

## **Abstract**

MUKHOPADHYAY, PRALAY. Exact Tests and Exact Confidence Intervals for the Ratio of Two Binomial Proportions. (Under the direction of Professor Roger L. Berger)

Testing for the ratio of binomial proportions (often called the relative risk) is quite common in clinical trials and epidemiology study or more generally in the pharmaceutical setting. Although this is an easy problem when we have large sample sizes, it becomes more challenging when sample sizes are small to moderate. In this type of situations asymptotic methods often lead to tests that are very liberal, i.e., have a very high Type I error. Hence one has to resort to exact methods. Although one can use Fisher's exact test if testing for unity relative risk, for the more general problem of testing for non-unity relative risk the only form of exact inference possible is by using exact unconditional tests. The standard exact unconditional test used for this problem is quite conservative, i.e., results in tests with very low power. We have proposed a test for this problem (based on the method suggested by Berger and Boos) which not only maintains the nominal size but is uniformly more powerful than the standard test (in most of the cases). A detailed comparison has been done between the two tests and various examples (from the pharmaceutical setting) have been used to compare the two methods.

Along with testing for the relative risk, researchers are also interested in obtaining

confidence intervals for this parameter. Again due to small sample sizes the asymptotic methods often result in intervals that have poor coverage. We compare the confidence intervals generated from inverting the standard exact test and the test that we are proposing. Since both these tests are exact they result in intervals that are guaranteed to maintain the nominal coverage. We show that the standard intervals are quite conservative and our intervals in general have shorter lengths and coverage probabilities closer to the nominal coverage.

Although exact tests are desirable, it is often hard to implement them in practice because of computational complexities. In the last Chapter we compare the performance of the two exact tests discussed earlier with an approximate test (based on the idea of Storer and Kim) and two Bayesian tests for the hypothesis of efficacy. The hypothesis of efficacy is a special case of the relative risk testing problem and is often used in vaccine studies. For this specific problem we see that the approximate and the Bayesian tests perform quite well in terms of maintaining a low Type I error and results in tests with high power. Also these tests have a nice practical appeal because of the ease with which they can be implemented.

**Exact Tests and Exact Confidence Intervals for the Ratio of Two  
Binomial Proportions**

by

**Pralay Mukhopadhyay**

A dissertation submitted in partial satisfaction of the  
requirements for the degree of  
Doctor of Philosophy

in

**STATISTICS**

in the

**GRADUATE SCHOOL**

at

**NC STATE UNIVERSITY**

**2003**

---

Professor Roger L. Berger  
Chair of Advisory Committee

---

Professor Sujit K. Ghosh

---

Professor Anastasios A. Tsiatis

---

Professor John Bishir

*To my parents, my brother and my sister in law.*

## Biography

Pralay Mukhopadhyay was born on September 30, 1976 in Calcutta, India. He received his B.Sc. with honors in Statistics, from the University of Calcutta, in 1998. After that he joined the graduate program in statistics at North Carolina State University. He received the M.S. degree in statistics in 2000 and continued in the Ph.D. program. During his graduate studies he worked as a student intern in SAS Institute and Quintiles. After receiving his doctoral degree he plans to work as a research biostatistician at Bristol Myers Squibb Company in Connecticut.

## Acknowledgements

I had a wonderful experience during my three years as a Ph.D. student at the department of Statistics in North Carolina State University. And I would like to take this opportunity to express my gratitude to a lot of people. Without their help none of these would have been possible.

First I would like to thank my adviser Dr. Roger Berger. I have been very fortunate to have him as my adviser. Not only is he a brilliant statistician but also a very understanding and caring person. There has never been a time he did not help me out with my questions (however stupid they may be) or encourage me, when things were not working as expected. Words can hardly express my gratitude for him.

I would like to thank Dr. Sujit Ghosh for his help and advice in parts of the thesis. His comments and expertise (especially in the third Chapter of the thesis) has been invaluable. He has been very encouraging and supportive and has always taken the extra step to help me with my questions. I would like to thank him deeply for all his help.

I would like to thank Dr. Anastasios Tsiatis for serving in my committee. I think he is one of the best teachers and researcher in our department and it was my privilege to have him as a committee member. I am grateful to him for all his help.

I would like to thank Dr. John Bishir from the Mathematics department for

serving as a committee member. His thoughts and suggestions have been very helpful. I am grateful that he took the time to serve in my committee.

I would like to thank the entire faculty and staff of the department of statistics at North Carolina State University. I have always felt at home in this department and my entire experience was very fulfilling. I want to take this opportunity to thank the following persons in particular. First in my list will be Dr. Sastry Pantula. He has always been a source of encouragement for me throughout the entire duration of my study. He has never ceased to amaze me by the way he keeps track of all the students in our department. I would like to thank Dr. Leonard Stefanski, Dr. Bibhuti Bhattacharya, and Dr. David Dickey for being such great teachers. I would also like to thank Dr. William Swallow for being the wonderful person he is and for always keeping his door open for the students. And last but not the least I would like to thank Janice Gaddy for being so efficient and caring.

I would like to thank Lovely Goyal for being such a great friend. I really admire her for coping up with all my idiosyncrasies. I would like to thank my friend Jimmy Doi for always being so helpful, especially with all my "out of the world" computing questions, known only to him and Terry.

And now it is time for me to thank the people who are dearest to me. They are my dad, Dr. Parimal Mukhopadhyay, my mom, Mrs. Manju Mukhopadhyay, my brother, Dr. Pabak Mukhopadhyay, and my sister in law Mrs. Jayita Mukhopadhyay.

Without them nothing in the world would have been possible for me let alone my education. They have been my strength and my motivation to move on in the most difficult of times. I want to thank God for letting me be a part of this family. May he continue to bless my family and me.



# Contents

|  |           |
|--|-----------|
| <b>List of Figures</b>   | <b>ix</b> |
| <b>List of Tables</b>  | <b>xi</b> |
| <b>1 Exact Unconditional Tests for the Ratio of Two Binomial Proportions</b>       | <b>1</b>  |
| 1.1 Introduction . . . . .   | 1         |
| 1.2 Statistical formulation of the problem . . . . .                               | 5         |
| 1.3 Data Sets . . . . .  | 12        |
| 1.3.1 Animal Toxicology Data . . . . .   | 12        |
| 1.3.2 Childhood Nephroblastoma Data . . . . .                                      | 13        |
| 1.3.3 Influenza Vaccine Data . . . . .   | 14        |
| 1.4 Exact Unconditional Test . . . . .   | 16        |
| 1.5 Confidence Interval p value Test . . . . .                                     | 20        |
| 1.5.1 Validity of the Confidence Interval p value Test . . . . .                   | 22        |
| 1.5.2 Construction of $100(1 - \beta)\%$ confidence set for $(P_1, P_2)$ . . . . . | 23        |
| 1.6 Rejection Region, Size and Power . . . . .                                     | 29        |
| 1.7 Results . . . . .  | 34        |
| 1.8 Data Sets Revisited . . . . .  | 59        |
| <b>2 Exact Confidence Intervals for the Ratio of Two Binomial Proportions</b>      | <b>71</b> |
| 2.1 Introduction . . . . .   | 71        |
| 2.2 Statistical formulation of the problem . . . . .                               | 76        |
| 2.3 Computing the exact confidence intervals . . . . .                             | 82        |
| 2.4 Results . . . . .  | 85        |
| 2.4.1 Length of the interval . . . . .   | 86        |
| 2.4.2 Coverage Probabilities . . . . .   | 95        |
| 2.5 Data Sets . . . . .  | 109       |

|          |   |            |
|----------|---|------------|
| <b>3</b> | <b>Exact, Approximate, and Bayesian Tests for the Hypothesis of Ef-</b> |            |
|          | <b>ficacy</b>   | <b>118</b> |
| 3.1      | Introduction . . . . .  | 118        |
| 3.2      | Exact, Approximate, and Bayesian tests . . . . .                        | 122        |
| 3.2.1    | Exact and Approximate tests . . . . .                                   | 123        |
| 3.2.2    | Bayesian tests . . . . .  | 127        |
| 3.3      | Size and Power . . . . .  | 136        |
| 3.4      | Results . . . . .   | 138        |
| 3.5      | Conclusion . . . . .  | 154        |
|          | <b>Bibliography</b>   | <b>165</b> |

# List of Figures

|     |   |     |
|-----|---|-----|
| 1.1 | The shaded area represents the $(1-\beta)$ confidence region $C_\beta$ . The diagonal is the line segment $P_1/P_2 = R$ . . . . .   | 26  |
| 1.2 | p value function plot for animal toxicology data: $N_1 = 350, x_1 = 212, N_2 = 77, x_2 = 37$ . . . . .  | 62  |
| 1.3 | Size function plots for $p_{sup}$ (Solid line), $p_\beta$ (Dotted line), and Asymptotic test (Broken line) when $(N_1, N_2) = (350, 77)$ . . . . .  | 63  |
| 1.4 | p value function plot for childhood nephroblastoma data: $N_1 = 88, x_1 = 83, N_2 = 76, x_2 = 69$ . . . . .   | 68  |
| 1.5 | Size function plots for $p_{sup}$ (Solid line), $p_\beta$ (Dotted line), and Asymptotic test (Broken line) when $(N_1, N_2) = (88, 76)$ . . . . .   | 69  |
| 1.6 | p value function plot for influenza vaccine data: $N_1 = 15, x_1 = 7, N_2 = 15, x_2 = 12$ . . . . .   | 70  |
| 1.7 | Size function plots for $p_{sup}$ (Solid line), $p_\beta$ (Dotted line), and Asymptotic test (Broken line) when $(N_1, N_2) = (15, 15)$ . Here the size functions for $p_{sup}$ and $p_\beta$ are identical . . . . . | 70  |
| 2.1 | Coverage probability plot for $R = 0.9$ and 2 when $(N_1, N_2) = (15, 15)$ : $p_{sup}$ =Solid Line, $p_\beta$ =Dotted Line. Note that for Ratio = 2 the coverage for both the intervals are identical . . . . .       | 100 |
| 2.2 | Coverage probability plot for $R = 3$ and 10 when $(N_1, N_2) = (15, 15)$ : $p_{sup}$ =Solid Line, $p_\beta$ =Dotted Line . . . . .   | 101 |
| 2.3 | Coverage probability plot when $(N_1, N_2) = (15, 15)$ , $P_2 = 0.3$ , Ratio= $P_1/P_2$ : $p_{sup}$ =Solid Line, $p_\beta$ =Dotted Line . . . . .   | 102 |
| 2.4 | Coverage probability plot for $R = 0.9$ and 2 when $(N_1, N_2) = (20, 20)$ : $p_{sup}$ =Solid Line, $p_\beta$ =Dotted Line. Note that for Ratio = 2 the coverage for both the intervals are identical . . . . .       | 103 |
| 2.5 | Coverage probability plot for $R = 3$ and 10 when $(N_1, N_2) = (20, 20)$ : $p_{sup}$ =Solid Line, $p_\beta$ =Dotted Line . . . . .   | 104 |
| 2.6 | Coverage probability plot when $(N_1, N_2) = (20, 20)$ , $P_2 = 0.3$ , Ratio= $P_1/P_2$ : $p_{sup}$ =Solid Line, $p_\beta$ =Dotted Line . . . . .   | 105 |

|      |  |     |
|------|--|-----|
| 2.7  | Coverage probability plot for $R = 0.9$ and 2 when $(N_1, N_2)=(10,20)$ :<br>$p_{sup}$ =Solid Line, $p_{\beta}$ =Dotted Line . . . . .   | 106 |
| 2.8  | Coverage probability plot for $R = 3$ and 10 when $(N_1, N_2)=(10,20)$ :<br>$p_{sup}$ =Solid Line, $p_{\beta}$ =Dotted Line. Note that for Ratio = 3 and 10 the<br>coverage for both the intervals are identical . . . . . | 107 |
| 2.9  | Coverage probability plot when $(N_1, N_2)=(10,20)$ , $P_2 = 0.3$ , Ratio= $P_1/P_2$ :<br>$p_{sup}$ =Solid Line, $p_{\beta}$ =Dotted Line . . . . .  | 108 |
| 2.10 | Coverage probability plot when $(N_1, N_2)=(350,77)$ : $p_{sup}$ =Solid Line,<br>$p_{\beta}$ =Dotted Line, Asymptotic test=Broken line . . . . .   | 113 |
| 2.11 | Coverage probability plot when $(N_1, N_2)=(88,76)$ : $p_{sup}$ =Solid Line,<br>$p_{\beta}$ =Dotted Line, Asymptotic test=Broken line . . . . .  | 114 |
| 2.12 | Coverage probability plot when $(N_1, N_2)=(15,15)$ : $p_{sup}$ =Solid Line,<br>$p_{\beta}$ =Dotted Line, Asymptotic test=Broken line . . . . .  | 115 |
| 3.1  | Size plots when $(N_1, N_2)=(20,20)$ . . . . .   | 156 |
| 3.2  | Size plots when $(N_1, N_2)=(50,50)$ . . . . .   | 156 |
| 3.3  | Size plots when $(N_1, N_2)=(75,75)$ . . . . .   | 157 |
| 3.4  | Size plots when $(N_1, N_2)=(100,100)$ . . . . .   | 157 |
| 3.5  | Size plots when $(N_1, N_2)=(20,40)$ . The size function for $p_{sup}$ and $p_{\beta}$<br>are identical in this case . . . . .   | 158 |
| 3.6  | Size plots when $(N_1, N_2)=(40,20)$ . The size function for $p_{sup}$ and $p_{\beta}$<br>are identical in this case . . . . .   | 158 |
| 3.7  | Size plots when $(N_1, N_2)=(50,100)$ . . . . .  | 159 |
| 3.8  | Size plots when $(N_1, N_2)=(100,50)$ . . . . .  | 159 |
| 3.9  | Size plot for the $LS$ test when $(N_1, N_2)=(200,50)$ . . . . .   | 160 |
| 3.10 | Size plot for the $LS$ test when $(N_1, N_2)=(50,200)$ . . . . .   | 160 |
| 3.11 | Size plot for the $LS$ test when $(N_1, N_2)=(200,200)$ . . . . .  | 161 |
| 3.12 | Size plot for the $LS$ test when $(N_1, N_2)=(1000,50)$ . . . . .  | 162 |
| 3.13 | Size plot for the $LS$ test when $(N_1, N_2)=(50,1000)$ . . . . .  | 162 |
| 3.14 | Distribution of the 95th percentile of $W$ when $(N_1, N_2)=(1000,50)$ and<br>$(P_1, P_2)=(0.01,0.01/0.9)$ . . . . .   | 163 |
| 3.15 | Distribution of the 95th percentile of $W$ when $(N_1, N_2)=(1000,50)$ and<br>$(P_1, P_2)=(0.8,0.8/0.9)$ . . . . .   | 163 |
| 3.16 | Distribution of the 95th percentile of $W$ when $(N_1, N_2)=(50,1000)$ and<br>$(P_1, P_2)=(0.01,0.01/0.9)$ . . . . .   | 164 |
| 3.17 | Distribution of the 95th percentile of $W$ when $(N_1, N_2)=(50,1000)$ and<br>$(P_1, P_2)=(0.8,0.8/0.9)$ . . . . .   | 164 |

# List of Tables

|      |   |    |
|------|---|----|
| 1.1  | $H_0 : R \leq 0.1$ versus $H_1 : R > 0.1$ : $N_1 < N_2$   | 37 |
| 1.2  | $H_0 : R \geq 0.1$ versus $H_1 : R < 0.1$ : $N_1 < N_2$   | 37 |
| 1.3  | $H_0 : R \leq 0.33$ versus $H_1 : R > 0.33$ : $N_1 < N_2$ | 38 |
| 1.4  | $H_0 : R \geq 0.33$ versus $H_1 : R < 0.33$ : $N_1 < N_2$ | 38 |
| 1.5  | $H_0 : R \leq 0.5$ versus $H_1 : R > 0.5$ : $N_1 < N_2$   | 39 |
| 1.6  | $H_0 : R \geq 0.5$ versus $H_1 : R < 0.5$ : $N_1 < N_2$   | 39 |
| 1.7  | $H_0 : R \leq 0.77$ versus $H_1 : R > 0.77$ : $N_1 < N_2$ | 41 |
| 1.8  | $H_0 : R \geq 0.77$ versus $H_1 : R < 0.77$ : $N_1 < N_2$ | 41 |
| 1.9  | $H_0 : R \leq 0.9$ versus $H_1 : R > 0.9$ : $N_1 < N_2$   | 42 |
| 1.10 | $H_0 : R \geq 0.9$ versus $H_1 : R < 0.9$ : $N_1 < N_2$   | 42 |
| 1.11 | $H_0 : R \leq 1.11$ versus $H_1 : R > 1.11$ : $N_1 < N_2$ | 43 |
| 1.12 | $H_0 : R \geq 1.11$ versus $H_1 : R < 1.11$ : $N_1 < N_2$ | 43 |
| 1.13 | $H_0 : R \leq 1.29$ versus $H_1 : R > 1.29$ : $N_1 < N_2$ | 45 |
| 1.14 | $H_0 : R \geq 1.29$ versus $H_1 : R < 1.29$ : $N_1 < N_2$ | 45 |
| 1.15 | $H_0 : R \leq 2$ versus $H_1 : R > 2$ : $N_1 < N_2$       | 46 |
| 1.16 | $H_0 : R \geq 2$ versus $H_1 : R < 2$ : $N_1 < N_2$       | 46 |
| 1.17 | $H_0 : R \leq 3$ versus $H_1 : R > 3$ : $N_1 < N_2$       | 47 |
| 1.18 | $H_0 : R \geq 3$ versus $H_1 : R < 3$ : $N_1 < N_2$       | 47 |
| 1.19 | $H_0 : R \leq 10$ versus $H_1 : R > 10$ : $N_1 < N_2$     | 48 |
| 1.20 | $H_0 : R \geq 10$ versus $H_1 : R < 10$ : $N_1 < N_2$     | 48 |
| 1.21 | $H_0 : R \leq 0.1$ versus $H_1 : R > 0.1$ : $N_1 = N_2$   | 49 |
| 1.22 | $H_0 : R \geq 0.1$ versus $H_1 : R < 0.1$ : $N_1 = N_2$   | 49 |
| 1.23 | $H_0 : R \leq 0.33$ versus $H_1 : R > 0.33$ : $N_1 = N_2$ | 50 |
| 1.24 | $H_0 : R \geq 0.33$ versus $H_1 : R < 0.33$ : $N_1 = N_2$ | 50 |
| 1.25 | $H_0 : R \leq 0.5$ versus $H_1 : R > 0.5$ : $N_1 = N_2$   | 52 |
| 1.26 | $H_0 : R \geq 0.5$ versus $H_1 : R < 0.5$ : $N_1 = N_2$   | 52 |
| 1.27 | $H_0 : R \leq 0.77$ versus $H_1 : R > 0.77$ : $N_1 = N_2$ | 53 |
| 1.28 | $H_0 : R \geq 0.77$ versus $H_1 : R < 0.77$ : $N_1 = N_2$ | 53 |
| 1.29 | $H_0 : R \leq 0.9$ versus $H_1 : R > 0.9$ : $N_1 = N_2$   | 54 |
| 1.30 | $H_0 : R \geq 0.9$ versus $H_1 : R < 0.9$ : $N_1 = N_2$   | 54 |

|      |   |     |
|------|---|-----|
| 1.31 | Summary of Results. . . . .   | 56  |
| 1.32 | Table depicting overall performance of $p_\beta$ over $p_{sup}$ for all the cases. .  | 58  |
| 1.33 | $H_0 : R \leq 1$ versus $H_1 : R > 1$ . . . . .   | 60  |
| 1.34 | Power for $p_{sup}$ and $p_\beta$ for the animal toxicology data . . . . .  | 64  |
| 1.35 | $H_0 : R \geq 1.15$ versus $H_1 : R < 1.15$ . . . . .   | 65  |
| 1.36 | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ . . . . .   | 67  |
| 2.1  | Comparison of the mean lengths for the two confidence intervals . . .   | 87  |
| 2.2  | Table summarizing number of sample points for which $p_\beta$ intervals are shorter by 0.1, longer by 0.1, or have an absolute difference less than 0.1 compared to $p_{sup}$ intervals . . . . . | 92  |
| 2.3  | Comparison of Confidence Intervals for $p_{sup}$ and $p_\beta$ when $(N_1, N_2) = (15, 15)$<br>- Cases where $p_\beta$ has a shorter interval length . . . . .                                    | 93  |
| 2.4  | Comparison of Confidence Intervals for $p_{sup}$ and $p_\beta$ when $(N_1, N_2) = (15, 15)$<br>- Cases where $p_{sup}$ has a shorter interval length . . . . .                                    | 93  |
| 2.5  | Comparison of the mean lengths for the two confidence intervals after eliminating the sample points with small control group response . . .   | 94  |
| 2.6  | Exact and asymptotic intervals for the animal toxicology data . . . .   | 110 |
| 2.7  | Exact and asymptotic intervals for the childhood nephroblastoma data  | 112 |
| 2.8  | Exact and asymptotic intervals for the influenza vaccine data . . . . .   | 116 |
| 3.1  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Size for all the five tests when $N_1 = N_2$ . . . . .  | 140 |
| 3.2  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Type II error for specific choices of $(P_1, P_2)$ when $N_1 = N_2$ . . . . .   | 142 |
| 3.3  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Power for specific choices of $(P_1, P_2)$ when $N_1 = N_2$ . . . . .   | 143 |
| 3.4  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Total error for specific choices of $(P_1, P_2)$ when $N_1 = N_2$ . . . . .   | 144 |
| 3.5  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Size for all the five tests when $N_1 \neq N_2$ . . . . .   | 145 |
| 3.6  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Type II error for specific choices of $(P_1, P_2)$ when $N_1 \neq N_2$ . . . . .  | 147 |
| 3.7  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Power for specific choices of $(P_1, P_2)$ when $N_1 \neq N_2$ . . . . .  | 148 |
| 3.8  | $H_0 : \pi \leq 0.1$ versus $H_1 : \pi > 0.1$ : Total error for specific choices of $(P_1, P_2)$ when $N_1 \neq N_2$ . . . . .  | 149 |

# Chapter 1

## Exact Unconditional Tests for the Ratio of Two Binomial Proportions

### 1.1 Introduction

If a comparison is made between two groups where the outcome of interest is dichotomous (such as a success and a failure), the ratio of proportions of success (or failures) is often of interest. Thus if we consider two groups, A and B, and,  $P_1$  and  $P_2$  denote the true proportions of successes from each group after conducting the experiment, the ratio  $(P_1/P_2)$  is often the parameter of interest. In different applications such as in epidemiology this quantity is often called the relative risk. In the next few paragraphs we will discuss how in different situations researchers are interested in the hypothesis testing problem involving the relative risk parameter.

Assessment of an equivalence between two groups is often an important problem in the pharmaceutical industry. When the patent of a brand-name drug has expired, other manufacturers can manufacture the same drug and sell it at a much lower cost. They have to ensure that the chemical composition of the generic drug is same as the original drug. The FDA usually do not require a new drug application submission for this type of generic products. Rather they are satisfied if the generic drug company can demonstrate bio-equivalence between the original drug and their version. Since it is impossible to establish that the two treatments are exactly same, often a hypothesis of interest is to test whether the difference is negligible. In this example let us denote by  $P_1$  and  $P_2$  the true response (say, cure rates) probabilities for the existing drug and the generic drug. Then the hypothesis of interest might be to test

$$H_0 : R \geq R_0 \text{ versus } H_1 : R < R_0,$$

where  $R = (P_1/P_2)$ . Usually  $R_0$  is chosen to be strictly greater than 1. This is known as the test of non-inferiority for the ratio of proportions.

In clinical trials, when a new drug is compared against an existing drug (placebo), often a question of interest is whether the new drug performs better than the existing drug (placebo). Thus one would be interested in knowing whether the proportion of success for the new drug is greater than the existing drug (placebo) by some particular quantity. This quantity is usually determined by the clinicians. Let  $P_1$  and  $P_2$  denote the true cure rates for the new drug and the existing drug (placebo). Thus the



hypothesis of interest will be

$$H_0 : R \leq R_0 \text{ versus } H_1 : R > R_0,$$

where  $R = (P_1/P_2)$ . Again a value of  $R_0$  strictly greater than 1 is chosen to establish superiority. This is known as the test of superiority for the ratio of proportions.

In vaccine efficacy trials the problem of interest is often to test whether two vaccines offer equal protection against a certain disease. If a new vaccine is tested against placebo the outcome of interest would be to show the incidence of disease among the patients treated with the new vaccine is significantly lower than those treated with placebo. For example Fries et al. (1993) had conducted a viral challenge study to compare incidence rates of influenza among subjects who received a new influenza vaccine and those who received a placebo injection. Because of the variable incidence rates of disease in the two groups this problem is often represented in terms of the relative risk. Since the outcome is dichotomous for both the treatment and the control groups, it is appropriate to consider two independent binomial trials. Let  $P_1$  and  $P_2$  denote the true disease incidence rates in the vaccine group and the control group. Usually the vaccine efficacy parameter, denoted by  $\pi$ , is defined as (1 - relative risk for the disease between the vaccine and the placebo group), i.e.,  $\pi = 1 - P_1/P_2$ . An important hypothesis of interest is to test

$$H_0 : \pi \leq \pi_0 \text{ versus } H_1 : \pi > \pi_0,$$

where  $\pi_0$  denotes the minimal level of efficacy expected for the new vaccine. This value can either be chosen to be 0 which is typically the case when establishing therapeutic efficacy of a drug. In designing a vaccine efficacy trial one however chooses a non-zero value of  $\pi_0$  to establish superiority of the new vaccine and justify the risk of vaccinating healthy subjects. This is known as the test of efficacy. In terms of relative risk the hypothesis may be rewritten as

$$H_0 : R \geq 1 - \pi_0 \text{ versus } H_1 : R < 1 - \pi_0.$$

This test of efficacy has been considered by various authors such as Chan (1998) and Kang & Chen (2000). In this problem it is always assumed that the vaccine is not worse than placebo, i.e.,  $P_1 \leq P_2$ .

In general however people would be interested in testing for the relative risk parameter for different values of  $R_0$  and in both directions of the alternative hypothesis ( $R > R_0$  or  $R < R_0$ ) depending on the problem of interest.

In section 1.2 we will describe how to formulate this problem statistically. In section 1.3 we will introduce three data sets obtained from various clinical trials. We shall use these data sets to show how asymptotic inference leads to poor results (in terms of maintaining nominal size) and establish the need for exact inference. In section 1.4 we will discuss the standard exact unconditional test used for this problem. In section 1.5 we shall talk about the test we are proposing, which we will call the confidence interval p value test. In section 1.6 we shall discuss how to compute the

rejection regions, size and power for both these tests. In section 1.7 we shall present in detail all the results for the rejection region comparisons. And in the last section, i.e., section 1.8 we shall revisit the data sets introduced in 1.3 and analyze them using the exact tests.

## 1.2 Statistical formulation of the problem

The problem in hand can be described as follows: Let  $X_1$  and  $X_2$  represent two independent response variables distributed as  $\text{bin}(N_1, P_1)$  and  $\text{bin}(N_2, P_2)$  respectively. Here  $N_1$  and  $N_2$  represents the sample sizes in each group, and  $P_1$ , and  $P_2$ , denotes the true response rates. If we denote by  $x_1$  and  $x_2$  the observed number of successes in each group, then the binomial probability mass function of  $X_1$  and  $X_2$  will be

$$\text{bin}(N_1, x_1, P_1) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1}, \quad x_1 = 0, \dots, N_1,$$

and

$$\text{bin}(N_2, x_2, P_2) = \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2}, \quad x_2 = 0, \dots, N_2.$$

The sample space of  $(X_1, X_2)$  will be denoted as  $\mathbf{S} = \{0, \dots, N_1\} \times \{0, \dots, N_2\}$ . Thus there are  $(N_1 + 1) \times (N_2 + 1)$  points in  $\mathbf{S}$ .

For observed outcomes  $(x_1, x_2)$  the data can be presented in a  $2 \times 2$  contingency table as

|         | Group A     | Group B     | Total                   |
|---------|-------------|-------------|-------------------------|
| Success | $x_1$       | $x_2$       | $x_1 + x_2$             |
| Failure | $N_1 - x_1$ | $N_2 - x_2$ | $N_1 + N_2 - x_1 - x_2$ |
| Total   | $N_1$       | $N_2$       | $N_1 + N_2$             |

For testing one of the hypotheses discussed in the previous section the most commonly used test statistic is

$$Z = \frac{\hat{P}_1 - R_0 \hat{P}_2}{\hat{\sigma}},$$

where

$$\hat{\sigma} = \sqrt{\frac{\tilde{P}_1(1 - \tilde{P}_1)}{N_1} + (R_0)^2 \frac{\tilde{P}_2(1 - \tilde{P}_2)}{N_2}}.$$

Here  $\hat{P}_1 = x_1/N_1$ ,  $\hat{P}_2 = x_2/N_2$ , and  $\tilde{P}_1$  and  $\tilde{P}_2$  are estimates of  $P_1$  and  $P_2$ . There are three ways in which  $P_1$  and  $P_2$  may be estimated.

1. They can be replaced by their observed values  $\hat{P}_1$  and  $\hat{P}_2$ .
2. They can be replaced by  $\tilde{P}_1$  and  $\tilde{P}_2$ , computed under the null hypothesis restriction,  $\frac{\tilde{P}_1}{\tilde{P}_2} = R_0$ , subject to the marginal total remaining equal to those observed.
3. They can be replaced by  $\tilde{P}_1$  and  $\tilde{P}_2$ , the maximum likelihood estimates of  $P_1$  and  $P_2$  under the null hypothesis restriction  $\frac{\tilde{P}_1}{\tilde{P}_2} = R_0$ .

Farrington & Manning (1990) have compared these three approaches and recommended the third approach. They showed that the confidence intervals generated by inverting this test has better coverage probabilities, i.e., coverage probabilities closer

to the nominal coverage when the test statistic is computed using the third approach. For our given problem we will consider the above  $Z$  statistic with  $\sigma$  estimated using the restricted maximum likelihood method.

The  $Z$  statistic is asymptotically distributed as a standard normal under the null hypothesis. Hence all the asymptotic theory goes through very nicely for larger sample sizes. So we can use the usual testing procedures to test the various hypotheses discussed in section 1.1. However for small to moderate samples the asymptotic theory does not hold properly. In fact in section 1.3 we will show how poorly the asymptotic test (assuming the test statistic follows a standard normal under the null hypothesis) can perform even for moderately large sample sizes. Here by poor performance we mean the actual size of the test exceeds the nominal size.

The exact distribution of  $Z$  depends on all possible outcomes of two binomial responses given the sample sizes  $N_1$  and  $N_2$ . Thus each outcome  $(x_1, x_2)$  corresponds to a  $2 \times 2$  table. Probability of a particular outcome is

$$P(X_1 = x_1, X_2 = x_2) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2}.$$

The observed exact significance level for a particular value of the parameter  $(P_1, P_2)$  is the sum of probabilities of the tables that are at least as extreme as the observed table. Thus if  $Z_{obs}$  is the value of the statistic for the observed table, then depending on the hypothesis of interest, the tail region will consist of all tables for which  $Z \leq Z_{obs}$  or  $Z \geq Z_{obs}$ . Let us assume that smaller value of the test statistic represents an extreme

observation. If we denote by  $Z(x_1^0, x_2^0)$  the value of  $Z$  for the observed table then we can write this probability as

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}. \quad (1.1)$$

Note that the value of the tail probability function (1.1) depends on the true values of the parameters  $P_1$  and  $P_2$  which we will call nuisance parameters. Various methods of removing nuisance parameters have been discussed by Basu (1977). One of them is conditioning on the sufficient statistics. For the special case when we are testing for the null hypothesis of equality of proportions ( $H_0 : P_1 = P_2 = P$ ), i.e.,  $R_0 = 1$ , the above probability function may be rewritten as

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} \binom{N_2}{x_2} P^{x_1 + x_2} (1 - P)^{N_1 + N_2 - x_1 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}. \quad (1.2)$$

Here the only unknown parameter is  $P$ . One can eliminate this parameter by conditioning on a sufficient statistic for  $P$ . For this special case the marginal total  $T = X_1 + X_2$  is a sufficient statistic for  $P$ . Thus conditioning on  $T$  will eliminate the parameter  $P$  in the calculation of the probability of  $Z \leq Z_{obs}$ . This approach leads to the conditional exact test and this particular test is referred to as Fisher's exact test (Fisher, 1935). Therefore for the special case of testing for the null hypothesis,  $H_0 : P_1 = P_2$ , one may use a conditional exact test. Fisher's exact test is widely used for this particular testing problem. However the main drawback of Fisher's test is

that it is based on the assumption that not only the sum of the total sample sizes are fixed but also the sum of the total responses are fixed as well. So in terms of a  $2 \times 2$  contingency table we are assuming that both the row totals and the column totals are fixed. The disadvantage behind this assumption is that this can lead to a distribution for the test statistic (under the null hypothesis) which is very discrete, especially for smaller sample sizes. This will lead to an overly conservative test where the true size of the test will be much lower than the nominal size.

Now suppose we are interested in testing the null hypothesis,  $H_0 : R = R_0$ . Then one can replace  $P_2$  with  $P_1/R_0$  and rewrite the tail probability function as

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} (P_1/R_0)^{x_2} (1 - P_1/R_0)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}.$$

For the general case when  $R_0 \neq 1$  although there is only one unknown parameter  $P_1$ , there is no simple sufficient statistics to condition on. It may be possible to come up with a sufficient statistic to condition on, but to the best of our knowledge there is no simple sufficient statistic (as in the case of  $R_0 = 1$ , Fisher's exact test) which does the trick. Hence for the more general testing problem of non-unity relative risk a conditional exact approach does not work. The only form of exact inference that can be used in this case is the unconditional exact approach.

In an unconditional approach the nuisance parameters are eliminated by actually

maximizing on the domain of the nuisance parameter, under the null hypothesis. So let us go back to our original problem of testing for  $H_0 : R \geq R_0$  (or  $H_0 : R \leq R_0$ ) and suppose we want to compute (1.1). If we use an unconditional approach we will try to eliminate  $P_1$  and  $P_2$  by actually maximizing over the domain of  $P_1$  and  $P_2$  under the null hypothesis. So what we will try to evaluate is,

$$\sup_{(P_1, P_2) \in H_0} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}. \quad (1.3)$$

The above expression is not that easy to compute as maximizing on a two dimensional set of points may be computationally quite challenging. However one can use certain results to show that evaluating this probability by maximizing over the entire domain of the null hypothesis, and maximizing on the boundary of the null hypothesis are equivalent. In other words one can show that for evaluating this probability, maximizing on a two dimensional set of points  $(P_1, P_2)$  such that  $P_1/P_2 \geq R_0$  (assuming the null hypothesis to be  $R \geq R_0$ ) and maximizing on the line  $P_1/P_2 = R_0$  (which is the boundary of the null hypothesis,  $R \geq R_0$ ) are equivalent. We will talk about this result in greater detail in section 1.4. For the time being let us assume that this is true. Then we can replace  $P_2$  by  $P_1/R_0$ . In that case  $P_1$  will lie between 0 and  $\min(1, R_0)$ . This is because  $(0 \leq P_1 \leq 1)$  and  $(0 \leq P_1/R_0 \leq 1)$ , which implies  $(0 \leq P_1 \leq \min(R_0, 1))$ . Let us denote the domain of  $P_1$  as  $D(R_0)$ . Then one can



rewrite (1.3) as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1-P_1)^{N_1-x_1} \binom{N_2}{x_2} (P_1/R_0)^{x_2} (1-P_1/R_0)^{N_2-x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}.$$

This expression has only a single unknown parameter  $P_1$  and we have to maximize on the domain of  $P_1$  under  $H_0$  (which is  $D(R_0)$ ) in order to evaluate it. This problem is computationally much easier compared to maximizing on a two dimensional set of points.

So for testing the null hypothesis of equality of proportions ( $H_0 : P_1 = P_2 = P$ ) one can use a conditional test, viz., Fisher's exact test. However for the general problem of testing the null hypothesis,  $H_0 : R = R_0$  (or  $H_0 : R \leq R_0$  or  $H_0 : R \geq R_0$ ), and  $R_0 \neq 1$ , the only way of performing exact inference is to use an unconditional exact test. Also for the test of equality of proportions various authors such as Suissa & Shuster (1985), Haber (1986), have studied the two approaches and came to the conclusion that the exact unconditional approach yields a more powerful test than Fisher's exact test.

In the next section we shall introduce three data sets that we will analyze using the asymptotic  $Z$  test. We will show the poor performance of the asymptotic test even for moderately large samples. In section 1.8 we will revisit these examples and analyze them using the exact unconditional approach.

## 1.3 Data Sets

We shall consider three data sets which were obtained from various clinical trial studies. We will test various hypotheses for the relative risk using the asymptotic test. Later on we will test the same hypotheses using the exact tests and the results will be discussed in detail in section 1.8. Here we will present the p value and the size of the asymptotic test. For this test we are assuming that the  $Z$  statistic (discussed in the previous section) follows a standard normal distribution under the null hypothesis.

### 1.3.1 Animal Toxicology Data

This data has been obtained from the StatXact Manual, Version 5. The data had been collected from an animal toxicology study where researchers were interested in testing a natural versus a synthetic chemical. The natural chemical was injected into 77 rats while the synthetic chemical had been injected into 350 rats. The natural chemical induced tumor in 37 of the 77 rats while the synthetic chemical induced tumor in 212 of 350 rats. The data can be presented in a  $2 \times 2$  table as follows:

| Response | Synthetic A | Natural | Total |
|----------|-------------|---------|-------|
| Tumor    | 212         | 37      | 249   |
| No Tumor | 138         | 40      | 178   |
| Total    | 350         | 77      | 427   |

One important hypothesis of interest in this case would be to test whether the synthetic chemical was more carcinogenic than the natural one. If we denote by  $P_1$

the true proportion of tumor rates for the synthetic chemical, and  $P_2$  as the true proportion of tumor rates for the natural chemical then one would be interested in testing

$$H_0 : P_1 \leq P_2 \text{ versus } H_1 : P_1 > P_2.$$

The p value for the asymptotic test in this case is 0.0218 which is highly significant at 5% level. Hence there is strong evidence in favor of the alternative hypothesis. However we will show in section 1.8 that the size of the test is 0.0899 which is much higher than the nominal 5% size. So we see that even for moderately large sample size (the total sample size being 427 in this case) the asymptotic test performs very poorly.

