

A COMPARISON OF HYPOTHESIS TESTING PROCEDURES FOR TWO  
POPULATION PROPORTIONS

by

MOLLY HORT

A.S., Garden City Community College, 2003

B.S., Kansas State University, 2005

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics  
College of Arts and Sciences

KANSAS STATE UNIVERSITY  
Manhattan, Kansas

2008

Approved by:

Major Professor  
Dr. John Boyer

## **Abstract**

It has been shown that the most straightforward approach to testing for the difference of two independent population proportions, called the Wald procedure, tends to declare differences too often. Because of this poor performance, various researchers have proposed simple adjustments to the Wald approach that tend to provide significance levels closer to the nominal. Additionally, several tests that take advantage of different methodologies have been proposed.

This paper extends the work of Tebbs and Roths (2008), who wrote an R program to compare confidence interval coverage for a variety of these procedures when used to estimate a contrast in two or more binomial parameters. Their program has been adapted to generate exact significance levels and power for the two parameter hypothesis testing situation.

Several combinations of binomial parameters and sample sizes are considered. Recommendations for a choice of procedure are made for practical situations.

## Table of Contents

Acknowledgements.....	iv
CHAPTER 1 - Introduction and the Wald Estimator .....	1
CHAPTER 2 - Alternatives to the Wald Interval .....	4
Limitations of Wald Approach and Reason for Studying Other Estimators .....	4
Alternatives.....	4
Laplace-Wald.....	4
Price-Bonett .....	4
Haldane and Jeffreys-Perks.....	5
CHAPTER 3 - Problems with Estimators.....	6
CHAPTER 4 - Tebbs/Roths Work.....	7
Confidence Interval Approach to Linear Combinations.....	7
Description of R Program.....	8
Converting to a Power Study for Difference in Two Proportions .....	8
Description of the Power Study .....	10
CHAPTER 5 - Exact Results of Power Study .....	12
Results for $p_1 = 0.0$ .....	12
Results for $p_1 = 0.10$ .....	13
Results for $p_1 = 0.20$ .....	13
Results for $p_1 = 0.30$ .....	14
Results for $p_1 = 0.40$ .....	14
Results for $p_1 = 0.50$ .....	14
CHAPTER 6 - Recommendations .....	16
References.....	17
Appendix A - R Program .....	19
Appendix B - Power Study Results .....	26

## **Acknowledgements**

I would first like to thank my major professor, Dr. John Boyer, for allowing me to do my master's report research under him. He has been extremely helpful and always willing to provide guidance when needed, even though I know it has been difficult to hold the department head position as well as work with a student's master's project. I feel very fortunate to have had the chance to work with Dr. Boyer.

I also would like to thank the members of my committee, Dr. James Higgins and Dr. Suzanne Dubnicka. They are both amazing instructors and I am glad they agreed to serve on my committee.

Lastly, my research would not have been possible without the work of Joshua M. Tebbs and Scott A. Roths. I extended the work done by them in two of their previous papers and could not have done so without their continued correspondence. They were always willing to answer the questions Dr. Boyer and I had while working on this report. Thanks to the both of them for all of the help they have provided.

## CHAPTER 1 - Introduction and the Wald Estimator

Binomial parameters are of interest in many different areas of statistical research. In particular, two sample problems are often investigated to see if there is a difference in two binomial proportions. One can use confidence intervals to estimate the difference in proportions or do a formal hypothesis test with the null hypothesis being that  $H_0: p_1 = p_2$ . Because of the widespread use of binomial parameters, various statisticians have developed confidence interval formulas and testing procedures for dealing with binomial random variables. Pertinent references regarding confidence interval construction include Brown, Cai, and DasGupta (2001), Reiczigel (2003), Zhou, Tsao and Qin (2004), and Brown and Li (2005).

The usual solution to this problem, which is taught in introductory statistics courses, is to use the asymptotic normality of the sample fractions and test a hypothesis about the difference between two parameters using maximum likelihood methods. First, suppose that  $X_1$  and  $X_2$  are independent binomial random variables such that  $X_1 \sim \text{Bin}(n_1, p_1)$  and  $X_2 \sim \text{Bin}(n_2, p_2)$ . Then, an approximate  $100(1-\alpha)\%$  confidence interval for the difference in proportions,  $p_1 - p_2$  is defined as

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad (1.1)$$

where the maximum likelihood estimator (MLE) of  $\hat{p}_i$  is  $\frac{x_i}{n_i}$ .

An equivalent test of  $H_0: p_1 = p_2$  versus  $H_a: p_1 \neq p_2$  would call for one to reject  $H_0$  if

$$z = \left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \right| \geq z_{\alpha/2} \quad (1.2)$$

where  $z_{\alpha/2}$  is the standard normal deviate.

This confidence interval and testing procedure was developed by Abraham Wald and is perhaps the most straightforward way to test two-tailed hypotheses about two binomial proportions (Agresti & Caffo, 2000). Clearly, there is a comparable version of this test for a one-tailed alternative.

To illustrate the two ideas given above, consider the example below (Anderson & Williams, 2008).

*Example:* A 2003 *New York Times*/CBS News poll sampled 523 adults who were planning a vacation during the next six months and found that 141 were expecting to travel by airplane. A similar survey question in a May 1993 *New York Times*/CBS News poll found that of 477 adults who were planning a vacation in the next six months, 81 were expecting to travel by airplane.

A 95% confidence interval for the difference in proportions for the two polls can be computed as follows:

$$\begin{aligned}\hat{p}_1 &= \frac{x_1}{n_1} = \frac{141}{523} = 0.27 & \hat{p}_2 &= \frac{x_2}{n_2} = \frac{81}{477} = 0.17 \\ (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= (0.27 - 0.17) \pm 1.96 \sqrt{\frac{0.27(1-0.27)}{523} + \frac{0.17(1-0.17)}{477}} \\ &= 0.10 \pm 0.0508 \\ &= (0.0490, 0.1506)\end{aligned}$$

Therefore, the difference in proportions is estimated to be between 0.0490 and 0.1506. In other words, the percentage of adults who were planning a vacation in the next six months was estimated to be between 4.9% and 15.06% more in 2003 than in 1993. From this confidence interval, it can be concluded with 95% confidence that there is a non-zero difference between the two population proportions.

Equivalently, one could perform a hypothesis test with this data to test the equality of population proportions. In that case, the hypotheses

$$H_0: p_1 = p_2 \quad \text{vs.} \quad H_a: p_1 \neq p_2$$

would be tested by generating the test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0.27 - 0.17}{\sqrt{\frac{0.27(1-0.27)}{523} + \frac{0.17(1-0.17)}{477}}} = 3.85$$

This results in a two-sided p-value  $< 0.0001$ . The null hypothesis is rejected at the 5% level of significance and a difference between the two population proportions is concluded. This is consistent with the results of the confidence interval approach above.

As mentioned previously, the aforementioned confidence interval is commonly referred to as the Wald interval. However, as will be discussed later in this paper, the Wald interval performs poorly under certain conditions (Agresti & Caffo, 2000). Because of this, others have developed alternative approaches to constructing binomial confidence intervals that seem to perform better under those conditions.

Chapter 2 of this paper begins with a discussion of the limitations of the Wald interval and suggests why there is a need to adjust confidence intervals to account for these limitations. Other proposed estimators will also be discussed in Chapter 2, including the Laplace-Wald, Price-Bonett, Haldane, and Jeffreys-Perks intervals. Chapter 3 will discuss the issues with the aforementioned confidence intervals, both computationally and performance-wise. Chapter 4 will present two new interval estimators proposed by Roths and Tebbs in their 2006 paper. It will also describe their computer program, which computes coverage probabilities for each of the estimators described. Finally, there will be an explanation of how to use their program to compute exact levels of significance and power in performing a two-tailed test of hypotheses about the difference between two independent binomial parameters. Chapter 5 will present the results of the power study. The paper will conclude in Chapter 6 with some recommendations and conclusions based on the power study.

## CHAPTER 2 - Alternatives to the Wald Interval

### Limitations of Wald Approach and Reason for Studying Other Estimators

It has been shown by Price and Bonett (2004) that confidence intervals found using the Wald approach discussed in Chapter 1 often do not give the desired nominal coverage.

Generally, the Wald approach behaves poorly when sample sizes are small, with the most dramatic results when each  $p_i$  is near zero or one. Moreover, Agresti and Caffo (2000) note that the coverage is often well below the nominal level, regardless of how large the sample size is.

Because of this lowered coverage probability, simple adjustments have been proposed by various researchers to obtain coverage probabilities closer to the nominal level.

### Alternatives

#### *Laplace-Wald*

The simplest adjustment to account for this lowered coverage probability with small sample sizes is to add a count of one to the successes and a count of one to the failures for each sample and then compute the usual Wald interval based on this new dataset. In other words,  $n_i$  and  $\hat{p}_i$  in equation (1.1) are replaced by  $n_i^* = n_i + 2$  and  $\hat{p}_i^* = (x_i + 1) / n_i^*$ . This interval seems to require a relatively minor adjustment but tends to improve coverage probabilities without making the computation more difficult. Adjustments of this sort were proposed by Wilson (1927). However, this paper uses the specific version described in Greenland (2001).

#### *Price-Bonett*

Another simple adjustment as mentioned in Roths and Tebbs (2008) was developed by Price and Bonett, who demonstrated that their interval holds closest to the correct coverage probability. In a fashion similar to the Laplace-Wald Interval, the Price-Bonnett modifications involves adding a count of  $2/k$  successes and a count of  $2/k$  failures for each sample, where  $k$  is the number of nonzero constants among  $c_1, c_2, \dots, c_k$  and  $c_1, c_2, \dots, c_k$  are the specified constants for the interval estimate of a linear combination  $\theta = c_1 p_1 + c_2 p_2 + \dots + c_k p_k$  of  $k$  binomial proportions. In our two sample case,  $k$  will be 2. Therefore,  $n_i$  and  $\hat{p}_i$  in (1.1) are replaced by  $n_i^* = n_i + 2$  and



$\hat{p}_i^* = (x_i + 1) / n_i^*$ . Clearly, this interval reduces to the Laplace-Wald interval when dealing with a two population problem, as is the focus of this paper. Because of this, any discussions into the Price-Bonett interval will be dropped hereafter.

