluated in Chapters 4 and 5.
References


Blackwelder, W. C. Proving the Null Hypothesis in Clinical Trials. Controlled Clinical Trials 1982 3: 345-353.


Dunnett, C. W., Gent, M. Significance Testing to Establish Equivalence Between Treatments, With Specific Reference to Data in the Form of 2x2 Tables. Biometrics 1977 33: 593-602.


Steinijans, V. W., Hauck, W. W., Diletti, E., Hauschke, D., Anderson, S. Effect of changing the bioequivalence range from (0.80, 1.20) to (0.80, 1.25) on the power and sample size. Int J Clin Pharmacol Ther Toxicol 1992 30(12): 571-575.