### 1.3.2 Childhood Nephroblastoma Data

This data was reported by Rodary et al. (1989). Here data is presented from a randomized clinical trial to compare two given types of treatments for childhood nephroblastoma. One is nephrectomy followed by post-operative radiotherapy. The other is pre-operative chemotherapy to reduce the tumor mass, followed by nephrectomy. The data can be presented in a  $2 \times 2$  table as follows:

| Response       | Chemo | Radio | Total |
|----------------|-------|-------|-------|
| Rupture free   | 83    | 69    | 152   |
| Ruptured tumor | 5     | 7     | 12    |
| Total          | 88    | 76    | 164   |

One hypothesis of interest will be to test whether radio therapy is non-inferior to chemo therapy, i.e.,

$$H_0 : R \geq R_0 \text{ versus } H_1 : R < R_0,$$

where  $R = \frac{P_1}{P_2}$ .  $P_1$  and  $P_2$  represents the rupture free rates for the chemo and radio population. Let us choose  $R_0 = 1.15$ . The asymptotic p value in this case is 0.0539 which is not significant at 5% level, but suggests some evidence against the null hypothesis. The size of this test is 0.06129 which again exceeds the 5% nominal level.

### 1.3.3 Influenza Vaccine Data

This data was reported by Fries et al. (1993). The study was conducted to evaluate protective efficacy of a Recombinant protein influenza A vaccine against wild type H1N1 virus challenge. Subjects were randomized to receive the experimental vaccine or placebo injection. After that they were challenged intranasally with influenza virus. All subjects were closely monitored for viral infection in the next 9 days and any clinical symptoms of fever, upper respiratory infection, recurrent cough or lower respiratory infection, and myalgia were reported. Again the data may be presented in a  $2 \times 2$  table as

| Disease      | Vaccine | Placebo | Total |
|--------------|---------|---------|-------|
| Infected     | 7       | 12      | 19    |
| Not Infected | 8       | 3       | 11    |
| Total        | 15      | 15      | 30    |

Here the hypothesis of interest was to test

$$H_0 : \pi \leq 0.1 \text{ versus } H_1 : \pi > 0.1.$$

Recall that  $\pi = 1 - \frac{P_1}{P_2}$ . The asymptotic p value for this test is 0.0636 which is not significant at 5% level. The size of this test is 0.0837.

In this section we saw how poorly the asymptotic test can perform, not only for very small sample sizes (like the influenza vaccine data) but even for larger sample sizes (like the animal toxicology data). Hence there is an important need for exact inference for this type of problems.

In sections 1.4 and 1.5 we will introduce two exact tests for testing the relative risk hypotheses.

1. The exact unconditional test.
2. The confidence interval p value test.

Test (1) is the standard exact unconditional test that is used in these situations for exact inference. Test (2), i.e., the confidence interval p value test is a modification of the standard exact test and is based on the method suggested by Berger & Boos (1994). Both these tests will be compared with respect to

- Size of the test.
- Power of the test.

## 1.4 Exact Unconditional Test

In section 1.2, we had discussed that for the exact unconditional test, the p value for a particular observation  $(x_1, x_2)$  can be computed by maximizing the probability function over the entire domain of the parameter space, under the null hypothesis. Now we will formally define the p value for the exact unconditional test as follows. Suppose we are interested in testing the null hypothesis  $H_0 : R \geq R_0$ . For a given  $N_1$ ,  $N_2$ , and  $R_0$ , the p value of the exact unconditional test for a particular sample point  $(x_1^0, x_2^0)$  can be defined as

$$\begin{aligned} & \sup_{\frac{P_1}{P_2} \geq R_0} P(Z(X_1, X_2) \leq Z(x_1^0, x_2^0) \mid P_1, P_2) \\ &= \sup_{\frac{P_1}{P_2} \geq R_0} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}. \end{aligned} \quad (1.4)$$

Let us denote this p value as  $p_{sup}(x_1^0, x_2^0)$ . For a level  $\alpha$  test one would reject the null hypothesis if  $p_{sup}(x_1^0, x_2^0) \leq \alpha$ . Hence forth we will just use  $p_{sup}$  to denote the p value for the observed sample points  $(x_1^0, x_2^0)$ . Now to compute  $p_{sup}$ , we have to maximize the above function over a two dimensional set of points, which is computationally very intensive. However, Sidik (1997) proposed certain conditions under which we can maximize on the boundary of  $H_0$  to compute  $p_{sup}$  instead of maximizing over the entire domain. Now if we are considering the null hypothesis  $H_0 : R \geq R_0$  then the

boundary of the null hypothesis will be  $R = R_0$ . Hence maximizing on this boundary will imply maximizing on the line  $P_1/P_2 = R_0$ , which is a much simpler problem than maximizing on the entire domain of the null hypothesis. Later on Kopit (1999) used these conditions to maximize over an interval when he was considering the problem of testing non-zero differences of proportion. Below we have stated the main conditions of Sidik's theorem.

**Sidik's Theorem:** *Let the families of density functions of  $X_1$  and  $X_2$  be  $f(x_1, \theta_1)$  and  $g(x_2, \theta_2)$ , where  $\theta_1$  and  $\theta_2$  are real valued parameters. Also let  $X_1$  and  $X_2$  be independent. Let  $R$  be a set such that if  $(x_1, x_2)$  is in  $R$ , then so are  $(x_1 - 1, x_2)$  and  $(x_1, x_2 + 1)$ . A set with such a property is said to be Bernard convex. If the family of distributions of  $X_1$  and  $X_2$  are stochastically increasing in  $\theta_1$  and  $\theta_2$  respectively, then*

$$P_{(\theta_1, \theta_2)}[(X_1, X_2) \in R] \leq P_{(\theta'_1, \theta'_2)}[(X_1, X_2) \in R]$$

*for all  $(\theta'_1, \theta'_2)$  such that  $\theta_1 \leq \theta'_1$  and  $\theta_2 \leq \theta'_2$ .*

For our problem,  $X_1$  and  $X_2$  are independent by assumption. Also since  $X_1$  and  $X_2$  are binomially distributed they are stochastically increasing in their proportion parameters. The last condition that needs verification is the Bernard convexity condition (Bernard, 1947). For the problem of non-zero difference of proportions, Kopit was able to justify theoretically that all the sample points lying in the rejection region of the test were Bernard convex. However since he was considering the simple Wald

statistic the mathematical complexities were much more easier to handle. For the  $Z$  statistic that we have considered, it would be very difficult to show mathematically that  $R$  is Bernard convex. However for all the rejection sets that we have considered (for both the tests), later on in this chapter, we have verified by enumeration the Bernard convexity condition. Hence using this theorem we can maximize on the line  $P_1/P_2 = R_0$ . In other words we can replace  $P_2$  with  $P_1/R_0$  and rewrite (1.4) as

$$\sup_{\frac{P_1}{P_2}=R_0} P(Z(X_1, X_2) \leq Z(x_1^0, x_2^0) \mid P_1). \quad (1.5)$$

Now maximizing on the line  $\frac{P_1}{P_2} = R_0$ , is equivalent to maximizing over  $P_1 \in D(R_0)$ .

So we can rewrite (1.5) as

$$\sup_{P_1 \in D(R_0)} P(Z(X_1, X_2) \leq Z(x_1^0, x_2^0) \mid P_1).$$

In terms of the relative risk, it will be defined as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1-x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2-x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}. \quad (1.6)$$

We will use (1.6) to compute  $p_{sup}$ .

The exact unconditional test can be computed in the following three steps.

1. Compute all possible values of the  $Z$  statistic for all tables and order them. Let

$Z_{obs}$  be the observed value of the  $Z$  statistic as obtained from the observed table.



The tail of the observed table includes all tables for which the  $Z$  statistics is less than or equal to  $Z_{obs}$  (or greater than or equal to, depending on the direction of the hypothesis).

2. For a given value of the nuisance parameter  $P_1$  the tail probability is calculated by summing up the probabilities of the tables more extreme than the observed one.
3. Repeat step 2 for all possible values in the domain of the nuisance parameter  $P_1$ .

The required p value which we are denoting as  $p_{sup}$  will be the maximum of all these tail probabilities. Since the domain of  $P_1$  is continuous it has been suggested by Chan to divide it into a fine grid of 1000 points and compute the tail probabilities for each of the 1000 equally divided values of the nuisance parameter and take the maximum of all these values. A grid size of 1000 points is considered to give reasonably accurate result for all practical purposes.

One point to note is that for the exact unconditional test the  $Z$  statistic itself is the ultimate source of ordering. Because once the sample points are ordered with respect to the  $Z$  statistic values, the tail probabilities will maintain the same ordering as the statistic. Thus for this test, the statistic itself is a sufficient source of ordering. This point would be of relevance once we talk about our proposed test. We will see

that this is not the case for the confidence interval p value test. We shall discuss it in detail when we talk about rejection regions for both these tests in section 1.6.

## 1.5 Confidence Interval p value Test

This test was first proposed by Berger & Boos (1994). We will apply this to the problem at hand. Before we discuss this test, let us talk about the main disadvantages of the standard exact unconditional test.

1. The standard test is computationally intensive. Since the whole domain of the nuisance parameter is considered for the maximization, the search for the supremum of  $P_1$  over the entire domain space might be quite cumbersome, especially when a fine grid is used.
2. Once the value of the parameter  $P_1$  is estimated (from the observed data), it would seem to be a waste of information to consider values of  $P_1$  that are entirely unsupported by the data.

Therefore it will make more sense to maximize on a subset of the entire domain of  $P_1$  which has a very high probability of containing the true value. In that case it will be computationally less intensive, as the maximization will take place on a restricted set. Since the subset has a very high chance of containing the true value, so it will not consider values of  $P_1$  that are completely unsupported by the data. This idea

answers Fisher's main concern regarding unconditional tests, which is looking at the worst possible scenario for replacing the nuisance parameter. Fisher had correctly pointed out that this might often lead to very conservative results.

The original idea of replacing the nuisance parameter by some other simpler method was first proposed by Storer & Kim (1990). Instead of maximizing on the entire domain, they replaced the nuisance parameter by its maximum likelihood estimate, given the observed data. However the problem with that approach was it did not satisfy the definition of a valid p value. The correct definition states that for any statistic  $p$ , if under the null hypothesis

$$P(p \leq \alpha) \leq \alpha, \forall \alpha \in [0, 1], \quad (1.7)$$

then  $p$  is a valid p value. In other words  $P(\text{Reject Null} \mid \text{Null}) \leq \alpha$ , which is the definition of a level  $\alpha$  test. The method proposed by Storer and Kim does not satisfy this definition. Hence the actual size of the test exceeds the nominal size. However Berger and Boos suggested a more sensible approach that improves on the original definition, and also generates a valid p value.

Now let us assume that  $p(x_1, x_2, \theta_0)$  is a valid p value for the sample point  $(x_1, x_2)$ , which may be calculated when the true value of  $\theta_0$ , the nuisance parameter, is known. For brevity we will just use the notation  $p(\theta_0)$  to denote this p value. If  $\theta_0$  is unknown, then a valid  $p$  value may be obtained by maximizing  $p(\theta)$  over the parameter space of  $\theta \in H_0$ . So if we define this p value as  $\sup_{\theta \in H_0} p(\theta)$  then it satisfies (1.7). And

this is exactly how we defined  $p_{sup}$  for our given problem in the last section. What Berger and Boos suggested was instead of maximizing on the entire parameter space of  $\theta \in H_0$ , maximize on a confidence set for  $\theta$ . So let us define  $C_\beta$  to be a  $100(1 - \beta)\%$  confidence set for the nuisance parameter  $\theta$ , under the null hypothesis. Then the confidence interval p value can be defined as

$$p_\beta = \sup_{\theta \in C_\beta} p(\theta) + \beta, \quad (1.8)$$

where  $\beta$  represents the penalizing factor added for maximizing on a restricted set. The value of  $\beta$  and the confidence set  $C_\beta$  has to be specified before looking at the data. The values of  $\beta$  suggested by Berger (1996) are 0.001 or 0.0001. So the confidence set suggested is a 99.9% or 99.99% set. Obviously for a level  $\alpha$  test,  $\beta$  must be less than  $\alpha$  to obtain a meaningful test.

Below we have given a proof for the validity of  $p_\beta$  as a p value. After that we will return to our problem of interest, i.e., testing for the hypothesis of relative risk and describe how we can construct the confidence interval p value test for this problem.

### 1.5.1 Validity of the Confidence Interval p value Test

**Lemma.** Suppose that for each  $\theta \in H_0$ ,  $p(\theta)$  satisfies  $P(p(\theta) \leq \alpha) \leq \alpha$  for any assumed known value of  $\theta$ , and,  $\forall \alpha \in (0, 1)$ . Let  $C_\beta$  satisfy  $Pr(\theta \in C_\beta) \geq 1 - \beta$ , under the null hypothesis. Let  $p_\beta$  be defined as in (1.8). Then  $p_\beta$  is a valid p value.

**Proof.** Let the null hypothesis be true, and let us denote the true but unknown value of  $\theta$  under the null hypothesis as  $\theta_0$ .

If  $\beta > \alpha$ , then since  $p_\beta$  is never smaller than  $\beta$ ,  $Pr(p_\beta \leq \alpha) = 0 \leq \alpha$ .

If  $\beta \leq \alpha$ , then

$$\begin{aligned}
 Pr(p_\beta \leq \alpha) &= Pr(p_\beta \leq \alpha, \theta_0 \in C_\beta) + Pr(p_\beta \leq \alpha, \theta_0 \in \bar{C}_\beta) \\
 &\leq Pr(p(\theta_0) + \beta \leq \alpha, \theta_0 \in C_\beta) + Pr(\theta_0 \in \bar{C}_\beta) \\
 &\leq Pr(p(\theta_0) \leq \alpha - \beta) + \beta \\
 &\leq \alpha - \beta + \beta = \alpha
 \end{aligned} \tag{1.9}$$

The first inequality follows because  $\sup_{\theta \in C_\beta} p(\theta) \geq p(\theta_0)$  when  $\theta_0 \in C_\beta$ .

Now we will discuss how to construct the  $100(1 - \beta)\%$  confidence set for  $(P_1, P_2)$  when we are considering the problem of testing for the relative risk.

### 1.5.2 Construction of $100(1 - \beta)\%$ confidence set for $(P_1, P_2)$

In order to maximize the p value function on the restricted set, we need to construct the  $100(1 - \beta)\%$  confidence set for  $(P_1, P_2)$ . We first construct two  $\sqrt{100(1 - \beta)\%}$  confidence intervals for  $P_1$  and  $P_2$  from the observed data. The product of this two

intervals will yield a  $100(1 - \beta)\%$  confidence region for  $(P_1, P_2)$ . Let us denote the intervals for  $P_1$  and  $P_2$  as  $(L_1, U_1)$ , and  $(L_2, U_2)$  respectively. The region formed by the intersection of  $[(L_1, U_1)] \times [(L_2, U_2)]$  and  $H_0$  will form the required  $100(1 - \beta)\%$  confidence region  $C_\beta$ , i.e.,

$$C_\beta = [(L_1, U_1)] \times [(L_2, U_2)] \cap H_0.$$

In Figure 1.1 we have plotted  $P_2$  against  $P_1$ . Note that the coordinates  $(L_1, L_2)$ ,  $(U_1, L_2)$ ,  $(U_1, U_2)$ , and  $(L_1, U_2)$  forms four corners of a rectangle. The region where this rectangle intersects with the null hypothesis  $H_0$  is our required  $100(1 - \beta)\%$  confidence region for  $(P_1, P_2)$ .

The confidence intervals for  $P_1$  and  $P_2$  might be constructed in various ways. We have chosen the Clopper and Pearson confidence intervals (Clopper & Pearson, 1934) for the binomial random variable. If  $X_1 \sim \text{bin}(N_1, P_1)$ , then the  $100(1 - \alpha)\%$  Clopper-Pearson confidence interval for  $P_1$  is given by

$$\left[ \frac{x_1}{x_1 + (N_1 - x_1 + 1)F_{2(N_1 - x_1 + 1), 2x_1, \beta/2}}, \frac{(x_1 + 1)F_{2(x_1 + 1), 2(N_1 - x_1), \beta/2}}{N_1 - x_1 + (x_1 + 1)F_{2(x_1 + 1), 2(N_1 - x_1), \beta/2}} \right].$$

Similarly, if  $X_2 \sim \text{bin}(N_2, P_2)$ , then the  $100(1 - \alpha)\%$  Clopper-Pearson confidence interval for  $P_2$  is given by

$$\left[ \frac{x_2}{x_2 + (N_2 - x_2 + 1)F_{2(N_2 - x_2 + 1), 2x_2, \alpha/2}}, \frac{(x_2 + 1)F_{2(x_2 + 1), 2(N_2 - x_2), \alpha/2}}{N_2 - x_2 + (x_2 + 1)F_{2(x_2 + 1), 2(N_2 - x_2), \alpha/2}} \right],$$

where  $F_{a,b,\alpha/2}$  is the upper  $100(\alpha/2)$ th percentile of an  $F$  distribution with  $a$  and  $b$  degrees of freedom. Note that when  $x_1$  (or  $x_2$ ) is 0 then the lower end point of the

interval for  $P_1$  (or  $P_2$ ) is set to 0. When  $x_1 = N_1$  (or  $x_2 = N_2$ ), then the upper end point of the interval for  $P_1$  (or  $P_2$ ) is set to 1.

For our problem we shall compute a 99.9% confidence set for  $(P_1, P_2)$ . Therefore we shall consider  $\beta$  to be 0.001 (as  $100(1 - \beta) = 99.9 \Rightarrow \beta = 0.001$ ). Since we need each of the confidence intervals to be  $\sqrt{99.9}\%$ ,

$$(1 - \alpha)/2 = \sqrt{(1 - \beta)}/2 = \sqrt{(1 - 0.001)}/2.$$

So

$$1 - \alpha = \sqrt{1 - 0.001} = \sqrt{0.999} \approx 0.9995,$$

$$\alpha = 0.0005.$$

Hence, in order to get a 99.9% confidence region for  $(P_1, P_2)$  we need to choose,  $\beta = 0.001$ , and hence  $\alpha = 0.0005$ , for each of the Clopper-Pearson confidence intervals for  $P_1$  and  $P_2$ . One important point to note is that although other confidence intervals may be used for constructing this confidence set, however we need to choose an exact interval so that the confidence region is guaranteed to have at least  $100(1 - \beta)\%$  coverage. This is important because recall that in (1.8) we are adding the term  $\beta$  to ensure that the actual size never exceeds the nominal  $\alpha$  level. However if the true coverage of the confidence set falls below  $100(1 - \beta)\%$  then in reality we are constructing a  $100(1 - \beta')\%$  confidence set, where  $\beta' > \beta$ . So from (1.9) what we have is

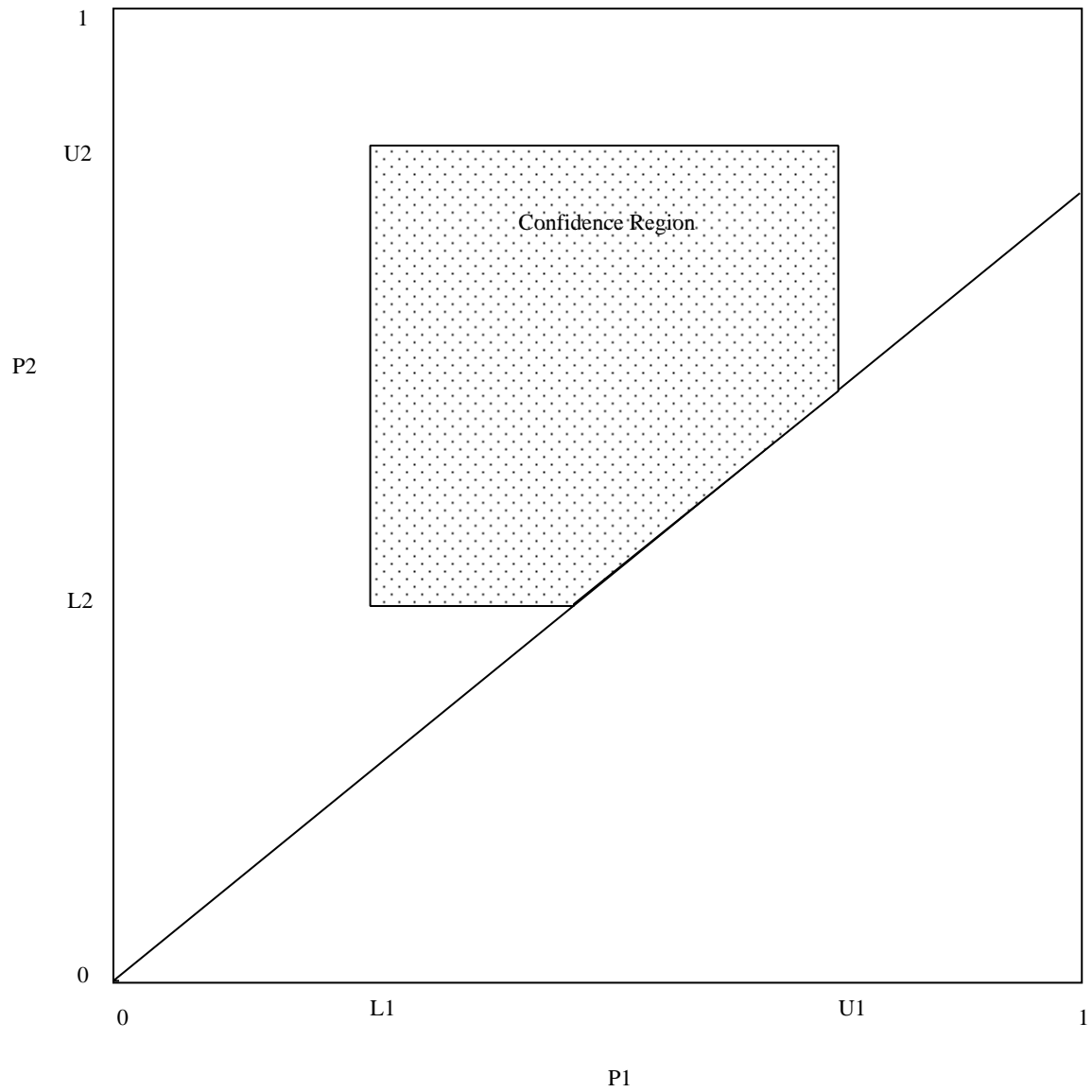


Figure 1.1: The shaded area represents the  $(1-\beta)$  confidence region  $C_\beta$ . The diagonal is the line segment  $P_1/P_2 = R$



$$P(p_\beta \leq \alpha) \leq \alpha - \beta + \beta'.$$

But  $\alpha - \beta + \beta' > \alpha$ . Therefore  $P(p_\beta \leq \alpha)$  need not be less than or equal to  $\alpha$ . Hence we are getting a level  $(\alpha - \beta + \beta')$  test instead of an  $\alpha$  level test, which is undesirable.

Now we are in a position to define the confidence interval p value for our given problem. Recall that in (1.4) we maximized on the entire domain of  $(P_1, P_2) \in H_0$  to define the exact unconditional p value. In (1.8) we showed how we can define a valid p value by actually maximizing on a  $100(1 - \beta)\%$  confidence set for  $(P_1, P_2) \in H_0$ . In this section we have described a method for constructing this confidence set which we are denoting as  $C_\beta$ . Hence, for testing the null hypothesis,  $H_0 : R \geq R_0$ , one can define the confidence interval p value for the observed sample points  $(x_1^0, x_2^0)$  as

$$\sup_{(P_1, P_2) \in C_\beta} \left\{ \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]} \right\} + \beta. \quad (1.10)$$

Let us denote by  $B$  the boundary of the null hypothesis i.e.,  $B = \{(P_1, P_2) : P_1/P_2 = R_0\}$ . Then by Siddik's theorem we can rewrite (1.10) as

$$\sup_{(P_1, P_1/R_0) \in C_\beta^1} \left\{ \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]} \right\} + \beta, \quad (1.11)$$

where  $C_\beta^1 = C_\beta \cap B$ . But  $(P_1, P_1/R_0) \in C_\beta^1$  implies  $P_1 \in C_\beta^*$ , where

$$C_\beta^* = \{\max(L_1, L_2 R_0), \min(U_1, U_2 R_0)\}$$

Hence we can rewrite (1.11) as

$$\sup_{P_1 \in C_\beta^*} \left\{ \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1-P_1)^{N_1-x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2-x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]} \right\} + \beta. \quad (1.12)$$

We will use (1.12) for computing the confidence interval p value,  $p_\beta$ . The main advantage of this test is that it is often the case that the p value function tends to have a sharp peak at an extreme point (lying outside the confidence set of the parameter). In this case the standard exact unconditional test will choose the global maximum over the entire domain of the nuisance parameter under the null hypothesis, and hence will use the value of the parameter which is very unlikely. However the confidence interval will maximize only in the  $100(1-\beta)\%$  confidence set for  $(P_1, P_2)$ . Therefore the p value obtained from the standard test will be more conservative than our test. However in situations where the global maximum lies within the  $100(1-\beta)\%$  confidence set, this test will exceed the p value of the standard test, only by the quantity  $\beta$ . So this test is less conservative in many situations and is guaranteed to perform almost as good as the standard test. Also since it is maximizing on a much smaller set, it is computationally less intensive. And since this test also generates a valid p value, it is an exact test and maintains the nominal size.

## 1.6 Rejection Region, Size and Power

In this section we shall discuss how to compute the rejection regions for each of the tests, and hence we shall talk about the size and power of the tests. We will start by defining the rejection region. For both the tests it consists of the set of sample points  $(x_1, x_2)$  for which the corresponding p values are less than or equal to  $\alpha$ .

1. For the exact unconditional test it can be defined as

$$R_{sup}(\alpha) = [(x_1, x_2) \mid p_{sup}(x_1, x_2) \leq \alpha].$$

2. For the Confidence interval p value test it can be defined as

$$R_{\beta}(\alpha) = [(x_1, x_2) \mid p_{\beta}(x_1, x_2) \leq \alpha].$$

One important point to note is that between these two tests only for the exact unconditional test we can use the test statistic to define the rejection region. For the confidence interval p value test, the p value is the ultimate source of ordering and not the test statistic. We will explain this in detail in the next few paragraphs.

The rejection region for a discrete distribution is calculated in the following way. Let us assume that the smaller value of a test statistic represent a more extreme table. First we calculate the test statistics for all possible values of the sample points and arrange them in an ascending order. Then we start with the smallest point and calculate the p value. If it is less than or equal to the nominal level  $\alpha$  then that

point is included in the rejection region. Again the next sample point with the next smallest value of the test statistic is considered, and the probability is calculated. This process is continued until the cumulative probability exceeds the nominal level. And the set of points for which the cumulative probability was less than or equal to  $\alpha$  are kept. Thus for addition of each point, the probability is evaluated for the same distribution with that particular point added. Since the probability is calculated for the same distribution, naturally cumulative addition of points will only lead to an increase in the probability value.

The cumulative probability with the addition of an extra point will be additive. This is the case for the exact unconditional test. This is because for each sample point added the probability is computed for the same probability model. However this is not necessarily the case for the confidence interval p value test. The reason for this is that with the addition of each sample point, the cumulative probability is evaluated at a different set. For example, when we are evaluating the confidence interval p value, the confidence interval changes, with the addition of each sample point  $(x_1, x_2)$ . This is because the confidence interval for the nuisance parameter is evaluated for that particular observed point each time a new point is added. Since the maximization is done on a different set each time, the values of the parameters  $(P_1, P_2)$  will keep on changing with the addition of each point. In other words the cumulative sample points will be computed each time on a different probability model. Hence in this case there is no guarantee that with the addition of a sample point this probability

will increase.

So for the exact unconditional test there is a one to one correspondence between the  $Z$  statistic and the p value. The test statistic  $Z$  is the ultimate source of ordering. However for the confidence interval p value test, the p values are the ultimate source of ordering and serve as the test statistic. In fact this is the reason why the test performs better than the standard test in many situations. When the rejection region for the standard test is exhausted, i.e., it consists of all the points for which the cumulative probability is less than or equal to the nominal level  $\alpha$ , there is always a possibility that you can find an extra point or two for which the probability evaluated at the restricted space yields a value less than or equal to the nominal level. However due to the additive nature of the probability calculated for the exact unconditional test, there is no way it can be included in the rejection region for that test.

Now we shall discuss the size and power calculations for each of the tests. First we will present the mathematical formula and then we will discuss how to compute them. Note that for the size computation again we will use the same Bernard convexity argument and evaluate the probability at the boundary of the null hypothesis.

1. For the exact unconditional test the size is defined as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1-P_1)^{N_1-x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1-\frac{P_1}{R_0}\right)^{N_2-x_2} I_{[(x_1, x_2) \in R_{sup}(\alpha)]}. \quad (1.13)$$

2. For the confidence interval p value test the size is defined as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2} I_{[(x_1, x_2) \in R_\beta(\alpha)]}. \quad (1.14)$$

Recall that  $D(R_0) = (0, \min(1, R_0))$  represents the domain of  $P_1$ . Similarly power of the test for a given value of  $(P_1, P_2)$  can be computed as

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[(x_1, x_2) \in R_i(\alpha)]}, \quad (1.15)$$

where  $i = sup$  or  $\beta$ .

All computations have been done using the SAS Software, Version 8.1. The rejection regions have been calculated in the following manner. First we calculate the  $Z$  statistic for all possible sample points. Thus if we have  $N_1$  observations in one group and  $N_2$  observations in another group then the total number of possible sample points are  $(N_1 + 1)(N_2 + 1)$ . So we compute the  $Z$  statistic for each pair of possible sample point and arrange them in an ascending order. Then for each of this sample point the p value is computed using (1.6) for  $p_{sup}$ , and (1.12) for  $p_\beta$ , as discussed earlier. For  $p_{sup}$  we compute the probabilities for each sample point until it reaches the nominal value  $\alpha$  (i.e.  $\leq \alpha$ ). We have considered  $\alpha = 0.05$ . But for computing  $p_\beta$  we calculate the probabilities until it reaches 0.3, and from there we select sample points for which the p value is less than or equal to 0.05. This is because, as discussed earlier the cumulative probabilities for this test is not necessarily additive with the addition of

sample points. Hence we calculate the probabilities until it reaches 0.3 and assume that after that the p value will not go below the nominal level of 0.05. And from all these sample points we select samples for which the p values are less than or equal to 0.05.

Once the rejection sets are computed we compute the size for each of the test by evaluating it on the size formula. In the next section we present results for these tests and compare their performance under different situations.

The test of efficacy is a special case of the test for the relative risk parameter where the values of  $\pi_0 = 1 - R_0$  are pretty much restricted, i.e., small non-zero values of  $\pi_0$  are considered (at least for all practical purposes). This parameter is usually considered for the vaccine efficacy application. We shall compare the performance of both exact, approximate and Bayesian tests for this application in the third chapter. For the test of efficacy the problem of interest is to test the alternative hypothesis  $H_1 : \pi > \pi_0$ , i.e.,  $H_1 : R < R_0$ . However when we consider the relative risk parameter in general, we need to consider all the situations in detail, i.e., we need to consider the hypothesis in both directions, values of  $R_0$ , between 0 and 1, as well as greater than 1. Also the unbalancedness of the samples needs to be taken in to account. In the next section we have done an exhaustive comparison of  $p_{sup}$  and  $p_\beta$  for the relative risk testing problem.

## 1.7 Results

Now we shall discuss in detail the results for the test of the hypothesis of relative risk. We shall consider the test in both directions, i.e., for both the alternatives  $H_1 : R > R_0$  and  $H_1 : R < R_0$ . We have tried to make an exhaustive comparison and have considered the following cases:

1.  $0 \leq R \leq 1, N_1 < N_2$ , alt:  $R < R_0$
2.  $0 \leq R \leq 1, N_1 < N_2$ , alt:  $R > R_0$
3.  $0 \leq R \leq 1, N_1 > N_2$ , alt:  $R < R_0$
4.  $0 \leq R \leq 1, N_1 > N_2$ , alt:  $R > R_0$
5.  $1 < R, N_1 < N_2$ , alt:  $R < R_0$
6.  $1 < R, N_1 < N_2$ , alt:  $R > R_0$
7.  $1 < R, N_1 > N_2$ , alt:  $R < R_0$
8.  $1 < R, N_1 > N_2$ , alt:  $R > R_0$
9.  $0 \leq R \leq 1, N_1 = N_2$ , alt:  $R < R_0$
10.  $0 \leq R \leq 1, N_1 = N_2$ , alt:  $R > R_0$
11.  $1 < R, N_1 = N_2$ , alt:  $R < R_0$



12.  $1 < R, N_1 = N_2$ , alt:  $R > R_0$

We will consider the level of the test to be 0.05. For each of these cases we shall compare the rejection sets for both the tests and we will compare the number of sample points in each of the rejection sets. If one of the rejection set is a complete subset of the other, for example if  $R_{sup} \subseteq R_\beta$ , then the latter will be uniformly more powerful. However if the rejection sets are overlapping, i.e.,  $R_{sup}^c \cap R_\beta \neq \phi$  and  $R_{sup} \cap R_\beta^c \neq \phi$ , then this argument will not hold. We shall show that for most of the situations we have considered the rejection set for  $p_\beta$  completely contains the rejection set for  $p_{sup}$  and hence  $p_\beta$  yields an uniformly more powerful test.

We will consider the following values of  $R_0$ . They are 0.1, 0.33, 0.5, 0.77, 0.9, 1.11, 1.29, 2, 3, and 10. Note that the values of  $R_0$  that are greater than 1 are reciprocals of the values that are less than 1. For example  $1.11 = 1/0.9$ . The reason for choosing these values of  $R_0$  is mainly because, owing to the symmetric nature of the test statistic, the sample points in the rejection region of a test for which  $0 \leq R \leq 1, N_1 < N_2$ , alt:  $R < R_0$  will be diagonally opposite to that of the test for  $1 \leq R, N_1 > N_2$ , alt:  $R > 1/R_0$ . For example if there are  $N$  points in the rejection set for one test and (2,3) is one of those points, then there will be  $N$  points in the rejection set for the other test and (3,2) will be one of the points. Therefore discussing 6 of the 12 cases will be sufficient for drawing conclusions. For all these cases if we say one test is better than the other it means that test is uniformly more powerful.

The sample sizes  $(N_1, N_2)$  that are considered are the following:

1. For balanced cases they are (10,10), (15,15), (30,30), (50,50), (75,75) and (100,100).
2. For unbalanced cases they are (10,20), (10,30), (10,40), (15,30), (15,45) and (15,60) respectively.

The different columns of each table may be interpreted as follows. The first two columns, i.e.,  $N_1$  and  $N_2$ , represent the two sample sizes. For columns three and four we have used the following notations to present the idea. If  $|A|$  represent the cardinality of the set  $A$  and  $|B|$  represent the cardinality of the set  $B$ , then  $|A - B|$  represents the number of points that are in set  $A$  but not in set  $B$ . Similarly  $|B - A|$  represent the number of points that are in set  $B$  but not in set  $A$ . We have used this idea to represent the number of sample points that are in the rejection set for one test but not in the rejection set for the other and vice versa. Hence we will denote columns three and four as  $|R_{sup} - R_\beta|$  and  $|R_\beta - R_{sup}|$ , respectively. The next two columns, i.e.,  $|R_{sup}|$  and  $|R_\beta|$  denote the actual number of sample points in the rejection sets for  $p_{sup}$  and  $p_\beta$ . And the last column, % Chg, represents the % increase (or decrease) in the size of the rejection set when  $p_\beta$  is used instead of  $p_{sup}$ , i.e.,

$$\% \text{Chg} = \frac{|R_\beta| - |R_{sup}|}{|R_{sup}|} \times 100.$$

From Table 1.1 we see that for the alternative  $R > 0.1$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10, 40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 2, 4,

Table 1.1:  $H_0 : R \leq 0.1$  versus  $H_1 : R > 0.1$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 2                     | 176         | 178         | 1.1   |
| 10    | 30    | 0                     | 4                     | 257         | 261         | 1.5   |
| 10    | 40    | 0                     | 9                     | 339         | 348         | 2.7   |
| 15    | 30    | 0                     | 5                     | 396         | 401         | 1.2   |
| 15    | 45    | 0                     | 12                    | 584         | 596         | 2.1   |
| 15    | 60    | 0                     | 21                    | 771         | 792         | 2.7   |

Table 1.2:  $H_0 : R \geq 0.1$  versus  $H_1 : R < 0.1$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 0           | 0           | 0     |
| 10    | 30    | 0                     | 0                     | 0           | 0           | 0     |
| 10    | 40    | 0                     | 0                     | 0           | 0           | 0     |
| 15    | 30    | 0                     | 0                     | 0           | 0           | 0     |
| 15    | 45    | 0                     | 0                     | 0           | 0           | 0     |
| 15    | 60    | 0                     | 0                     | 0           | 0           | 0     |

9, 5, 12, and 21 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$  and  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . The mean % increase in the rejection set size is 1.9%.

By the symmetric nature of the test, when the alternative is  $R < 10$  and when  $(N_1, N_2)$  are (20,10), (30, 10), (40,10), (30,15), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 2, 4, 9, 5, 12, and 21 points respectively.