### ***Haldane and Jeffreys-Perks***

The estimators that are called the Haldane and Jeffreys-Perks estimators result from a Bayesian approach to the problem. One applies a prior distribution to each of the  $p_i$ 's which has a Beta distribution with both parameters equal to  $\alpha$  (thus resulting in a density which is symmetric about  $1/2$ ), and so the joint posterior distribution of the  $P_i$ 's, given  $X_1, X_2, \dots, X_k = x_1, x_2, \dots, x_k$  can be shown to be

$$f(p_1, p_2, \dots, p_k | x_1, x_2, \dots, x_k; \alpha) = \prod_{i=1}^k \frac{p_i^{x_i + \alpha - 1} (1 - p_i)^{n_i - x_i + \alpha - 1}}{\beta(x_i + \alpha, n_i - x_i + \alpha)} \quad (2.1)$$

for  $0 < p_i < 1$ , where  $\beta(x_i + \alpha, n_i - x_i + \alpha) = \frac{\Gamma(x_i + \alpha)\Gamma(n_i - x_i + \alpha)}{\Gamma(n_i + 2\alpha)}$ . One then transforms the

density by considering  $\theta_1 = p_1 - p_2$  and  $\theta_2 = p_1 + p_2$ . In this context,  $\theta_2$  is a nuisance parameter to be dealt with. Following the approach of Beal (1987),  $\theta_2$  is replaced by its posterior mean

$\frac{x_1 + \alpha}{n_1 + 2\alpha} + \frac{x_2 + \alpha}{n_2 + 2\alpha}$ , and a resulting  $100(1-\alpha)\%$  confidence interval is determined from the one-

dimensional posterior for  $\theta_1$ .

It can be shown that the random variable  $\theta_1$  is asymptotically normal, although with a complicated variance. This asymptotic normality is then used to generate a  $100(1-\alpha)\%$  confidence interval estimate of  $\theta_1$ .

The special case of  $\alpha = 0$  results in the Haldane interval. One should note that in this particular case, the value of  $\theta_2$  that is used is the sum of the two maximum likelihood estimators of  $p_1$  and  $p_2$  respectively. Similarly, the special case of applying  $\alpha = 1/2$  results in the Jeffreys-Perks interval.

## CHAPTER 3 - Problems with Estimators

The Wald Interval discussed in Chapter 1 requires very little in the way of computation and is usually the preferred method when teaching introductory statistics students. Though computationally straightforward, this interval generally is not recommended unless sample sizes are very large due to the unsatisfactory coverage probability. As noted previously, a greater problem exists when the population proportions are near zero or one (Roths & Tebbs, 2006). Another problem with the Wald interval is that, because of how the interval is constructed, it is entirely possible to obtain interval limits outside of the interval  $[-1, 1]$ . This makes no sense when dealing with proportions, and thus poses a problem. Additionally, Roths and Tebbs (2008) have shown that the Wald interval is often anticonservative.

The modification proposed by Greenland (2001) is very slight and does not complicate the computations. However, like the Wald interval, it may still generate an interval that makes no sense because of values outside of the interval  $[-1, 1]$ . The Laplace-Wald modification also tends to be a little conservative, resulting in significance levels below the nominal level.

The Haldane interval seems to be anticonservative according to the analysis performed by Roths and Tebbs (2008). Due to the complicated computations and anticonservative nature of the Haldane interval, it appears that this interval would not be a good choice under the conditions cited by Roths and Tebbs. However, those conditions are reduced to a two population case in this paper to investigate the Haldane interval further.

The last interval discussed thus far was the Jeffreys-Perks interval. The Jeffreys-Perks interval seems to not have as many problems as the others. It has been shown by Roths and Tebbs (2008) that the Jeffreys-Perks interval rivals the Laplace-Wald interval. It is, however, considerably more computationally complicated than the simple adjustment needed in the Laplace-Wald interval.

Because of the computational and performance issues identified above for the estimators described thus far, Roths and Tebbs (2006) recently developed two closely related alternatives to deal with this problem. Chapter 4 will describe these new estimators in detail, as well as provide a simplification to the two-population case that is dealt with in this paper.

## CHAPTER 4 - Tebbs/Roths Work

### Confidence Interval Approach to Linear Combinations

In Chapter 2, a description of the procedure for constructing the Haldane and Jeffreys-Perks intervals was given. The specific  $\alpha$  value needed to produce the Haldane interval from Equation (2.1) is  $\alpha = 0$ , and  $\alpha = \frac{1}{2}$  is needed to produce the Jeffreys-Perks interval. Specifying an  $\alpha$  value in these two intervals could create a problem if the  $\alpha$  value specified constitutes a poor choice. In this case, the posterior distribution could be far from what the true posterior distribution should be. Consequently, the posterior mean estimate could be incorrect, and hence, the confidence interval as well. To account for this, Roths and Tebbs (2006) have developed a parametric empirical-Bayesian approach that may work better in such situations. They have developed two different approaches to this problem, one using an estimate of  $\alpha$  based on Maximum-Likelihood (MLE) methods, and the other using Method of Moments (MOM). It should be noted that both of these approaches are motivated by the work performed by Beal (1987).

The MLE approach involves first noting that the joint marginal distribution for  $x_1$  and  $x_2$  depends on  $\alpha$ ; that is

$$f(x_1, x_2 | \alpha) = \binom{n_1}{x_1} \frac{\beta(x_1 + \alpha, n_1 - x_1 + \alpha)}{\beta(\alpha, \alpha)} \cdot \binom{n_2}{x_2} \frac{\beta(x_2 + \alpha, n_2 - x_2 + \alpha)}{\beta(\alpha, \alpha)}.$$

One then treats  $\alpha$  as the unknown parameter here, and  $\alpha$  can be estimated using conventional methods. Roths and Tebbs (2006) estimate  $\alpha$  by maximum likelihood methods, solving

$$\frac{\partial}{\partial \alpha} f(x_1, x_2 | \alpha) = 0$$

for  $\alpha$  using numerical methods. The resulting estimator, computed in exactly the same way as the Haldane and Jeffreys-Perks versions, is called the Empirical Bayesian MLE (EBMLE) estimator (Roths & Tebbs, 2006). They note that on rare occasions, the value of  $\alpha$  produced by this method may be infinite and so they provide a remedy for that situation.

In a similar fashion, they used the method of moments technique to find estimates of  $\alpha$ . Since the means of  $x_1$  and  $x_2$  are both free of  $\alpha$ , it involves equating theoretical second moments

of  $x_1$  and  $x_2$  to the observed second moments and solving for  $\alpha$ . The resulting estimator they labeled Empirical Bayesian MOM (EBMOM).

### **Description of R Program**

Roths and Tebbs (2008) developed an R program to compute confidence interval limits for a linear combination  $\theta = c_1 p_1 + c_2 p_2 + \dots + c_k p_k$  using all intervals discussed in this paper. In addition to the intervals, this program also reports exact coverage probability, exact mean length of the confidence interval, and the conditional mean length ratio for any linear combination of population proportions and any collection of sample sizes.

In their work, Roth and Tebbs (2006) describe their reasons for using the conditional mean length ratio provided by the R program. Desirable confidence intervals are often precise and cover the difference  $p_1 - p_2$  with a probability at the nominal level or above. However, not all confidence intervals obtain this. In those cases, there is a question of whether or not it is worth having a narrower confidence interval that does not cover this difference. This could lead the researcher astray.

In light of this problem, Roth and Tebbs (2006) discussed a method that calls for the researcher to compute the mean length of the confidence interval for those cases where  $p_1 - p_2$  is included ( $\mu_I$ ) and to compute the mean length for those cases where  $p_1 - p_2$  is excluded from the interval ( $\mu_E$ ). Then, the conditional mean length ratio is defined as  $\mu_I/\mu_E$ , for which smaller values are desired. If a larger value occurs (greater than one), this indicates the mean length is smaller when  $p_1 - p_2$  is excluded and larger when  $p_1 - p_2$  is included, which is not desirable to a researcher.

The R program, in its entirety, is in Appendix A of this paper. In addition, this program can be found at <http://www.stat.sc.edu/~tebbs/index.htm>.

### **Converting to a Power Study for Difference in Two Proportions**

Since there has already been significant work regarding confidence interval construction and consideration of which interval works best in certain situations, the focus here is not on confidence intervals, but instead on size and power in a hypothesis test about the difference between two independent binomial parameters. The confidence intervals described thus far can be used as methods to investigate power in such a hypothesis test. The power study focuses on

the null hypothesis  $H_0: p_1 - p_2 = 0$ , and the two-sided alternative  $H_a: p_1 - p_2 \neq 0$ . It has been widely taught in introductory statistics courses that this two sided hypothesis can be tested using a confidence interval around  $p_1 - p_2$ , and rejecting  $H_0$  if and only if the confidence interval for  $p_1 - p_2$  fails to contain zero. Therefore, the R program developed by Roths and Tebbs (2008) can be used with any contrast  $c_1, c_2, \dots, c_k$ .

Restricting attention to the two sample case,  $(c_1, c_2) = (1, -1)$  is used and hence confidence intervals for  $p_1 - p_2$  are constructed based on the methods described in Chapter 2.

The exact coverage probability can then be obtained from the R program, and for those cases where  $p_1$  and  $p_2$  are equal, one can subtract this probability from one to obtain the size of the test.

To obtain power at alternative parameter configurations,  $p_1$  and  $p_2$  are set to differing values. For example, if  $p_1$  is set at 0.5 and  $p_2$  is set at 0.4, then  $p_1 - p_2 = 0.1$ . The program is versatile enough that it allows us to obtain coverage probabilities as if  $p_1 - p_2 = 0$ . Therefore, a type II error is committed here if the resulting interval covers zero, hence the coverage probability is the type II error probability. Therefore, to obtain the power of the test, one needs only to subtract this coverage probability from one. In this fashion, the confidence interval approaches discussed in Chapter 2 may be used to investigate both size and power of the various testing procedures to determine which, if any, produce the desired or optimum results in a multitude of situations.

A description of how to use the R program developed by Roths and Tebbs (2008) follows as it is used in this paper. Although it is written in a more general context, attention is restricted to the use of two sample problems as described in this paper. Once the code has been pasted into R, the function needed is:

```
> results(c(c1,c2),c(n1,n2),c(p1,p2),conflevel,method)
```

As mentioned in the preceding paragraph,  $c_1 = 1$  and  $c_2 = -1$  for the two sample case. Any specified combination of sample sizes  $n_1$  and  $n_2$  are permitted, as well as any combination of population proportions  $p_1$  and  $p_2$ . In the code, `conflevel` is the desired confidence level for the intervals, and `method` is chosen using the following 7 choices:

- 1 – Wald
- 2 – Laplace-Wald
- 3 – Price-Bonett
- 4 – Haldane
- 5 – Jeffreys-Perks
- 6 – EBMLE
- 7 – EBMOM

The program outputs include the exact coverage probability and mean length.

A sample of the output from a run of this program is provided below. The example used assumes two independent binomial random variables  $X_1$  and  $X_2$  such that  $X_1 \sim \text{Bin}(10, 0.2)$  and  $X_2 \sim \text{Bin}(25, 0.7)$ .

```
> results(c(1,-1),c(10,25),c(0.2,0.7),0.95,5)
[1] 0.1817852 0.5692779 1.1952525
```

As noted, this finds, for the Jeffreys-Perks procedure, the exact coverage probability, the exact mean length of the confidence interval, and the conditional mean length ratio, as discussed in this chapter. It is imperative to note that the output from this program includes coverage probabilities of  $p_1 - p_2 = 0$  instead of the true value of  $p_1 - p_2$ , which is  $0.2 - 0.7 = -0.5$  in this example.

### **Description of the Power Study**

The power study involved the use of four different combinations of sample sizes. These combinations were chosen to determine the effect of sample size on power in each of the methods discussed. The four sample size combinations  $(n_1, n_2)$  used were: (10, 10), (10, 25), (25, 10), and (25, 25).

To investigate the effect of proportions on power, a variety of combinations of  $p_1$  and  $p_2$  were used for each of the sample sizes listed above. The value of  $p_1$  was fixed at 0.0 and  $p_2$  was varied from 0.0 to 1.00 by 0.05. In addition,  $p_1$  was fixed at 0.1, 0.2, 0.3, 0.4, and 0.5 while varying  $p_2$  in the same fashion as just mentioned. It is interesting to note why a fixed value of  $p_1 = 0.6$  through  $p_1 = 1.00$  was not used. This is because the power curves for the latter combinations of  $(p_1, p_2)$  would be exactly the same as the power curves for the former combinations of  $(p_1, p_2)$ . This is due to the binomial distribution being used in all calculations so

that, whether looking at the proportion of successes or the proportion of failures, the resulting power would be equivalent.

In all power calculations, a confidence level of 0.95 was chosen, implying, of course, that the corresponding tests of hypotheses use a significance level of  $\alpha = 0.05$ . Certainly, any confidence level could be used to investigate power and the program could be run to see the results.

## CHAPTER 5 - Exact Results of Power Study

The previous section indicated which combinations of  $n_1$ ,  $n_2$ ,  $p_1$ , and  $p_2$  values were used in this study. Given the values stated, there are 24 power curve graphs included in this paper. All of these graphs can be found in Appendix B. The results are described below, according to the fixed value of  $p_1$ . However, one might first note a few properties to look for.

In order to compare power curves, one needs to first note the size of the test. The size of a test is the probability of rejecting the null hypothesis when the null hypothesis is true. This is also known as the type I error rate. Therefore, as mentioned in Chapter 4, one would assess this by looking at the rejection probability where  $p_1 = p_2$ .

If the size is approximately the same as the nominal level of 0.05, the power curves are suitable to compare. Ideally, one would want the size to equal the nominal level, but this is often not the case. As will be seen later, some methods have a size well above the nominal, which indicates that particular method would not be an optimal choice for analysis purposes.

### Results for $p_1 = 0.0$

In all of the different sample size configurations when  $p_1 = p_2 = 0.0$ , it is easy to see that the size of the test is zero. This makes sense because, with  $p_1 = p_2 = 0.0$ , there are no successes in the dataset. Therefore,  $p_1 - p_2 = 0$  and the resulting confidence intervals will all contain zero, thus giving coverage probabilities of one and a size of zero.

Another observation that can be seen from the graphs deals with comparing a small sample size to a relatively larger sample size. As one would expect, by comparing Figure B-1 to Figure B-4, power tends to increase as sample sizes increase. This is the case for all methods discussed in this paper.

It is difficult to compare the six described methods in terms of power when dealing with population proportions that are both zero. For this reason, other results with larger proportions will now be discussed.



### **Results for $p_1 = 0.10$**

Figures B-5 through B-8 provide some interesting information. Figure B-5 shows that the Wald approach has a significance level closer to the nominal level. In contrast, the other interval approaches have a smaller significance level, making them more conservative tests. This significance level discrepancy raises a few questions. Although it appears that the Wald approach results in greater power and the others provide less power, the discrepancy may be due to the conservativeness of those other tests.

Looking at all of the figures relating to  $p_1 = 0.10$ , it appears that the Haldane method results in high power, even while being slightly conservative. As already mentioned, the Wald approach leads to higher power but also tends to be anticonservative. The performance of the Haldane method can be seen best in Figure B-6. The significance level of the test is below the nominal level, but it has higher power than the majority of the other tests. This is consistent across all combinations of sample sizes. Therefore, the Haldane method would be recommended when one population proportion is near 0.10.

### **Results for $p_1 = 0.20$**

Looking at Figures B-9 through B-12, it appears that the Wald approach results in the greatest power. However, upon closer inspection, it can be seen that this approach has a size of nearly two times what it should be. This means that the Wald approach tends to reject the null hypothesis a lot more often, even when it is true. This could cause problems in many applications. Because of this, it appears that the Wald approach is not suitable in maintaining a desired nominal level of coverage. The figures also show that the other testing approaches seem to work well in maintaining the nominal level.

The graphs show that the Haldane method provides higher power, regardless of sample sizes. However, the Haldane method tends to be anticonservative in most sample size configurations. It can be seen that the Jeffreys-Perks approach results in significance levels slightly lower than the nominal level in all sample size configurations. Therefore, this method provides good power while not risking a higher type I error rate. The EBMLE method isn't too far from the Jeffreys-Perks. In a situation where one proportion is near 0.2, the Jeffreys-Perks approach would be recommended, with the EBMLE approach as a reasonable alternative.

### **Results for $p_1 = 0.30$**

Figures B-13 to B-16 provide a similar conclusion as to which method to suggest. Comparing the significance levels for all methods, it is found that the EBMLE approach gives levels very close to the nominal level, being slightly conservative. The Jeffreys-Perks method tends to give high power, but is a little anticonservative.

In addition to having significance levels close to what is desired, the power of the EBMLE test is fairly high. A few of the graphs seem to suggest the Haldane approach would result in greater power, but this is likely the consequence of an anticonservative test. Therefore, it appears the EBMLE procedure would be recommended when one population proportion is near 0.30.

### **Results for $p_1 = 0.40$**

When  $p_1 = 0.40$ , if one is just looking at significance levels, the EBMLE procedure seems to provide the closest to nominal significance level. However, by looking at Figures B-17 to B-20, it is apparent that the EBMLE procedure results in somewhat smaller power than some of the other comparable procedures.

Therefore, if one is to be entirely concerned about the type I error rate and keeping that rate at the nominal level, the EBMLE approach would be suggested. If the type I error rate isn't as much of a concern, the Jeffreys-Perks procedure would be a better fit. This is because the significance level is very close to the nominal level but the power is higher than what is produced by the EBMLE procedure. This would be a matter of personal preference and the decision would be made in accordance with the researcher's objectives.

### **Results for $p_1 = 0.50$**

Looking at Figures B-21 to B-23, it appears that the Laplace-Wald, Haldane, Jeffreys-Perks and EBMLE methods hold the highest power without compromising the type I error rate. One can also see by looking at Figure B-21 that the EBMOM procedure has a type I error rate very close to the four previously mentioned tests but gives lower power. In contrast, the Wald approach, again, tends to be an anticonservative test.

Upon closer inspection, one can see that the Haldane and Jeffreys-Perks approaches tend to produce somewhat anticonservative tests as compared to the EBMLE and Laplace-Wald tests.

Additionally, in Figure B-22, the EBMLE approach is somewhat superior to Laplace-Wald in terms of power. Both of them have similar significance levels, but the EBMLE approach results in greater power.

For all of the reasons discussed above, the EBMLE method was shown to produce greater power and closest to nominal significance levels. Therefore, in situations where  $p_1 = 0.5$ , the EBMLE method would be suggested.

Another note can be made by looking at these graphs. In Figure B-24, all of the curves are very similar. This suggests that, as the  $n_i$ 's increase when  $p_1$  is near 0.5, all methods tend to give similar power curves. In this case, for higher sample sizes, one could use any of the six methods and obtain a test with high power, assuming one population proportion is near 0.5.

## CHAPTER 6 - Recommendations

There are several noteworthy conclusions that come from this power study. First, it appears that the Haldane approach results in higher power without compromising the type I error rate when one population proportion was near zero. As one population proportion nears 0.2, it appears the Jeffreys-Perks procedure or the EBMLE procedure would result in higher power. Finally, the EBMLE procedure tends to provide the highest power when one proportion is near 0.5 without resulting in an anticonservative test.

Using the symmetry argument introduced in Chapter 4, it should also be noted that the procedures above work well for higher values of the proportion as well. The Haldane approach would be best suited for a population proportion near one, whereas the Jeffreys-Perks or EBMLE procedures would work well when a population proportion is near 0.8, and the EBMLE procedure working well with a population proportion around 0.5 or 0.6.

The power study also suggested, as was hypothesized, that the Wald procedure fails to achieve the desired significance level. This is especially apparent when one population proportion is near zero or one.

Finally, one might make note about what effect sample size has on power and significance levels. It is worthwhile to investigate what happens when the two populations have differing sample sizes. As it turns out, this doesn't have a great effect on the power of the test. This effect can be seen by looking at all of the figures in Appendix B.

Lastly, because one may not know what  $p_1$  and  $p_2$  are in a smaller sample size situation, an overall recommendation based on the findings in this report is given. The EBMLE procedure most often results in the highest power without ending up with an anticonservative test. Even in those situations where some of the other methods seem better, the EBMLE procedure is comparable, as it holds the desired significance level and has similar power. Therefore, the EBMLE procedure is recommended, as it gives high power without compromising the type I error rate.

## References

- 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*, 280-288.
- Anderson, S., & Williams, T. A. (2008). *Statistics for Business and Economics*. Thomson South-Western.
- Beal, S. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, *73*, 941-950.
- Brown, L., Cai, T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*, 101-117.
- Brown, L., & Li, X. (2005). Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference*, *130*, 359-375.
- Greenland, S. (2001). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, *55*, 172.
- Price, R., & Bonett, D. (2004). An improved confidence interval for a linear function of binomial proportions. *Computational Statistics and Data Analysis*, *45*, 449-456.
- Reiczigel, J. (2003). Confidence intervals for the binomial parameter: Some new considerations. *Statistics in Medicine*, *22*, 611-621.
- Roths, S., & Tebbs, J. (2006). Revisiting beal's confidence intervals for the difference of two binomial proportions. *Communications in Statistics: Theory and Methods*, *35*(9), 1593-1609.

- Roths, S., & Tebbs, J. (2008). New large-sample confidence intervals for a linear combination of binomial parameters. *Journal of Statistical Planning and Inference*, 138(6), 1884-1893.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.
- Zhou, X., Tsao, M., & Qin, Q. (2004). New intervals for the difference between two independent binomial proportions. *Journal of Statistical Planning and Inference*, 123, 97-115.