However for Table 1.2 there are no points in the rejection set for both the tests for the given values of  $(N_1, N_2)$  when the alternative is  $R < 0.1$ . Similarly when  $N_1$  and  $N_2$  are switched and the alternative is  $R > 10$  there are no points in the rejection set for both the tests. This is not surprising owing to the fact that the sample sizes

Table 1.3:  $H_0 : R \leq 0.33$  versus  $H_1 : R > 0.33$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 12                    | 122         | 134         | 9.8   |
| 10    | 30    | 0                     | 4                     | 201         | 203         | 2.0   |
| 10    | 40    | 0                     | 14                    | 255         | 269         | 5.4   |
| 15    | 30    | 0                     | 23                    | 292         | 315         | 7.8   |
| 15    | 45    | 0                     | 8                     | 470         | 478         | 1.7   |
| 15    | 60    | 0                     | 35                    | 600         | 635         | 5.8   |

Table 1.4:  $H_0 : R \geq 0.33$  versus  $H_1 : R < 0.33$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 7           | 7           | 0     |
| 10    | 30    | 0                     | 0                     | 10          | 10          | 0     |
| 10    | 40    | 0                     | 0                     | 13          | 13          | 0     |
| 15    | 30    | 0                     | 0                     | 24          | 24          | 0     |
| 15    | 45    | 1                     | 0                     | 36          | 35          | -2.7  |
| 15    | 60    | 0                     | 0                     | 46          | 46          | 0     |

are quite small.

From Table 1.3 we see that for the alternative  $R > 0.33$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10,40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 12, 4, 14, 23, 8, and 35 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$  and therefore  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . The mean % increase in the rejection set size is by 5.4%.

By the symmetric nature of the test, when the alternative is  $R < 3$ , and when  $(N_1, N_2)$  are (20,10), (30,10), (40,10), (30,15), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 12, 4, 14, 23, 8, and 35 points respectively.

Table 1.5:  $H_0 : R \leq 0.5$  versus  $H_1 : R > 0.5$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 3                     | 108         | 111         | 2.7   |
| 10    | 30    | 0                     | 24                    | 144         | 168         | 16.7  |
| 10    | 40    | 0                     | 4                     | 224         | 228         | 1.8   |
| 15    | 30    | 0                     | 8                     | 258         | 266         | 3.1   |
| 15    | 45    | 0                     | 43                    | 353         | 396         | 12.2  |
| 15    | 60    | 0                     | 3                     | 538         | 541         | 0.6   |

Table 1.6:  $H_0 : R \geq 0.5$  versus  $H_1 : R < 0.5$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 5                     | 13          | 18          | 38.4  |
| 10    | 30    | 0                     | 6                     | 20          | 26          | 30    |
| 10    | 40    | 0                     | 7                     | 28          | 35          | 25    |
| 15    | 30    | 0                     | 6                     | 44          | 50          | 13.6  |
| 15    | 45    | 0                     | 9                     | 68          | 77          | 13.3  |
| 15    | 60    | 0                     | 10                    | 91          | 101         | 11    |

However for Table 1.4, except for (15,45) both the rejection sets are identical. When  $N_1$  and  $N_2$  are switched and the alternative is  $R > 3$  similar inference can be drawn. Note that because of this one case where  $p_{sup}$  has one extra point in its rejection set there is a mean decrease in the rejection set size by 0.5 %.

From Table 1.5 we see that for the alternative  $R > 0.5$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10, 40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 3, 24, 4, 8, 43, and 3 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$  and thus  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . Here there is a mean % increase of 6.1 % in the rejection set size.

By the symmetric nature of the test, when the alternative is  $R < 2$  and when  $(N_1, N_2)$  are (20,10), (30,10), (40,10), (30,15), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 3, 24, 4, 8, 43, and 3 points respectively.

From Table 1.6 we see that for the alternative  $R < 0.5$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10, 40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 5, 6, 7, 6, 9, and 10 points respectively. Again in all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$  and hence  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . The mean % increase in rejection set size is 21.9 %.

By the symmetric nature of the test, when the alternative is  $R > 2$  and when  $(N_1, N_2)$  are (20,10), (30,10), (40,10), (30,15), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 5, 6, 7, 6, 9, and 10 points respectively.

From Table 1.7 we see that for the alternative  $R > 0.77$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10,40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 1, 8, 27, 2, 17, and 54 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ , and  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . The mean % increase in the rejection set size is 5.5%.

By the symmetric nature of the test, when the alternative is  $R < 1.29$  and when  $(N_1, N_2)$  are (20,10), (30,10), (40,10), (30,15), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 1, 8, 27, 2, 17, and 54 points respectively.

Table 1.7:  $H_0 : R \leq 0.77$  versus  $H_1 : R > 0.77$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 1                     | 81          | 82          | 1.2   |
| 10    | 30    | 0                     | 8                     | 113         | 121         | 7.1   |
| 10    | 40    | 0                     | 27                    | 134         | 161         | 2     |
| 15    | 30    | 0                     | 2                     | 196         | 198         | 1     |
| 15    | 45    | 0                     | 17                    | 281         | 298         | 6     |
| 15    | 60    | 0                     | 54                    | 344         | 398         | 15.7  |

Table 1.8:  $H_0 : R \geq 0.77$  versus  $H_1 : R < 0.77$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 4                     | 33          | 37          | 12.1  |
| 10    | 30    | 0                     | 11                    | 49          | 59          | 20.4  |
| 10    | 40    | 0                     | 11                    | 67          | 78          | 16.4  |
| 15    | 30    | 0                     | 2                     | 98          | 100         | 2     |
| 15    | 45    | 0                     | 2                     | 153         | 155         | 1.3   |
| 15    | 60    | 1                     | 0                     | 208         | 207         | -.4   |

From Table 1.8 we see that for the alternative  $R < 0.77$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10,40), (15,30), (15,45),  $p_\beta$  performs better than  $p_{sup}$  by 4, 11, 11, 2, and 2 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ . However when  $(N_1, N_2)$  is (15, 60) the rejection set for  $p_{sup}$  contains one extra point. The mean % increase is 8.6%.

By the symmetric nature of the test, when the alternative is  $R > 1.29$  and when  $(N_1, N_2)$  are (20,10), (30, 10), (40,10), (30,15), (45,15) and (60,15) similar inference can be drawn.

From Table 1.9 we see that for the alternative  $R > 0.9$ , when  $(N_1, N_2)$  are (10,30), (10,40), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 4, 17, 10, and 38 points

Table 1.9:  $H_0 : R \leq 0.9$  versus  $H_1 : R > 0.9$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 70          | 70          | 0     |
| 10    | 30    | 0                     | 4                     | 100         | 104         | 4     |
| 10    | 40    | 0                     | 17                    | 121         | 138         | 14    |
| 15    | 30    | 2                     | 3                     | 167         | 168         | 1.8   |
| 15    | 45    | 0                     | 10                    | 247         | 257         | 4     |
| 15    | 60    | 0                     | 38                    | 306         | 344         | 12.4  |

Table 1.10:  $H_0 : R \geq 0.9$  versus  $H_1 : R < 0.9$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 4                     | 42          | 46          | 9.5   |
| 10    | 30    | 0                     | 11                    | 64          | 75          | 17.2  |
| 10    | 40    | 0                     | 16                    | 87          | 103         | 18.4  |
| 15    | 30    | 0                     | 17                    | 107         | 124         | 15.9  |
| 15    | 45    | 0                     | 25                    | 166         | 191         | 15.1  |
| 15    | 60    | 0                     | 40                    | 222         | 262         | 18.1  |

respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$  and hence  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . The mean % increase is by 12.41%.

By the symmetric nature of the test, when the alternative is  $R < 1.1$  and when  $(N_1, N_2)$  are (30, 10), (40,10), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 4, 17, 10, and 38 points respectively.

From Table 1.10 we see that for the alternative  $R < 0.9$ , when  $(N_1, N_2)$  are (10,20), (10,30), (10, 40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 4, 11, 16, 17, 25, and 40 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ .



Table 1.11:  $H_0 : R \leq 1.11$  versus  $H_1 : R > 1.11$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 54          | 54          | 0     |
| 10    | 30    | 0                     | 2                     | 80          | 82          | 2.5   |
| 10    | 40    | 0                     | 5                     | 103         | 108         | 4.9   |
| 15    | 30    | 0                     | 1                     | 132         | 133         | 0.8   |
| 15    | 45    | 0                     | 2                     | 202         | 204         | 1     |
| 15    | 60    | 0                     | 15                    | 259         | 274         | 5.8   |

Table 1.12:  $H_0 : R \geq 1.11$  versus  $H_1 : R < 1.11$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 72          | 72          | 0     |
| 10    | 30    | 0                     | 2                     | 108         | 110         | 1.8   |
| 10    | 40    | 0                     | 5                     | 144         | 149         | 3.5   |
| 15    | 30    | 1                     | 1                     | 174         | 174         | 0     |
| 15    | 45    | 0                     | 2                     | 265         | 267         | 0.8   |
| 15    | 60    | 0                     | 4                     | 360         | 364         | 1.1   |

By the symmetric nature of the test, when the alternative is  $R > 1.1$  and when  $(N_1, N_2)$  are (20,10), (30, 10), (40,10), (30,15), (45,15) and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 4, 11, 16, 17, 25, and 40 points respectively.

For the next set of tables we will consider the cases where  $N_1 < N_2$  and  $R_0 > 1$  for both the alternatives, i.e., for  $R < R_0$  and  $R > R_0$ .

From Table 1.11 we see that for the alternative  $R > 1.11$ , when  $(N_1, N_2)$  are (10,30), (10,40), (15,30), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 2, 5, 1, 2, and 15 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ . The mean % increase in the rejection sets in this case is 2.5%.

By the symmetric nature of the test, when the alternative is  $R < 0.9$  and when  $(N_1, N_2)$  are (30,10), (40,10), (30,15), (45,15), and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 2, 5, 1, 2, and 15 points respectively.

From Table 1.12 we see that for the alternative  $R < 1.11$ , when  $(N_1, N_2)$  are (10,30), (10,40), (15,45) and (15,60),  $p_\beta$  performs better than  $p_{sup}$  by 2, 5, 2, and 4 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ , and  $p_\beta$  will generate a more powerful test than  $p_{sup}$ . However for  $(N_1, N_2) = (15,30)$  the rejection sets overlap, i.e., both  $p_{sup}$  and  $p_\beta$  have one sample point in their rejection set that is not present in the other's set. The mean % increase is only 1.5%.

Again by the symmetric nature of the test, when the alternative is  $R > 0.9$  and when  $(N_1, N_2)$  are (30,10), (40,10), (30,15), (45,15) and (60,15), inference can be drawn likewise.

From Table 1.13 we see that for the alternative  $R > 1.29$ , when  $(N_1, N_2)$  are (10, 40), (15,30), and (15, 60),  $p_\beta$  performs better than  $p_{sup}$  by 2, 1, and 3 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$  and the rejection sets are not overlapping. One should note that the improvement is slowly decreasing, i.e., both the rejection sets are almost same. Here the mean % increase is only 0.74%.

Table 1.13:  $H_0 : R \leq 1.29$  versus  $H_1 : R > 1.29$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 42          | 42          | 0     |
| 10    | 30    | 0                     | 0                     | 70          | 70          | 0     |
| 10    | 40    | 0                     | 2                     | 91          | 93          | 2.2   |
| 15    | 30    | 0                     | 1                     | 108         | 109         | 0.9   |
| 15    | 45    | 0                     | 0                     | 173         | 173         | 0     |
| 15    | 60    | 0                     | 3                     | 229         | 232         | 1.3   |

Table 1.14:  $H_0 : R \geq 1.29$  versus  $H_1 : R < 1.29$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 1                     | 86          | 87          | 1.1   |
| 10    | 30    | 1                     | 0                     | 136         | 135         | -0.7  |
| 10    | 40    | 2                     | 0                     | 186         | 184         | -1.1  |
| 15    | 30    | 0                     | 0                     | 207         | 207         | 0     |
| 15    | 45    | 0                     | 2                     | 318         | 320         | 0.6   |
| 15    | 60    | 0                     | 3                     | 432         | 435         | 0.7   |

By the symmetric nature of the test, when the alternative is  $R < .77$  and when  $(N_1, N_2)$  are (40,10), (30,15), and (60,15),  $p_\beta$  performs better than  $p_{sup}$  by 2, 1, and 3 points respectively.

From Table 1.14 we see that for the alternative  $R < 1.29$ , in out of the six cases in only three cases there is a slight improvement. And in two cases  $p_{sup}$  has more points in the rejection set than  $p_\beta$ . Thus one would notice that the significant improvement that we have seen when  $R_0$  was between 0 and 1 is no more. The further we are moving from 1 the margin of improvement is decreasing. In this case the % increase has fallen to 0.09% showing almost no improvement.

Table 1.15:  $H_0 : R \leq 2$  versus  $H_1 : R > 2$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 1                     | 0                     | 24          | 23          | -4.1  |
| 10    | 30    | 0                     | 0                     | 39          | 39          | 0     |
| 10    | 40    | 0                     | 0                     | 55          | 55          | 0     |
| 15    | 30    | 1                     | 0                     | 61          | 60          | -1.6  |
| 15    | 45    | 0                     | 1                     | 100         | 101         | 1     |
| 15    | 60    | 0                     | 4                     | 135         | 139         | 3     |

Table 1.16:  $H_0 : R \geq 2$  versus  $H_1 : R < 2$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 124         | 124         | 0     |
| 10    | 30    | 2                     | 0                     | 195         | 193         | -1    |
| 10    | 40    | 0                     | 1                     | 262         | 263         | 0.3   |
| 15    | 30    | 0                     | 0                     | 287         | 287         | 0     |
| 15    | 45    | 0                     | 0                     | 443         | 443         | 0     |
| 15    | 60    | 0                     | 0                     | 602         | 602         | 0     |

By the symmetric nature of the test, when the alternative is  $R > .77$  and when  $(N_1, N_2)$  are (20,10), (30,10), (40,10), (30,15), (45,15) and (60,15) the performance will be same.

Again for both Tables 1.15 and 1.16 there has been no remarkable difference in the two rejection sets. They are almost identical with a few points extra one way or the other. Here there is actually a mean % decrease in the rejection set size. For Table 1.15 the % decrease is by 0.29% and for Table 1.16 it is by 0.12%. For the reverse case when  $N_2 > N_1$  and  $R_0 = 0.5$  the results will be same as well.

For Table 1.17 the rejection sets were same for both the tests except for (10, 40) and (15,60) where  $p_{sup}$  has slightly more points in its rejection region. For Table 1.18

Table 1.17:  $H_0 : R \leq 3$  versus  $H_1 : R > 3$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 12          | 12          | 0     |
| 10    | 30    | 0                     | 0                     | 23          | 23          | 0     |
| 10    | 40    | 2                     | 0                     | 34          | 32          | -5.8  |
| 15    | 30    | 0                     | 0                     | 35          | 35          | 0     |
| 15    | 45    | 0                     | 0                     | 60          | 60          | 0     |
| 15    | 60    | 1                     | 0                     | 87          | 86          | -1.1  |

Table 1.18:  $H_0 : R \geq 3$  versus  $H_1 : R < 3$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 150         | 150         | 0     |
| 10    | 30    | 1                     | 0                     | 233         | 232         | -0.4  |
| 10    | 40    | 0                     | 0                     | 313         | 313         | 0     |
| 15    | 30    | 2                     | 2                     | 341         | 341         | 0     |
| 15    | 45    | 1                     | 0                     | 523         | 522         | 0.1   |
| 15    | 60    | 1                     | 0                     | 706         | 705         | -0.1  |

also either the rejection sets are same or there is a slight improvement for  $p_{sup}$ . When  $(N_1, N_2) = (15, 30)$  the rejection sets overlap. In both these cases there is a mean % decrease by 1.2% and 0.1% respectively.

For Table 1.19 the rejection sets are almost identical in the last three cases, and identical in the first three. However note that since the rejection set itself contains so few points so an improvement by only one point for  $p_{sup}$  in the case of (15,30) and (15,45) has made the % Chg, -33.3%, and -12.5%. Here the mean % decrease is by 4.9%. For Table 1.20,  $p_\beta$  showed a slight improvement over  $p_{sup}$  for  $(N_1, N_2)$  equal to (10,20), (15,30) and (15,60). Here actually the % change has been positive, i.e., an increase by 1.3%.

Table 1.19:  $H_0 : R \leq 10$  versus  $H_1 : R > 10$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 0                     | 0           | 0           | 0     |
| 10    | 30    | 0                     | 0                     | 2           | 2           | 0     |
| 10    | 40    | 0                     | 0                     | 4           | 4           | 0     |
| 15    | 30    | 1                     | 0                     | 3           | 2           | -33.3 |
| 15    | 45    | 1                     | 0                     | 8           | 7           | -12.5 |
| 15    | 60    | 0                     | 2                     | 12          | 14          | 16.6  |

Table 1.20:  $H_0 : R \geq 10$  versus  $H_1 : R < 10$ :  $N_1 < N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 20    | 0                     | 1                     | 188         | 189         | 0.5   |
| 10    | 30    | 0                     | 0                     | 290         | 290         | 0     |
| 10    | 40    | 0                     | 0                     | 391         | 391         | 0     |
| 15    | 30    | 0                     | 4                     | 418         | 422         | 1     |
| 15    | 45    | 0                     | 0                     | 644         | 644         | 0     |
| 15    | 60    | 0                     | 1                     | 867         | 866         | -0.10 |

One very interesting point to note would be when  $N_1 < N_2$ , and  $0 \leq R_0 \leq 1$ ,  $p_\beta$  performs significantly better than  $p_{sup}$  almost always for both directions of the test. Thus by symmetry when  $N_1 > N_2$  and  $R_0 > 1$ ,  $p_\beta$  will perform significantly better than  $p_{sup}$ . However we have seen that for  $N_1 < N_2$  and  $R_0 > 1$  there is not much difference in the performances of the two tests. In fact sometimes  $p_{sup}$  seem to perform better marginally. Therefore for the situation when  $N_1 > N_2$  and  $0 \leq R_0 \leq 1$  there will not be much difference in the performance of  $p_{sup}$  and  $p_\beta$  in terms of power of the test.

Next we will consider the cases for which  $N_1 = N_2$ , i.e., the balanced cases. Here we shall only consider for values of  $R_0=0.1, 0.33, 0.5, 0.77$  and  $0.9$  in both

Table 1.21:  $H_0 : R \leq 0.1$  versus  $H_1 : R > 0.1$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 93          | 93          | 0     |
| 15    | 15    | 0                     | 0                     | 207         | 207         | 0     |
| 30    | 30    | 0                     | 2                     | 823         | 825         | 0.2   |
| 50    | 50    | 0                     | 10                    | 2291        | 2301        | 0.4   |
| 75    | 75    | 0                     | 28                    | 5168        | 5196        | 0.5   |
| 100   | 100   | 0                     | 47                    | 9210        | 9257        | 0.5   |

Table 1.22:  $H_0 : R \geq 0.1$  versus  $H_1 : R < 0.1$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 0           | 0           | 0     |
| 15    | 15    | 0                     | 0                     | 0           | 0           | 0     |
| 30    | 30    | 0                     | 0                     | 4           | 4           | 0     |
| 50    | 50    | 0                     | 1                     | 33          | 34          | 3     |
| 75    | 75    | 1                     | 9                     | 94          | 103         | 8.5   |
| 100   | 100   | 1                     | 7                     | 216         | 222         | 2.7   |

directions. Owing to the symmetry all the reciprocal values of  $R_0$ , i.e.,  $1/0.1$  etc. will be automatically considered.

From Table 1.21 we see that for the alternative  $R > 0.1$ , when  $(N_1, N_2)$  are (30,30), (50,50), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 2, 10, 28, and 47 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ , and  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . However the mean % increase in the rejection set size is only 0.26 %. This is because even though there was a gain of as much as 47 points,  $|R_{sup}|$  and  $|R_\beta|$  were quite high as well.

Table 1.23:  $H_0 : R \leq 0.33$  versus  $H_1 : R > 0.33$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 71          | 71          | 0     |
| 15    | 15    | 0                     | 1                     | 161         | 161         | 0.6   |
| 30    | 30    | 0                     | 0                     | 674         | 674         | 0     |
| 50    | 50    | 1                     | 1                     | 1913        | 1913        | 0     |
| 75    | 75    | 0                     | 28                    | 4334        | 4362        | 0.6   |
| 100   | 100   | 0                     | 49                    | 7774        | 7823        | 0.6   |

Table 1.24:  $H_0 : R \geq 0.33$  versus  $H_1 : R < 0.33$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 4           | 12          | 0     |
| 15    | 15    | 0                     | 0                     | 12          | 12          | 0     |
| 30    | 30    | 0                     | 1                     | 70          | 71          | 1.4   |
| 50    | 50    | 0                     | 3                     | 235         | 238         | 1.3   |
| 75    | 75    | 1                     | 6                     | 591         | 596         | 0.8   |
| 100   | 100   | 3                     | 7                     | 1122        | 1126        | 0.4   |

By the symmetric nature of the test, when the alternative is  $R < 10$ , and when  $(N_1, N_2)$  are (30,30), (50,50), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 2, 10, 28, and 47 points respectively.

For Table 1.22 we see that for the alternative  $R < 0.1$ , when  $(N_1, N_2)$  is (50,50) there is a gain of 1 point for  $p_\beta$ . For (75,75) and (100,100) although there is a gain of 9 and 7 points but the rejection sets are overlapping, as  $p_{sup}$  has one point that is not in  $p_\beta$ . For the other 3 cases the rejection sets are identical. There is a mean % increase of 1.86 %.

From Table 1.23 we see that for the alternative  $R > 0.33$ , when  $(N_1, N_2)$  are (15,15), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 1, 28, and 49 points



respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ , and  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . For (50,50) the rejection regions overlap as both of them have an extra point. The mean % increase in the rejection set size was only 0.30 %.

By the symmetric nature of the test, when the alternative is  $R < 3$ , and when  $(N_1, N_2)$  are (15,15), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 1, 28, and 49 points respectively.

For Table 1.24 we see that for the alternative  $R < 0.33$ , when  $(N_1, N_2)$  are (30,30) and (50,50) there is a gain of 1 and 3 points for  $p_\beta$ . For (10,10) and (15,15) the rejection sets were identical, and for (75,75), and (100,100) they were overlapping. There is a mean % increase of 0.64 %.

From Table 1.25 we see that for the alternative  $R > 0.5$ , when  $(N_1, N_2)$  are (30,30), (50,50), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 5, 14, 23, and 33 points respectively. In all these cases the rejection set for  $p_{sup}$  is a complete subset of the rejection set for  $p_\beta$ , and  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . The mean % increase in the rejection set size was only 0.43 %.

By the symmetric nature of the test, when the alternative is  $R < 2$ , and when  $(N_1, N_2)$  are (30,30), (50,50), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 5, 14, 23, and 33 points respectively.

Table 1.25:  $H_0 : R \leq 0.5$  versus  $H_1 : R > 0.5$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 55          | 55          | 0     |
| 15    | 15    | 0                     | 0                     | 133         | 133         | 0     |
| 30    | 30    | 0                     | 5                     | 568         | 573         | 0.8   |
| 50    | 50    | 0                     | 14                    | 1637        | 1651        | 0.8   |
| 75    | 75    | 0                     | 23                    | 3772        | 3795        | 0.6   |
| 100   | 100   | 0                     | 33                    | 6809        | 6842        | 0.4   |

Table 1.26:  $H_0 : R \geq 0.5$  versus  $H_1 : R < 0.5$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 7           | 7           | 0     |
| 15    | 15    | 0                     | 0                     | 23          | 23          | 0     |
| 30    | 30    | 3                     | 0                     | 132         | 129         | -2.2  |
| 50    | 50    | 0                     | 4                     | 406         | 410         | 1     |
| 75    | 75    | 0                     | 13                    | 981         | 994         | 1.3   |
| 100   | 100   | 0                     | 5                     | 1855        | 1860        | 0.3   |

For Table 1.26, we see that for the alternative  $R < 0.5$ , when  $(N_1, N_2)$  are (50,50), (75,75), and (100,100),  $p_\beta$  performs better than  $p_{sup}$  by 4, 13, and 5 points respectively. However for (30,30)  $p_{sup}$  has 3 extra points in its rejection region. The mean % increase in the rejection set size was only 0.06 %.

For Table 1.27 we see that when the sample sizes are 30,  $p_\beta$  gains four more extra points compared to  $p_{sup}$ . As usual for smaller sample sizes like 10 or 15, there is no improvement. For (75,75) and (100,100) the rejection sets overlap. For table 1.28, we see that for sample sizes of 10, 30, and 50,  $p_\beta$  has gained 1, 5 and 7 points respectively. For sample size of 75,  $p_{sup}$  actually performs better. And for (100,100) the rejection regions overlap. The mean % increase in the rejection set size was only 0.54 %.

Table 1.27:  $H_0 : R \leq 0.77$  versus  $H_1 : R > 0.77$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 38          | 38          | 0     |
| 15    | 15    | 1                     | 0                     | 97          | 96          | -1    |
| 30    | 30    | 0                     | 4                     | 425         | 429         | 1     |
| 50    | 50    | 0                     | 0                     | 1265        | 1265        | 0     |
| 75    | 75    | 1                     | 4                     | 2949        | 2952        | 0.1   |
| 100   | 100   | 1                     | 3                     | 5359        | 5361        | 0     |

Table 1.28:  $H_0 : R \geq 0.77$  versus  $H_1 : R < 0.77$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 18          | 18          | 0     |
| 15    | 15    | 0                     | 1                     | 47          | 48          | 2.1   |
| 30    | 30    | 0                     | 5                     | 226         | 231         | 2.2   |
| 50    | 50    | 0                     | 7                     | 690         | 697         | -1    |
| 75    | 75    | 3                     | 0                     | 1679        | 1676        | -.2   |
| 100   | 100   | 2                     | 4                     | 3079        | 3081        | 0.1   |

From Table 1.29 we see that for the alternative  $R > 0.9$ , either the rejection sets overlap, or  $p_{sup}$  performs slightly better. The results will be same for the alternative  $R < 1.1$  Here there was actually a mean % decrease in the rejection set size by 0.48 %.

For Table 1.30, we see that for the alternative  $R < 0.9$ , when  $(N_1, N_2)$  are (30,30), (50,50) and (75,75),  $p_\beta$  performs better than  $p_{sup}$  by 4, 33, and 15 points respectively. For sample sizes of 10 and 15 the rejection regions were identical. And for (100,100) they overlapped. There was a mean % increase in the rejection set size by 1.04 %.

One can summarize the results as follows. When  $N_1 < N_2$ , and  $0 \leq R_0 \leq 1$  (or  $N_1 > N_2$ , and  $R_0 > 1$ ),  $p_\beta$  shows remarkable improvement over  $p_{sup}$ . In these

Table 1.29:  $H_0 : R \leq 0.9$  versus  $H_1 : R > 0.9$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 32          | 32          | 0     |
| 15    | 15    | 1                     | 0                     | 81          | 80          | -1.2  |
| 30    | 30    | 0                     | 0                     | 366         | 366         | 0     |
| 50    | 50    | 2                     | 1                     | 1090        | 1089        | -0.1  |
| 75    | 75    | 4                     | 4                     | 2559        | 2559        | 0     |
| 100   | 100   | 4                     | 0                     | 4675        | 4671        | -0.1  |

Table 1.30:  $H_0 : R \geq 0.9$  versus  $H_1 : R < 0.9$ :  $N_1 = N_2$ 

| $N_1$ | $N_2$ | $ R_{sup} - R_\beta $ | $ R_\beta - R_{sup} $ | $ R_{sup} $ | $ R_\beta $ | % Chg |
|-------|-------|-----------------------|-----------------------|-------------|-------------|-------|
| 10    | 10    | 0                     | 0                     | 24          | 24          | 0     |
| 15    | 15    | 0                     | 0                     | 60          | 60          | 0     |
| 30    | 30    | 0                     | 4                     | 276         | 280         | 1.4   |
| 50    | 50    | 0                     | 33                    | 813         | 846         | 4.1   |
| 75    | 75    | 0                     | 15                    | 1996        | 2011        | 0.8   |
| 100   | 100   | 1                     | 3                     | 3685        | 3687        | 0.1   |

cases we have seen the size plots to be quite skewed leaving the opportunity for improvement when we use the confidence interval method to obtain the p values. Illustrative examples later will help in explaining the point more clearly. However when  $N_1 < N_2$ , and,  $R_0 > 1$  (or  $N_1 > N_2$ , and,  $0 \leq R_0 \leq 1$ ), there is not much difference between the two rejection sets. Especially as the value of  $R$  increases (or decreases for the other case) the improvement starts decreasing. We have seen that in both these cases, the  $p_{sup}$  size plot function tend to be very close to the nominal level for all values of the parameter, leaving little room to add points in the rejection set, and hence little room for improvement. However for the more common applications in the pharmaceutical area (for example, testing for non-inferiority or efficacy) people

are more interested in values of  $R_0$  closer to 1.

For balanced samples although we have seen improvement, but the improvement is more for larger sample sizes. And as the numbers suggest, it is not as remarkable as for the unbalanced case where  $N_1$  was less than  $N_2$ , and  $R_0$  was between 0 and 1. In balanced cases the size plot function tend to be more closer to the nominal size without many sharp peaks and often the confidence set for the nuisance parameter contains the actual supremum. Therefore in these cases,  $p_\beta$  exceeds  $p_{sup}$  by the penalizing term of  $\beta$ .

Table 1.31 summarizes all the results. In out of the total 180 cases that we have compared, we have seen that in 95 cases  $p_{sup}$  is a complete subset of  $p_\beta$ . Therefore in 52.8% of the cases  $p_\beta$  will generate an uniformly more powerful test than  $p_{sup}$ . In 55 cases, the rejection sets were identical, i.e., in 30.5 % of the cases, the rejection sets were identical. Out of the remaining 30 cases, in 19 of the cases  $p_\beta$  is a complete subset of  $p_{sup}$ , and for the rest of the 11 cases the rejection sets overlapped. Thus, in only 10.5% of the cases  $p_{sup}$  provides a better test. And in the remaining 6.2% of the cases the rejection sets overlapped. Also another interesting point to note is that, among the 19 cases where  $p_{sup}$  had a larger rejection set, there were only two cases where the improvement was by 3 points, three cases where the improvement was by 2 points, and in the remaining 14 cases it was just by 1 point. This shows that for all practical purposes they were almost identical, and the tests will have almost

Table 1.31: Summary of Results.

|   | Number of Cases | % of cases |
|---|-----------------|------------|
| $R_{sup}$ is a complete subset of $R_\beta$ | 95              | 52.8%      |
| $R_\beta$ is a complete subset of $R_{sup}$ | 19              | 10.5%      |
| $R_{sup}$ and $R_\beta$ are identical       | 55              | 30.5%      |
| $R_{sup}$ and $R_\beta$ overlaps            | 11              | 6.2%       |
| Total                                       | 180             | 100%       |

the same power. In contrast when ever we saw an improvement for  $p_\beta$  it was usually much more prominent and there was a considerable gain in power. So in 83.3 % of the cases  $p_\beta$  is at least as good as  $p_{sup}$ . And only in 10.5 % of the cases it is worse. What we conclude from our findings is that the test we have proposed for this problem is uniformly better than  $p_{sup}$  almost always.

Table 1.32 summarizes the over all performance of  $p_\beta$  over  $p_{sup}$ , for all the 12 possible testing situations considered. In the next section, i.e., section 1.8, we will re visit the data sets introduced in section 1.3, and see how the exact methods perform.

At this point it might be interesting to note that one can actually define an approximate test based on the confidence interval method that is guaranteed to give

an uniformly more powerful test, always. Recall that for  $p_\beta$ , we are adding the extra term  $\beta$ , so that it becomes a level  $\alpha$  test. However if one is willing to be satisfied with a level  $\alpha + \beta$  test instead of a level  $\alpha$  test, i.e., not add the term  $\beta$ , then the rejection set for  $p_\beta$  will be guaranteed to contain at least as many points as  $p_{sup}$ . Hence  $p_\beta$  will be always uniformly more powerful than  $p_{sup}$ . Since the values of  $\beta$  chosen are very small quantities like 0.001 or 0.0001, the size of this test will never exceed 0.051 or 0.0501. So for all practical purposes this will be almost as good as a level  $\alpha$  test. Also compared to other approximate tests (where there is no guarantee as to how much the size might exceed the nominal size), this test will be much more useful.

Table 1.32: Table depicting overall performance of  $p_\beta$  over  $p_{sup}$  for all the cases.

|    |                   |             |                   |   |
|----|-------------------|-------------|-------------------|---|
| 1  | $0 \leq R \leq 1$ | $N_1 < N_2$ | alt: $R \leq R_0$ | Performs better almost always   |
| 2  | $0 \leq R \leq 1$ | $N_1 < N_2$ | alt: $R \geq R_0$ | Performs better almost always   |
| 3  | $0 \leq R \leq 1$ | $N_1 > N_2$ | alt: $R \leq R_0$ | Performs similar or marginally worse  |
| 4  | $0 \leq R \leq 1$ | $N_1 > N_2$ | alt: $R \geq R_0$ | Performs similar or marginally worse  |
| 5  | $1 \leq R$        | $N_1 < N_2$ | alt: $R \leq R_0$ | Performs similar or marginally worse  |
| 6  | $1 \leq R$        | $N_1 < N_2$ | alt: $R \geq R_0$ | Performs similar or marginally worse  |
| 7  | $1 \leq R$        | $N_1 > N_2$ | alt: $R \leq R_0$ | Performs better almost always   |
| 8  | $1 \leq R$        | $N_1 > N_2$ | alt: $R \geq R_0$ | Performs better almost always   |
| 9  | $0 \leq R \leq 1$ | $N_1 = N_2$ | alt: $R \leq R_0$ | Performs better for larger sample sizes.<br>Shows improvement when $R_0$ is close to 1. |
| 10 | $0 \leq R \leq 1$ | $N_1 = N_2$ | alt: $R \geq R_0$ | Performs better for larger sample sizes.<br>Shows improvement when $R_0$ is close to 1. |
| 11 | $1 \leq R$        | $N_1 = N_2$ | alt: $R \leq R_0$ | Performs better for larger sample sizes.<br>Shows improvement when $R_0$ is close to 1. |
| 12 | $1 \leq R$        | $N_1 = N_2$ | alt: $R \geq R_0$ | Performs better for larger sample sizes.<br>Shows improvement when $R_0$ is close to 1. |



## 1.8 Data Sets Revisited

In this section we will analyze the data sets introduced in section 1.3 using the exact unconditional tests. Recall that in section 1.3, we had presented the p value and size for the asymptotic test. We had shown that the asymptotic test performed rather poorly (in terms of maintaining the nominal size). Although we have not discussed in 1.3 how to compute the p value and size for the asymptotic test, it is very similar to calculating the p value or size for any of the exact tests. If we denote the p value for the level  $\alpha$  asymptotic test as  $p_{as}$ , we can define it as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} (P_1/R_0)^{x_2} (1 - P_1/R_0)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z_\alpha]},$$

where  $D(R_0) = [0, \min(1, R_0)]$  is the domain of  $P_1$  and  $Z_\alpha$  is the  $\alpha$ th percentile of a standard normal distribution.

Also let us denote the rejection set for this test as  $R_{as}$ . Then for a level  $\alpha$  test,  $R_{as}$  will consist of the set of points  $(x_1, x_2)$  such that  $p_{as}(x_1, x_2) \leq \alpha$ , i.e.,

$$R_{as} = [(x_1, x_2) \mid p_{as}(x_1, x_2) \leq \alpha]$$

Using (1.13) or (1.14) and replacing the rejection set  $R_{sup}$  or  $R_\beta$  with  $R_{as}$  one can compute the size of the asymptotic test. Similarly we can use (1.15) to compute the power for the asymptotic test (replacing  $R_{sup}$  or  $R_\beta$  with  $R_{as}$ ).

**Animal Toxicology Data:** Here the toxicity of two different kinds of chemicals (the synthetic A, and the natural chemical) were compared. The synthetic A induced

Table 1.33:  $H_0 : R \leq 1$  versus  $H_1 : R > 1$ 

| Test      | p value | Significant at 5% | Size   |
|-----------|---------|-------------------|--------|
| $p_{sup}$ | 0.0809  | No                | 0.0363 |
| $p_\beta$ | 0.0246  | Yes               | 0.0489 |
| $p_{as}$  | 0.0218  | Yes               | 0.0899 |

tumor on 212 of the 350 rats while the natural chemical induced tumor on 37 of the 77 rats. We were interested in testing

$$H_0 : P_1 \leq P_2 \text{ versus } H_1 : P_1 > P_2$$

where,  $P_1$  and  $P_2$ , represents the true tumor rates for the synthetic A population and the natural population. So in terms of the relative risk we are interested in testing the null hypothesis  $R \leq 1$ . We tested the given hypothesis at  $\alpha = 0.05$  level of significance. In Table 1.33 we have presented the p value for all the three test (the two exact and the asymptotic test) and the size of the tests.

From Table 1.33 we see that both  $p_\beta$  and  $p_{as}$  are significant at 5% level, where as  $p_{sup}$  is not significant. At this point it would be interesting to look at the p value function plot to see what is going on. Note that one can obtain the p value function by plotting

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]} \quad (1.16)$$

against  $P_1$  over the entire domain of  $P_1$  (which is between 0 and  $\min(1, R_0)$ ).

In Figure 1.2, the two vertical dotted line represent the confidence interval for

the parameter  $P_1$ . Recall that the p value for the standard test  $p_{sup}$  is obtained by maximizing (1.16) over the entire domain of  $P_1$ . Whereas the confidence interval p value  $p_\beta$  is obtained by maximizing on the confidence set of the parameter. This is a good example which shows why  $p_\beta$  often performs better than  $p_{sup}$ . We see that this function takes a sharp peak at the extreme end of  $P_1$ . Thus, if one maximizes this function over the entire domain of the parameter space we will get the p value to be 0.0809. However if we consider maximizing on the confidence set of  $(P_1, P_2)$  we see that it is equivalent to maximizing on that narrow interval (0.512, 0.679) denoted by the two vertical lines. And within this region the maximum for the p value function comes out to be 0.0236. Adding the term  $\beta = 0.001$  makes the p value 0.0246. Therefore we see that how conservative the result can be if we use the standard exact unconditional test.