## Appendix A - R Program

```
#####  
  
## R code for Tebbs and Roths (2007) ##  
## "New large-sample confidence intervals for a linear combination of binomial ##  
## proportions" ##  
## Journal of Statistical Planning and Inference, in press. ##  
## Date: 23 October 2006 ##  
## Revised: 21 May 2007 ##  
## Please email Scott Roths (sar320@psu.edu) if you have any questions. ##  
#####  
  
gram = function(x)  
{  
# takes matrix of independent columns and returns matrix of orthonormal columns  
# algorithm from Christensen  
r = length(x[,1])  
y = matrix(1, nr=length(x[,1]), nc=r)  
y[,1] = x[,1] / sqrt(sum(x[,1]^2))  
temp = x[,2] - sum(x[,2]*y[,1])*y[,1]  
y[,2] = temp / sqrt(sum(temp^2))  
if(r>2)  
{  
for(s in 3:r)  
{  
temp = x[,s] - apply(apply(x[,s]*y[,1:(s-1)],2,sum)*t(y[,1:(s-1)]),2,sum)  
y[,s] = temp / sqrt(sum(temp^2))  
}  
}  
return(y)  
}
```

```

#####
ortho.comp = function(x)
{
# takes a row vector and returns basis of orthogonal rows
# 'normalized' such that M'M = (x'x)*I
# g-by-g identity matrix
g = length(x)
M = matrix(0, nr=g, nc=g)
M[row(M) == col(M)] = 1
# replace appropriate row with x
i = min((1:g)[x!=0])
M[i,] = x
# orthogonalize M
M = sqrt(sum(x^2))*gram(t(M))
return(t(M))
}
#####

limits = function(c,n,y,conf=.95,type=4,n.old=NULL,y.old=NULL)
{
# returns limits for single experiment with at least 2 groups
quad.form = function() # local to limits
{
# returns truncated roots of c2x^2+c1x+c0=0
discr = c1^2-4*c2*c0
if(discr<0) return(c(sum(c*y/n),sum(c*y/n)))
else return(c(max(sum(c[c<0]),(-c1-sqrt(discr))/2/c2),min(sum(c[c>0]),(-
c1+sqrt(discr))/2/c2)))
}
dln.bb = function(alpha) # local to 'limits'
{
# evaluates derivative of log-likelihood of Y

```



```

alpha = alpha*rep(1,g) # makes alpha a vector
temp1 = rbind(2*alpha,2*alpha,y.old+alpha,n.old-y.old+alpha)
temp2 = rbind(alpha,alpha,n.old+2*alpha,n.old+2*alpha)
sum(apply(temp1,2,digamma)) - sum(apply(temp2,2,digamma))
}
is.good = function() # local to 'limits'
{
# returns 1 if root exists between 0.005 and 1000, -1 otherwise
-sign(dln.bb(0.005)*dln.bb(1000))
}
g = length(c) # number of binomial groups
chi = qchisq(conf,1) # critical value
if(type==1 || type==2 || type==3)
{
if(type==2) # Laplace-Wald
{
# adds one success and one failure
n = n+2
y = y+1
}
if(type==3) # Price Bonett
{
k = length(c[c!=0]) # number of nonzero coefficients in c
n = n+4/k # adds 4/k trials
y = y+2/k # adds 2/k successes
}
moe = sqrt(chi*(sum((c/n)^2*y) - sum((c*y/n)^2/n)))
l = max(sum(c[c<0]),sum(c*y/n)-moe)
u = min(sum(c[c>0]),sum(c*y/n)+moe)
return(c(l,u))
}
}

```

```

else # data-driven formulas
{
L = ortho.comp(c) # completes orthogonal rows
mult = sum(c^2) # multiple st L'L = mult*Identity
if(is.null(n.old))
{
# previous information not incorporated
n.old = n
y.old = y
}
if(type==4) alpha = 0 # extended Haldane limits
if(type==5) alpha = .5 # extended Jeffreys-Perks limits
if(type==6) # extended MLE limits
{
# if root exists, assign that to alpha, else -1
alpha = ifelse(is.good()==1, uniroot(dln.bb,c(.005,1000),tol=10**(-5))$root, -1)
# if all success counts are endpoints, assign 0 to alpha
if(sum((n.old-y.old)*y.old)==0) alpha = 0
}
if(type==7) # extended MOM limits
{
if(sum(abs(rep(n.old[1],g)-n.old))==0) # equal sample sizes
{
denom = g*n.old[1]^2+g*n.old[1]-4*sum(y.old^2)
alpha = ifelse(denom==0, -1, max(0,2*sum(y.old^2)-g*n.old[1]^2/denom))
}
else # unequal sample sizes
{
denom = 4*y.old^2-n.old^2-n.old
alpha = ifelse(prod(denom)==0, -1, max(0,mean((n.old^2-2*y.old^2)/denom)))
}
}
}

```

```

}
if(alpha<0) a = apply(L,1,sum)/2 # if alpha is infinite
else a = L%%((y+alpha)/(n+2*alpha)) # if alpha is finite
a1 = sum(c*y/n) # point estimator
a = as.matrix(a[2:g]) # nuisance parameters
L = as.matrix(t(L)[,2:g]) # coefficients of nuisance parameters
# coefficients of the quadratic
c2 = 1+chi*sum(c^4/n)/mult^2
c1 = -2*a1-chi*sum(c^3/n)/mult+2*chi*sum(c^3/n*L%%a)/mult^2
c0 = a1^2-chi*sum(c^2/n*L%%a)/mult+chi*sum(c^2/n*L%%a^2)/mult^2
return(quad.form())
}
}
#####
results = function(c,n,p,conf,type)
{
# Returns the exact coverage probability and mean length
# c - coefficients of interest
# n - sample sizes
# p - population proportions
# conf - confidence coefficient
# type - specifies which interval formula
# (1) Wald
# (2) Laplace-Wald
# (3) Price-Bonett
# (4) Haldane
# (5) Jeffreys-Perks
# (6) MLE
# (7) MOM
wts = function(x)
{

```

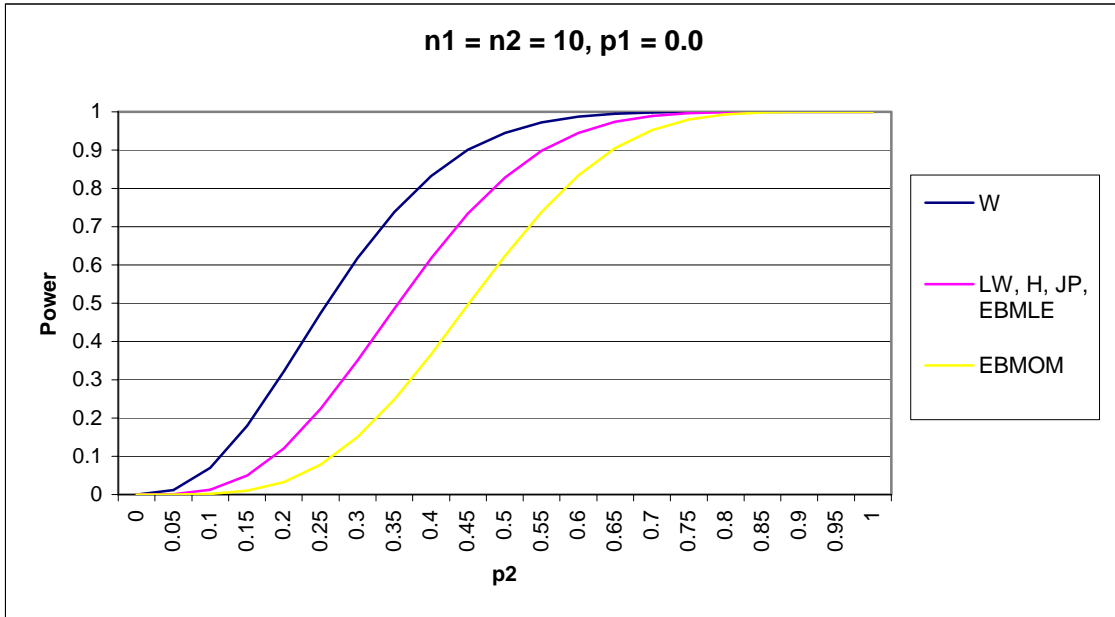
```

# local to 'results'
# returns cover indicator and width weighted by prob
lmt = limits(c,n,x,conf,type)
pmf = prod(choose(n,x)*p^x*(1-p)^(n-x))
cov = ifelse(lmt[1]<=a1 && a1<=lmt[2], 1, 0)
wid = lmt[2]-lmt[1]
wid.i = cov*wid # width for intervals covering parameter
return(pmf*c(cov,wid,wid.i))
}
# considering just groups of interest
g = length(c)
nonz = (1:g)[c!=0]
k = length(nonz)
c = c[nonz]
n = n[nonz]
p = p[nonz]
a1 = 0 # linear combination of interest
N = prod(n+1) # number of values of y in support
M = matrix(nr=N, nc=k)
M[,1] = rep(0:n[1], length.out=N, each=1)
for(i in 2:k)
{
# creates matrix of y row vectors
M[,i] = rep(0:n[i], length.out=N, each=prod(n[1:(i-1)]+1))
}
temp = apply(apply(M,1,wts),1,sum) # computes temporary results
wid.e = (temp[2]-temp[3])/(1-temp[1])
wid.i = temp[3]/temp[1]
temp[3] = wid.i/wid.e
temp
}

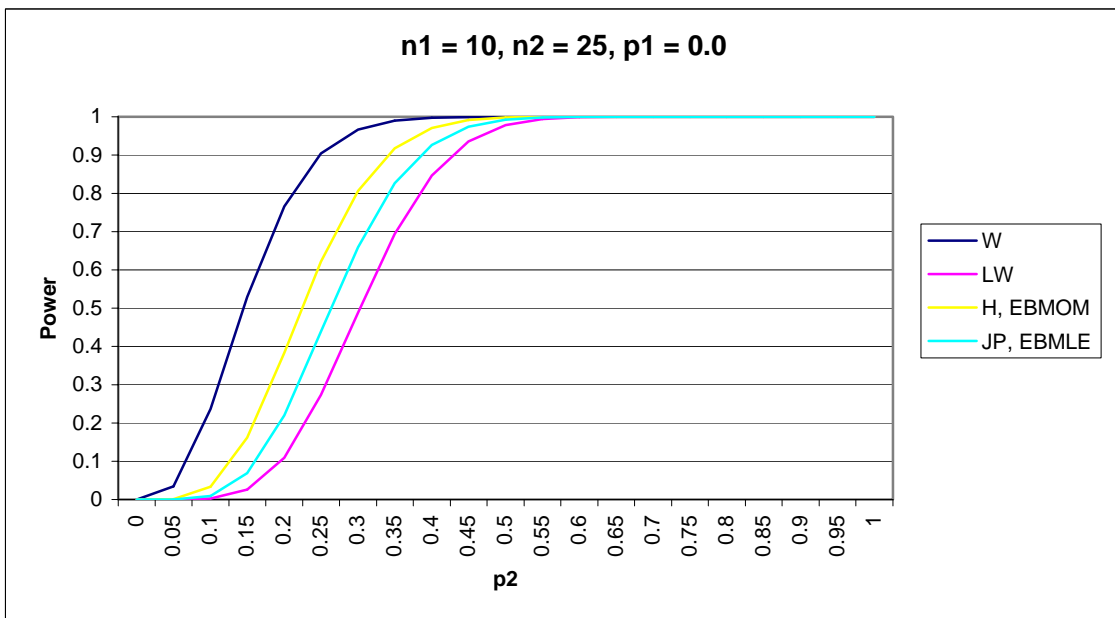
```

```
#####  
limits.all = function(c,n,y,conf)  
{  
# returns all six confidence intervals  
ret = matrix(nr=6,nc=2,dimnames=list(c("Wald", "Laplace",  
"Price-Bonett", "Haldane", "Jeffreys-Perks ", "EBMLE"),c("Lower", "Upper"))  
for(i in 1:6) ret[i,] = limits(c,n,y,conf,type=i)  
round(ret,4)  
}
```