In Figure 1.3 we have plotted the size plot function for both of the exact tests and the asymptotic test. The size plot function for  $p_{sup}$  may be obtained by plotting

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} (P_1/R_0)^{x_2} (1 - P_1/R_0)^{N_2 - x_2} I_{[I(x_1, x_2) \in R_{sup}(\alpha)]}$$

against  $P_1$  over the entire domain of  $P_1$ . Similarly we can obtain the size plot function for the other two tests by replacing  $R_{sup}(\alpha)$  with the corresponding rejection sets. Notice how close the size plot function for  $p_\beta$  (dotted line) is to the nominal 5% level,

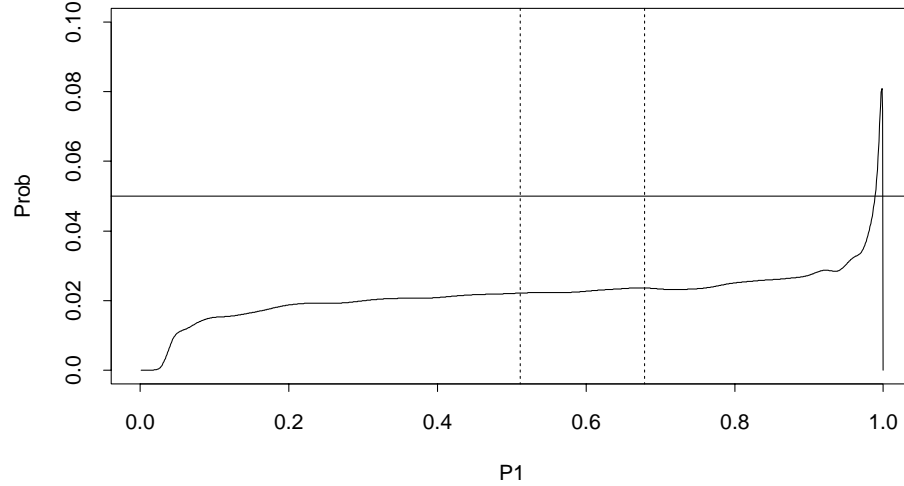


Figure 1.2: p value function plot for animal toxicology data:  $N_1 = 350, x_1 = 212, N_2 = 77, x_2 = 37$

compared to  $p_{sup}$ . This again illustrates the fact that there are so many extra points in the rejection set for  $p_\beta$  that are not in the rejection set for  $p_{sup}$ . Infact in this case the number of points in the  $p_{sup}$  rejection set is 10848 while for the  $p_\beta$  rejection set is 11454. Therefore there is a 5.58% increase in the rejection set size. Obviously for this example  $p_\beta$  is uniformly more powerful than  $p_{sup}$ . Also notice how the size function plot for the asymptotic test (broken line) exceeds the 5% nominal level for values of  $P_1$  above 0.55. This shows how unreliable the asymptotic test is even for moderately large sample sizes.

We used (1.15) for computing the power for  $p_{sup}$  and  $p_\beta$  for values of  $(P_1, P_2)$  in

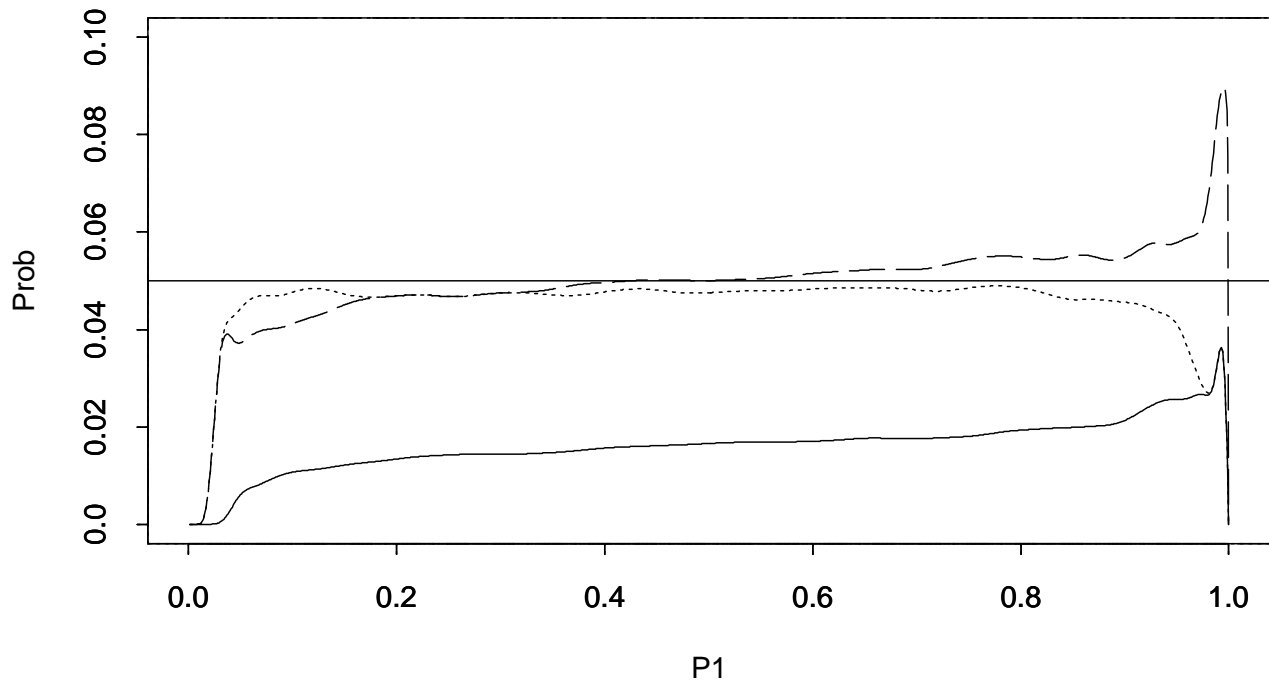


Figure 1.3: Size function plots for  $p_{sup}$  (Solid line),  $p_{\beta}$  (Dotted line), and Asymptotic test (Broken line) when  $(N_1, N_2) = (350, 77)$

Table 1.34: Power for  $p_{sup}$  and  $p_\beta$  for the animal toxicology data

| $P_1$ | $P_2$ | Power for $p_{sup}$ | Power for $p_\beta$ | % Change |
|-------|-------|---------------------|---------------------|----------|
| 0.3   | 0.2   | 0.350               | 0.559               | 59.7     |
| 0.6   | 0.5   | 0.302               | 0.476               | 57.6     |
| 0.4   | 0.2   | 0.908               | 0.968               | 6.6      |

the alternative hypothesis. The results are presented in Table 1.34. The last column (% Change) denotes the % increase (or decrease) in power of  $p_\beta$  compared to  $p_{sup}$ .

From Table 1.34 we see that there is an increase in power by as much as 60% for  $p_\beta$  compared to  $p_{sup}$ , which is quite remarkable.

**Childhood Nephroblastoma Data:** This data was obtained from a clinical trial studying two different kinds of treatment for childhood nephroblastoma. They were chemo therapy and radio therapy. For the chemo group, out of 88 patients, 83 had a rupture free tumor rate. For the radio group, out of 76 patients, 69 had a rupture free response. We were interested in testing whether radio therapy was non-inferior to chemo. And value of  $R_0$  chosen was 1.15. So the hypothesis of interest was,

$$H_0 : R \geq 1.15 \text{ versus } H_1 : R < 1.15.$$

In Table 1.35 we have presented the p values and the size for the three tests. Again the hypothesis was tested at 5% level of significance. We see that although all the three tests are not significant at 5% level, but both  $p_\beta$  and  $p_{as}$  are very close to being significant.

Now let us take a closer look at the p value function plot (Figure 1.4) for this test.

Table 1.35:  $H_0 : R \geq 1.15$  versus  $H_1 : R < 1.15$ 

| Test      | p value | Significant at 5% | Size   |
|-----------|---------|-------------------|--------|
| $p_{sup}$ | 0.04954 | No                | 0.0806 |
| $p_\beta$ | 0.0542  | No, Some evidence | 0.0489 |
| $p_{as}$  | 0.0539  | No, Some evidence | 0.0613 |

From this plot we see that again if we take a global maximum over  $P_1$  the p value comes out to be 0.0806. But if we maximize on the confidence set for  $(P_1, P_2)$  and therefore take the maximum within the confidence interval (0.838, 0.994) to calculate the p value, it comes out to be 0.0532. Adding the term  $\beta = 0.001$  makes the p value 0.0542. So both these examples illustrate why the confidence interval test is less conservative and can often result in a p value much lower than the standard p value.

In Figure 1.5 we present the size function plot for all the three tests. From the plot one can see that the size functions are almost identical for both the exact tests. This again illustrates the fact that for balanced (or nearly balanced) samples, the rejection sets often tend to be identical. We have seen that in the balanced cases the p value function tend to be closer to the nominal level and often the confidence interval for the parameter contains the global maximum. So for all those cases  $p_{sup}$  and  $p_\beta$  will be identical (only differing by the quantity  $\beta$ , which is negligible). For this example the number of points in the rejection set for  $p_{sup}$  was 3071 while that for  $p_\beta$  was 3069. Notice that the size plot function for the asymptotic test (broken

line) does quite well in this case only exceeding the nominal  $\alpha = 0.05$  line slightly (except for values of  $P_1$  between 0.05 and 0.09). The actual size of the asymptotic test is 0.0613.

Since the size plots for both the exact tests are almost identical, we did not calculate the power in this case knowing they would be almost same for both the test. Also the fact that the size functions are so close to the nominal level shows that both the test will have good power.

**Influenza Vaccine Data:** In this study the subjects were randomized to receive an experimental vaccine for influenza, or placebo. All subjects were monitored for viral infections. The data showed that for the vaccine group, 7 out of 15 patients got infected, and for the placebo group, 12 out of 15 patients got infected. We were interested in testing

$$H_0 : \pi \leq 0.1 \text{ versus } H_1 : \pi > 0.1$$

where  $\pi = 1 - P_1/P_2$ , i.e.,  $\pi = 1 - R$ .

Again we have presented the p value and the size for all the three tests. All tests were conducted at 5% level of significance. From the results in Table 1.36 we see that none of the tests are significant. Also the size for  $p_{sup}$  and  $p_\beta$  are same. In fact the rejection sets are identical in this case.

Let us take a closer look at the p value function plot (Figure 1.6). From this plot we see that the maximum of the p value function lies within the confidence interval



Table 1.36:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ 

| Test      | p-value | Significant at 5% | Size    |
|-----------|---------|-------------------|---------|
| $p_{sup}$ | 0.0856  | No                | 0.04331 |
| $p_\beta$ | 0.0866  | No                | 0.04331 |
| $p_{as}$  | 0.0636  | No                | 0.08371 |

of the nuisance parameter (0.296, 0.868). So there is no gain in using  $p_\beta$  instead of  $p_{sup}$ . Here  $p_\beta = 0.0866$ , and  $p_{sup} = 0.0856$ . The reason for the difference is due to the fact that the penalizing term of  $\beta = 0.001$  is being added because of the restricted maximization.

Figure 1.7 presents the size plot function for all the three tests. Since the rejection sets for both the exact tests are identical their size plot functions are identical as well. Here the number of points in the rejection set for both the exact tests were 60. Also notice how poorly the asymptotic test (Broken line) performs owing to the small sample size.

**Conclusion:** There are a couple of important things to note from these three examples. The main point of course is that  $p_\beta$  will perform much better than  $p_{sup}$  in many situations, however even when it does not perform better, it will be almost as good as  $p_{sup}$ . Second point to note is that when sample sizes are moderate to large (like the first two examples), the confidence interval becomes much narrower, and hence if one maximizes on that interval the chances of improvement (getting a lower p value) will be higher. From Figure 1.6 we see when the sample sizes are small

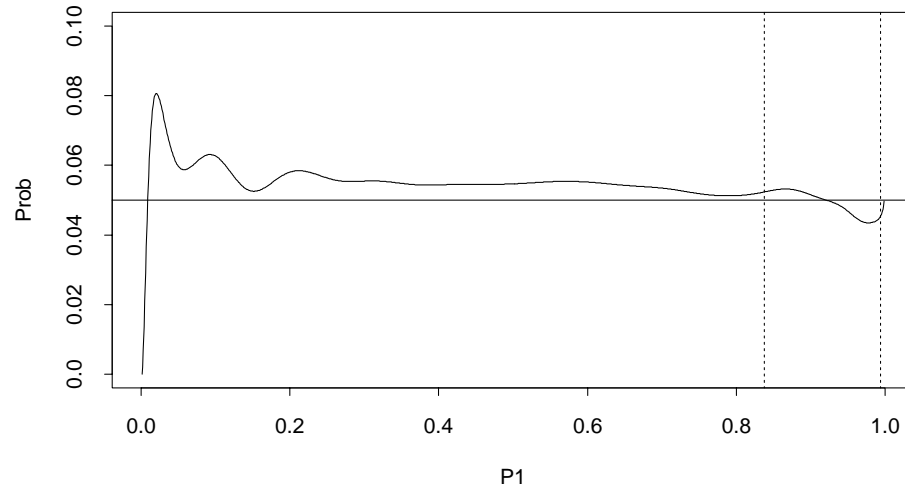


Figure 1.4: p value function plot for childhood nephroblastoma data:  $N_1 = 88, x_1 = 83, N_2 = 76, x_2 = 69$

the confidence interval is much wider and hence the actual maximum of the p value function lies within the interval. So  $p_\beta$  has a better chance of showing improvement for larger sample sizes. Therefore these three examples illustrate the fact that whenever  $p_\beta$  shows an improvement, the improvement is quite remarkable. However it is guaranteed to perform almost as good as the standard exact test.

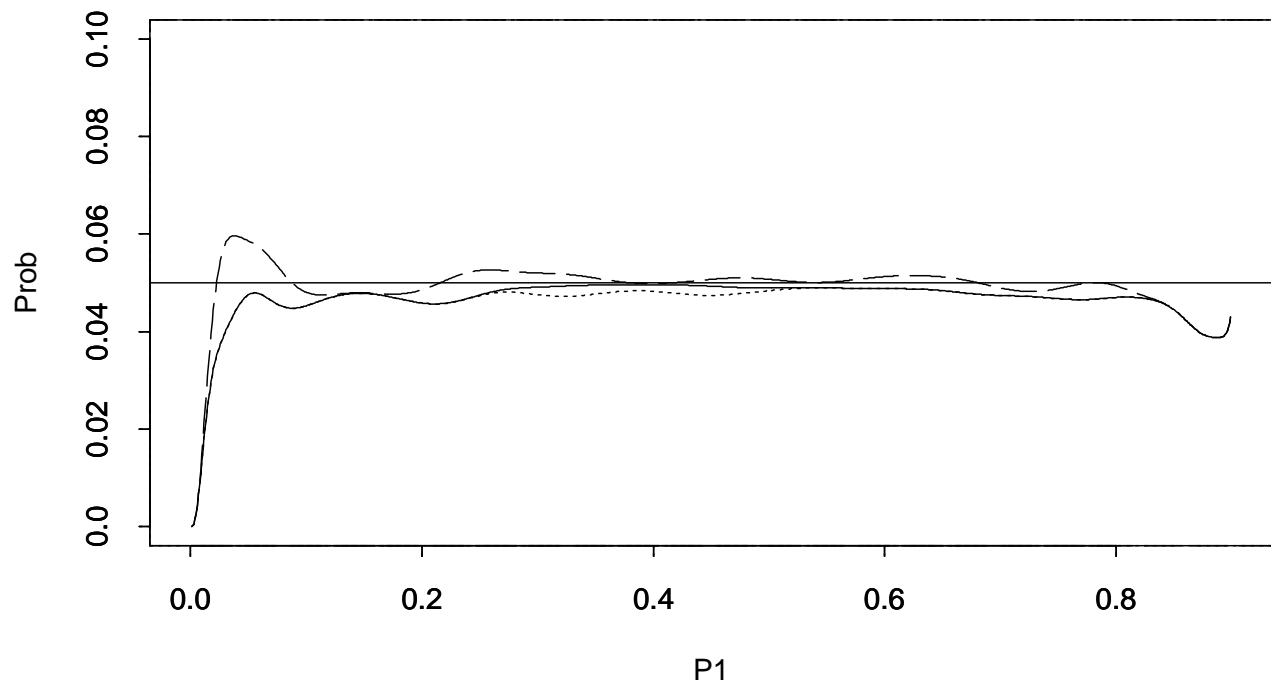


Figure 1.5: Size function plots for  $p_{sup}$  (Solid line),  $p_{\beta}$  (Dotted line), and Asymptotic test (Broken line) when  $(N_1, N_2) = (88, 76)$

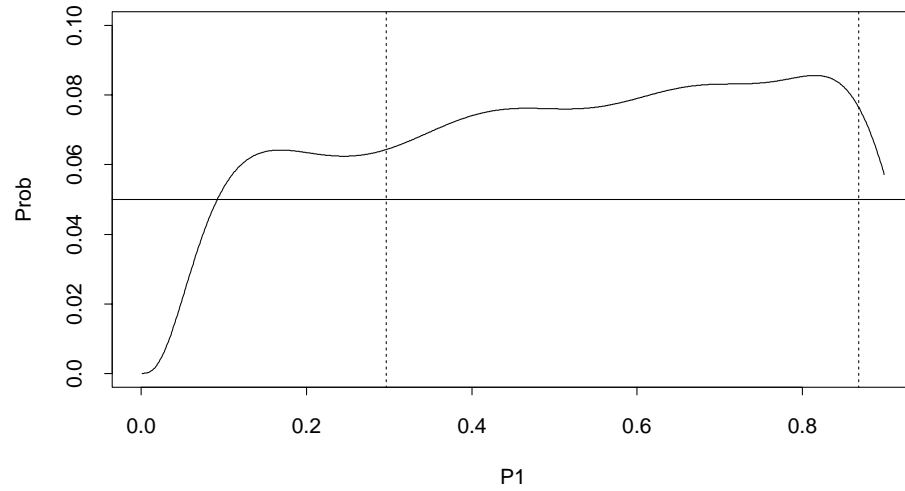


Figure 1.6: p value function plot for influenza vaccine data:  $N_1 = 15, x_1 = 7, N_2 = 15, x_2 = 12$

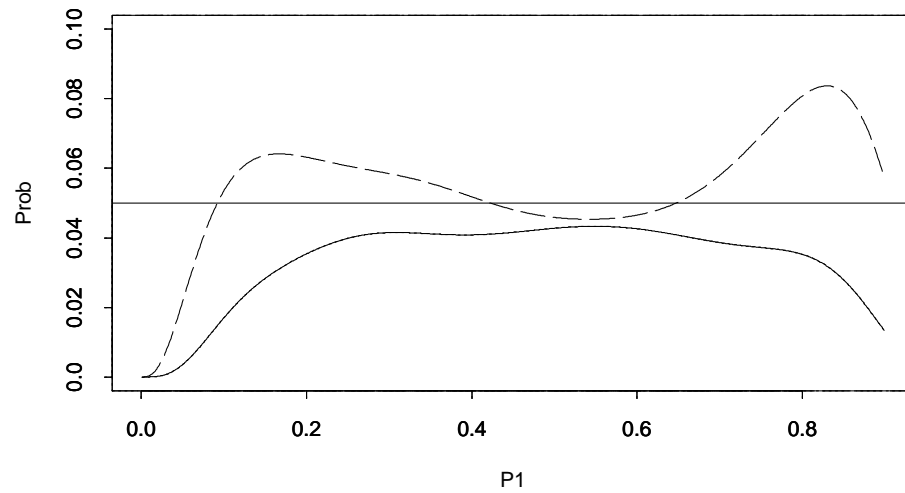


Figure 1.7: Size function plots for  $p_{sup}$  (Solid line),  $p_{\beta}$  (Dotted line), and Asymptotic test (Broken line) when  $(N_1, N_2) = (15, 15)$ . Here the size functions for  $p_{sup}$  and  $p_{\beta}$  are identical

## Chapter 2

# Exact Confidence Intervals for the Ratio of Two Binomial Proportions

### 2.1 Introduction

The relative risk parameter is often of interest in various applications such as epidemiology, clinical trials, public health, social sciences, etc. For example in various public health applications people are interested in knowing whether certain pollutants in the air might increase the chances of a disease by two fold. In clinical trials one of the main goal is to test whether the new drug performs better than an existing drug (or placebo) for curing a certain disease. The problem can be set up as (say), whether the new drug has a cure rate of 1.5 times the existing drug (or placebo). Or if we are evaluating the disease incidence rate (more appropriate in vaccine trials)

then people would be interested in testing whether the new treatment reduces the chances of occurrence of the disease by (say) 50%.

Note that in all these applications the problem may be set up in the following way. We have two groups, say the treatment group, and the placebo group. Our responses are dichotomous (such as, a success or a failure). Hence the data may be presented in a  $2 \times 2$  contingency table. If we denote by  $P_1$ , and  $P_2$ , the true proportion of success in each of the groups, then we can define the relative risk as  $R = P_1/P_2$ . For the various applications, mentioned above, confidence intervals for the relative risk are typically provided along with the estimate. This gives a good idea about the bounds within which the true value of the relative risk parameter  $R$  will lie.

Since the responses in each group are binary, one can use the binomial distribution to model the data. So in essence we are trying to construct confidence interval in the discrete distribution setting. This has been an old problem of interest, and to the best of our knowledge, there has not been any right or wrong method of doing this. One simple minded approach would be to ignore the discreteness completely, i.e., assume asymptotic normality and construct asymptotic intervals. The problem with this approach is that there is no guarantee that the coverage probability of the interval will be at least above a nominal level. The coverage probability gives the chance of an interval containing the true value of the parameter when the experiment is repeated for a large number of times. Typically one wants the coverage probability

to be at least greater than or equal to a certain nominal coverage (such as 95%). The problem with the asymptotic intervals are they may be very liberal and the coverage probability may fall far short of the nominal coverage.

The confidence intervals that are computed based on the actual distribution (instead of any asymptotic normality assumption) are typically called "Exact Confidence Intervals". These intervals may be constructed in such a manner that the coverage can be guaranteed to be at least  $100(1 - \alpha)\%$ . The usual method of constructing such intervals is to invert an equal tailed test. In many applications, such as in clinical trials or vaccine trials, or more generally in the pharmaceutical setting this guaranteeing of nominal coverage is a desirable property. However the main problem with the exact intervals is that they tend to be very conservative, especially for smaller sample sizes. By more conservative, we mean that the coverage probability is strictly greater than the nominal coverage and the lengths of the interval are very wide. Various authors have come up with different solutions to this problem. Vollset (1993), and Agresti & Coull (1998) suggested using approximate intervals based on the normal approximation. This approach gives shorter length, but does not guarantee nominal coverage. Another approach was to look for less conservative exact intervals. Authors such as Sterne (1954), Crow (1956), and, Blyth & Still (1983) studied intervals for the binomial distribution. Similarly Crow & Gardner (1959), Casella (1986) , and Casella & Roberts (1989) studied the problem in terms of the Poisson distribution. Their goal was to suggest various numerical process that gives intervals with shorter

length, yet maintaining the nominal coverage.

Our motive in this chapter will however be different. We will construct the confidence intervals by inverting two equal tailed one sided test. Chan & Zhang (1999) used this method to construct confidence intervals for the difference of proportion,  $\delta = P_1 - P_2$ , by inverting the standard exact unconditional test (denoted as  $p_{sup}$  in Chapter 1). Later Agresti & Min (2001) discussed the problem of constructing exact intervals for various parameters such as the difference, ratio and the odds ratio. Although people have studied in more details the odds ratio (see Agresti & Min, 2002), to the best of our knowledge no has studied the problem of constructing exact confidence intervals for the ratio of proportions in depth (especially for the tests that we are considering). Our goal is to show that for the relative risk parameter, one can construct improved exact confidence intervals by inverting the confidence interval p value test (denoted as  $p_\beta$  in Chapter 1) instead of the standard test. Recall that in Chapter 1 we had shown that the confidence interval p value test was superior to the exact unconditional test and gave a more powerful test. In this Chapter we will try to compare the confidence intervals generated by these two tests in terms of length and coverage probability. We will show that for small to moderate sample sizes the confidence intervals generated by  $p_\beta$  have shorter lengths and coverage probabilities closer to the nominal coverage.

Although we said that we will be inverting two one sided test to obtain the confi-



dence interval, it might be interesting to note that Agresti & Min (2001) have already shown that inverting a single two sided test resulted in less conservative intervals rather than inverting two one sided test. However since the goal of this chapter is to compare intervals generated by two different exact tests, we will argue that since this fact is true in general, the improvement (if any) that we will see by inverting  $p_\beta$  instead of  $p_{sup}$  will still be there had we considered a single two sided test.

One important point to mention is that, recall that in Chapter 1, we had explained how for the general case of testing for non-unity relative risk Fisher's exact test (Fisher, 1935) can not be computed. And hence people have to resort to only exact unconditional tests if they are to do exact inference for this problem. Extending the same idea we can see that one can not compute confidence intervals using Fisher's exact test but has to resort to confidence intervals that we are going to discuss. What we often see in practice is that people tend to present p values for the Fisher's exact test and asymptotic confidence intervals. However this is not a good thing to do because you are using two different methods to get the test and confidence interval and although one guarantees that the size of the test will not exceed the nominal size, the intervals does not guarantee that the coverage will be greater than or equal to the nominal coverage.

In section 2.2 we shall discuss how to formulate the problem statistically, and how to construct the exact confidence intervals. In section 2.3 we shall discuss how to

actually compute the confidence intervals. The results will be presented in section 2.4. And lastly in section 2.5 we shall apply the two confidence interval methods to the three data sets introduced in Chapter 1.

## 2.2 Statistical formulation of the problem

The problem in hand can be described as follows: Let  $X_1$  and  $X_2$  represent two independent response variables distributed as  $\text{bin}(N_1, P_1)$ , and  $\text{bin}(N_2, P_2)$ , respectively. Here  $N_1$  and  $N_2$  represents the sample sizes in each group, and  $P_1$ , and  $P_2$ , denotes the true response rates. If we denote by  $x_1$  and  $x_2$ , the observed number of successes in each group, then the binomial probability mass function of  $X_1$  and  $X_2$  will be

$$\text{bin}(N_1, x_1, P_1) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1}, \quad x_1 = 0, \dots, N_1,$$

and

$$\text{bin}(N_2, x_2, P_2) = \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2}, \quad x_2 = 0, \dots, N_2.$$

The sample space of  $(X_1, X_2)$  will be denoted as  $\mathbf{S} = \{0, \dots, N_1\} \times \{0, \dots, N_2\}$ . Thus there are  $(N_1 + 1) \times (N_2 + 1)$  points in  $\mathbf{S}$ .

Now consider testing the null hypothesis  $H_0 : R = R_0$ , and let  $Z$  be the test statistic used to formulate this test. The exact distribution of  $Z$  depends on all possible outcomes of two binomial responses given the sample sizes  $N_1$  and  $N_2$ . Probability

of a particular outcome is

$$Pr(X_1 = x_1, X_2 = x_2) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2}.$$

Replacing  $P_2$  by  $P_1/R_0$  (under the null hypothesis) we have

$$Pr(X_1 = x_1, X_2 = x_2) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} (P_1/R_0)^{x_2} (1 - (P_1/R_0))^{N_2 - x_2}. \quad (2.1)$$

Now we are in a position to construct the exact confidence interval for  $R$ . We will construct a two sided  $100(1 - \alpha)\%$  confidence interval for the true value of the relative risk parameter  $R$  by inverting two single one sided  $\alpha/2$  level test. This method guarantees an interval that will have at least  $100(1 - \alpha)\%$  coverage. Let us denote this interval by  $(R_L, R_U)$ . Here the two bounds are obtained by inverting the two one sided test. The upper bound  $R_U$  may be derived as follows. Consider the one sided hypothesis

$$H_0 : R = R_0 \text{ versus } H_1 : R < R_0, \quad (2.2)$$

for a pre-specified value of  $R_0$ . For this test we will reject the null hypothesis for a lower value of the test statistic. For a given sample point  $(x_1, x_2)$ , a test statistic  $Z$ , and a given value of  $R_0$ , let us denote the p value of an exact test for testing this hypothesis as  $p(x_1, x_2 \mid Z, R_0)$ . Since the test we are considering is an exact test, the size of the test will never exceed the nominal size. So for an  $\alpha/2$  level test, the null

hypothesis in (2.2) will be rejected if

$$p(x_1, x_2 \mid Z, R_0) \leq \alpha/2.$$

Now using the duality of hypothesis testing and interval estimation (Casella & Berger, 2002), one can define a  $100(1 - \alpha/2)\%$  confidence set for  $R$  based on the outcome  $(x_1, x_2)$  as

$$C_{\alpha/2}(x_1, x_2) = (R : p(x_1, x_2 \mid Z, R) > \alpha/2),$$

where

$$P(R \in C_{\alpha/2}(x_1, x_2)) \geq 1 - \alpha/2.$$

The confidence set  $C_{\alpha/2}(x_1, x_2)$  is an unique interval of the form  $[0, R_U)$ , if the tail probability function  $p(x_1, x_2 \mid Z, R)$  is a monotone decreasing function in  $R$ . In this case the upper bound  $R_U$  will be the value of  $R$  for which  $p(x_1, x_2 \mid Z, R) = \alpha/2$ . But if this p value function is not monotone in  $R$  then the confidence set might consist of multiple disjoint intervals. So to be on the conservative side, one can define the upper bound  $R_U$  as

$$R_U = \sup C_{\alpha/2}(x_1, x_2). \quad (2.3)$$

The interval  $[0, R_U)$  will contain the value  $R$  with at least  $(1 - \alpha/2)$  probability. Similarly if we consider the hypothesis

$$H_0 : R = R_0 \text{ versus } H_1 : R > R_0, \quad (2.4)$$

then one can define the lower interval using the same kind of argument. Here we will reject the null hypothesis for large values of the test statistic  $Z$ . One can define  $C_{\alpha/2}(x_1, x_2)$  in the same way. Note that here if we assume that  $p(x_1, x_2 \mid Z, R)$  is a monotone increasing function in  $R$  then the lower bound of the interval,  $R_L$  will be the value of  $R$  for which  $p(x_1, x_2 \mid Z, R) = \alpha/2$ . But if this p value function is not monotone in  $R$  then the confidence set might consist of multiple disjoint intervals. So to be on the conservative side one can define the lower bound  $R_L$  as

$$R_L = \inf C_{\alpha/2}(x_1, x_2) \quad (2.5)$$

Again the interval  $(R_L, \infty]$  will contain the value  $R$  with at least  $(1 - \alpha/2)$  probability.

Therefore we have

$$P(R_L < R < R_U) = 1 - Pr(R \leq R_L) - Pr(R \geq R_U) \geq 1 - \alpha/2 - \alpha/2 = 1 - \alpha$$

Now that we have described how we will construct the intervals the next step would be to describe the test statistics  $Z$  that we will use, and the exact tests that we will invert to get the confidence intervals. The test statistics suggested by Miettinen & Nurminen (1985), and Farrington & Manning (1990) for testing any of the hypothesis (2.2) or (2.4) is

$$Z = \frac{\hat{P}_1 - R_0 \hat{P}_2}{\hat{\sigma}},$$

where

$$\hat{\sigma} = \sqrt{\frac{\tilde{P}_1(1 - \tilde{P}_1)}{N_1} + (R_0)^2 \frac{\tilde{P}_2(1 - \tilde{P}_2)}{N_2}}. \quad (2.6)$$

Here  $\hat{P}_1 = x_1/N_1$ ,  $\hat{P}_2 = x_2/N_2$ , and  $\tilde{P}_1$  and  $\tilde{P}_2$  are the maximum likelihood estimates of  $P_1$  and  $P_2$  obtained under the null hypothesis restriction  $\frac{\tilde{P}_1}{\tilde{P}_2} = R_0$ .

We shall use this statistic for constructing the confidence intervals. However our main objective is to obtain confidence intervals from two different kinds of test. The standard exact unconditional test proposed by Chan (1998) and the confidence interval p value test proposed by Berger & Boos (1994). Although it will be a good idea to compare these two confidence intervals for different test statistics, we will be using only the test statistic mentioned above. Since both tests are exact, the intervals generated will be exact as well.

In Chapter 1 we have discussed in detail about unconditional tests and mentioned how the standard unconditional test and the confidence interval p value test are computed. Recall that for the standard exact test the p value is obtained by maximizing on the entire domain of the nuisance parameter space, while the confidence interval p value was obtained by maximizing on a confidence set for the nuisance parameter. We had defined both the p values for testing the alternative hypothesis  $H_1 : R < R_0$  as follows. The standard exact test p value,  $p_{sup}$  was defined as,

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}, \quad (2.7)$$

where  $D(R_0) = [0, \min(R_0, 1)]$  is the domain of  $P_1$ . The confidence interval p value  $p_\beta$  was defined as

$$\sup_{P_1 \in C_\beta^*} \left\{ \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1-P_1)^{N_1-x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2-x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]} \right\} + \beta, \quad (2.8)$$

where

$$C_\beta^* = \{\max(L_1, L_2 R_0), \min(U_1, U_2 R_0)\}$$

Recall that here  $(L_1, U_1)$  and  $(L_2, U_2)$  were the  $\sqrt{100(1-\beta)\%}$  confidence intervals for  $P_1$  and  $P_2$ , and  $C_\beta^*$  was the  $100(1-\beta)\%$  confidence set for  $(P_1, P_2)$ . For more details about the construction of the confidence interval p value (as well as the standard p value) please refer to Chapter 1. The main difference between these two p values were that for the standard p value we maximized on the entire domain of the parameter space to eliminate the nuisance parameter while for the confidence interval p value method we had maximized on a restricted set (the  $100(1-\beta)\%$  confidence set for  $(P_1, P_2)$ ) instead of the entire domain of  $(P_1, P_2)$ .

Since both the standard test and the confidence interval tests are exact tests their size will not exceed the nominal size. Hence inverting these two tests will result in confidence intervals that will maintain at least the nominal coverage.

## 2.3 Computing the exact confidence intervals

In this section we are going to discuss the main steps involved in computing the exact confidence intervals. Suppose we want to compute the upper bound  $R_U$ , for the interval. Then one has to invert the one sided test (2.2). For constructing the upper bound of a  $100(1 - \alpha)\%$  confidence interval for a particular outcome  $(x_1, x_2)$ , when the sample sizes are  $(N_1, N_2)$ , one may use the following steps.

1. Compute the p value over a series of values of the parameter  $R$  (covering the entire range of  $R$ ) and only keep those values of  $R$  for which the p values are greater than or equal to  $\alpha/2$ .
2. Choose the maximum of the value of  $R$  in step 2 for which the p value was greater than or equal to  $\alpha/2$ .

The main problem in implementing this algorithm in practice is that Step 2 will be computationally very intensive. This is because the entire range of  $R$  is  $[0, \infty)$  and if one tries to cover the entire range of  $R$  it will be a very tedious task. However it does take care of one technicality. Recall that while discussing about the construction of these intervals we had mentioned that there is no guarantee that the p value function will be monotone decreasing (or increasing, as the case may be) in  $R$ . In other words say if we start from zero and keep on computing the p values until we reach a value of  $R$  for which the p value is less than or equal to  $\alpha/2$ , there is no guarantee that



higher values of  $R$  will generate p values that are less than or equal to  $\alpha/2$ . It is always theoretically possible that there will be another value of  $R$  (or a set of values of  $R$ ) for which the p value function is greater than  $\alpha/2$ . That is why we defined the upper bound of the interval as the supremum over the set of  $R$  for which the p value was less than or equal to  $\alpha/2$  (see 2.3). If we follow this algorithm since we are considering the entire range of  $R$  this problem is automatically taken care of.

We see that if we try to compute the confidence interval very rigorously it will be computationally very intensive and will not have much practical appeal. So instead this is the approach we implemented. Suppose we want to construct the upper bound of the interval for a given observed response  $(x_1, x_2)$ . And let us denote by  $\hat{R} = \hat{P}_1/\hat{P}_2$  the maximum likelihood estimate of  $R$  for sample sizes  $(N_1, N_2)$  and observed response  $(x_1, x_2)$ .

1. First compute the p value for the exact test ( $p_{sup}$  or  $p_\beta$ ) for  $R = \hat{R}$ .
2. Consider a very fine grid of values of  $R$  starting from  $R = \hat{R}$  and increasing in value. Compute the p value over this fine grid of points until the p value reaches  $\alpha'$ , where  $\alpha'$  is less than  $\alpha/2$ . Only keep those values of  $R$  for which the p values are greater than or equal to  $\alpha/2$ .
3. Choose the maximum of the value of  $R$  (amongst the set of values kept in step 2) for which the p value was greater than or equal to  $\alpha/2$ . This will give  $R_U$ , the upper bound of the interval.

So what are the main differences between the original algorithm and our algorithm.

1. Instead of starting from  $R = 0$  we are starting from  $\hat{R}$ . This is a reasonable thing to do because by doing this we are only assuming that the confidence interval will contain the maximum likelihood estimate of the true value of  $R$ .
2. We are assuming that if we consider values of  $R$  until the p value falls to  $\alpha'$  that should be sufficient, i.e., we are assuming that once it has reached  $\alpha'$  after that the p value function will not increase again to  $\alpha/2$ . Although there is nothing to guarantee that this will be the case but this seems to be a very reasonable thing to do in practice. We had done a thorough investigation regarding the monotonicity of the p values with respect to  $R$ . In all the cases we have considered we found that the p value was monotonic decreasing in  $R$  (when we are considering the upper bound) and there was no fluctuation what so ever. For implementing this in the construction of a 95% confidence interval we chose  $\alpha'$  to be 0.01. Note that here  $\alpha/2$  will be 0.025. So what we are saying is that consider the set of  $R$  for which the p value reaches 0.01, and from that set choose the maximum value of  $R$  for which the p value was greater than or equal to 0.025.

For implementing this method first we computed the p values for both the standard test ( $p_{sup}$ ) and the confidence interval test ( $p_\beta$ ). We considered a grid size of 300 equally divided spaces over which we maximized to eliminate the nuisance parameter.

Recall that in Chapter 1 we had suggested to take a grid size of 1000. But owing to the computational complexity we considered 300 to be a reasonable compromise (between accuracy and computing time). The values of  $R$  considered were in the increments of 0.01. For example if we start from  $\hat{R} = 5$ , then we consider values of 5.01, 5.02,...,etc. So our intervals will be accurate to two decimal places. Although we have not talked about the construction of the lower bound  $R_L$ , it is very similar except for the fact that here we start from  $R = \hat{R}$ , and continue moving downwards, i.e., choose smaller values of  $R$ . Of course here the p value will tend to decrease with decrease in  $R$  and in step 2 we shall consider the lowest value of  $R$  for which the p value was greater than or equal to 0.025.

In the next section we will present the results for both the confidence intervals. As we have stated earlier the results will be compared in terms of the length of the confidence interval and the coverage probabilities.

## 2.4 Results

In this section we will present the results from comparing the two different exact intervals. We will divide this section into two parts. In the first part we will present the lengths of the confidence intervals, and in the second part present the coverage probabilities. We have chosen a combination of balanced and unbalanced samples. The sample sizes we will be discussing are:

- (10,10), (15,15), (20,20), (30,30), and (40,40) for the balanced case.
- (10,20), (10,40), (40,10), and (20,40) for the unbalanced case.

All confidence intervals constructed are 95% intervals. They have been obtained by inverting two 0.025 level test (as discussed in section 2.2). All computations have been done using the SAS Software, Version 8.1, and StatXact, Version 5.

### 2.4.1 Length of the interval

When people compare confidence intervals they generally talk about two aspects of the interval. First is the length, and second is the coverage of the interval. If we consider  $(R_L, R_U)$  to be a confidence interval for  $R$ , then the length of the interval will be  $R_U - R_L$ . In Table 2.1 we are presenting values for the mean length of the confidence intervals. By mean length we are simply referring to the arithmetic mean of the lengths of all possible intervals when the sample sizes are  $(N_1, N_2)$ . Now if the sample sizes are  $(N_1, N_2)$ , then there are  $(N_1 + 1)(N_2 + 1)$  possible pairs of observation that can be obtained. And the average of the length of the confidence intervals of all these points will give the mean length of the interval.

One important point to note is that since these intervals are guaranteed to maintain at least the nominal coverage, for the cases when the control group has 0 responses, i.e., when the number of success for the  $N_2$  group is 0, the upper bound of the interval will be infinity. For computing the mean length we will discard these

Table 2.1: Comparison of the mean lengths for the two confidence intervals

| $N_1$ | $N_2$ | Mean length for $p_{sup}$ CI | Mean length for $p_\beta$ CI | Difference in length |
|-------|-------|------------------------------|------------------------------|----------------------|
| 10    | 10    | 19.77                        | 19.80                        | 0.03                 |
| 15    | 15    | 20.11                        | 19.66                        | 0.45                 |
| 20    | 20    | 20.52                        | 19.74                        | 0.78                 |
| 30    | 30    | 21.42                        | 19.98                        | 1.44                 |
| 40    | 40    | 22.12                        | 20.90                        | 1.22                 |
| 10    | 40    | 20.89                        | 21.21                        | 0.32                 |
| 40    | 10    | 21.07                        | 19.65                        | 1.42                 |
| 20    | 40    | 21.00                        | 20.12                        | 0.88                 |

cases. So if our sample sizes are  $(N_1, N_2)$ , we will only use  $(N_1 + 1)(N_2 + 1) - (N_2 + 1)$  points in calculating the mean length. Also if the number of responses in the treatment group is 0, i.e., the number of success for the  $N_1$  group is 0, then the lower bound of the interval will be 0.

Table 2.1 shows the mean lengths for the two confidence intervals for the different sample sizes along with their absolute difference. Note that except for the case of (10,10) and (10,40) in all the other cases the  $p_\beta$  interval has a shorter mean length compared to the  $p_{sup}$  interval. And for both these cases when  $p_{sup}$  interval has a shorter mean length, the absolute difference is only 0.03 and 0.32.

Now it would be interesting to look into more details in the individual cases and see for which sample points the lengths of one interval is shorter than the other. From what we have seen from the various cases, the general pattern is that for a specific sample size of  $(N_1, N_2)$ , out of the  $(N_1 + 1)(N_2 + 1)$  possible cases, a large number

of them have almost the same interval length. And the remaining points are divided between the cases where the  $p_\beta$  lengths were shorter than  $p_{sup}$  and vice versa. Let us assume that for all the cases where the difference in the two interval lengths is not more than 0.1, the interval lengths are approximately same. In Table 2.2, for the different  $(N_1, N_2)$  pairs, we have summarized the number of sample points for which the  $p_\beta$  interval lengths were shorter than  $p_{sup}$  intervals, number of points where they were approximately same, and number of points where  $p_{sup}$  interval lengths were shorter than  $p_\beta$  intervals. Later we will look in to details in some of the cases and try to detect any general pattern of behavior.

We have used the following notations in Table 2.2. Let us denote by  $L(p_{sup})$  the length of the  $p_{sup}$  interval and by  $L(p_\beta)$  the length of the  $p_\beta$  interval.  $L(p_{sup}) - L(p_\beta) > 0.1$  means length of the  $p_{sup}$  interval is greater than  $p_\beta$  by at least 0.1. Similarly  $L(p_\beta) - L(p_{sup}) > 0.1$  means length of the  $p_\beta$  interval is greater than  $p_{sup}$  by at least 0.1. And  $|L(p_{sup}) - L(p_\beta)| < 0.1$  means that the absolute difference between the two lengths is less than 0.1. The first column  $(N_1, N_2)$  of Table 2.2 denote the sample sizes in each group. The second column  $(N_1 + 1)(N_2 + 1)$  denote the total number of possible sample points. The third column shows all the three cases (length of  $p_\beta$  being shorter, approximately equal to, or greater than  $p_{sup}$ ). And in the last column we are presenting the number (and % relative to the total number of points) of sample points in each of the three cases.

Note that from Table 2.2 we see that for a greater number of sample points (in each case) the interval lengths are approximately same. In fact if we average across all the  $(N_1, N_2)$  cases considered we have for 65.5% cases the absolute difference between the interval lengths are less than 0.1. In 27.2% of the cases length of the  $p_\beta$  intervals were shorter than the  $p_{sup}$  intervals by at least 0.1. And in the remaining 7.3% of the cases the  $p_{sup}$  intervals were shorter than the  $p_\beta$  intervals by at least 0.1.

Now let us take a closer look at some of the specific cases. Let us consider (15,15) as an example. In Table 2.3 we have presented 10 sample points where the lengths of the  $p_\beta$  intervals were shorter than the  $p_{sup}$  intervals. And in Table 2.4, 10 sample points where the lengths of the  $p_{sup}$  intervals were shorter than the  $p_\beta$  intervals have been presented. Note that the 10 cases presented in these two tables are the 10 most extreme cases, i.e., the ones with the maximum difference in the interval lengths. One interesting point to note is that for all the cases where the  $p_\beta$  intervals were shorter than the  $p_{sup}$  intervals, the actual lengths for both the intervals were much shorter (see Table 2.3) in comparison to the actual lengths for the cases where the  $p_{sup}$  intervals were shorter than the  $p_\beta$  intervals (see Table 2.4). For example consider the case in Table 2.3 where we have the sample sizes to be (15,15) and the observed sample points  $(x_1, x_2)$  to be (14,2). In this case length for the  $p_{sup}$  interval is 75.02, while that for  $p_\beta$  is 47.32. Now let us look at Table 2.4 where again  $(N_1, N_2) = (15, 15)$  and  $(x_1, x_2) = (15, 1)$ . Here the length for the  $p_{sup}$  interval is 589.83, while that for  $p_\beta$  is 602.08. We can see the big difference in the actual lengths for the two cases.

Also if we look across both the tables we find a pattern, i.e., we see that in general when ever lengths for the  $p_{sup}$  interval was shorter than the  $p_{\beta}$  interval the actual lengths for both the intervals were quite high in comparison to the other case where  $p_{\beta}$  intervals were shorter than the  $p_{sup}$  ones. To drive this point home we define a standardized measure where we divide the difference in the two lengths by the average of the lengths of the two confidence intervals. Let us call this standardized measure as Relative Change. Then one can define the Relative Change as

$$2 \times \frac{|L(p_{sup}) - L(p_{\beta})|}{L(p_{sup}) + L(p_{\beta})}.$$

We see that for all the cases in Table 2.3, the Relative Changes are closer to 0.40, while for Table 2.4 it is only 0.02. These numbers clearly indicate that when ever the  $p_{\beta}$  intervals are shorter, the actual lengths of both the intervals are not that high, making the Relative Change much higher compared to the cases where  $p_{sup}$  intervals are shorter. Now this is a trend that we have observed in general, i.e., for all the different pairs of sample sizes  $(N_1, N_2)$  that we have considered we have seen this pattern. So what does this mean from a practical point of view. If some one is reporting confidence intervals with very large lengths, then it does not really matter if there is a slight improvement. For example let us look at the case in Table 2.4 where we have the sample sizes to be 15 each and the observed points to be (15,1). It will not be very informative if we say that  $R$  will lie between (3.13,592.96) instead of (3.11,605.20) as the upper bound is so high for both the cases. Now let us look at



the case where  $(N_1, N_2) = (15, 15)$  and  $(x_1, x_2) = (14, 2)$  (see Table 2.3). Here it does help if some one says  $R$  lies between  $(2.37, 49.69)$  instead of  $(2.38, 77.41)$ . In this case it is much more informative to learn that the upper bound for  $R$  is 49.69 instead of 77.41.

Another important point to note is that for all the sample points in Table 2.4 the control group has only a single response, i.e., number of successes for the  $N_2$  group is only one. Although from Table 2.3 we see that for the other situation also (when  $p_\beta$  intervals are shorter) the responses in the control group are quite small but it is definitely not as extreme as in Table 2.3. This is the same trend we have observed in all the cases we have studied, i.e.,  $p_{sup}$  intervals tend to be shorter in these extreme cases when the response in the control group is only one. Since the intervals are exact intervals, to maintain the right coverage, for these extreme cases the interval lengths tend to be very high. That is why when ever we see an improvement from the  $p_{sup}$  intervals, the actual lengths of both the intervals are very high, and hence the Relative Changes are very small. However if we come up with some applications where we do expect the control group to have very small responses then the standard method certainly gives shorter intervals compared to the  $p_\beta$  intervals. For example let us consider the responses to be the cure rates for a certain drug, and placebo, and the drug has been proven to be very successful. Also let us assume that the chances of recovery without actual treatment is extremely low. In that case one can expect very low response from the placebo group, and high response from the treatment group.

Table 2.2: Table summarizing number of sample points for which  $p_\beta$  intervals are shorter by 0.1, longer by 0.1, or have an absolute difference less than 0.1 compared to  $p_{sup}$  intervals

| $(N_1, N_2)$ | $(N_1 + 1)(N_2 + 1)$ | $L(p_{sup})$ vs. $L(p_\beta)$   | Number (%) of points               |
|--------------|----------------------|---|------------------------------------|
| (10,10)      | 121                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 12(9.9)<br>19(15.7)<br>90(74.4)    |
| (15,15)      | 256                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 46(18)<br>22(8.6)<br>188(73.4)     |
| (20,20)      | 441                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 90(20.4)<br>30(6.8)<br>321(72.8)   |
| (30,30)      | 961                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 332(34.5)<br>37(3.9)<br>592(61.6)  |
| (40,40)      | 1681                 | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 534(31.8)<br>47(2.8)<br>1100(65.4) |
| (10,20)      | 231                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 18(7.8)<br>31(13.4)<br>182(78.8)   |
| (10,40)      | 451                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 267(59.2)<br>44(9.8)<br>140(31.0)  |
| (40,10)      | 451                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 142(31.5)<br>29(6.4)<br>280(62.1)  |
| (20,40)      | 861                  | $L(p_{sup}) - L(p_\beta) > 0.1$<br>$L(p_\beta) - L(p_{sup}) > 0.1$<br>$ L(p_{sup}) - L(p_\beta)  < 0.1$ | 270(31.4)<br>27(3.1)<br>564(65.5)  |

Table 2.3: Comparison of Confidence Intervals for  $p_{sup}$  and  $p_\beta$  when  $(N_1, N_2)=(15, 15)$   
- Cases where  $p_\beta$  has a shorter interval length

| $X_1$ | $X_2$ | $CI(p_{sup})$ | $CI(p_\beta)$ | $L(p_{sup})$ | $L(p_\beta)$ | Relative Change |
|-------|-------|---------------|---------------|--------------|--------------|-----------------|
| 11    | 4     | (1.16, 14.47) | (1.16, 8.72)  | 13.31        | 7.56         | 0.5504          |
| 10    | 2     | (.35, 54.05)  | (1.49, 45.76) | 52.70        | 44.28        | 0.1737          |
| 15    | 3     | (2.08, 32.85) | (2.07, 23.26) | 30.78        | 21.19        | 0.3688          |
| 13    | 3     | (1.79, 27.75) | (1.79, 17.61) | 25.95        | 15.81        | 0.4473          |
| 13    | 2     | (2.07, 55.55) | (2.07, 44.58) | 53.47        | 42.50        | 0.2190          |
| 14    | 3     | (1.98, 32.85) | (1.97, 19.45) | 30.86        | 17.47        | 0.5123          |
| 12    | 2     | (1.94, 54.05) | (1.81, 40.21) | 52.11        | 38.39        | 0.2937          |
| 15    | 2     | (2.47, 77.41) | (2.46, 60.98) | 74.93        | 58.52        | 0.2973          |
| 11    | 2     | (1.74, 54.05) | (1.61, 36.22) | 52.31        | 34.60        | 0.3949          |
| 14    | 2     | (2.38, 77.41) | (2.37, 49.69) | 75.02        | 47.31        | 0.4361          |
| 15    | 2     | (2.47, 77.41) | (2.46, 60.98) | 74.93        | 58.52        | 0.3949          |
| 11    | 2     | (1.74, 54.05) | (1.61, 36.22) | 52.31        | 34.60        | 0.4361          |
| 14    | 2     | (2.38, 77.41) | (2.37, 49.69) | 75.02        | 47.31        | 0.4361          |

Table 2.4: Comparison of Confidence Intervals for  $p_{sup}$  and  $p_\beta$  when  $(N_1, N_2)=(15, 15)$   
- Cases where  $p_{sup}$  has a shorter interval length

| $X_1$ | $X_2$ | $CI(p_{sup})$  | $CI(p_\beta)$  | $L(p_{sup})$ | $L(p_\beta)$ | Relative Change |
|-------|-------|----------------|----------------|--------------|--------------|-----------------|
| 6     | 1     | (1.04, 157.64) | (1.02, 160.90) | 156.60       | 159.88       | 0.0207          |
| 7     | 1     | (1.25, 187.47) | (1.30, 191.34) | 186.22       | 190.04       | 0.0205          |
| 8     | 1     | (1.36, 218.87) | (1.61, 223.39) | 217.51       | 221.78       | 0.0205          |
| 9     | 1     | (1.81, 251.95) | (1.83, 257.15) | 250.13       | 255.32       | 0.0202          |
| 10    | 1     | (2.07, 286.94) | (2.09, 292.87) | 284.86       | 290.77       | 0.0205          |
| 11    | 1     | (2.36, 324.22) | (2.35, 330.92) | 321.86       | 328.56       | 0.0204          |
| 12    | 1     | (2.47, 364.47) | (2.47, 372.00) | 362.00       | 369.00       | 0.0204          |
| 13    | 1     | (2.87, 409.10) | (2.86, 417.55) | 406.23       | 414.69       | 0.0203          |
| 14    | 1     | (3.05, 462.11) | (3.04, 471.65) | 459.05       | 468.60       | 0.0203          |
| 15    | 1     | (3.13, 592.96) | (3.11, 605.20) | 589.83       | 602.08       | 0.0205          |
| 13    | 1     | (2.87, 409.10) | (2.86, 417.55) | 406.23       | 414.69       | 0.0203          |
| 14    | 1     | (3.05, 462.11) | (3.04, 471.65) | 459.05       | 468.60       | 0.0203          |
| 15    | 1     | (3.13, 592.96) | (3.11, 605.20) | 589.83       | 602.08       | 0.0205          |

Table 2.5: Comparison of the mean lengths for the two confidence intervals after eliminating the sample points with small control group response

| $N_1$ | $N_2$ | Mean length for $p_{sup}$ CI | Mean length for $p_\beta$ CI | Difference in Length |
|-------|-------|------------------------------|------------------------------|----------------------|
| 10    | 10    | 11.75                        | 9.80                         | 1.95                 |
| 15    | 15    | 8.30                         | 7.57                         | 0.73                 |
| 20    | 20    | 7.22                         | 6.17                         | 1.05                 |
| 30    | 30    | 6.86                         | 5.08                         | 1.78                 |
| 40    | 40    | 5.86                         | 5.05                         | 0.81                 |
| 10    | 40    | 5.68                         | 5.86                         | 0.18                 |
| 40    | 10    | 11.84                        | 9.44                         | 2.40                 |
| 20    | 40    | 6.28                         | 5.09                         | 1.19                 |

Although we have seen that for larger sample sizes the  $p_\beta$  interval shows improvement in lesser extreme cases (both treatment and control group responses not very high or low) the improvement for  $p_{sup}$  intervals have always been for the extreme cases only.

In Table 2.5 we recalculated the mean lengths for both the intervals after eliminating some of the extreme cases (where the control group response is only one). Here we have discarded all the cases where the upper bound  $R_U$  for the interval is greater than 200. Now we see that the improvement for the  $p_\beta$  intervals are more prominent and they have a much shorter mean length. In fact in only one of the cases where the observed points are (10,40) that the  $p_{sup}$  interval has a shorter mean length. And the difference is only by 0.18.

### 2.4.2 Coverage Probabilities

For the interval  $(R_L, R_U)$  of  $R$  one can define the coverage probability as the probability of the interval containing the true value of  $R$ . It is always desirable to have the coverage probability to be at least greater than or equal to a certain nominal value. For example if we are constructing 95% confidence intervals, it is desirable to have the coverage probability to be at least 0.95, i.e.,  $P(R_L < R < R_U) \geq 0.95$ . In section 2.3 we have discussed that the way we have constructed the confidence intervals, the coverage will never fall below a chosen nominal coverage. However the problem with exact intervals are that, they tend to be very conservative, i.e., although the coverage will not go below the nominal coverage, their coverage probability values are closer to one. In order to maintain a very high coverage the confidence interval has to be very wide.

As a compromise what we want is intervals with shorter length and coverage not falling below the nominal coverage, but as close to the nominal coverage as possible. One can look at the problem as trying to reduce the confidence interval length after controlling for the coverage. So in our comparison of coverage probabilities, if one interval is closer to the nominal coverage (95% in our examples) then it is better. In this section we will mainly present coverage probability plots. The best way to compare coverage probabilities of two interval is through plots, as they give the bigger picture (rather than presenting numbers such as mean coverage probability values).

We will be presenting two different kinds of coverage probability plot. One we will call the smooth plot, and the other the the jagged plot. The main difference is as follows. The smooth plot reflects the coverage probability for a fixed value of  $R$ . The jagged plot reflects the coverage probability for a set of values of  $R$ . For example, we know that  $R = P_1/P_2$ . So if one decides to fix  $P_2$ , and vary across the values of  $P_1$  then we will obtain a series of values of  $R$ . And if we plot the coverage probabilities for each of the values of  $R$  (in the same plot), it will take a very jagged appearance.

At this point it will be appropriate to tell a little bit about how to actually get the plots. For the smooth plot, if we have sample sizes  $(N_1, N_2)$  and we want to get a coverage probability plot for  $R$ , first we will obtain all the sample points  $(x_1, x_2)$  for which the corresponding interval contains  $R$ . So one can define this set of points as

$$A_R = \{(x_1, x_2) : R_L(x_1, x_2) \leq R \leq R_U(x_1, x_2)\},$$

where  $R_L(x_1, x_2)$ , and  $R_U(x_1, x_2)$  denotes the corresponding confidence intervals for  $R$  when the sample points are  $(x_1, x_2)$ . Once we obtain  $A_R$  one can compute the coverage probabilities by evaluating the binomial probabilities (under the null hypothesis) for the set of points in  $A_R$ . Thus one can compute these probabilities for a particular value of  $(P_1, P_2 = P_1/R_0)$ , as

$$\sum_{A_R} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2}. \quad (2.9)$$

The smooth plot may be obtained by plotting the probabilities obtained from (2.9)

against the values of  $P_1$ .

As we mentioned earlier the jagged plots represent coverage probabilities for a set of values of  $R$ . For example suppose we fix  $P_2 = 0.33$ , and vary over  $P_1$  between 0 and 1. Then the range of  $R$  will lie between 0 and 3. In order to obtain the jagged plots one has to first evaluate  $A_R$  for a series of values of  $R$  between 0 and 3. The final plot is obtained by plotting the probabilities (obtained for each distinct  $A_R$ ) against the corresponding  $P_1$ . Since with the change of value of  $P_1$  (we have a new value of  $R$  as  $R = P_1/P_2$ ,  $P_2$  being fixed over here) , we are calculating the probability over a different  $A_R$ , which brings the jagged appearance in the plots. In order to maintain accuracy one has to plot the jagged plots over a very fine grid for the values of  $P_1$ . For our plots we have considered a grid size of 0.001 and divided the range of  $P_1$  (which is 0 to 1) in to 1000 equally spaced points, i.e., 0, 0.001,...,1. We evaluate  $A_R$  for all these different values of  $R(0/P_2, 0.001/P_2, ..., 1/P_2)$ , and compute the probability from (2.9) for each of this  $A_R$ . This probabilities when plotted against  $P_1$  gives the jagged plot.

In Figures 2.1 - 2.9 we have presented the coverage probability plots for  $(N_1, N_2) = (15,15), (20,20)$ , and  $(10,20)$ . For the smooth plots we have considered values of  $R=0.9, 2, 3$ , and 10. For the jagged plots we have fixed  $P_2=0.3$ , and varied over  $P_1$  between 0 and 1. Therefore for these plots  $R$  will vary between 0 and 3.33.

**(15,15):** In Figures 2.1, and 2.2, we have plotted the coverage probabilities for values

of  $R = 0.9, 2, 3$ , and  $10$ . We see that for  $R = 0.9, 3$ , and  $10$ , the coverage probabilities for the  $p_\beta$  interval are closer to the nominal coverage. In these 3 cases the  $A_R$  set for  $p_\beta$  is a strict subset of the  $A_R$  set for  $p_{sup}$ . For  $R = 2$   $A_R$  for both the tests are identical. In Figure 2.3, the jagged plot presents the coverage probability for  $R$  ranging from 0 to 3.33. We see that the coverage for the  $p_\beta$  interval (the dotted line) tends to be closer to the nominal 95% line compared to the  $p_{sup}$  interval coverage (the solid line).

**(20,20):** In Figures 2.4, and 2.5, again we see that for  $R = 0.9, 3$ , and  $10$  the coverage probabilities for the  $p_\beta$  interval is closer to the nominal coverage, while for  $R = 2$  both of them have the same coverage. Again for these 3 cases (where  $R = 0.9, 3$ , and  $10$ )  $A_R$  for  $p_\beta$  is a complete subset of  $p_{sup}$  while for  $R = 2$  they are identical. In Figure 2.6, the jagged plot presents the coverage probability for  $R$ , where the value of  $R$  ranges from 0 to 3.33. We see that the coverage for the  $p_\beta$  interval (the dotted line) tend to be closer to the nominal 95% line compared to the  $p_{sup}$  interval coverage (the solid line) over the entire parameter space of  $P_1$ .

**(10,20):** From Figures 2.7, and 2.8, we see that for  $R = 3$ , and  $10$  the coverage probabilities for both the intervals are identical (which means that the  $A_R$  sets are identical as well). Although for  $R = 0.9$ , there is an improvement for  $p_\beta$  (in terms of being closer to the nominal coverage), the roles are reversed for  $R = 2$ , i.e. in the former case  $A_R$  for  $p_\beta$  is a subset for  $p_{sup}$  and vice versa. From the jagged plot (Figure



2.9) we see that there is not much improvement and both the solid line (for  $p_{sup}$ ) and the dotted line (for  $p_\beta$ ) is very close to each other. There is some improvement for the  $p_\beta$  interval around values of  $P_1$  between 0.2, and 0.3.

We have plotted the coverage probabilities for all the sample sizes  $(N_1, N_2)$  that we had considered in section 2.4.1. We have seen that in general the coverage probabilities for the  $p_\beta$  interval tend to be closer to the nominal coverage. Again the improvement increases with increase in sample size. For very small sample sizes like (10,10) or (10,20), there is not much improvement. There are also some cases where the  $p_{sup}$  interval tends to be closer to the nominal coverage, but these are much fewer in number.

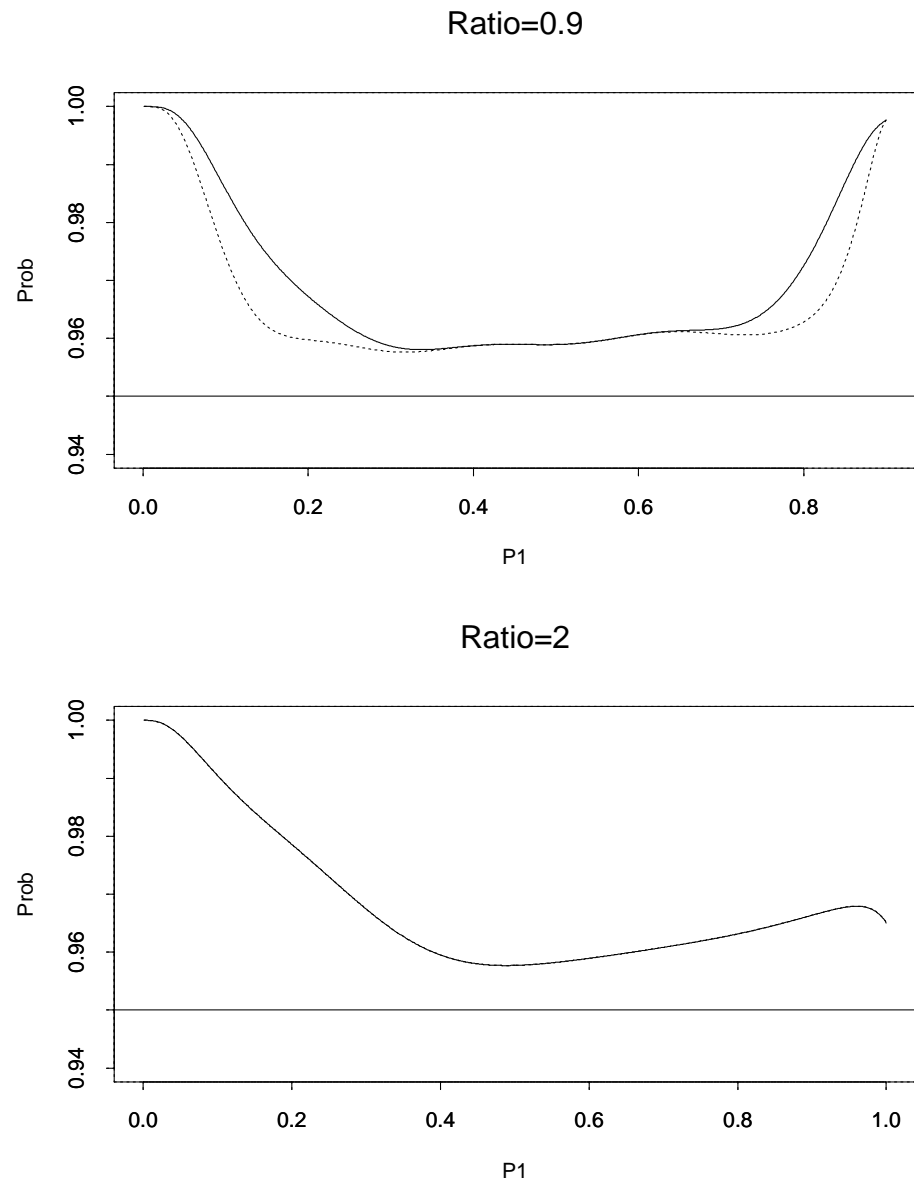


Figure 2.1: Coverage probability plot for  $R = 0.9$  and  $2$  when  $(N_1, N_2) = (15, 15)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line. Note that for Ratio =  $2$  the coverage for both the intervals are identical

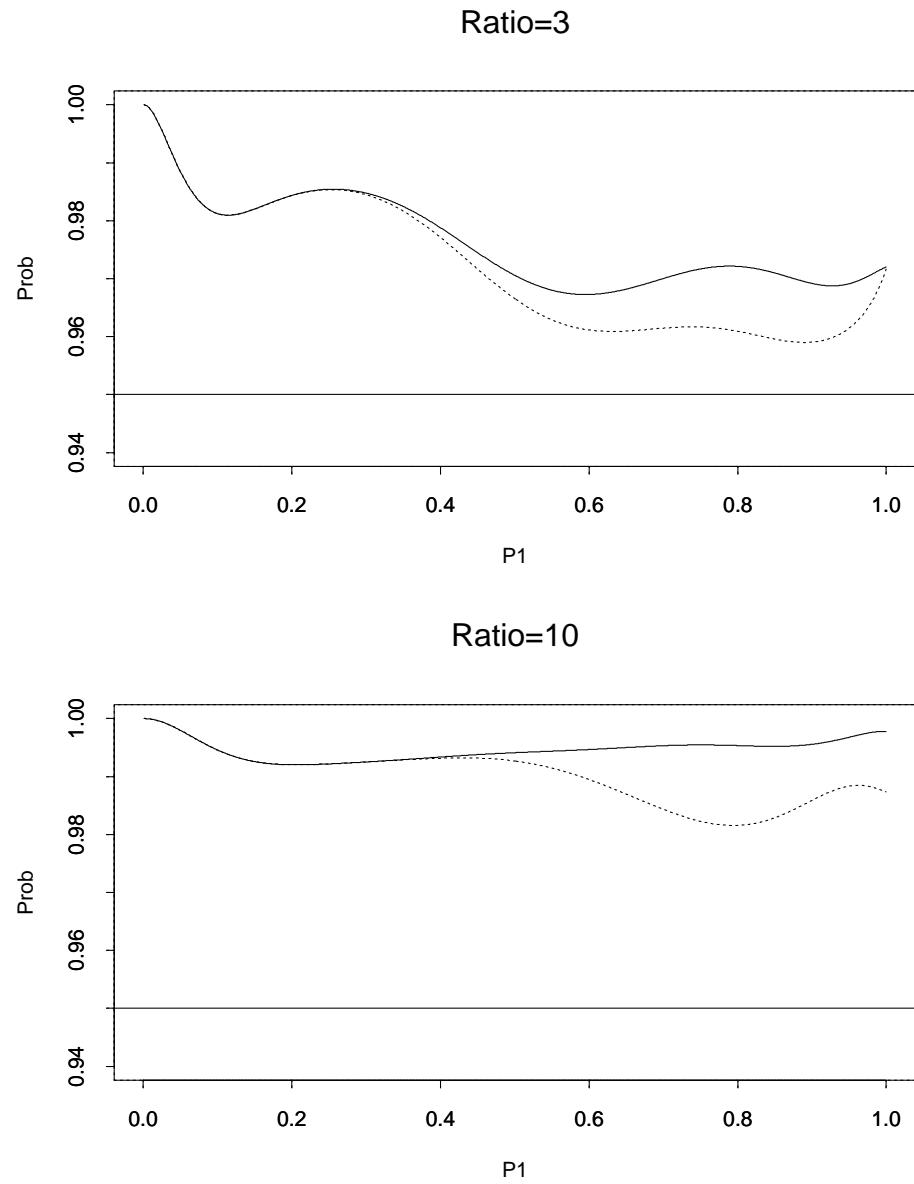


Figure 2.2: Coverage probability plot for  $R = 3$  and  $10$  when  $(N_1, N_2)=(15,15)$ :  
 $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line

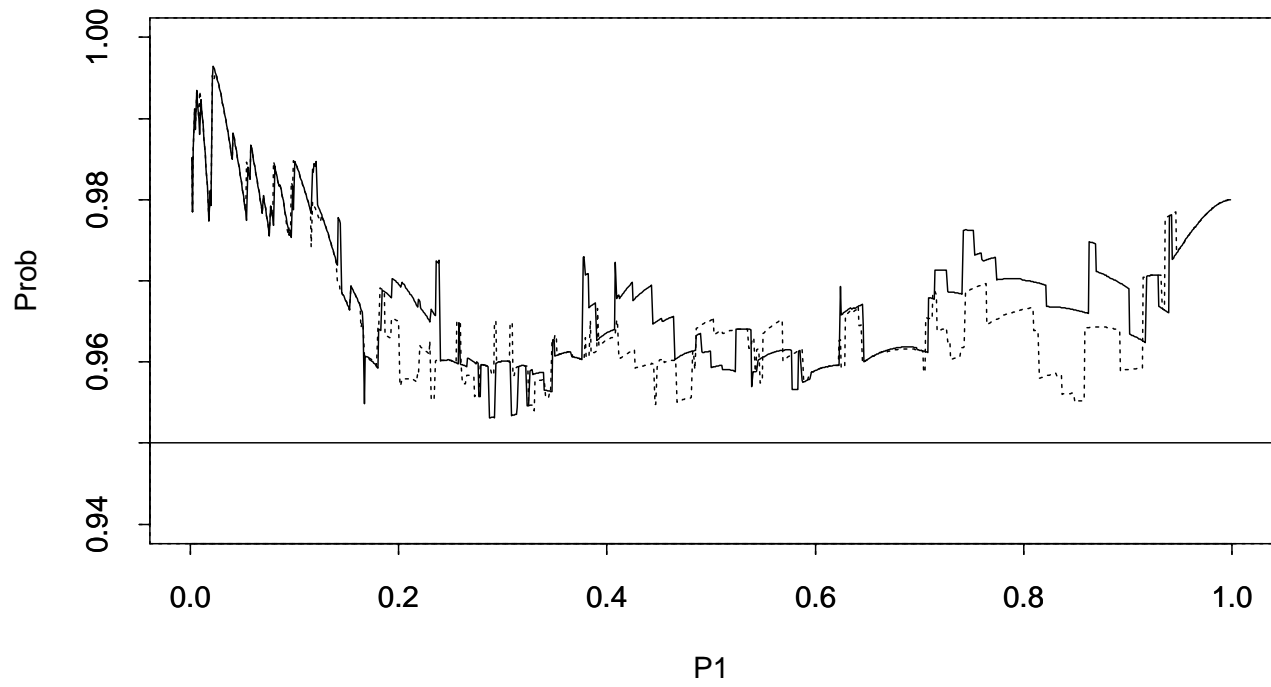


Figure 2.3: Coverage probability plot when  $(N_1, N_2)=(15,15)$ ,  $P_2 = 0.3$ , Ratio= $P_1/P_2$ :  
 $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line

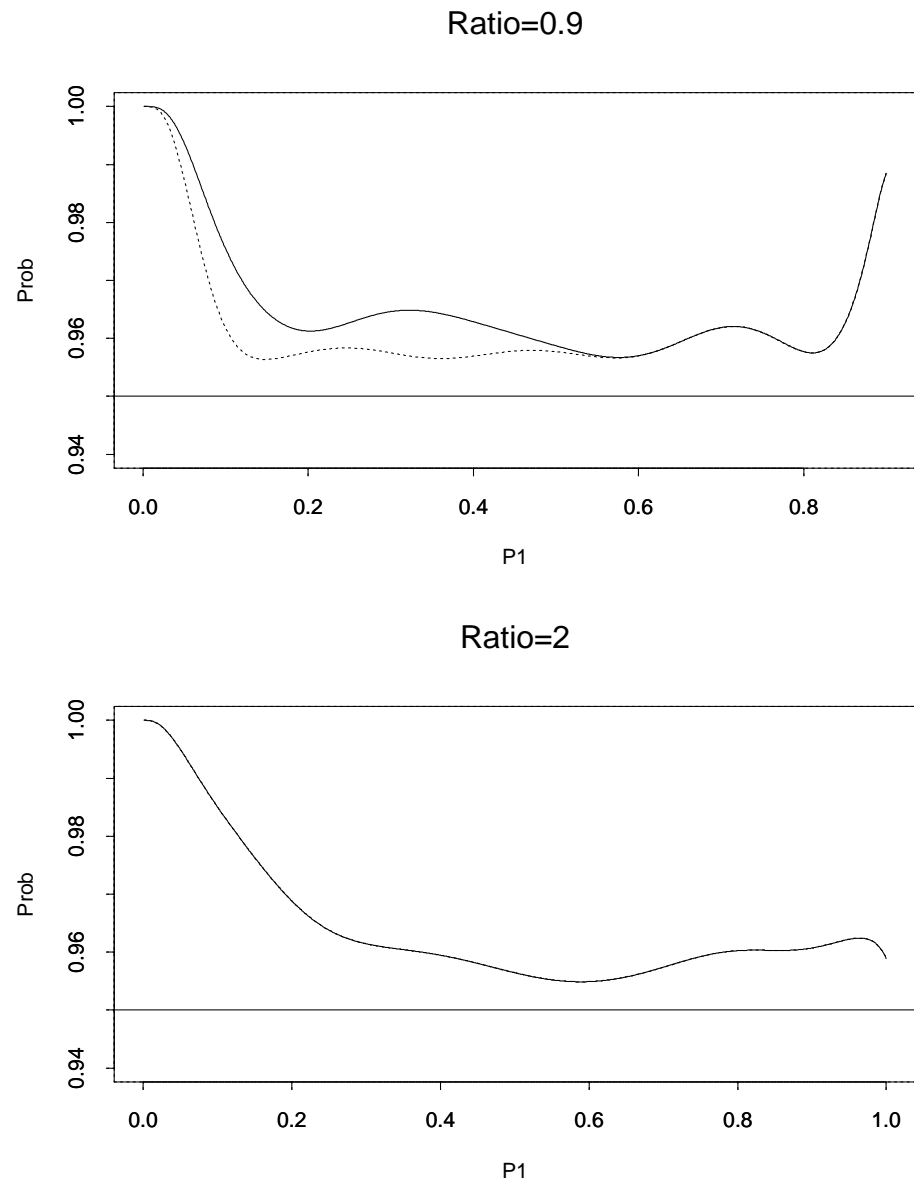


Figure 2.4: Coverage probability plot for  $R = 0.9$  and  $2$  when  $(N_1, N_2) = (20, 20)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line. Note that for Ratio =  $2$  the coverage for both the intervals are identical