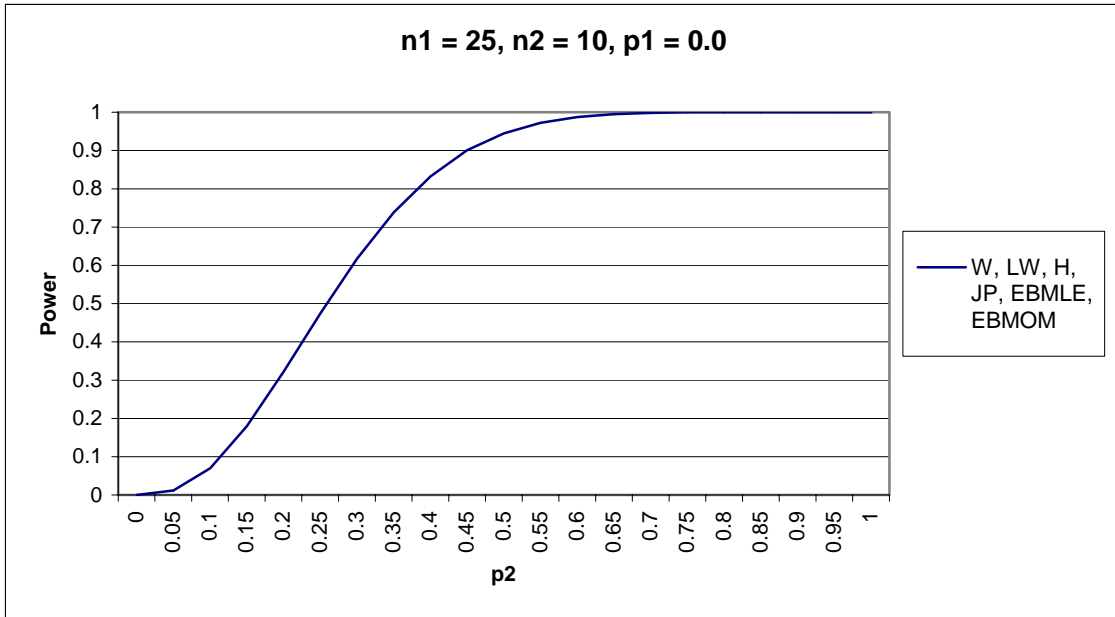
## Appendix B - Power Study Results



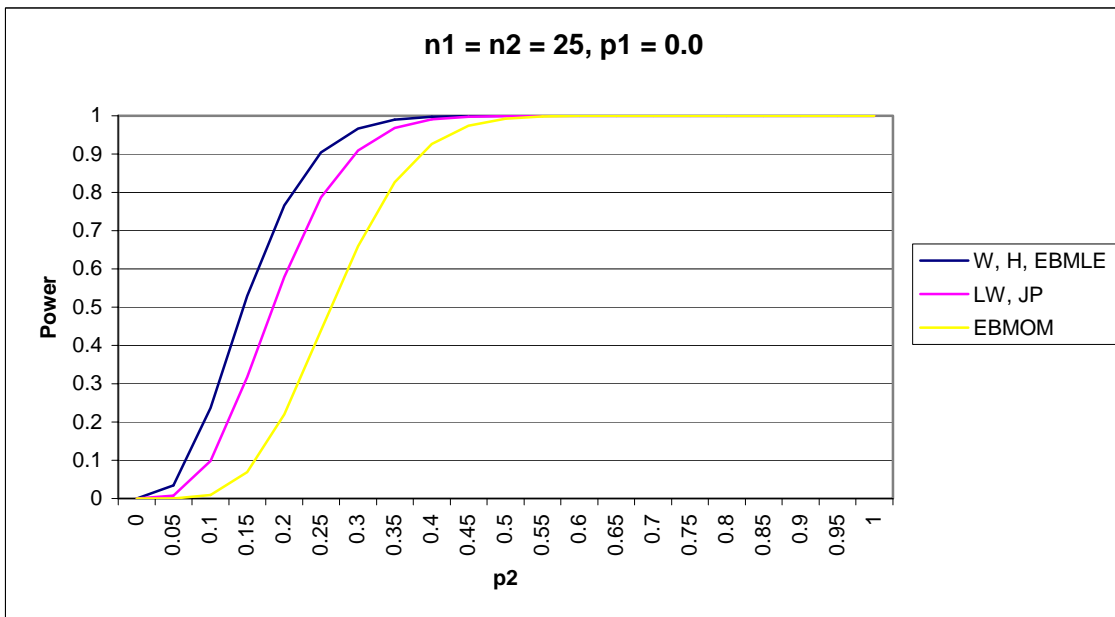
**Figure B-1**



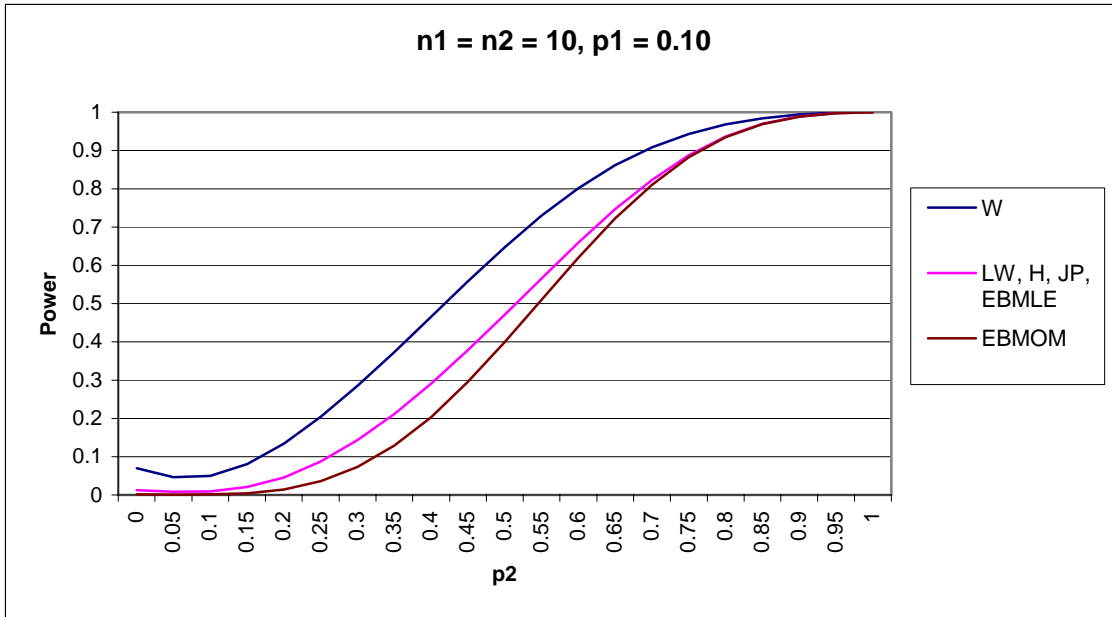
**Figure B-2**



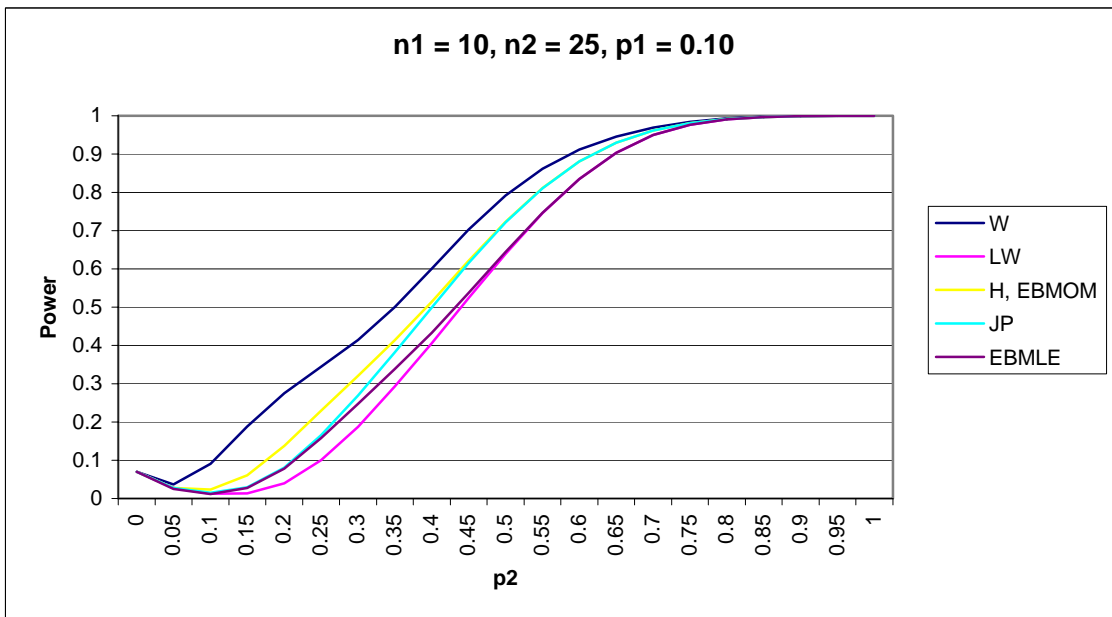
**Figure B-3**



**Figure B-4**

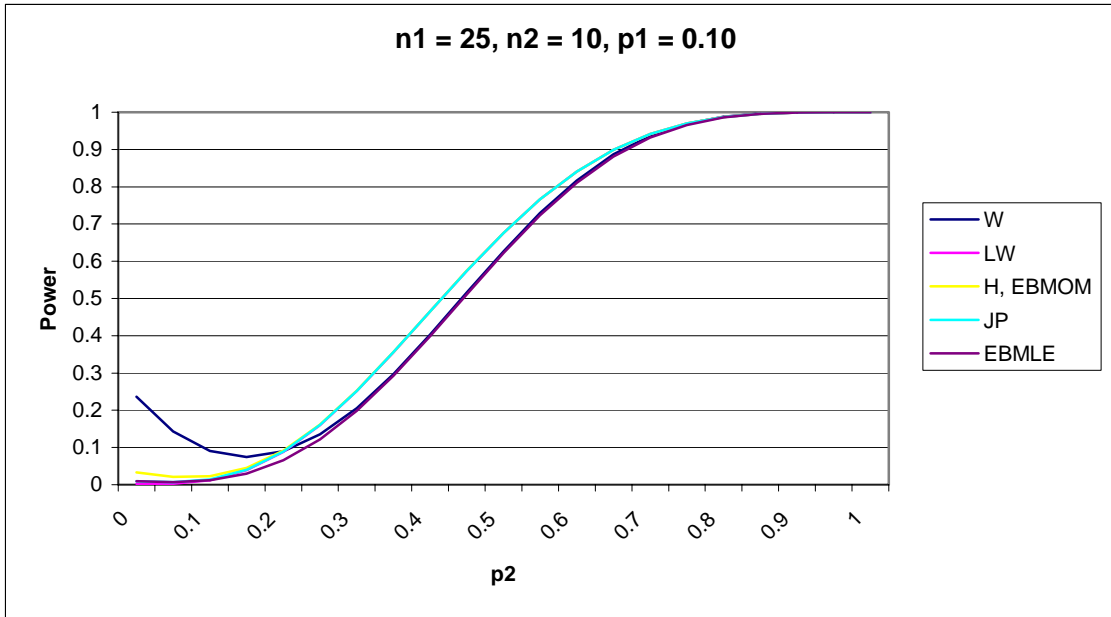