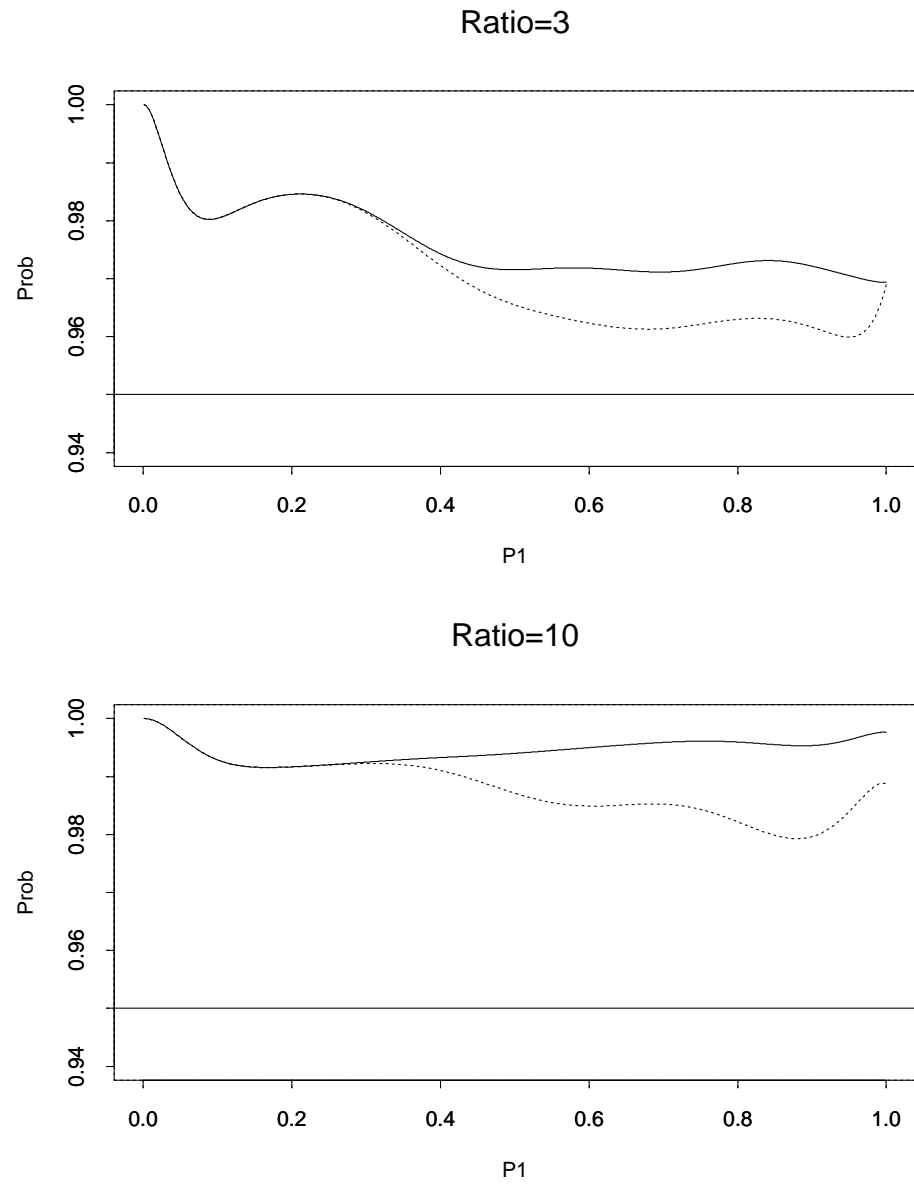


Figure 2.5: Coverage probability plot for  $R = 3$  and  $10$  when  $(N_1, N_2) = (20, 20)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line

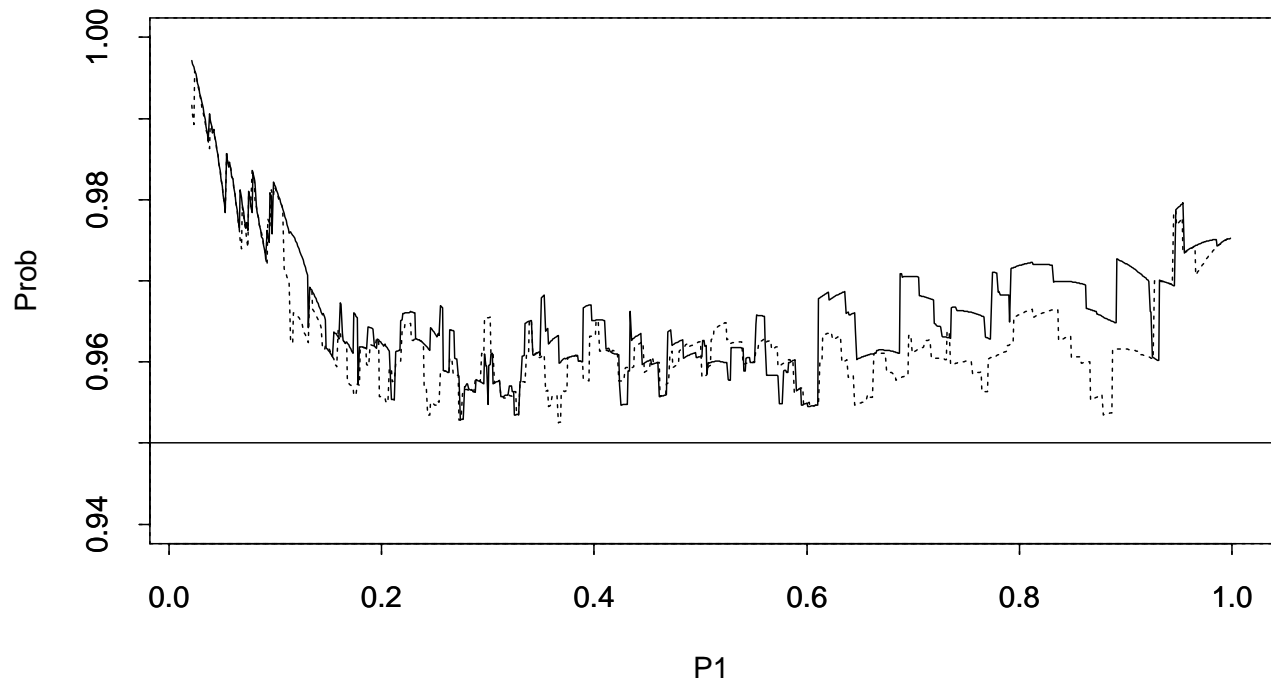


Figure 2.6: Coverage probability plot when  $(N_1, N_2)=(20,20)$ ,  $P_2 = 0.3$ , Ratio= $P_1/P_2$ :  
 $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line

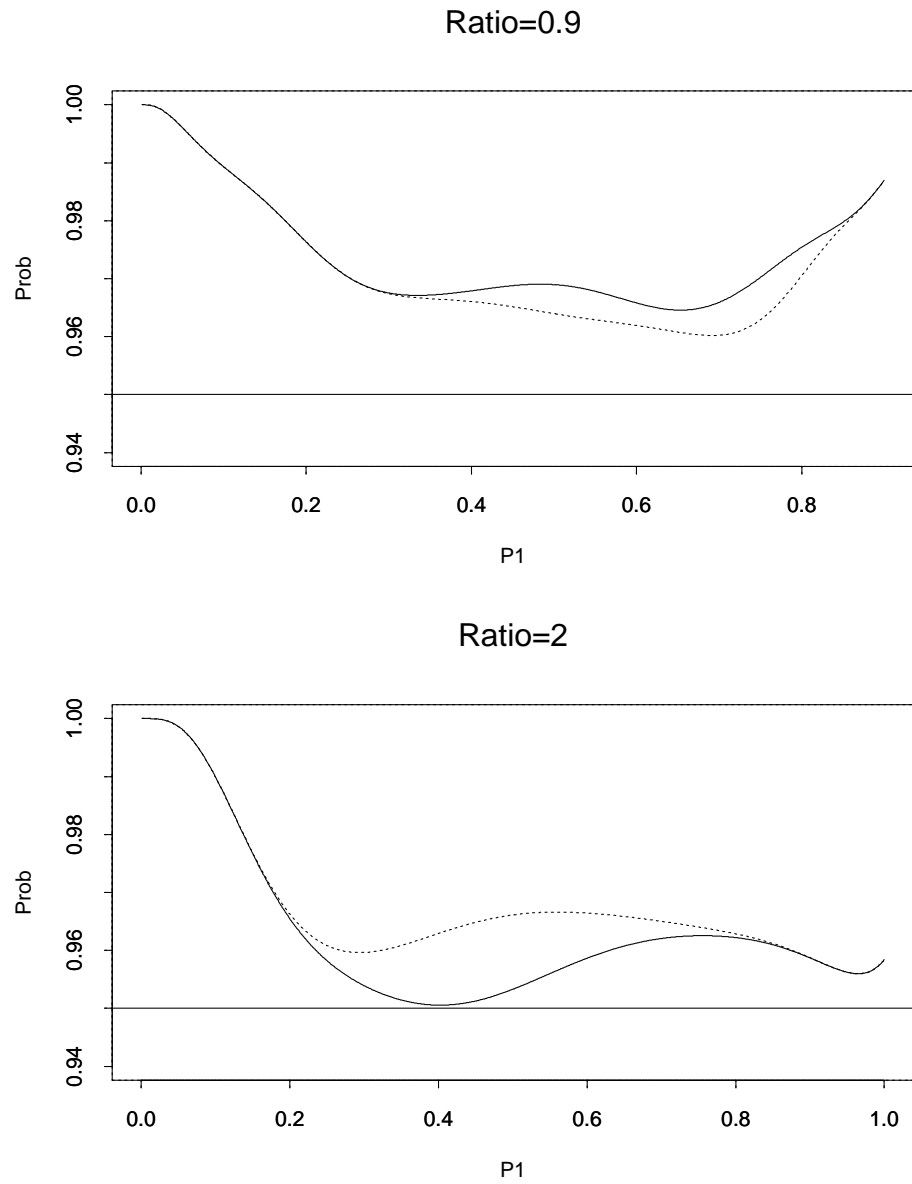


Figure 2.7: Coverage probability plot for  $R = 0.9$  and  $2$  when  $(N_1, N_2) = (10, 20)$ :  
 $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line



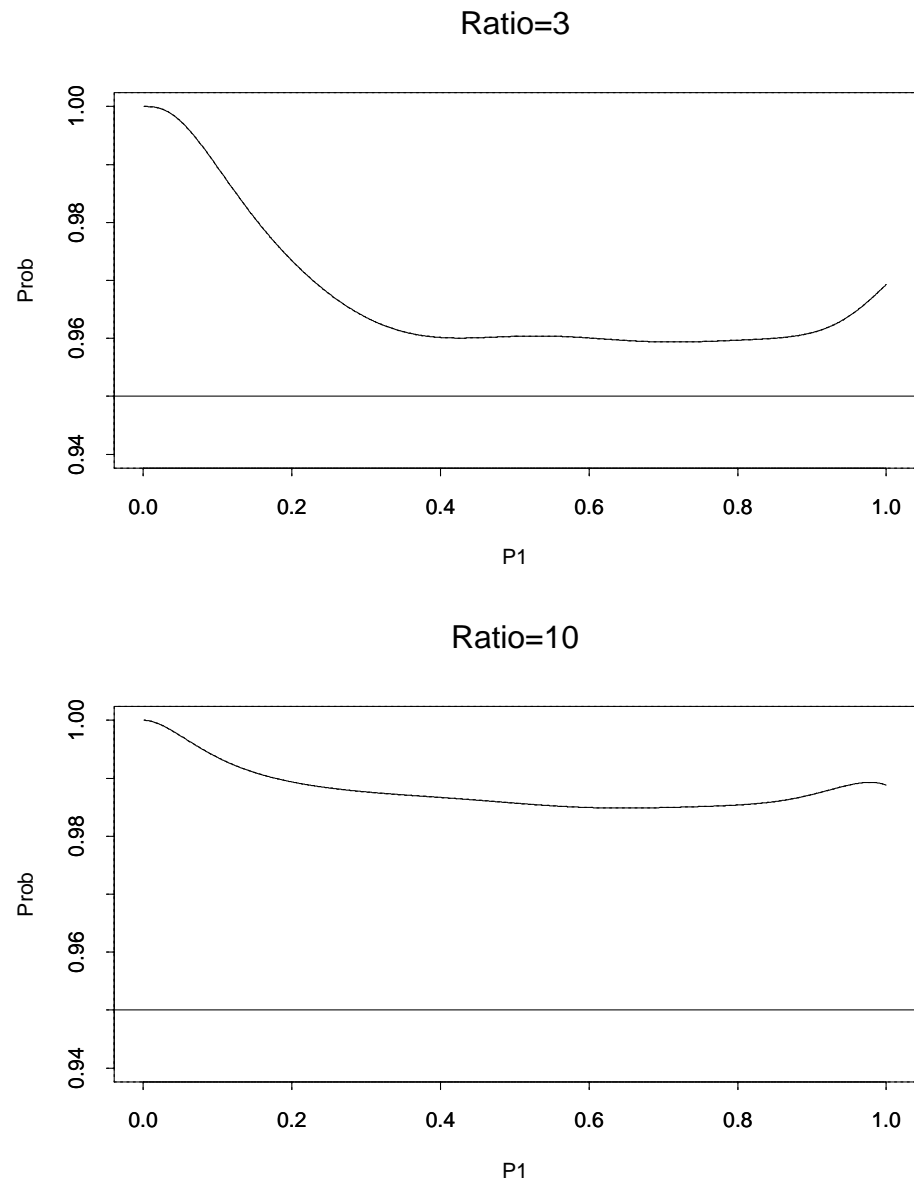


Figure 2.8: Coverage probability plot for  $R = 3$  and  $10$  when  $(N_1, N_2) = (10, 20)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line. Note that for Ratio = 3 and 10 the coverage for both the intervals are identical

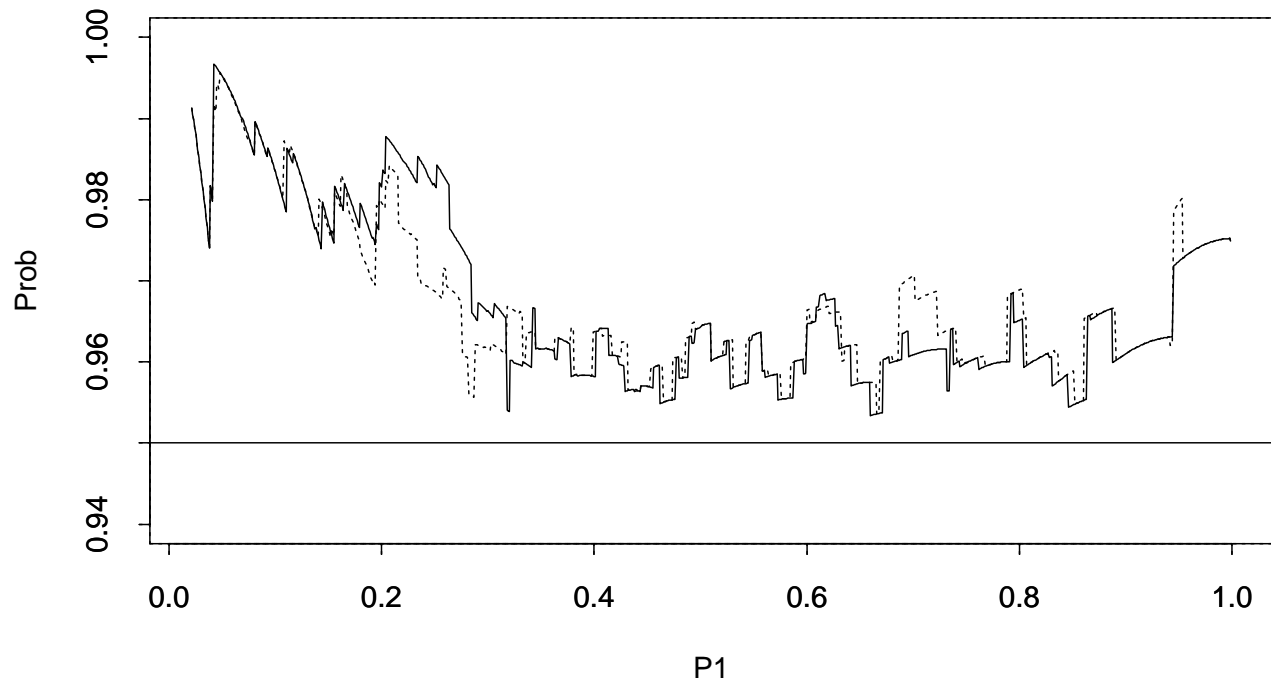


Figure 2.9: Coverage probability plot when  $(N_1, N_2)=(10,20)$ ,  $P_2 = 0.3$ , Ratio= $P_1/P_2$ :  
 $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line

## 2.5 Data Sets

We shall consider three data sets that we had formally introduced in Chapter 1. These data sets are from three different applications. The first which we called the Animal Toxicology data is from a study to compare the toxicity of two different chemicals. The second data set, which we are calling the Childhood Nephroblastoma data is from a clinical trial to compare two different kinds of treatment for childhood nephroblastoma. And the third data, which we are calling the Influenza Vaccine data set is from a vaccine trial where patients were randomized to receive either the influenza vaccine, or placebo. We will construct the confidence intervals for the two exact intervals and an asymptotic interval. The asymptotic intervals are constructed by assuming that the test statistic  $Z$  (see 2.6) follows a standard normal distribution under the null hypothesis. Then one can obtain the large sample confidence intervals for  $R$  as follows. We know that

$$P[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}] \approx 1 - \alpha,$$

where  $Z_{\alpha/2}$  denote the  $\alpha/2$ th percentile of a standard normal distribution. Hence,

$$P[-Z_{\alpha/2} \leq \frac{\hat{P}_1 - R\hat{P}_2}{\hat{\sigma}} \leq Z_{\alpha/2}] \approx 1 - \alpha.$$

Solving for  $R$  we have,

Table 2.6: Exact and asymptotic intervals for the animal toxicology data

|                            |                |
|----------------------------|----------------|
| CI for $p_{sup}$           | (0.9852,1.905) |
| CI for $p_\beta$           | (1.002,1.655)  |
| CI for the asymptotic test | (1.006,1.647)  |

$$P\left(\frac{-Z_{\alpha/2}\hat{\sigma} + \hat{P}_1}{\hat{P}_2} \leq R \leq \frac{Z_{\alpha/2}\hat{\sigma} + \hat{P}_1}{\hat{P}_2}\right) \approx 1 - \alpha.$$

We will present the confidence intervals and coverage probability plots for all the three tests. All the intervals constructed are 95% intervals.

**Animal Toxicology Data:** In this study the toxicity of two different kinds of chemicals (the synthetic A, and the natural chemical) were compared. The synthetic A induced tumor on 212 of the 350 rats, while the natural chemical induced tumor on 37 of the 77 rats. In Chapter 1 we were interested in testing

$$H_0 : P_1 \leq P_2 \text{ versus } H_1 : P_1 > P_2$$

where,  $P_1$  and  $P_2$ , represents the true tumor rates for the synthetic A population, and the natural population. So in terms of the relative risk we are testing the null hypothesis of  $R \leq 1$ . It will also be appropriate to construct confidence intervals for  $R$ . In Table 2.6 we have presented the confidence intervals from the  $p_{sup}$  test,  $p_\beta$  test, and the asymptotic test.

Note that the confidence interval for  $p_\beta$  is not only smaller than the  $p_{sup}$  interval, but it is completely contained within the  $p_{sup}$  interval. Also 1 lies within the  $p_{sup}$

interval but not within the  $p_\beta$  interval, implying that the  $p_\beta$  method is suggesting the toxicity for the synthetic A chemical is higher than the natural chemical. The standard exact interval fails to detect this. Although we see that the asymptotic interval is very similar to the  $p_\beta$  interval, and also suggests higher toxicity for the synthetic A chemical, (as 1 is not contained in the interval) but if we look at the coverage probability plot (Figure 2.10) we see that its coverage has fallen far short of the nominal coverage at values of  $P_1$  near 0 and 1. The coverage probability for the  $p_\beta$  interval is much closer to the nominal coverage than the  $p_{sup}$  interval in this case. This indicates that the set  $A_R$  for  $p_\beta$  is a subset of the  $A_R$  for  $p_{sup}$ .

**Childhood Nephroblastoma Data:** This data were obtained from a clinical trial studying two different kinds of treatment for childhood nephroblastoma. They were pre-operative chemo therapy followed by nephrectomy, and radio therapy followed by post-operative nephrectomy. For the chemo group, out of 88 patients, 83 had a rupture free tumor. For the radio group, out of 76 patients, 69 had a rupture free tumor. We were interested in testing whether radio therapy was non-inferior to chemo therapy and the value of  $R_0$  chosen was 1.15. So the hypothesis of interest was,

$$H_0 : R \geq 1.15 \text{ versus } H_1 : R < 1.15.$$

In Table 2.7 we are presenting the exact and asymptotic confidence intervals. Again we see that the  $p_\beta$  confidence interval is completely contained within the  $p_{sup}$  interval. Both the exact intervals contain 1.15 and does not suggest non-inferiority.

Table 2.7: Exact and asymptotic intervals for the childhood nephroblastoma data

|                            |                |
|----------------------------|----------------|
| CI for $p_{sup}$           | (0.9465,1.161) |
| CI for $p_{\beta}$         | (0.9476,1.154) |
| CI for the asymptotic test | (0.9481,1.156) |

However one should interpret this with caution. The upper bound of the  $p_{\beta}$  interval is marginally above the 1.15 mark. Also recall from Chapter 1 that the p value for the confidence interval test  $p_{\beta}$  was 0.054 showing some evidence against the null, while the standard test p value was 0.080 indicating no evidence at all (at 5% significance level). So the results from the confidence interval p value  $p_{\beta}$  and the exact interval derived from it definitely shows some evidence in favor of non-inferiority although it is not statistically significant at the 5% level. Further investigation is warranted in this case. Again we see that the standard method fails to detect this.

The coverage probability plot (Figure 2.11) again gives a similar picture as in the animal toxicology data. We see that the coverage probability for the asymptotic test goes slightly below the 0.95 line. Overall it does not do too bad. The coverage probability plot for  $p_{\beta}$  is again closer to the nominal coverage compared to  $p_{sup}$ . This implies that the set  $A_R$  for  $p_{\beta}$  is a subset of the  $A_R$  for  $p_{sup}$ .

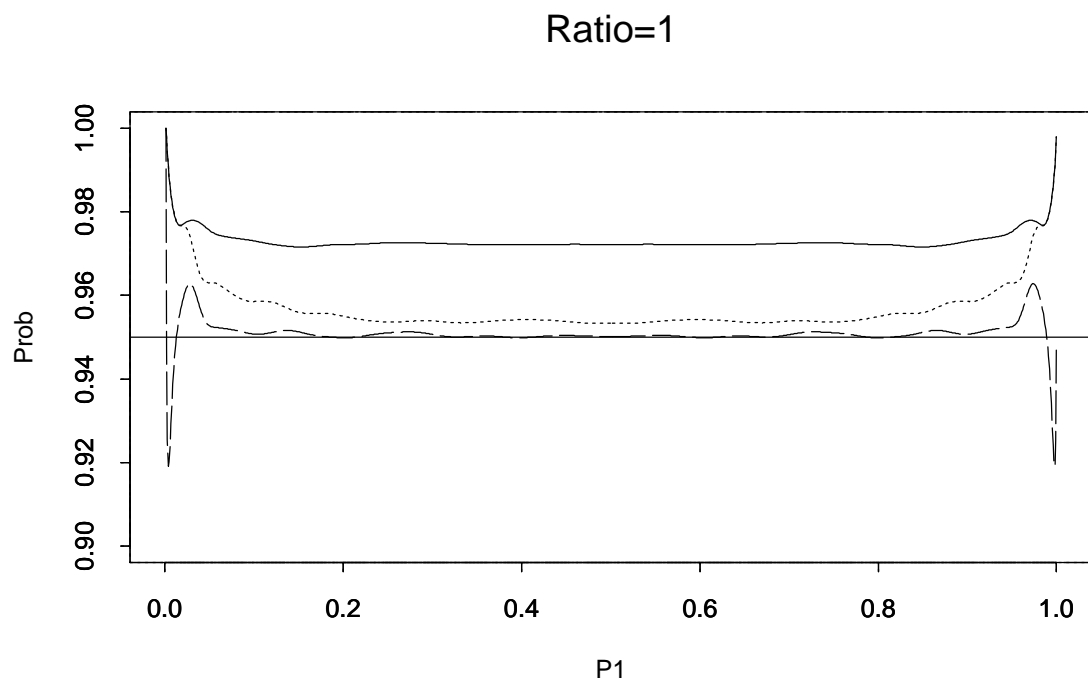


Figure 2.10: Coverage probability plot when  $(N_1, N_2)=(350, 77)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line, Asymptotic test=Broken line

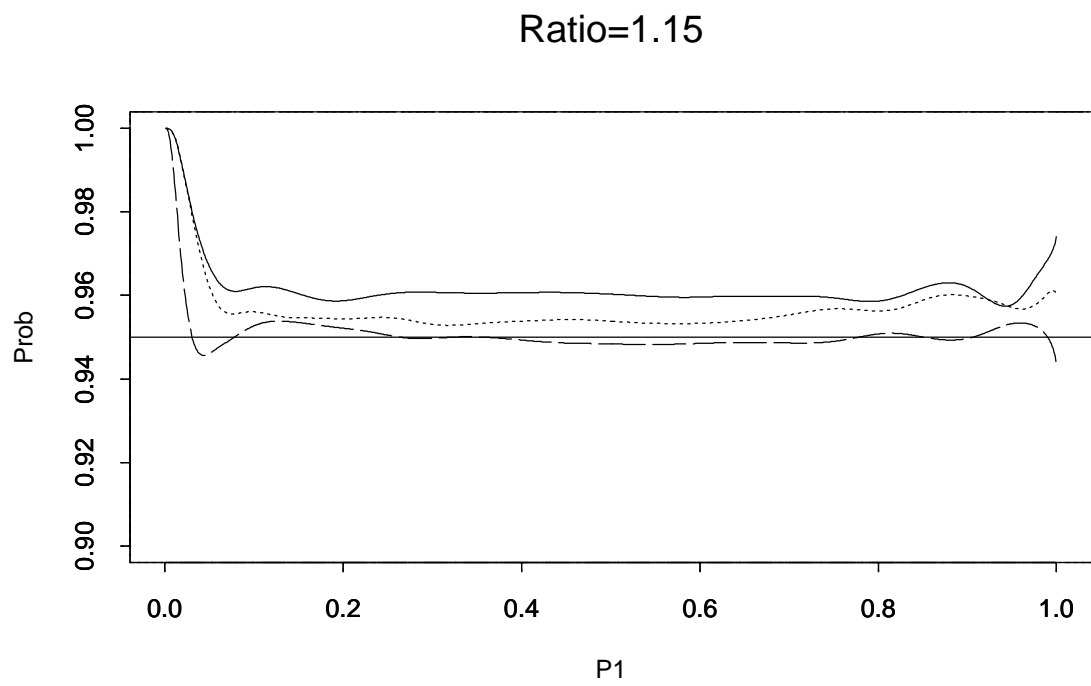


Figure 2.11: Coverage probability plot when  $(N_1, N_2)=(88, 76)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line, Asymptotic test=Broken line



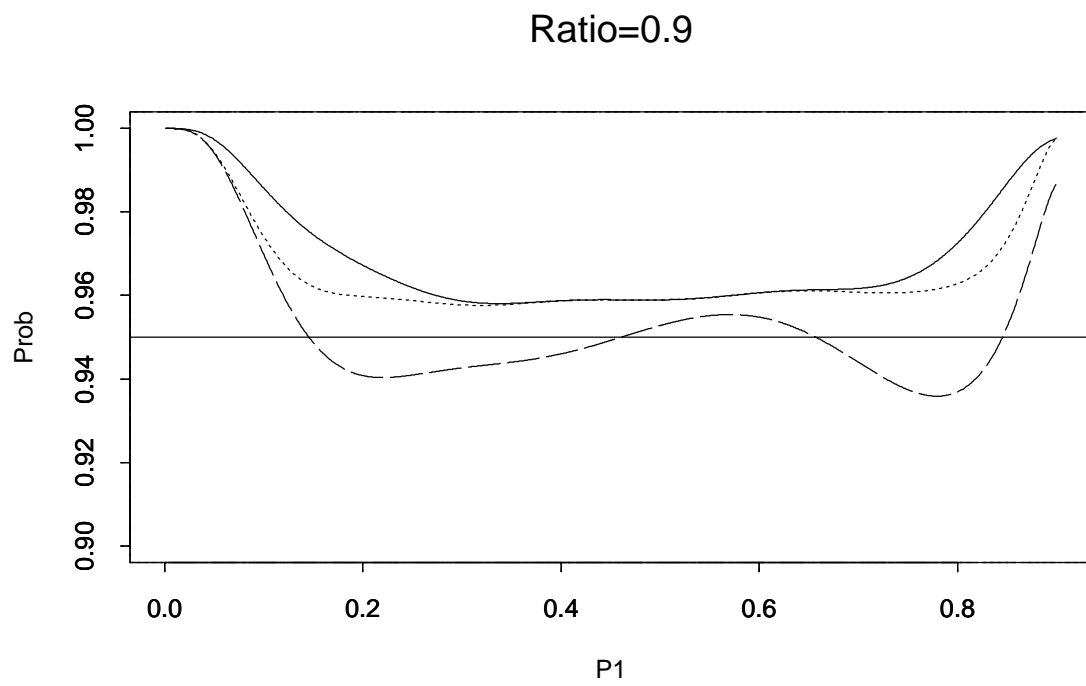


Figure 2.12: Coverage probability plot when  $(N_1, N_2)=(15,15)$ :  $p_{sup}$ =Solid Line,  $p_{\beta}$ =Dotted Line, Asymptotic test=Broken line

Table 2.8: Exact and asymptotic intervals for the influenza vaccine data

|                            |                |
|----------------------------|----------------|
| CI for $p_{sup}$           | (0.2608,1.037) |
| CI for $p_{\beta}$         | (0.2608,1.040) |
| CI for the asymptotic test | (0.2692,1.029) |

**Influenza Vaccine Data:** In this study the subjects were randomized to receive an experimental vaccine for influenza, or placebo. All subjects were monitored for symptoms of viral infection. The data showed that for the vaccine group, 7 out of 15 patients got infected, and for the placebo group, 12 out of 15 patients got infected. We were interested in testing

$$H_0 : \pi \leq 0.1 \text{ versus } H_1 : \pi > 0.1$$

where  $\pi = 1 - R$  was the efficacy parameter. Again we have presented the confidence intervals from the three tests in Table 2.8.

Here we see that the confidence intervals for both the exact tests are almost same, the upper bound for the  $p_{\beta}$  interval being slightly higher than the  $p_{sup}$  interval. This is not surprising because the sample sizes are very small. In this situations we have seen that the improvement occurs in lesser number of sample points compared to cases where we have larger sample sizes. The mean lengths still tend to be shorter. Recall from Table 2.1 that the mean length for the  $p_{\beta}$  interval was shorter than the  $p_{sup}$  interval when  $(N_1, N_2)$  was (15,15).

The coverage probability plot for  $R = 0.9$  (see Figure 2.12) shows that the  $p_{\beta}$

interval has coverage closer to the nominal coverage compared to the  $p_{sup}$  interval. Which means that the set  $A_R$  for  $p_\beta$  is a subset of the  $A_R$  for  $p_{sup}$ . Again note that the coverage for the asymptotic test has fallen far short of the nominal coverage. However this is not surprising as in this case the sample sizes are very small and the asymptotic assumption is inappropriate.

**Conclusion:** In this chapter we have made a detailed comparison between two different exact intervals. In various applications, (especially like the ones we talked about in this section) it is desirable to have confidence intervals whose coverage never falls below the nominal coverage. Although the asymptotic intervals have shorter interval lengths it fails to maintain the right coverage, even for moderately large sample sizes. Because of the discreteness of the problem the standard exact method results in a very conservative interval. We compared the standard interval with the  $p_\beta$  confidence interval (obtained by inverting the exact test proposed by Berger and Boos) in terms of length and coverage probability. We have seen from the various results and examples that our interval in general tends to have shorter confidence length and have coverage probabilities closer to the nominal coverage compared to the standard method. Also this method is not hard to implement in practice. So for the kind of problems we have talked about where exact inference is desirable the  $p_\beta$  confidence interval out performs the standard exact confidence interval  $p_{sup}$ .

## Chapter 3

# Exact, Approximate, and Bayesian Tests for the Hypothesis of Efficacy

### 3.1 Introduction

In most clinical trials one of the primary goal is to compare a new treatment to a standard treatment or placebo. Also it is common to have the primary outcome designed as a binary response. For example if the primary efficacy end point in an Oncology study is shrinkage of tumor then people are often interested in testing whether there is a difference in the shrinkage rate between patients in the treatment group and the control group. So one can define shrinkage of tumor in this case as a success and look at the response as a binary outcome, i.e., a success (if there is shrinkage) or a failure (if there is no shrinkage). So let us define  $P_1$  to be the true

proportion of success rate for the treatment group. And  $P_2$  as the true proportion of success rate for the control group. Then one can often use the relative risk parameter  $R$  (defined as  $P_1/P_2$ ) to compare the two groups.

In Chapters 1 and 2 we have covered in great details the problem of testing for the relative risk parameter along with construction of confidence intervals. For the kind of applications that we had in mind (such as testing for superiority or non-inferiority of a treatment) we had argued that it was desirable to use exact tests and exact confidence intervals instead of asymptotic ones. In most of these applications accuracy is of the utmost importance when analyzing the results and hence the need for exact inference. However we had also shown that the construction of exact tests or intervals is not only computationally challenging but they may also lead to overly conservative results.

In this Chapter our main goal would be to focus on some non-exact methods both from a frequentist and a Bayesian perspective. We will concentrate on the application of vaccine efficacy studies. So our main focus will be to develop and compare non-exact tests with the two exact tests (discussed in Chapters 1 & 2) with regards to this particular application. We will consider an approximate test (from the frequentist sense) and two different Bayesian tests. We will compare the frequentist properties of the Bayesian tests. Of course unlike the exact methods these tests do not guarantee that the size of the test will not exceed the nominal size. What would be of interest

is to see how poorly they are performing. So even if they exceed the nominal size by a little bit and there is a significant gain in power then it may be worthwhile to use them in practice. We will also show how easy it is to construct the approximate or the Bayesian tests compared to the exact tests.

In vaccine efficacy studies researchers are usually interested in testing whether the new vaccine is more effective in resisting a certain disease compared to placebo. Healthy subjects are randomized to receive either the vaccine or the placebo. So there is always the risk of injecting healthy subjects with the vaccine. And hence people are interested in determining the efficacy of the vaccine and whether it is worth injecting healthy subjects with the vaccine. One reasonable endpoint of interest will be occurrence of the disease within a period of time. Let us define  $P_1$  to be the true disease incidence rate for the vaccine group and  $P_2$  to be the true disease incidence rate for the placebo group. Then one can define the vaccine efficacy parameter  $\pi$  as

$$\pi = 1 - P_1/P_2.$$

We will assume  $P_1 \leq P_2$ , i.e., the vaccine is at least as good as placebo. The new vaccine will have 100% efficacy if  $P_1 = 0$  and will have no efficacy if  $P_1 = P_2$ .

People are often interested in testing the hypothesis of efficacy which is

$$H_0 : \pi \leq \pi_0 \text{ versus } H_1 : \pi > \pi_0, \tag{3.1}$$

for a pre-specified value of  $\pi_0$ . In terms of the relative risk, hypothesis (3.1) can be written as

$$H_0 : R \geq R_0 \text{ versus } H_1 : R < R_0, \quad (3.2)$$

where  $R_0 = 1 - \pi_0$ .

For establishing therapeutic efficacy of a drug one might choose  $\pi_0$  to be 0. However for vaccine trials usually small non-zero values of  $\pi_0$  are chosen. For example  $\pi_0 = 0.1, 0.2$ , etc. The parameter  $\pi$  may be interpreted as follows. If  $\pi = 0.1$  it means that the new vaccine will reduce the chances of occurrence of the disease by 10%. So for the alternative in (3.1) it will mean that the vaccine will reduce the chances of occurrence of the disease by over 10%.

We have shown in Chapter 1 that for testing such a hypothesis as in (3.1) it is more appropriate to use an exact test instead of asymptotic tests. Also we had established that the only form of exact inference possible was by using exact unconditional tests. In this Chapter we want to compare these exact methods with approximate and Bayesian tests. In section 3.2 we will define the two exact tests, the approximate unconditional test, and the two Bayesian tests that we are going to use. In section 3.3 we will describe how to compute the size and power for each of the tests. In section 3.4 we will present the results from comparing all the five tests. And lastly in section 3.5 we will summarize our findings and present an overall conclusion.

## 3.2 Exact, Approximate, and Bayesian tests

The problem in hand can be described as follows: Let  $X_1$  and  $X_2$  represent two independent response variables distributed as  $\text{bin}(N_1, P_1)$  and  $\text{bin}(N_2, P_2)$  respectively. Here  $N_1$  and  $N_2$  represents the sample sizes in each group, and  $P_1$ , and  $P_2$ , denotes the true response rates. If we denote by  $x_1$  and  $x_2$  the observed number of successes in each group, then the binomial probability mass function of  $X_1$  and  $X_2$  will be

$$\text{bin}(N_1, x_1, P_1) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1}, \quad x_1 = 0, \dots, N_1,$$

and

$$\text{bin}(N_2, x_2, P_2) = \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2}, \quad x_2 = 0, \dots, N_2.$$

The sample space of  $(X_1, X_2)$  will be denoted as  $\mathbf{S} = \{0, \dots, N_1\} \times \{0, \dots, N_2\}$ . Thus there are  $(N_1 + 1) \times (N_2 + 1)$  points in  $\mathbf{S}$ .

In section 3.2.1 we will talk about the two exact tests and the approximate unconditional test. Note that all the three tests are developed from a frequentist point of view. In section 3.2.2 we will talk about the development of the Bayesian tests which are based on completely different assumptions all together. Later in section 3.4 when we compare all these tests we will use a common platform, viz., size and power of the test. Thus in order to have a common ground for comparison we will study the frequentist properties of the Bayesian tests.



### 3.2.1 Exact and Approximate tests

For testing the hypothesis of efficacy discussed in the previous section the most commonly used test statistic (see Farrington & Manning, 1990) is

$$Z = \frac{\hat{P}_1 - (1 - \pi_0)\hat{P}_2}{\hat{\sigma}},$$

where

$$\hat{\sigma} = \sqrt{\frac{\tilde{P}_1(1 - \tilde{P}_1)}{N_1} + (1 - \pi_0)^2 \frac{\tilde{P}_2(1 - \tilde{P}_2)}{N_2}}.$$

Here  $\hat{P}_1 = x_1/N_1$ ,  $\hat{P}_2 = x_2/N_2$ , and  $\tilde{P}_1$  and  $\tilde{P}_2$  are the maximum likelihood estimates of  $P_1$  and  $P_2$  obtained under the null hypothesis restriction  $\frac{\tilde{P}_1}{\tilde{P}_2} = R_0$ .

The exact distribution of  $Z$  depends on all possible outcomes of two binomial responses given the sample sizes  $N_1$  and  $N_2$ . Thus each outcome  $(x_1, x_2)$  corresponds to a  $2 \times 2$  table. Probability of a particular outcome is

$$P(X_1 = x_1, X_2 = x_2) = \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2}.$$

The observed exact significance level for a particular value of the parameter  $(P_1, P_2)$  is the sum of probabilities of the tables that are at least as extreme as the observed table. Thus if  $Z_{obs}$  is the value of the statistic for the observed table, then the tail region will consist of all tables for which  $Z \leq Z_{obs}$ . If we denote by  $Z(x_1^0, x_2^0)$  the value of  $Z$  for the observed table then we can write this probability as

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}. \quad (3.3)$$

Recall from Chapter 1 that we had explained how the computation of (3.3) involves the problem of the elimination of nuisance parameters and why the conditional approach does not work. Then we had went on to show how we can use the unconditional approach to eliminate the nuisance parameters by maximizing on the domain of the nuisance parameters. Then we talked about using the Bernard convexity condition (Bernard, 1947) to actually maximize on the boundary of the null hypothesis instead of the entire null space. All this led to the formation of the standard exact unconditional test proposed by Chan (1998) which was denoted by  $p_{sup}$ . And it was defined as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}, \quad (3.4)$$

where  $D(R_0) = [0, \min(R_0, 1)]$  is the domain of  $P_1$ .

Next we had introduced the Berger-Boos approach (Berger & Boos, 1994) for this problem which suggested taking the maximum on a confidence set for the parameters  $(P_1, P_2)$  instead of maximizing on the entire domain of the parameter space. We had

used  $p_\beta$  to denote the p value for this test. And this was defined as

$$\sup_{P_1 \in C_\beta^*} \left\{ \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1-P_1)^{N_1-x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1-\frac{P_1}{R_0}\right)^{N_2-x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]} \right\} + \beta, \quad (3.5)$$

where

$$C_\beta^* = \{\max(L_1, L_2 R_0), \min(U_1, U_2 R_0)\}.$$

Recall that here  $(L_1, U_1)$  and  $(L_2, U_2)$  were the  $\sqrt{100(1-\beta)\%}$  confidence intervals for  $P_1$  and  $P_2$ , and  $C_\beta^*$  was the  $100(1-\beta)\%$  confidence set for  $(P_1, P_2)$ .

Both these tests were clearly discussed in Chapter 1 and hence we will not repeat all the details. However to summarize things we had shown that both these tests were exact tests and were guaranteed to maintain the nominal size. We had also shown the confidence interval test  $p_\beta$  was superior to the standard test  $p_{sup}$  in terms of power and computational ease. The third test that we will talk about is what we will call the Approximate unconditional test. This test is based on the idea proposed by Storer & Kim (1990). Later Kang & Chen (2001) studied this for the hypothesis of efficacy. However their results had some computational error and we thought that a more thorough investigation was appropriate.

The idea behind this test is very simple. Note that in the computation of the exact unconditional test the main problem is the elimination of the nuisance parameter. In the computation of (3.4) or (3.5) we are eliminating the nuisance parameter by maximizing on the domain of the parameter space. However a more simple minded

approach will be to replace the nuisance parameter by its maximum likelihood estimate. This will lead to what we are calling the Approximate unconditional test. Let us denote the p value for this test as  $p_{au}$ . Then one can define this as

$$p_{au} = \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} \tilde{P}_1^{x_1} (1-\tilde{P}_1)^{N_1-x_1} \binom{N_2}{x_2} (\tilde{P}_1/R_0)^{x_2} (1-\tilde{P}_1/R_0)^{N_2-x_2} I_{[Z(x_1, x_2) \leq Z(x_1^0, x_2^0)]}, \quad (3.6)$$

where  $\tilde{P}_1$  is the maximum likelihood estimate of  $P_1$  obtained under the null hypothesis restriction  $\frac{\tilde{P}_1}{\tilde{P}_2} = R_0$ .

The main advantage of this test is that owing to its simplistic nature this test is very easy to implement in practice. Since we are completely eliminating the maximization step the computation becomes much easier. The main disadvantage however is that this test does not give a valid p value. The definition of a valid p value is that for any statistic  $p$ , if under the null hypothesis

$$P(p \leq \alpha) \leq \alpha, \forall \alpha \in [0, 1], \quad (3.7)$$

then  $p$  is a valid p value. In other words  $P(\text{Reject Null} \mid \text{Null}) \leq \alpha$ , which is the definition of a level  $\alpha$  test. This test does not satisfy (3.7). Hence the actual size of the test may exceed the nominal size.

The three tests that we talked about, i.e.,  $p_{sup}$ ,  $p_\beta$ , and  $p_{au}$  are derived from a frequentist sense. The goal of these tests is to fix the type I error so that it does not exceed a certain nominal level and reduce the type II error, i.e., maximize the power

of the test (as power = 1 - type II error). In section 3.2.2 we will talk about two different tests derived from a Bayesian perspective. Here the tests are not constructed so as to fix the type I error and maximize on power. So it is a bit difficult to compare the Bayesian tests with the frequentist tests individually in terms of size and power. Later in section 3.4 we will try to introduce some kind of total error type of concept (which will be the sum of the size and the type II error) so that we can do a more meaningful comparison.

### 3.2.2 Bayesian tests

For Bayesian inference one would assume that  $P_1$  and  $P_2$  follow a distribution instead of assuming that they are fixed parameters. Since the values of  $P_1$  and  $P_2$  lie between 0 and 1 it is reasonable to assume that they follow a beta distribution. So let us assume apriori that  $P_1$  and  $P_2$  follow a beta distribution with certain known parameters. The motivation behind this assumption is that since  $X_1$  and  $X_2$  follow binomial distributions, the beta distribution will serve as a conjugate prior, i.e., the posterior distribution of  $P_1$  and  $P_2$  will also be a beta distribution. Since Bayesian inference is based on the posterior distribution, a known form of posterior distribution from which we can easily draw samples will make the problem much simpler.

For this problem our main interest lies in the ratio  $P_1/P_2$  and we will be more interested in the distribution of  $P_1/P_2$  rather than the joint distribution of  $P_1$  and

$P_2$ . But as the posterior distributions of  $P_1$  and  $P_2$  are independent beta we can draw samples from  $P_1$  and  $P_2$  and take the ratio to simulate from the distribution of  $P_1/P_2$ . Once this is done we can make any statistical inference from samples drawn from the distribution of  $P_1/P_2$ . Let us assume

$$P_1 \sim \text{beta}(\alpha_1, \beta_1) \text{ and } P_2 \sim \text{beta}(\alpha_2, \beta_2),$$

where  $(\alpha_1, \beta_1)$  and  $(\alpha_2, \beta_2)$  are known and  $P_1$  and  $P_2$  are independent. Hence the joint prior density can be written (up to proportionality constant) as

$$\pi(P_1, P_2) \propto P_1^{\alpha_1-1}(1-P_1)^{\beta_1-1}P_2^{\alpha_2-1}(1-P_2)^{\beta_2-1}.$$

We also know that

$$X_1 \sim \text{bin}(N_1, P_1) \text{ and } X_2 \sim \text{bin}(N_2, P_2),$$

and  $X_1$  and  $X_2$  are independent. Hence the likelihood function can be written as

$$L(P_1, P_2|X_1 = x_1, X_2 = x_2) \propto P_1^{x_1}(1-P_1)^{N_1-x_1}P_2^{x_2}(1-P_2)^{N_2-x_2}.$$

The posterior distribution will be proportional to the product of the likelihood times the prior density. Hence the posterior density can be obtained as

$$\pi(P_1, P_2|X_1 = x_1, X_2 = x_2) \propto L(P_1, P_2|X_1 = x_1, X_2 = x_2)\pi(P_1, P_2). \quad (3.8)$$

So now we can rewrite (3.8) as

$$\pi(P_1, P_2|X_1 = x_1, X_2 = x_2) \propto P_1^{x_1+\alpha_1-1}(1-P_1)^{N_1-x_1+\beta_1-1}P_2^{x_2+\alpha_2-1}(1-P_2)^{N_2-x_2+\beta_2-1}.$$

Because of independence we have

$$P_1|X_1 \sim \text{beta}(\alpha_1^*, \beta_1^*),$$

$$P_2|X_2 \sim \text{beta}(\alpha_2^*, \beta_2^*),$$

where  $(\alpha_i^*, \beta_i^*) = (X_i + \alpha_i, N_i - X_i + \beta_i), i = 1, 2$ . So we see that the posterior distribution of  $P_1$  and  $P_2$  will be a beta as well.

Now we can draw Monte Carlo samples from the distributions of  $P_1|X_1$  and  $P_2|X_2$  and hence simulate the posterior distribution of  $R = P_1/P_2$ . To do this we first draw  $l=1, \dots, B$  samples from  $P_1|X_1$  and another  $l=1, \dots, B$  samples from  $P_2|X_2$ . Note that  $B$  can be chosen as large as we wish so as to reduce the Monte Carlo error. For our purpose we find  $B = 1000$  to be sufficient to ensure that the error does not go above 0.75 in most of the cases. Let us denote the  $l^{th}$  sample as  $(P_i|X_i)^l, i = 1, 2$ . Then we have

$$(P_i|X_i)^l \sim \text{beta}(\alpha_i^*, \beta_i^*), i = 1, 2, l = 1, \dots, B.$$

Hence we can generate  $l = 1, \dots, B$  samples from the distribution of  $R = P_1/P_2$ . So the  $l^{th}$  sample  $R^l$  can be assumed to come from the posterior distribution of  $R$  and can be written as

$$R^l = \frac{(P_1|X_1)^l}{(P_2|X_2)^l} \sim (R|X_1, X_2).$$

Similarly one can draw samples from the prior distribution of  $R$ . The  $l^{th}$  sample  $R_{pr}^l$  can be assumed to come from the prior distribution of  $R$  and can be written as

$$R_{pr}^l = \frac{P_1^l}{P_2^l} \sim R.$$

An important aspect of the Bayesian inference lies in the choice of the values of the prior distribution parameters. One can choose a non-informative prior, assuming that there is no previous information about the problem what so ever. For our problem  $(\alpha_i, \beta_i) = (1/2, 1/2)$ ,  $i=1,2$ , will be a non-informative prior. This is often referred to as Jeffreys prior (Jeffreys, 1935). Other reasonable choices will be values of  $(1,1)$  or  $(2,2)$  (see Agresti & Caffo, 2000). The choice of  $\text{beta}(1,1)$  as the prior distribution for  $P_1$  and  $P_2$  is equivalent to choosing an uniform prior. The nice thing about choosing an uniform prior is that in this case we will have  $\pi(P_1, P_2) \propto 1$  and hence from (3.8) we will have

$$\pi(P_1, P_2 | X_1 = x_1, X_2 = x_2) \propto L(P_1, P_2 | X_1 = x_1, X_2 = x_2).$$

This means that drawing inference from the posterior distribution and drawing inference from the likelihood ratio will be equivalent. Since we will be comparing the Bayesian tests with frequentist tests this somewhat provides a common ground. So for our purpose we will assume apriori that  $P_1$  and  $P_2$  have a  $\text{beta}(1,1)$  distribution.

Note that since the posterior distribution of  $P_1$  and  $P_2$  are independent  $\text{beta}(X_1 + \alpha_1, N_1 - X_1 + \beta_1)$  and  $\text{beta}(X_2 + \alpha_2, N_2 - X_2 + \beta_2)$ , if we have reasonably large  $(N_1, N_2)$  it will not really matter much in practice whether we choose  $(\alpha_i, \beta_i) = (1/2, 1/2)$ ,  $(1,1)$ , or a  $(2,2)$ . But if we have really small sample sizes or  $(x_i = 0, N_i)$  then one has to be more careful and there is going to be a difference in results between choosing an informative prior and a non-informative prior.



Now we will derive the two Bayesian tests. We will call the first test as the Lindley Smith ( $LS$ ) test as this idea of testing was originally suggested by Lindley-Smith. The second test uses the Bayes Factor as the test statistic and we will call this the Bayes Factor ( $BF$ ) test.

### **Lindley Smith ( $LS$ ) test**

The idea behind this test is quite simple. Since we can simulate from the posterior distribution of  $R$  we can numerically compute the  $R_{1-\alpha}$ th percentile value of the distribution. So if we are testing for the alternative  $H_1 : R \leq R_0$  at  $\alpha = 0.05$  level of significance we will reject the null hypothesis if  $R_{0.95} \leq R_0$ . Else we will fail to reject the null hypothesis.

So in practice this is how we will construct the test. Suppose we have the data at hand to be  $(N_1, x_1)$ , and  $(N_2, x_2)$ , where  $N_1, N_2$  are the sample sizes in each group and  $x_1$  is the observed number of responses for the treatment group and  $x_2$  is the observed number of responses for the placebo group. And we want to test the hypothesis of efficacy (3.1) at 5% level of significance. In terms of the relative risk the alternative will be  $H_1 : R < R_0$ . So we follow the steps discussed previously and draw  $B$  samples from the posterior distribution of  $R$ . From these  $B$  samples we compute the 95th percentile  $R_{0.95}$ . If the value of the 95th percentile is less than or equal to  $R_0$  we reject the null hypothesis.

### Bayes Factor ( $BF$ ) test

Bayes Factor is one of the most popular methods used in Bayesian hypothesis testing problem. If one has to make an analogy the Bayes Factor is the Bayesian equivalent of the Likelihood ratio. Here we will try to present the main idea.

Let us assume that we have two hypothesis  $H_0$  and  $H_1$  and suppose the data  $D$  in hand is assumed to have come from either of the two hypothesis with probabilities  $P(D|H_0)$  and  $P(D|H_1)$ . Also let us assume that without looking at the data one has prior belief about each of the two hypothesis. Let the prior probability of occurrence of any of the hypothesis be  $P(H_0)$  and  $P(H_1)$ . Now the data will produce posterior probability for each hypothesis. Let they be  $P(H_0|D)$  and  $P(H_1|D)$ .

The Bayes Factor ( $BF$ ) is defined to be the ratio of the posterior odds of observing  $H_1$  to the prior odds of observing  $H_1$ . So

$$BF = \frac{P(H_1|D)/P(H_0|D)}{P(H_1)/P(H_0)}. \quad (3.9)$$

### Interpreting the $BF$

Kass & Raftery (1995) presents a nice discussion on the interpretation of  $BF$ . The  $BF$  is a summary of evidence provided by the data in favor of one of the hypothesis  $H_0$  or  $H_1$ . They have provided a chart for interpreting the  $BF$ .

1. If  $1 \leq BF \leq 3.2$  then the evidence against  $H_0$  is not worth mentioning.

2. If  $3.2 < BF \leq 10$  then the evidence against  $H_0$  is substantial.
3. If  $10 < BF \leq 100$  then the evidence against  $H_0$  is strong.
4. If  $100 < BF$  then the evidence against  $H_0$  is decisive.

Often people use the logarithmic scale at base 10 to interpret the  $BF$ . In that scale one can interpret the  $BF$  as follows.

1. If  $0 \leq \log_{10}(BF) \leq 1/2$  then the evidence against  $H_0$  is not worth mentioning.
2. If  $1/2 < \log_{10}(BF) \leq 1$  then the evidence against  $H_0$  is substantial.
3. If  $1 < \log_{10}(BF) \leq 2$  then the evidence against  $H_0$  is strong.
4. If  $2 < \log_{10}(BF)$  then the evidence against  $H_0$  is decisive.

### Computing the $BF$

Now that we have discussed what the  $BF$  is, the next step would be to determine how to compute it for our problem and hence how to construct the  $BF$  test. Recall that we have

$$BF = \frac{P(H_1|D)P(H_0)}{P(H_0|D)P(H_1)}. \quad (3.10)$$

Suppose we are testing the hypothesis of efficacy. In terms of relative risk we will be testing (3.2). So  $P(H_0|D) = P(R \geq R_0|D)$ . Since we can draw  $l = 1, \dots, B$  samples from the posterior distribution of  $R$ , we can estimate this probability by calculating

$$\frac{\sum_{l=1}^B I[R^l \geq R_0]}{B}$$

Using the same logic one can evaluate the  $BF$  for our problem using (3.10) as

$$\frac{\sum_{l=1}^B I[R^l < R_0] \sum_{l=1}^B I[R_{pr}^l \geq R_0]}{\sum_{l=1}^B I[R^l \geq R_0] \sum_{l=1}^B I[R_{pr}^l < R_0]}. \quad (3.11)$$

We will use (3.11) to estimate the  $BF$ .

### **The $BF$ test**

Now we are in a position to construct the  $BF$  test. For our test we will choose 30 to be the cutoff point. So we will construct our test as follows. For testing the alternative hypothesis  $H_1 : \pi > \pi_0$  if we compute the BF using (3.11) then we will reject the null hypothesis if  $BF > 30$ . In logarithmic scale, we will reject the null hypothesis if  $\log_{10}(BF) > 1.48$ .

There is no specific reason for choosing 30 apart from the fact that it gives a reasonable single number which we can assign as a cutoff point. This is because we are trying to construct a test that can be compared from a frequentist sense as well as a Bayesian sense. The decision rule suggested by Kass and Raftery (which they adopted from Jeffreys) tells us clearly when to base our decision in favor of the null hypothesis or the alternative. But if we are to decide between rejecting or not rejecting the hypothesis it probably makes more sense to have a definite cut off point. Based on this scale (and trying to find a single cutoff that makes the size of the test closest to the nominal size) for the test of efficacy we found 30 to be a reasonable choice. Other values that we tried were 10 and 50. We saw that for a cutoff of 10 the

test was very liberal and for a cutoff of 50 it was very conservative.

Note that from our computations what we found was that there is no single best optimum number that helps in maintaining the size as close to the nominal size as possible. What one can do is for each test find a cutoff for which the size of the test will not exceed the nominal size. So what we are suggesting is that for all reasonable values of  $(N_1, N_2)$  and  $\pi_0$  determine a cutoff  $K$ , such that

$$P[BF > K|H_0] \leq \alpha.$$

This will guarantee a level  $\alpha$  test from a frequentist sense. However one has to come up with some kind of a table similar to a  $Z$  table or an  $F$  table and it may not hold much practical appeal. But if some one chooses to do it (need not be very thorough and can do it for a reasonable number of sample sizes and values of  $\pi_0$ ) then it will be possible to come up with an exact test which is very easy to implement in practice. But if we choose some arbitrary number like 30 there will not be any guarantee that the size of the test will remain below the nominal level. Similarly one can find a  $K$  for which the  $LS$  test satisfies

$$P[R_{1-\alpha} < KR_0|H_0] \leq \alpha.$$

In that case the  $LS$  test will also yield a level  $\alpha$  test.

Another point to note is that the  $BF$  in itself has an appeal of its own. Because the  $BF$  actually tells us whether or not to favor one decision over the other. For

example a large value of the  $BF$  shows strong evidence for  $H_1$  where as a small value of the  $BF$  shows strong evidence for  $H_0$ . This argument about having evidence for either one of the hypothesis does not hold for a p value. The p value either rejects the null hypothesis or fails to reject the null hypothesis. But failing to reject the null hypothesis does not mean it shows evidence in favor of the null hypothesis.

In the next section we will talk about how to compute the size and power for each of the tests discussed in this section.

### 3.3 Size and Power

In Chapter 1 we had discussed about the size and power computations for the two exact tests  $p_{sup}$  and  $p_{\beta}$ . It will be very similar for the other three tests as well. First we need to compute the rejection sets for all the tests. Suppose we are testing at level  $\alpha$  for the hypothesis of efficacy for a pre-specified value of  $\pi_0 = 1 - R_0$ .

1. For the exact unconditional test it can be defined as

$$R_{sup}(\alpha) = [(x_1, x_2) \mid p_{sup}(x_1, x_2) \leq \alpha].$$

2. For the confidence interval p value test it can be defined as

$$R_{\beta}(\alpha) = [(x_1, x_2) \mid p_{\beta}(x_1, x_2) \leq \alpha].$$

3. For the approximate unconditional test it can be defined as

$$R_{au}(\alpha) = [(x_1, x_2) \mid p_{au}(x_1, x_2) \leq \alpha].$$

4. For the Lindley-Smith ( $LS$ ) test it can be defined as

$$R_{LS} = [(x_1, x_2) \mid R_{1-\alpha}(x_1, x_2) \leq R_0],$$

where  $R_{1-\alpha}(x_1, x_2)$  is the  $100(1-\alpha)$ th percentile of the posterior distribution of  $R$  when the observed responses are  $(x_1, x_2)$  and the sample sizes are  $(N_1, N_2)$ .

5. For the Bayes Factor ( $BF$ ) test it can be defined as

$$R_{BF} = [(x_1, x_2) \mid BF(x_1, x_2) > 30],$$

where  $BF(x_1, x_2)$  is the Bayes factor when the observed responses are  $(x_1, x_2)$  and the sample sizes are  $(N_1, N_2)$ .

Note that for the  $BF$  test if one wants to construct an exact level  $\alpha$  test then one can choose a cutoff depending on the value of  $\pi_0$  and sample sizes  $(N_1, N_2)$  such that the size never exceeds the nominal size. Let us denote by  $K(N_1, N_2, \pi_0)$  such a cutoff point. Then we will define the rejection set for the  $BF$  test as

$$R_{BF} = [(x_1, x_2) \mid BF(x_1, x_2) > K(N_1, N_2, \pi_0)].$$

Once we have computed the rejection sets, the size and power for any of the tests may be obtained by using one of the following formulae.

1. The size of the test can be computed as

$$\sup_{P_1 \in D(R_0)} \sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} \left(\frac{P_1}{R_0}\right)^{x_2} \left(1 - \frac{P_1}{R_0}\right)^{N_2 - x_2} I_{[(x_1, x_2) \in R_i(\alpha)]}, \quad (3.12)$$

where  $i = \text{sup}$  or  $\beta$  or  $au$  or  $BF$  or  $LS$ .

2. The power of the test for a given value of  $(P_1, P_2)$  can be computed as

$$\sum_{x_1=0}^{N_1} \sum_{x_2=0}^{N_2} \binom{N_1}{x_1} P_1^{x_1} (1 - P_1)^{N_1 - x_1} \binom{N_2}{x_2} P_2^{x_2} (1 - P_2)^{N_2 - x_2} I_{[(x_1, x_2) \in R_i(\alpha)]}, \quad (3.13)$$

where  $i = \text{sup}$  or  $\beta$  or  $au$  or  $BF$  or  $LS$ .

In the next section we will compare the performance of the five tests in terms of size and power for the hypothesis of efficacy. We do not expect the approximate or the Bayesian tests to always maintain the nominal size. However if the sizes for these tests do not exceed the nominal size by too much and they have good power then it may be worthwhile to use them in practice.

### 3.4 Results

The test of efficacy is a special case of the general testing problem for the relative risk that was discussed in Chapter 1. However this test is widely used to evaluate the



efficacy of vaccines in vaccine trials. As we discussed earlier the values of the efficacy parameter considered are small non-zero values. We have considered several choices of  $\pi_0$ . However we will only present the results for  $\pi_0 = 0.1$ . In all the cases the results were quite similar. Again like Chapter 1 we will consider two different cases. The first case will be when the sample sizes  $(N_1, N_2)$  are balanced, i.e.,  $N_1 = N_2$ . The other case will be when they are unbalanced, i.e.,  $N_1 < N_2$  or  $N_1 > N_2$ .

The sample sizes  $(N_1, N_2)$  that are considered are the following:

1. For balanced cases they are (20,20), (50,50), (75,75) and (100,100).
2. For the unbalanced cases they are (20,40), (40,20), (50,100), and (100,50) respectively.

All tests will be considered at  $\alpha = 0.05$  level. Note that for the Bayesian tests we will be drawing  $B = 1000$  samples from the prior and the posterior distribution of  $R$  (see section 3.2.2). We will present the actual size and power for each of the five tests and the size function plots. Also we will present the total error (sum of the size and type II error).

### Balanced Case

In Table 3.1 we are presenting the actual size for each of the five tests for testing the hypothesis of efficacy (3.1) with  $\pi_0 = 0.1$  when  $N_1 = N_2$ . The first column  $(N_1, N_2)$  denotes the sample sizes and the other five columns represents each of the tests.

Table 3.1:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Size for all the five tests when  $N_1 = N_2$ 

| $(N_1, N_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|-----------|-----------|----------|--------|--------|
| (20,20)      | 0.0439    | 0.0446    | 0.0533   | 0.0500 | 0.0373 |
| (50,50)      | 0.0318    | 0.0487    | 0.0577   | 0.0522 | 0.0424 |
| (75,75)      | 0.0466    | 0.0481    | 0.0529   | 0.0524 | 0.0407 |
| (100,100)    | 0.0490    | 0.0488    | 0.0531   | 0.0515 | 0.0427 |

From Table 3.1 we see that all the three non-exact tests are performing quite well. Although they are above the nominal 5% level in some of the cases but they are only marginally above it. Note that the worst case is for the approximate unconditional test  $p_{au}$  when  $(N_1, N_2) = (50, 50)$ . In this case the size is 0.0577. The good thing is that both the approximate unconditional and the Bayes test using Lindley-Smith approach have size very close to the nominal 0.05 size in all the cases. Also the size for the Bayes test  $BF$  never exceeds the nominal size. However between these five tests the  $BF$  is the most conservative one. Except for the case of (50,50) where  $p_{sup}$  has the lowest size, in all the other cases  $BF$  had the lowest size. But there is always room for improvement by choosing the optimum cutoff point. As we have discussed earlier although we had chosen an arbitrary cutoff value of 30 one can evaluate the cutoff for a specific hypothesis so as to obtain a level  $\alpha$  test. And that may result in a size function which is closer to the nominal 0.05 line.

Another important point to note is that in only two out of the 20 cases the size has actually fallen below 0.04. The first is for (20,20) where the  $BF$  test has size of 0.0373. The other case is (50,50) when the  $p_{sup}$  test has size of 0.0318. This is one situation where we see how conservative the standard exact test might get even for moderately large samples.

Figures 3.1 - 3.4 shows the size function plot for all the five tests for each of the sample sizes considered in Table 3.1 (see Chapter 1 regarding computation of the size plot). From Figure 3.1 (when  $(N_1, N_2) = (20, 20)$ ) we see that among all the five tests the performance of the  $BF$  test is worst followed by  $p_{sup}$  and  $p_\beta$ .  $p_{au}$  and  $LS$  does a much better job although  $p_{au}$  exceeds the 0.05 nominal size slightly. For sample sizes of 50 (see Figure 3.2)  $p_{sup}$  does the worst job followed by  $BF$ . The other three tests have size functions quite close to each other. Again for sample sizes of 75 (see Figure 3.3) we see that  $BF$  and  $p_{sup}$  does the worst job. Performance for the other three tests are comparable. However for (100,100), (see Figure 3.4) except for  $BF$  all the other four tests does a good job. From all the four cases we see that  $p_\beta$ ,  $p_{au}$  and  $LS$  seem to do a reasonably good job while  $p_{sup}$  and especially  $BF$  performs quite poorly. Overall for balanced samples we have seen reasonably good results (in terms of size function being close to the nominal 0.05 line) for all the tests. We had done the computations for other values of  $\pi_0$  such as 0.23 and 0.3 (not shown here) and various other choices of  $(N_1, N_2)$  and in all the cases found similar results.

Table 3.2:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Type II error for specific choices of  $(P_1, P_2)$  when  $N_1 = N_2$

| $(N_1, N_2)$ | $(P_1, P_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|--------------|-----------|-----------|----------|--------|--------|
| (20,20)      | (0.2,0.5)    | 0.5405    | 0.4692    | 0.4591   | 0.4591 | 0.5091 |
| (20,20)      | (0.3,0.5)    | 0.7900    | 0.7555    | 0.7535   | 0.7535 | 0.7863 |
| (50,50)      | (0.2,0.5)    | 0.1727    | 0.1207    | 0.1207   | 0.1207 | 0.1385 |
| (50,50)      | (0.3,0.5)    | 0.6056    | 0.5173    | 0.5163   | 0.5183 | 0.5420 |
| (75,75)      | (0.2,0.5)    | 0.0394    | 0.0350    | 0.0325   | 0.0321 | 0.0431 |
| (75,75)      | (0.3,0.5)    | 0.3991    | 0.3759    | 0.3631   | 0.3733 | 0.4123 |
| (100,100)    | (0.2,0.5)    | 0.0095    | 0.0092    | 0.0088   | 0.0084 | 0.0116 |
| (100,100)    | (0.3,0.5)    | 0.2697    | 0.2643    | 0.2535   | 0.2660 | 0.3043 |

In Tables 3.2 and 3.3 we are presenting the type II errors and power for all the five tests for particular values of  $(P_1, P_2)$  lying in the alternative hypothesis. Type II error is just 1 - power. We have used (3.13) to compute the power and subtracted it from 1 to get the type II error. We have considered values of  $(P_1, P_2) = (0.2, 0.5)$  and  $(0.3, 0.5)$ . Main reason for choosing these values are often it will be expected that the vaccine group will have low disease incidence rate and it will be much lower than the placebo group. So for our application this may be a reasonable choice.

From Table 3.2 we see that  $p_{sup}$  and  $BF$  has the highest type II error amongst all the five tests (for all the cases considered here). Hence they will have the lowest

Table 3.3:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Power for specific choices of  $(P_1, P_2)$  when  $N_1 = N_2$

| $(N_1, N_2)$ | $(P_1, P_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|--------------|-----------|-----------|----------|--------|--------|
| (20,20)      | (0.2,0.5)    | 0.4595    | 0.5308    | 0.5409   | 0.5409 | 0.4900 |
| (20,20)      | (0.3,0.5)    | 0.2100    | 0.2444    | 0.2465   | 0.2465 | 0.2137 |
| (50,50)      | (0.2,0.5)    | 0.8273    | 0.8793    | 0.8793   | 0.8792 | 0.8615 |
| (50,50)      | (0.3,0.5)    | 0.3943    | 0.4827    | 0.4837   | 0.4817 | 0.4580 |
| (75,75)      | (0.2,0.5)    | 0.9606    | 0.9650    | 0.9675   | 0.9679 | 0.9569 |
| (75,75)      | (0.3,0.5)    | 0.6009    | 0.3759    | 0.3631   | 0.3733 | 0.4123 |
| (100,100)    | (0.2,0.5)    | 0.9905    | 0.9908    | 0.9912   | 0.9916 | 0.9884 |
| (100,100)    | (0.3,0.5)    | 0.7303    | 0.7357    | 0.7465   | 0.7340 | 0.6957 |

power (see Table 3.3). However one point to note is that the cases where  $BF$  has the highest type II error (the last four cases) it is only slightly higher than  $p_{sup}$ , the maximum difference being about 0.035 (when  $(N_1, N_2) = (100,100)$  and  $(P_1, P_2) = (0.3,0.5)$ ). But for the first four cases where  $p_{sup}$  has the highest type II error the maximum difference is 0.063 (when  $(N_1, N_2) = (50,50)$  and  $(P_1, P_2) = (0.3,0.5)$ ).  $p_\beta$ ,  $p_{au}$ , and  $LS$  all have very similar type II errors and power.

In Table 3.4 we present the total error, which is the sum of the size of the test and the type II error evaluated at the specified values of  $P_1$  and  $P_2$ . As we had mentioned earlier it would be appropriate to look at the total error as the Bayesian

Table 3.4:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Total error for specific choices of  $(P_1, P_2)$  when  $N_1 = N_2$

| $(N_1, N_2)$ | $(P_1, P_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|--------------|-----------|-----------|----------|--------|--------|
| (20,20)      | (0.2,0.5)    | 0.5844    | 0.5138    | 0.5124   | 0.5091 | 0.5464 |
| (20,20)      | (0.3,0.5)    | 0.8339    | 0.8001    | 0.8068   | 0.8035 | 0.8236 |
| (50,50)      | (0.2,0.5)    | 0.2045    | 0.1694    | 0.1784   | 0.1729 | 0.1809 |
| (50,50)      | (0.3,0.5)    | 0.6375    | 0.5660    | 0.5740   | 0.5705 | 0.5844 |
| (75,75)      | (0.2,0.5)    | 0.0860    | 0.0831    | 0.0854   | 0.0845 | 0.0832 |
| (75,75)      | (0.3,0.5)    | 0.4457    | 0.4240    | 0.4160   | 0.4257 | 0.4524 |
| (100,100)    | (0.2,0.5)    | 0.0585    | 0.058     | 0.0619   | 0.0599 | 0.0543 |
| (100,100)    | (0.3,0.5)    | 0.3187    | 0.3131    | 0.3066   | 0.3175 | 0.3470 |

tests are not guaranteed to maintain the nominal size. From Table 3.4 we see that the Bayesian test  $LS$  has one of the lowest total error, comparable with the exact test  $p_\beta$ . The standard exact test  $p_{sup}$  has the largest total error in most of the cases. From the cases we have considered we see that for balanced sample sizes  $p_\beta$ ,  $p_{au}$  and  $LS$  outperforms  $p_{sup}$  and  $BF$ . Between these five tests  $BF$  results in the most conservative test followed by  $p_{sup}$ . Next we will consider the cases when the sample sizes are unbalanced.

### Unbalanced Case

Throughout our study on exact tests we have seen how important it is to consider

Table 3.5:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Size for all the five tests when  $N_1 \neq N_2$ 

| $(N_1, N_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|-----------|-----------|----------|--------|--------|
| (20,40)      | 0.0487    | 0.0487    | 0.0487   | 0.0549 | 0.0397 |
| (40,20)      | 0.0485    | 0.0485    | 0.0514   | 0.0517 | 0.0475 |
| (50,100)     | 0.0383    | 0.0481    | 0.0503   | 0.0565 | 0.0424 |
| (100,50)     | 0.0492    | 0.0490    | 0.0525   | 0.0672 | 0.0453 |

the unbalancedness of sample sizes in the treatment group and the control group. In Chapter 1 we had seen how unequal sample sizes in the two arms resulted in drastic improvement in power for the confidence interval p value test  $p_\beta$ . Again while comparing the non-exact tests to exact tests we saw some interesting results because of the unbalancedness. First we will present the same Tables as we did in the balanced cases.

Table 3.5 presents the size for the unbalanced samples that we are considering. Between the two exact tests, except for (50,100) both of them have almost the same size. For (50,100) we see quite an improvement for  $p_\beta$ .  $p_{au}$  performs reasonably well in all the four cases. Although  $BF$  seems a bit conservative but it always maintains the correct level over here. The thing that is interesting to note is that in the last case  $LS$  exceeds the 0.05 level by quite a bit. If we look at the size for the  $LS$  test when sample sizes are (100,50) we see that it is above 0.067.

In Figures 3.5 - 3.8 we present the size function plots for all the five tests for each of the sample pairs considered. From Figure 3.5 when sample sizes are (20,40) we see that except  $BF$  all the other four tests does a good job. Note that here the size function for  $p_{sup}$  and  $p_{\beta}$  are identical. Also from Figure 3.6 when sample sizes are (40,20) we see that  $BF$  performs poorly. Again the size functions for both the exact tests are identical. In this case both  $p_{au}$  and  $LS$  do relatively well compared to the exact tests. From Figure 3.7 (when sample sizes are (50,100)) we see that there is a significant gain in using  $p_{\beta}$ ,  $p_{au}$  or  $LS$  instead of  $p_{sup}$  or  $BF$ . Both of them (especially  $p_{sup}$ ) yields very poor results in this case. And lastly for the case of  $(N_1, N_2) = (100, 50)$  (see Figure 3.8)  $p_{\beta}$  and  $p_{au}$  are the definite winners. Again  $p_{sup}$  and  $BF$  gives the worst results. The Bayesian test  $LS$  turns out to be very liberal, especially for values of  $P_1$  near 0. The size for this test is 0.067 which is some thing to worry about. The size function plot gradually falls and after values of  $P_1$  near 0.3 it more or less stays near the nominal line. Evidently it is not doing too bad but it does show that things might get worse. In fact from the results that we have presented this is the first time we saw the size going above 0.06.