**Figure B-5**

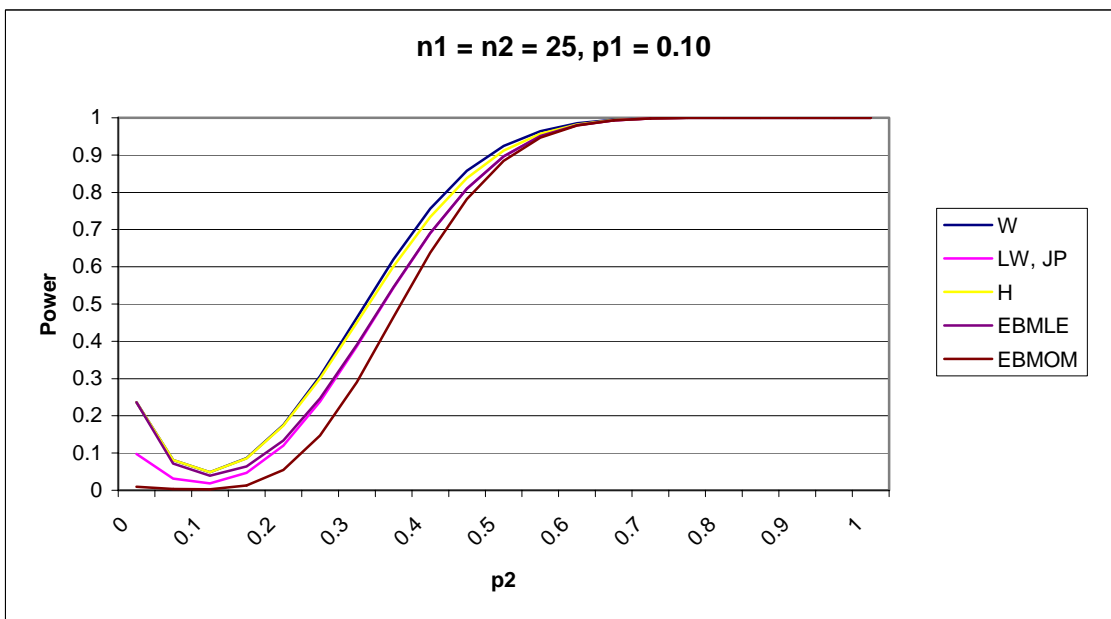


**Figure B-6**

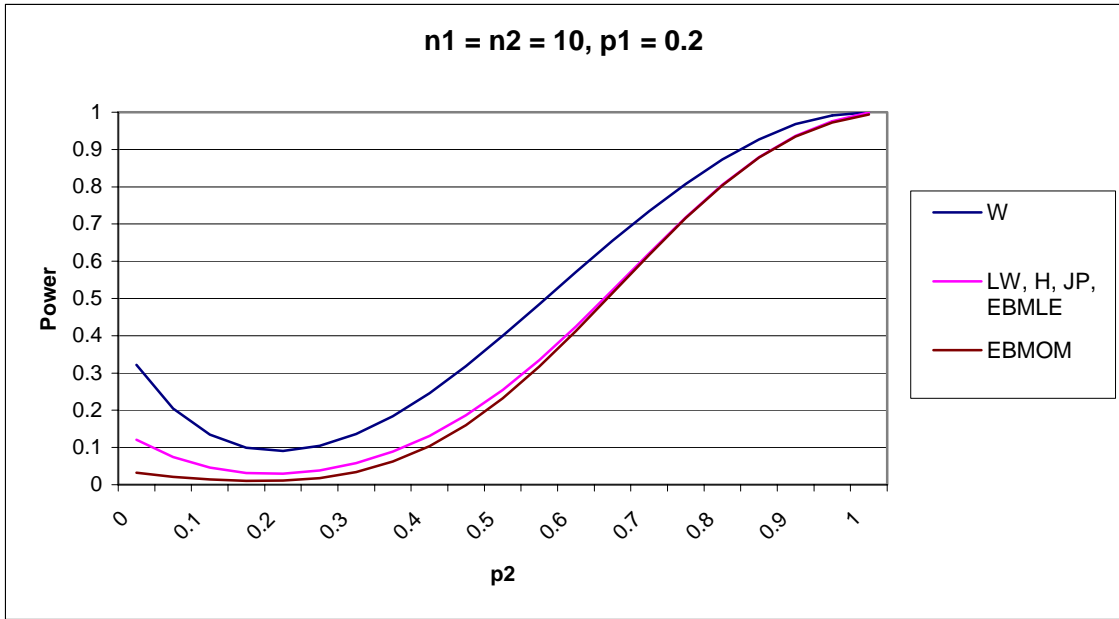




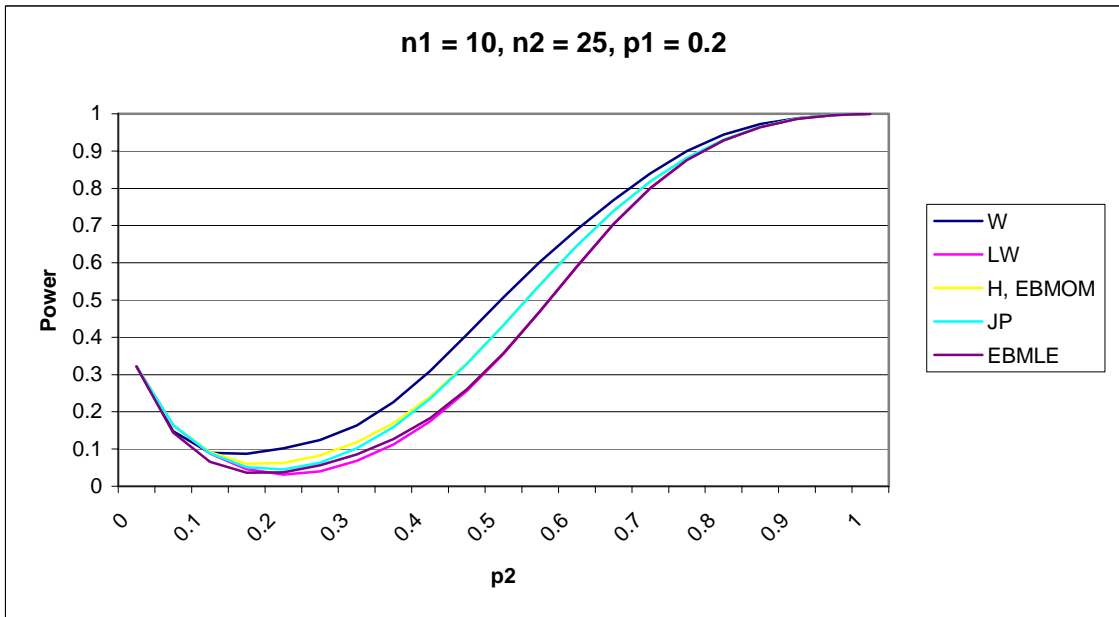
**Figure B-7**



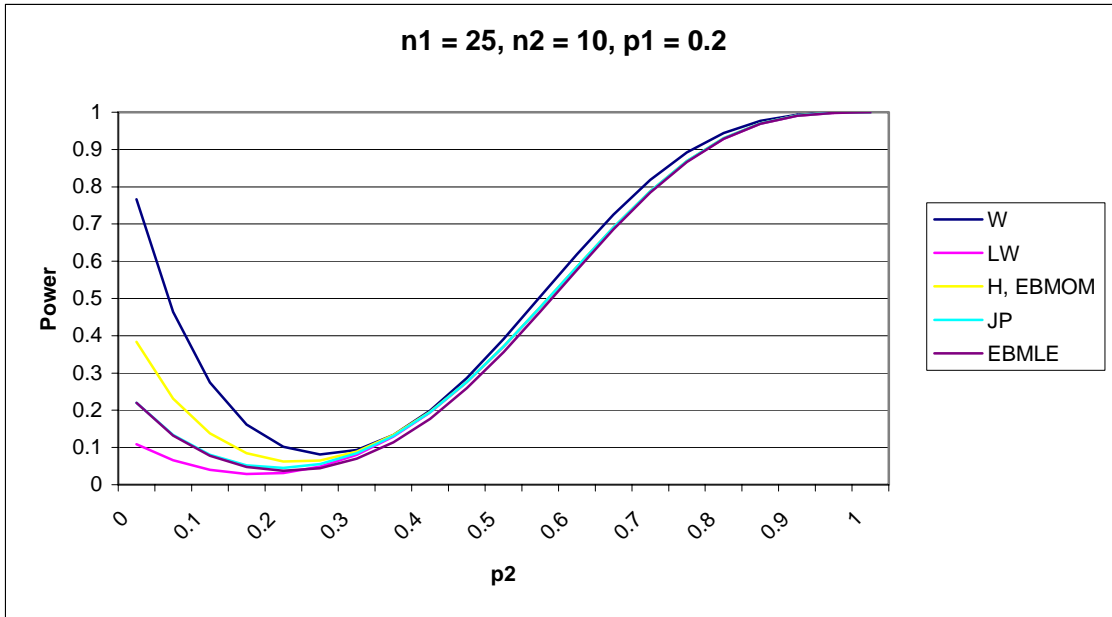
**Figure B-8**



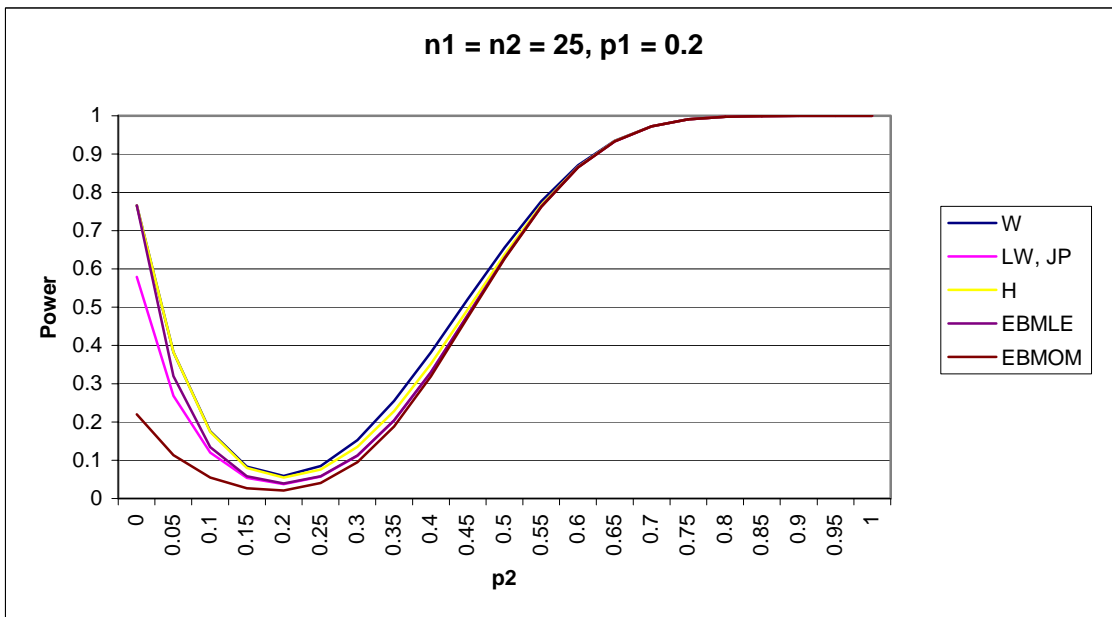