Again we did further comparisons with more sample size choices and different values of  $\pi_0$ . Except for the Bayesian test  $LS$  things performed reasonably well for the hypothesis of efficacy problem when we are testing for the alternative  $H_1 : \pi > \pi_0$  for small non-zero values of  $\pi_0$ . The problem that we found for the  $LS$  test was that as the unbalancedness increased (especially when  $N_1 < N_2$ ) the size of the test



Table 3.6:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Type II error for specific choices of  $(P_1, P_2)$  when  $N_1 \neq N_2$

| $(N_1, N_2)$ | $(P_1, P_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|--------------|-----------|-----------|----------|--------|--------|
| (20,40)      | (0.2,0.5)    | 0.3633    | 0.3633    | 0.3633   | 0.3634 | 0.4193 |
| (20,40)      | (0.3,0.5)    | 0.6981    | 0.6981    | 0.6981   | 0.6981 | 0.7504 |
| (40,20)      | (0.2,0.5)    | 0.3348    | 0.3348    | 0.3319   | 0.3404 | 0.3546 |
| (40,20)      | (0.3,0.5)    | 0.6894    | 0.6894    | 0.6743   | 0.6841 | 0.7111 |
| (50,100)     | (0.2,0.5)    | 0.0908    | 0.0579    | 0.0579   | 0.0603 | 0.0788 |
| (50,100)     | (0.3,0.5)    | 0.5170    | 0.4307    | 0.4290   | 0.4332 | 0.4930 |
| (100,50)     | (0.2,0.5)    | 0.0583    | 0.0487    | 0.0464   | 0.0450 | 0.0564 |
| (100,50)     | (0.3,0.5)    | 0.4367    | 0.4019    | 0.3992   | 0.3949 | 0.4441 |

increased drastically. We will discuss this in detail once we have presented all the results for the unbalanced cases, i.e., the Type II error, power, and the total error.

From Table 3.6 we see that when  $(N_1, N_2)$  is (20,40), (40,20) and (100,50) except for  $BF$  all the tests have similar type II error. The type II error for  $BF$  is slightly higher in these cases (especially when  $(P_1, P_2)=(0.3,0.5)$ ). For (50,100) the type II error for  $p_\beta$ ,  $p_{au}$  and  $LS$  are much lower than  $p_{sup}$  and  $BF$ . Which means that they will have higher power compared to  $p_{sup}$  and  $BF$ . Table 3.7 shows the actual values of power for the given choices of  $(P_1, P_2)$ .

Finally in Table 3.8 we present the total error for all the five tests. We see that

Table 3.7:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Power for specific choices of  $(P_1, P_2)$  when  $N_1 \neq N_2$

| $(N_1, N_2)$ | $(P_1, P_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|--------------|-----------|-----------|----------|--------|--------|
| (20,40)      | (0.2,0.5)    | 0.6367    | 0.6367    | 0.6367   | 0.6366 | 0.5807 |
| (20,40)      | (0.3,0.5)    | 0.6981    | 0.6981    | 0.6981   | 0.6981 | 0.7504 |
| (40,20)      | (0.2,0.5)    | 0.6652    | 0.6652    | 0.6681   | 0.6596 | 0.6454 |
| (40,20)      | (0.3,0.5)    | 0.3106    | 0.3106    | 0.3257   | 0.3159 | 0.2889 |
| (50,100)     | (0.2,0.5)    | 0.9091    | 0.9421    | 0.9421   | 0.9397 | 0.9212 |
| (50,100)     | (0.3,0.5)    | 0.4830    | 0.5693    | 0.5710   | 0.5668 | 0.5070 |
| (100,50)     | (0.2,0.5)    | 0.9416    | 0.9513    | 0.9536   | 0.9550 | 0.9436 |
| (100,50)     | (0.3,0.5)    | 0.5633    | 0.5983    | 0.6008   | 0.6051 | 0.5559 |

the total error for the Bayesian test  $LS$  is comparable to the other three frequentist tests. For most of the cases the total error for  $BF$  is higher than the other four tests followed by  $p_{sup}$ .

From our analysis we have come to the following conclusion. For this specific problem of testing for efficacy the approximate unconditional test  $p_{au}$  and the Bayesian test  $LS$  gives reasonably good results. Except for highly unbalanced samples we did not see the size go up too much above the 0.05 level and power of the tests were very much comparable to the exact methods. For the Bayesian test  $LS$  we have seen that highly unbalanced samples causes serious problems and one should use this test

Table 3.8:  $H_0 : \pi \leq 0.1$  versus  $H_1 : \pi > 0.1$ : Total error for specific choices of  $(P_1, P_2)$  when  $N_1 \neq N_2$

| $(N_1, N_2)$ | $(P_1, P_2)$ | $p_{sup}$ | $p_\beta$ | $p_{au}$ | $LS$   | $BF$   |
|--------------|--------------|-----------|-----------|----------|--------|--------|
| (20,40)      | (0.2,0.5)    | 0.4120    | 0.4120    | 0.4120   | 0.4183 | 0.4590 |
| (20,40)      | (0.3,0.5)    | 0.7468    | 0.7468    | 0.7468   | 0.7468 | 0.7901 |
| (40,20)      | (0.2,0.5)    | 0.3833    | 0.3833    | 0.3833   | 0.3921 | 0.4021 |
| (40,20)      | (0.3,0.5)    | 0.7379    | 0.7379    | 0.7257   | 0.7358 | 0.7586 |
| (50,100)     | (0.2,0.5)    | 0.1291    | 0.1060    | 0.1082   | 0.1168 | 0.1212 |
| (50,100)     | (0.3,0.5)    | 0.5553    | 0.4788    | 0.4793   | 0.4897 | 0.5354 |
| (100,50)     | (0.2,0.5)    | 0.1075    | 0.0977    | 0.0989   | 0.1122 | 0.0988 |
| (100,50)     | (0.3,0.5)    | 0.4859    | 0.4509    | 0.4517   | 0.4621 | 0.4894 |

with caution. We shall discuss this shortly. In fact for more balanced samples we can definitely conclude that amongst the five tests the standard exact test  $p_{sup}$  and the  $BF$  test are the most conservative ones. In terms of computational ease  $p_{sup}$  is the worst followed by  $p_\beta$  and  $p_{au}$ . The two Bayesian tests will be very easy to implement in practice. The confidence interval p value test  $p_\beta$  performed very nicely. For most of the cases it had one of the highest power and the size never goes above the nominal level. One of the reason for doing this comparison was to show that  $p_\beta$  should be a definite choice between exact, approximate, asymptotic and Bayesian methods. It maintains the nominal size plus yields very good results in terms of power.

The appeal for the approximate or the Bayesian tests lies in the fact that they can be very easily implemented in practice. The computation time is reduced significantly if we use the non-exact tests. In fact while presenting the non-exact  $p$  values it is always possible to present the size of the test along with it. This would be useful in the sense that if it is performing very poorly then one can resort to exact methods. But it will not be always possible to present the size for the exact tests because of computing time.

There is one important point that the readers must be aware of at this time. Throughout our discussion in this Chapter we have restricted ourselves to the specific problem of testing for efficacy. Unlike Chapter 1 we did not consider the general testing problem for the relative risk where we have to take in to account all possible values of  $R$  and test for both the alternatives  $H_1 : R < R_0$  and  $H_1 : R > R_0$ . The reason for this is because the non-exact methods are not at all reliable for the general testing problem. We had looked at several cases and found out that the results were quite poor. As we change the alternative and move across different values of  $R$  the size of the test can go up by quite a bit above the nominal size. So we do not recommend people use these tests for the general testing problem for relative risk. If maintaining the right size is of importance one has hardly any choice but to resort to exact methods. Again confidence intervals constructed from either  $p_{au}$  or the Bayesian tests will result in shorter intervals but will have poor coverage. It is always possible to construct a level  $\alpha$  test by using the Bayes Factor test if one is

willing to tabulate the cutoff points so that the size never exceeds the nominal size. But this will be a tedious task and may not have much practical appeal.

Another thing we should keep in mind is that we have tried to use the Bayesian tests in a frequentist sense and compared its size and type II error. However there is nothing wrong to just use this test from a Bayesian perspective and use it for the general testing problem for the relative risk. For example what ever be the value of  $R$  or the direction of the test one can still compute the Bayes Factor and interpret it.

### Highly Unbalanced Samples

As we have mentioned earlier unbalancedness caused problems for the Bayes  $LS$  test. Recall that when the sample sizes were  $(100,50)$  and  $(50,100)$  we saw the size to be 0.0672 and 0.0565. We investigated this in more detail by considering some highly unbalanced pairs, viz.,  $(200,50)$ ,  $(50,200)$ ,  $(1000,50)$ ,  $(50,1000)$ . In Figures 3.9, 3.10, and 3.11 we are presenting the size function plots for the  $LS$  test when the sample sizes are  $(200,50)$ ,  $(50,200)$ , and  $(200,200)$ . We see that for the unbalanced cases the size functions are either highly skewed to the left  $(200,50)$  or highly skewed to the right  $(50,200)$ . The reason for presenting the  $(200,200)$  plot is to show the readers that it still performs well when the samples are balanced.

Also we notice that when ever  $(N_1 < N_2)$  it takes a sharper peak compared to when  $(N_1 > N_2)$ . In Figures 3.12 and 3.13 we considered the very extreme cases of

(1000,50) and (50,1000). We see that for sample sizes (1000,50) the size goes up to 1 for very small values of  $P_1$  close to 0 and falls sharply. For (50,1000) although it is hardly as bad, still it goes up to almost 0.08.

Recall that for the *LS* test we would reject the null hypothesis when  $R_{0.95} < R_0$ . So if one plots the distribution of the 95th percentile of the relative risk  $R$  one may be able to see what is going on. Now if we define  $W = P_1 - R_0 P_2$  then we can write the hypothesis of efficacy in terms of  $W$ . After simplification this reduces to testing for the alternative  $H_1 : W < 0$ . In this case we will reject the null hypothesis if  $W_{0.95} < 0$  (where  $W_{0.95}$  is the upper 95th percentile of the distribution of  $W$ ). We actually plotted the distribution of the 95th percentile of  $W$  for the (1000,50) and the (50,1000) cases when the true values of  $(P_1, P_2 = P_1/R_0)$  are (0.01,0.01/0.9) and (0.8,0.8/0.9). Recall that here  $R_0$  was 0.9. The reason for choosing  $P_1 = 0.01$  and 0.8 was because these were the regions where we saw the sharp peaks in Figures 3.12 and 3.13.

Figures 3.14 and 3.15 gives the histogram plots for the distribution of the 95th percentile of  $W = P_1 - R_0 P_2$  when  $(N_1, N_2) = (1000, 50)$  and  $(P_1, P_2 = P_1/R_0) = (0.01, 0.01/0.9)$  and  $(0.8, 0.8/0.9)$ . We see from Figure 3.14 that in less than 15% of the cases the values are less than 0. This matches with the results from the actual size plot. From Figure 3.12 we see that although the size is 1 for  $P_1$  very close to 0, but it takes a sharp plunge and for values of  $P_1$  near 0.01 the probability is close to

0.15. From Figure 3.15 we see that around 4% of the time the values are less than 0. This again collaborates with the size plot for (1000,50). We do not see the size of the test to be very high near the region of  $P_1 = 0.8$  (see Figure 3.12). It is about 0.04. Again if we look at Figures 3.16 and 3.17 we can find a similar explanation. In Figure 3.16 we see that all the values are greater than 0. This matches with the fact that the size is 0 for  $(N_1, N_2) = (50, 1000)$  for values of  $P_1$  near 0.01 (see Figure 3.13). Again from Figure 3.17 we see that around 6-7% of the time the values are less than 0. Which again accounts for the probability being close to 0.065 for this case (near values of  $P_1 = 0.8$ , see Figure 3.13).

One important point to note is that since for our application we will expect very small values of  $P_1$ , i.e., a low disease incidence rate for the vaccine group one should be more worried about the case when  $N_1 \gg N_2$ . And there is a good possibility of seeing such unbalancedness in practice. For example if we know that the vaccine is quite effective one would design the study in such a way so as to have more subjects in the vaccine group rather than in placebo.

From these examples we see that one has to be very cautious in using the Bayes test derived from the Lindley-Smith approach for highly unbalanced sample sizes. In this case a better alternative would be to use the Bayes Factor test.

## 3.5 Conclusion

In this Chapter we dealt with the problem of testing for efficacy which is a special case of the general testing problem for the relative risk. In Chapters 1 and 2 we had considered the general testing problem and interval estimation for the relative risk parameter in depth. We had shown how the asymptotic methods failed to bring the desired results in terms of maintaining the correct size and why exact tests were necessary.

So Chapter 3 may sound as a contradiction to what ever we have said previously. Because here we have been endorsing non-exact methods. However that was not at all the point of this Chapter. From Chapters 1 and 2 it should be clear to the readers that although exact tests maintain the correct size but it does not come without a price. First of all it is difficult to implement in practice (mainly because of the computational complexities). Secondly it can be very conservative (especially the standard exact test  $p_{sup}$ ). For all these reasons we thought that if there were other alternatives that could be used at least for some specific application it might be worthwhile to compare them with the exact methods. And that is precisely what we did in this Chapter. We compared the two exact tests with an approximate test and two Bayesian tests. The appeal for the non-exact tests lies in their simplicity of implementation. From our results what we can conclude is that  $p_\beta$ ,  $p_{au}$ , and  $LS$  are all very useful tests for this application. Both  $p_{sup}$  and  $BF$  performed rather poorly.



While using the non-exact tests we recommend that one present the size of the test along with the p value so that it can be determined how accurate (or inaccurate) the results are.

The Bayesian tests are based on completely different assumptions and it will be unfair to just rely on a comparison from a frequentist sense. But even from a frequentist point of view the *LS* test performed quite well. As we have mentioned earlier none of the non-exact methods should be used for the general hypothesis testing problem for the relative risk. Also confidence intervals computed by inverting these tests will often have poor coverage (lower than the nominal coverage). But for the specific testing problem that we have considered they can serve as a very useful tool.

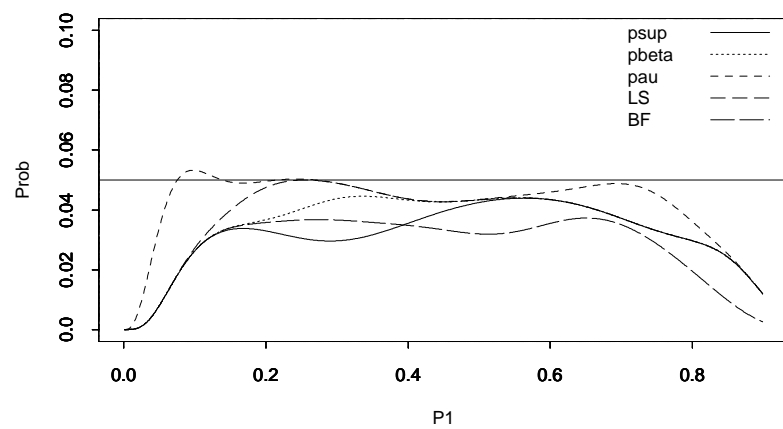


Figure 3.1: Size plots when  $(N_1, N_2)=(20,20)$

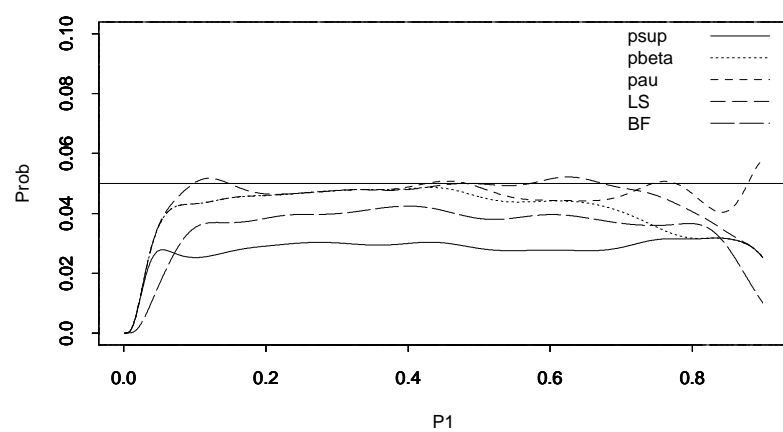


Figure 3.2: Size plots when  $(N_1, N_2)=(50,50)$

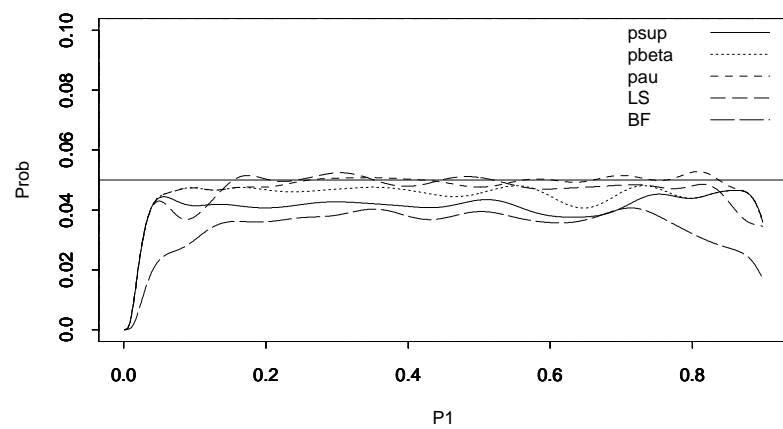


Figure 3.3: Size plots when  $(N_1, N_2)=(75,75)$

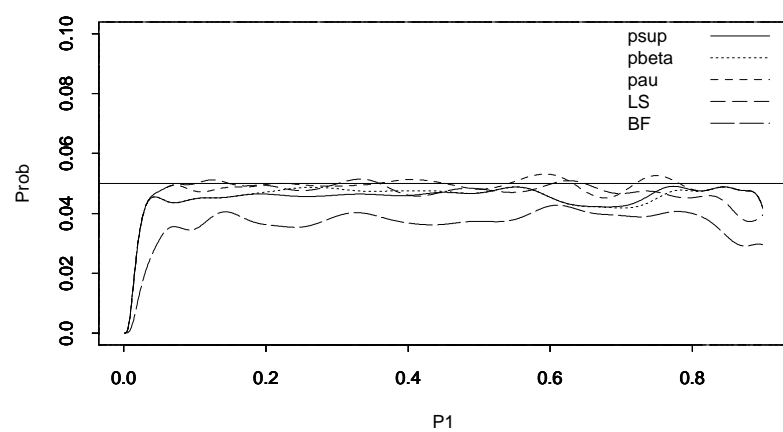


Figure 3.4: Size plots when  $(N_1, N_2)=(100,100)$

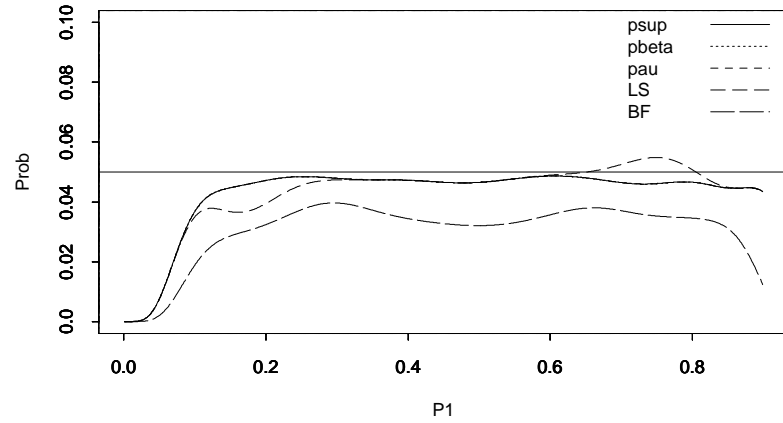


Figure 3.5: Size plots when  $(N_1, N_2)=(20,40)$ . The size function for  $p_{sup}$  and  $p_{\beta}$  are identical in this case

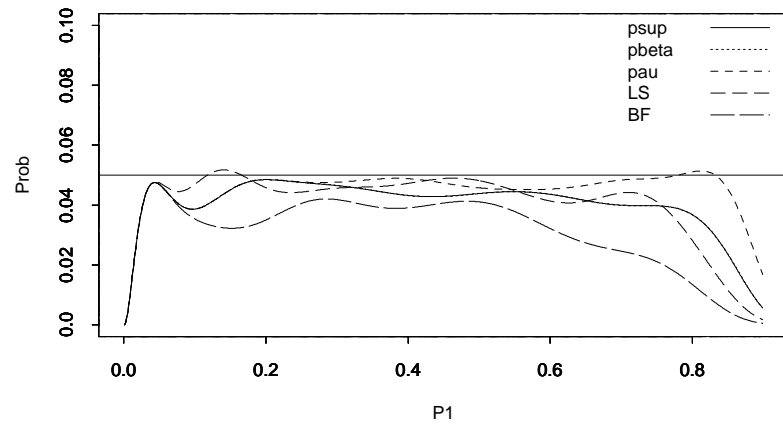


Figure 3.6: Size plots when  $(N_1, N_2)=(40,20)$ . The size function for  $p_{sup}$  and  $p_{\beta}$  are identical in this case

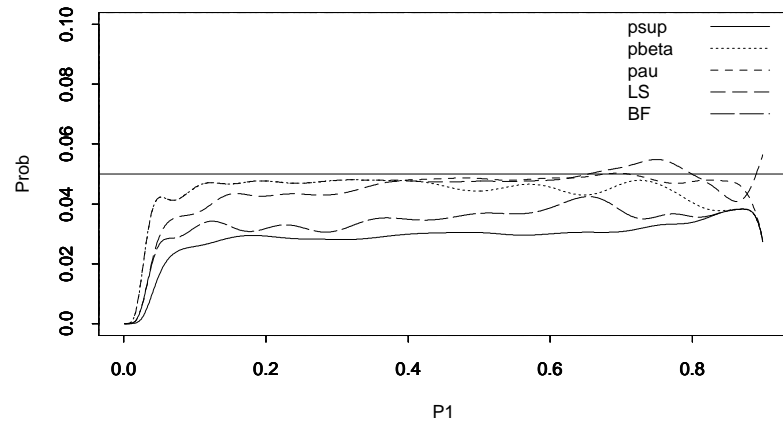


Figure 3.7: Size plots when  $(N_1, N_2)=(50,100)$

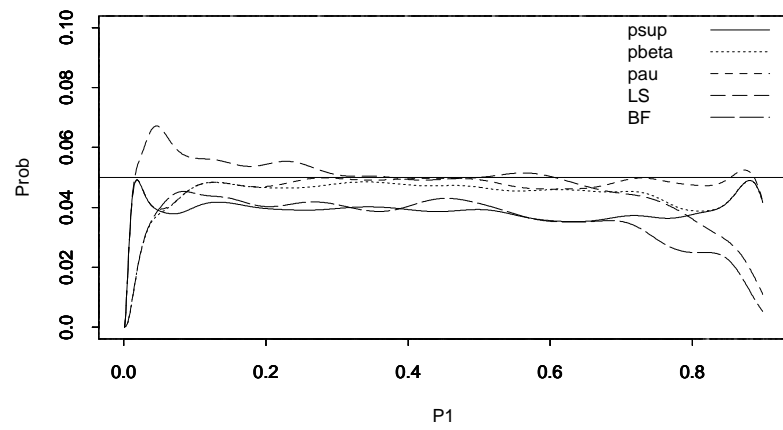


Figure 3.8: Size plots when  $(N_1, N_2)=(100,50)$

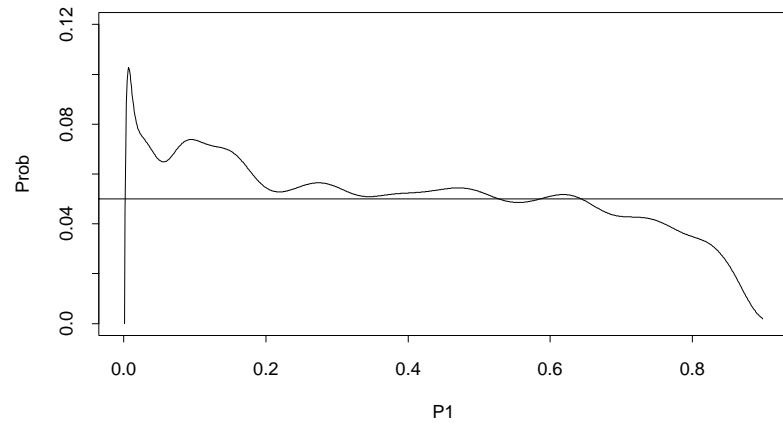


Figure 3.9: Size plot for the  $LS$  test when  $(N_1, N_2)=(200,50)$

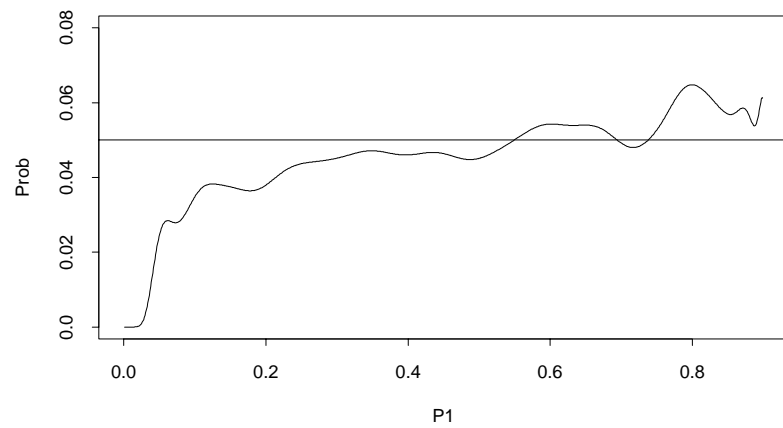


Figure 3.10: Size plot for the  $LS$  test when  $(N_1, N_2)=(50,200)$

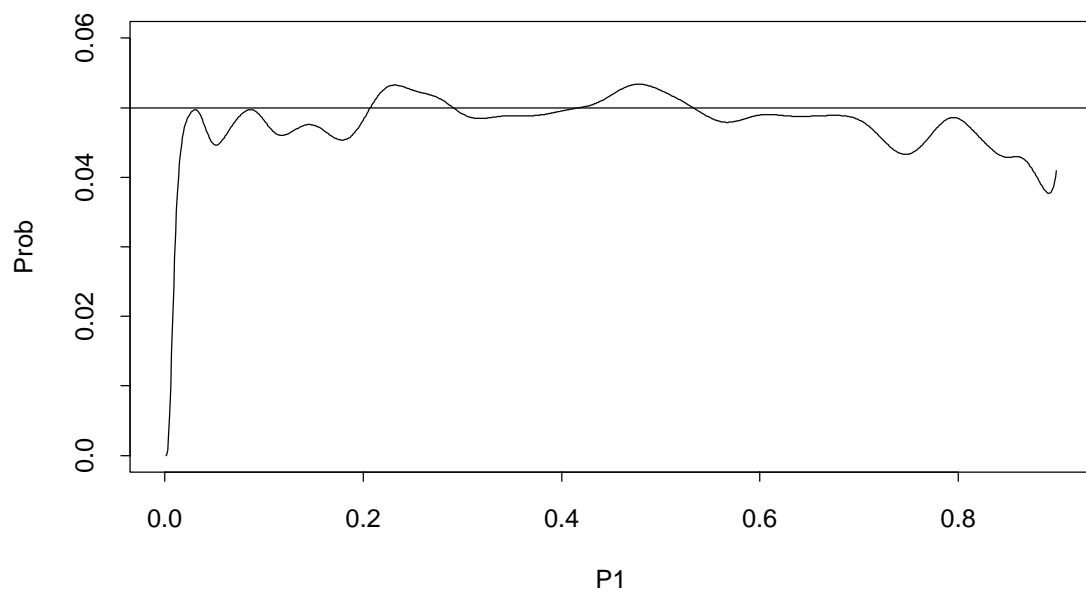


Figure 3.11: Size plot for the  $LS$  test when  $(N_1, N_2)=(200,200)$

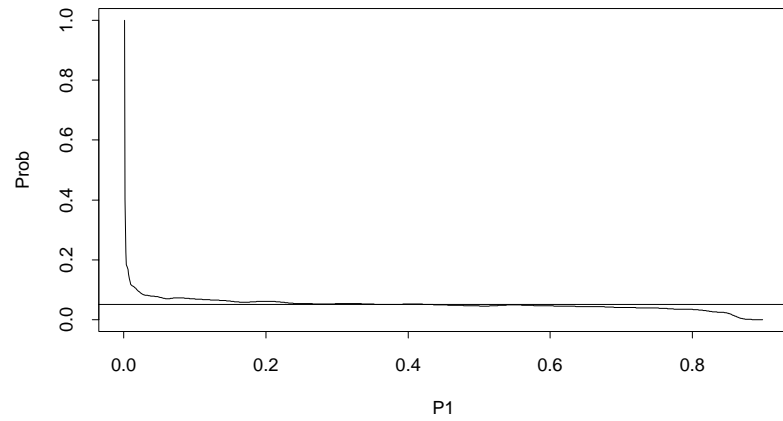


Figure 3.12: Size plot for the  $LS$  test when  $(N_1, N_2) = (1000, 50)$

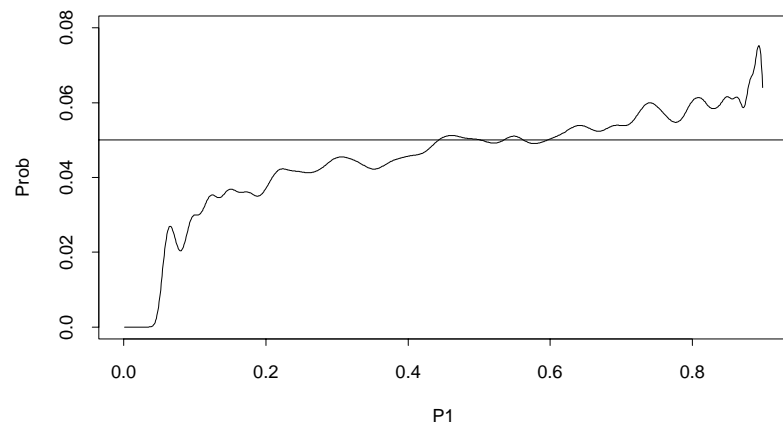


Figure 3.13: Size plot for the  $LS$  test when  $(N_1, N_2) = (50, 1000)$



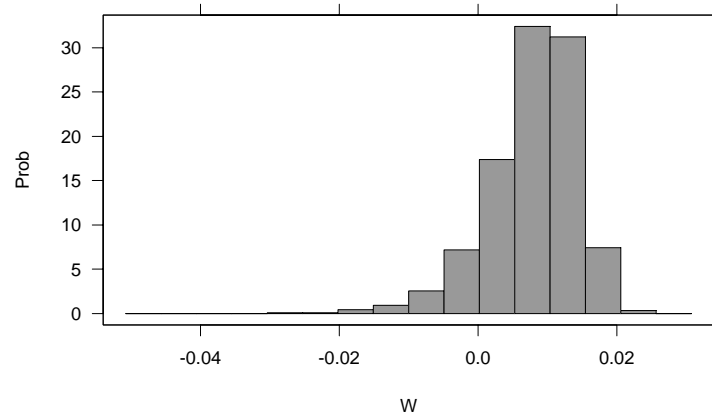


Figure 3.14: Distribution of the 95th percentile of  $W$  when  $(N_1, N_2)=(1000,50)$  and  $(P_1, P_2)=(0.01,0.01/0.9)$

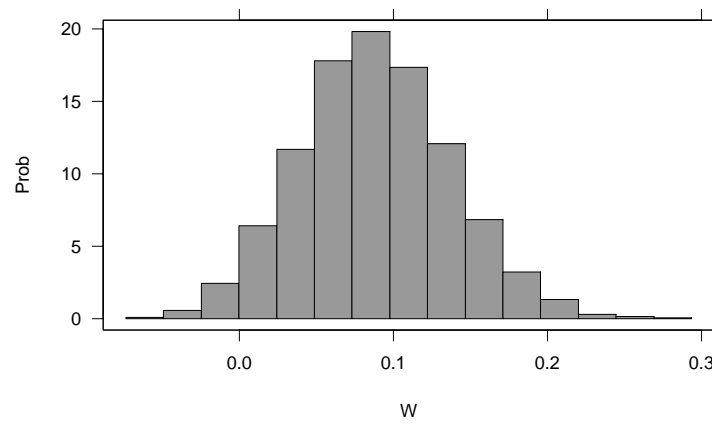


Figure 3.15: Distribution of the 95th percentile of  $W$  when  $(N_1, N_2)=(1000,50)$  and  $(P_1, P_2)=(0.8,0.8/0.9)$

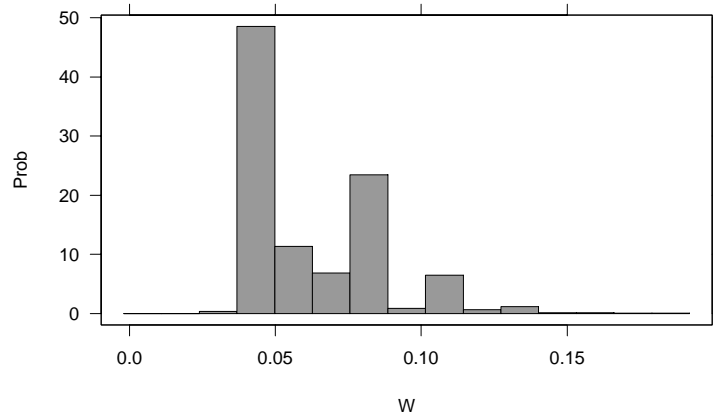


Figure 3.16: Distribution of the 95th percentile of  $W$  when  $(N_1, N_2)=(50,1000)$  and  $(P_1, P_2)=(0.01, 0.01/0.9)$

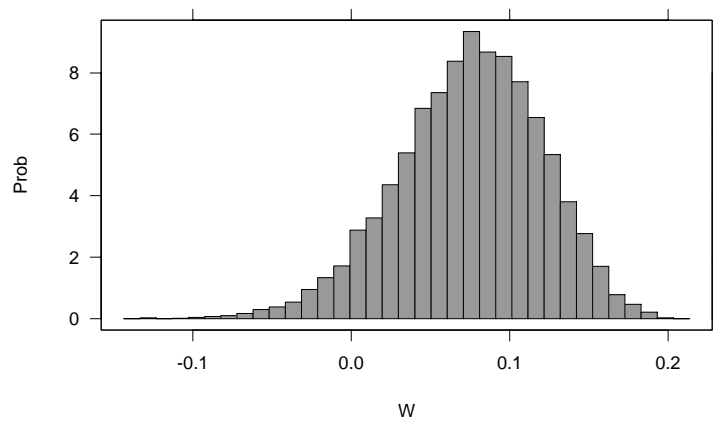


Figure 3.17: Distribution of the 95th percentile of  $W$  when  $(N_1, N_2)=(50,1000)$  and  $(P_1, P_2)=(0.8, 0.8/0.9)$

# Bibliography

Agresti, A. & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* **54**(4), 280–288.

Agresti, A. & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.

Agresti, A. & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**(3), 963–971.

Agresti, A. & Min, Y. (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics* **3**(3), 379–386.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association* **72**, 355–366.

Berger, R. L. (1996). More powerful tests from confidence interval  $p$  values. *The American Statistician* **50**, 314–318.

- Berger, R. L. & Boos, D. D. (1994).  $p$  values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**, 1012–1016.
- Bernard, G. (1947). Significance tests for  $2 \times 2$  tables. *Biometrika* **34**, 123–138.
- Blyth, C. R. & Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association* **78**, 108–116.
- Casella, G. (1986). Refining binomial confidence intervals. *The Canadian Journal of Statistics* **14**, 113–129.
- Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Duxbury Press.
- Casella, G. & Robert, C. (1989). Refining Poisson confidence intervals. *The Canadian Journal of Statistics* **17**, 45–57.
- Chan, I. S. F. (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine* **17**, 1403–1413.
- Chan, I. S. F. & Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55**, 1202–1209.
- Clopper, C. & Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.
- Crow, E. (1956). Confidence intervals for a proportion. *Biometrika* **43**, 423–435.

- Crow, E. & Gardner, R. (1956). Confidence intervals for the expectation of a poisson variable. *Biometrika* **46**, 441–453.
- Farrington, C. P. & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **9**, 1447–1454.
- Fisher, R. A. (1935). The design of experiment. *Oliver and Boyd, Edinburgh*.
- Fries, et al. (1993). Safety and immunogenicity of a recombinant protein influenza a vaccine in adult human volunteers and protective efficacy against wild-type h1n1 virus challenge. *Journal of Infectious Diseases* **167**, 593–601.
- Haber, M. (1986). An exact unconditional test for the 2?2 comparative trial. *Psychological Bulletin* **99**, 129–132.
- Jeffreys, H. (1935). Some tests of significance treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society* **31**, 203–222.
- Kang, S. H. & Chen, J. J. (2000). An approximate unconditional test of non-inferiority between two proportions. *Statistics in Medicine* **19**, 2089–2100.
- Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.
- Kopit, J. (1999). A more powerful exact test for a practical difference between two binomial proportions. *Doctoral Dissertation: North Carolina State University*.

- Miettinen, O. & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4**, 213–226.
- Rodary, C., Com-Nougue, C. & Tournade, M.-F. (1989). How to establish equivalence between treatments: A one-sided clinical trial in paediatric oncology. *Statistics in Medicine* **8**, 593–598.
- SAS (2002). SAS, Version 8.1, Cary, NC.
- Sidik, K. (1997). Exact unconditional tests for discrete data. *Doctoral Dissertation: North Carolina State University*.
- StatXact Manual (2002). StatXact, Version 5, Cambridge, MA.
- Sterne, T. (1954). Some remarks on confidence or fiducial limits. *Biometrika* **41**, 275–278.
- Storer, B. E. & Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association* **85**, 146–155.
- Suissa, S. & Shuster, J. J. (1985). Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society, Series A, General* **148**, 317–327.
- Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine* **12**, 809–824.