**Figure B-9**



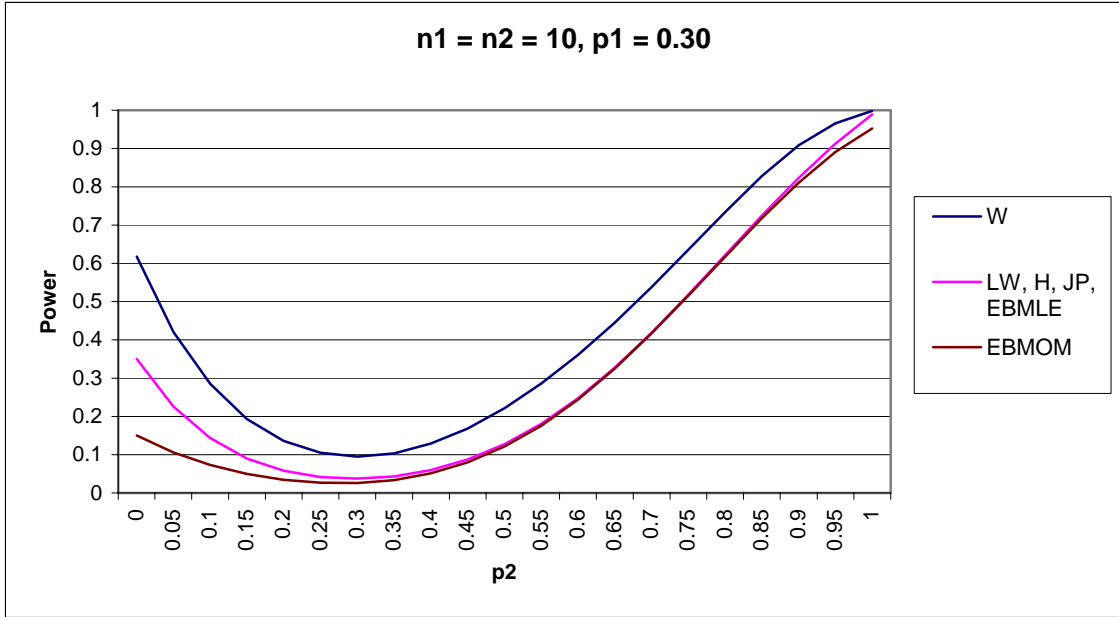
**Figure B-10**



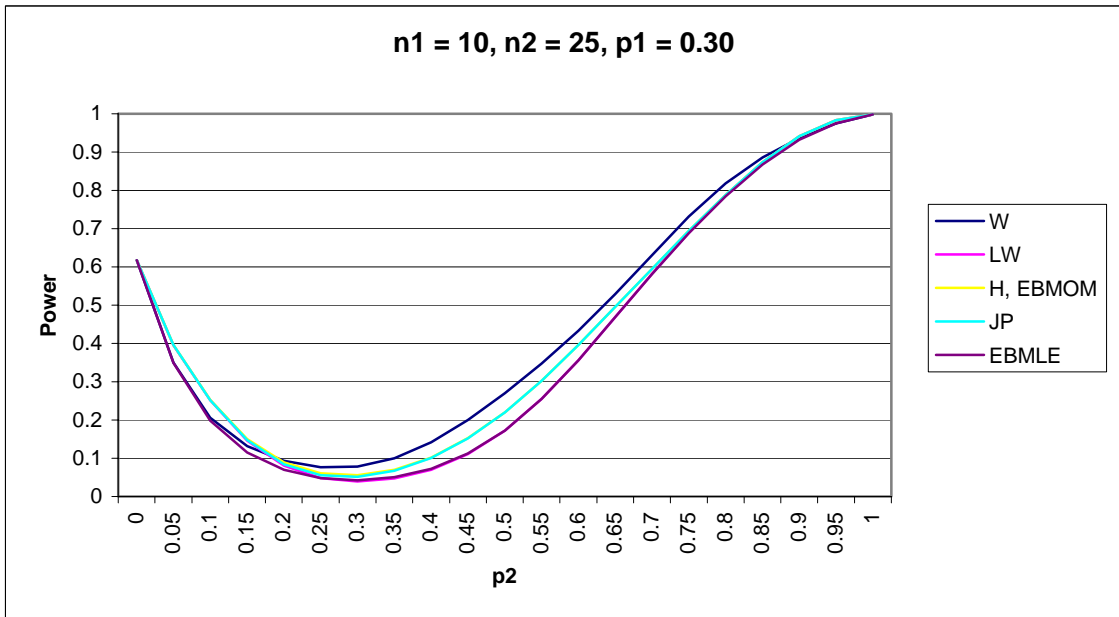
**Figure B-11**



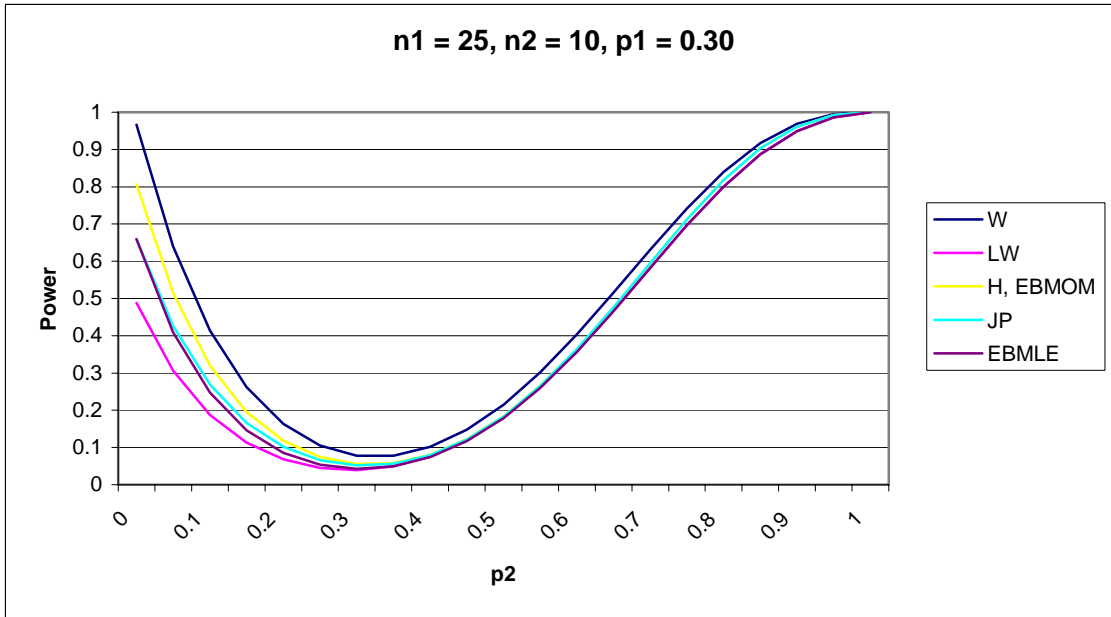
**Figure B-12**



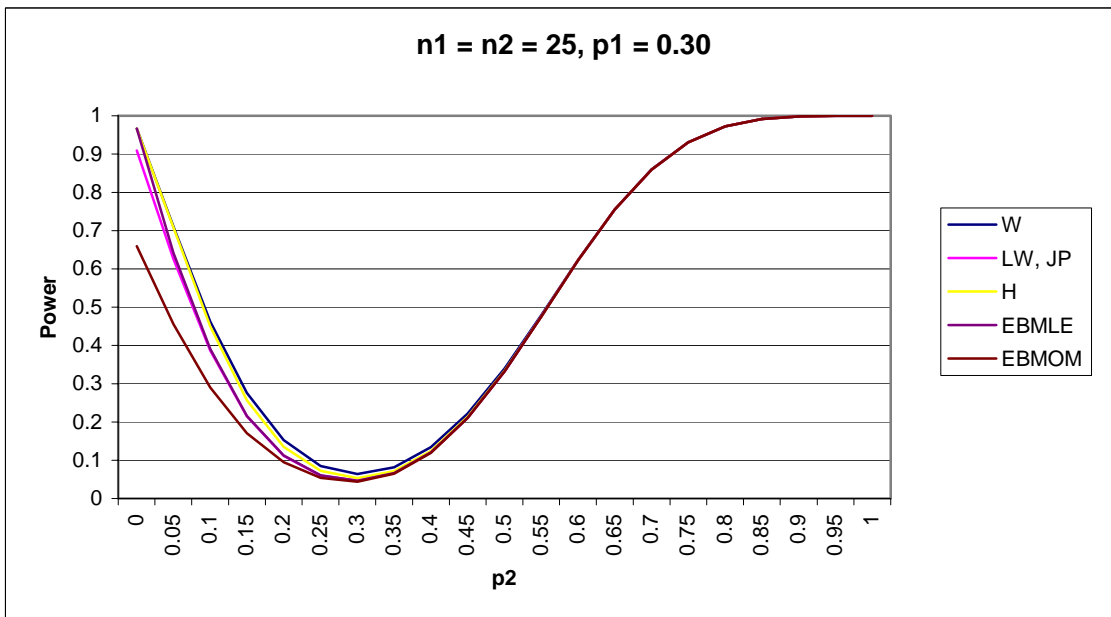
**Figure B-13**



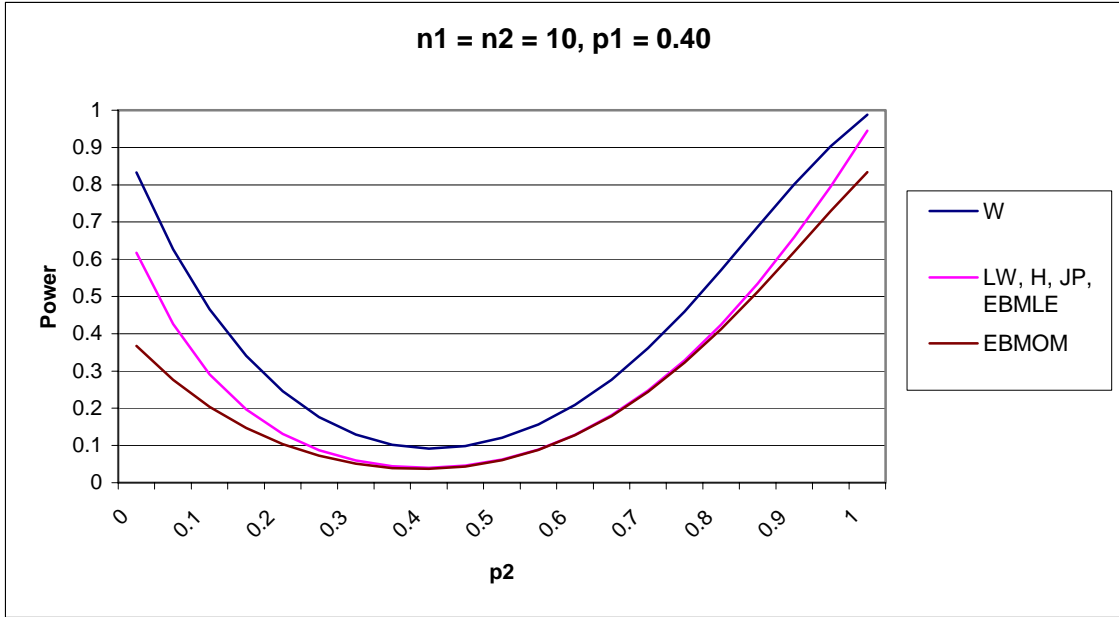
**Figure B-14**



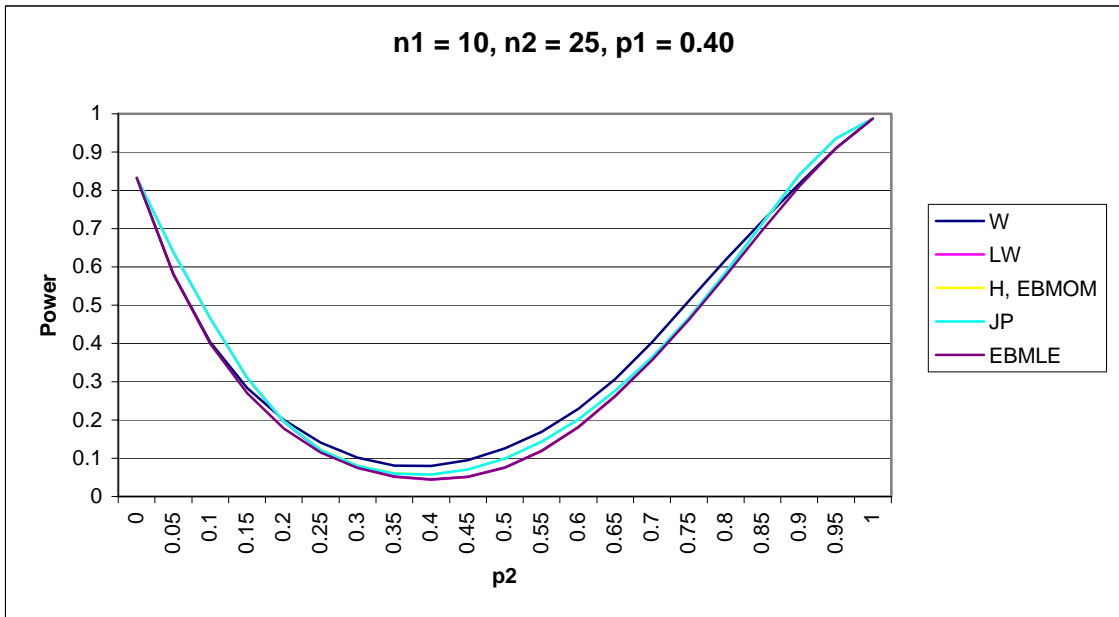
**Figure B-15**



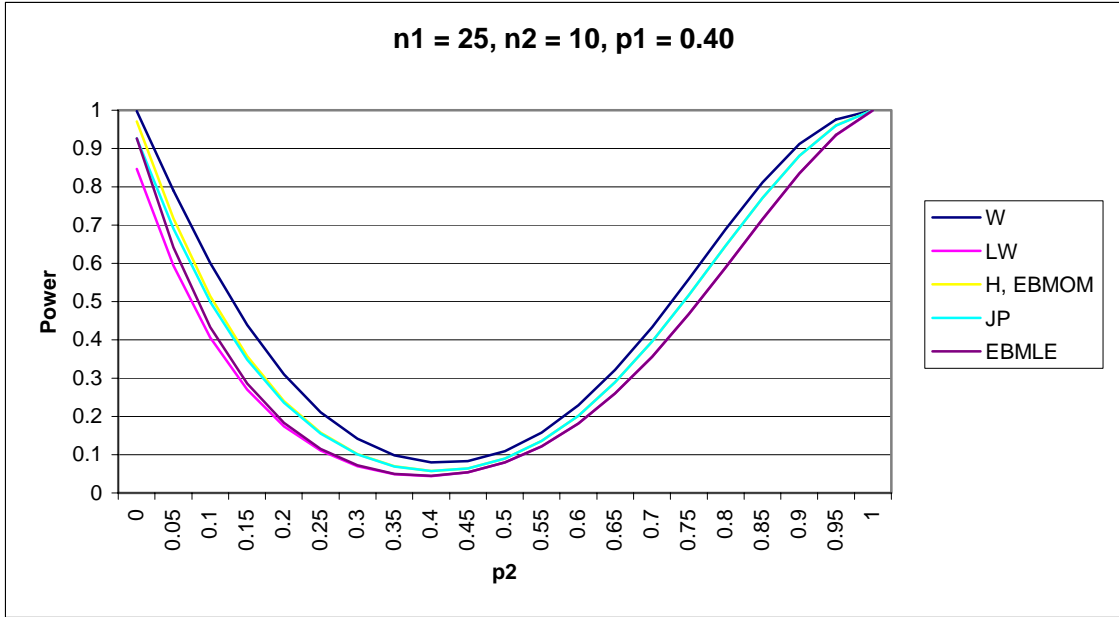
**Figure B-16**



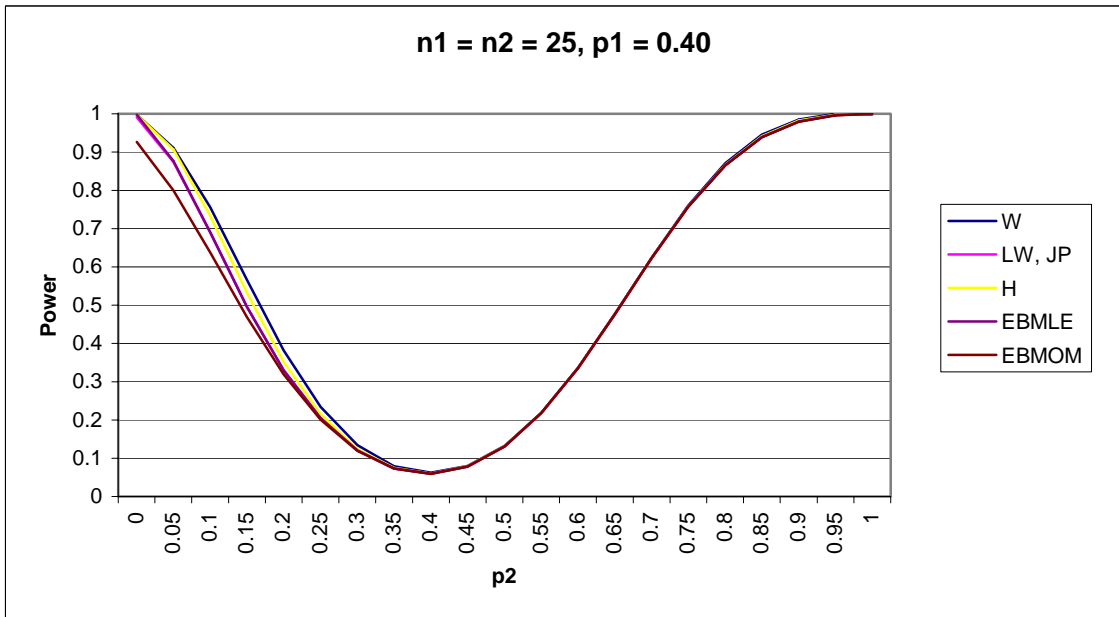
**Figure B-17**



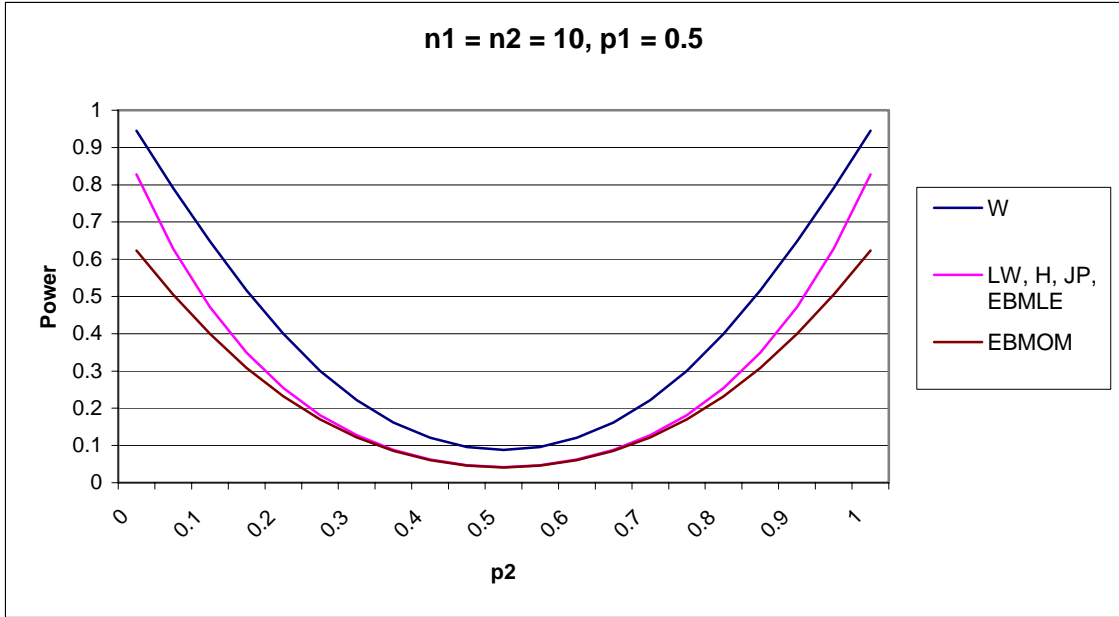
**Figure B-18**



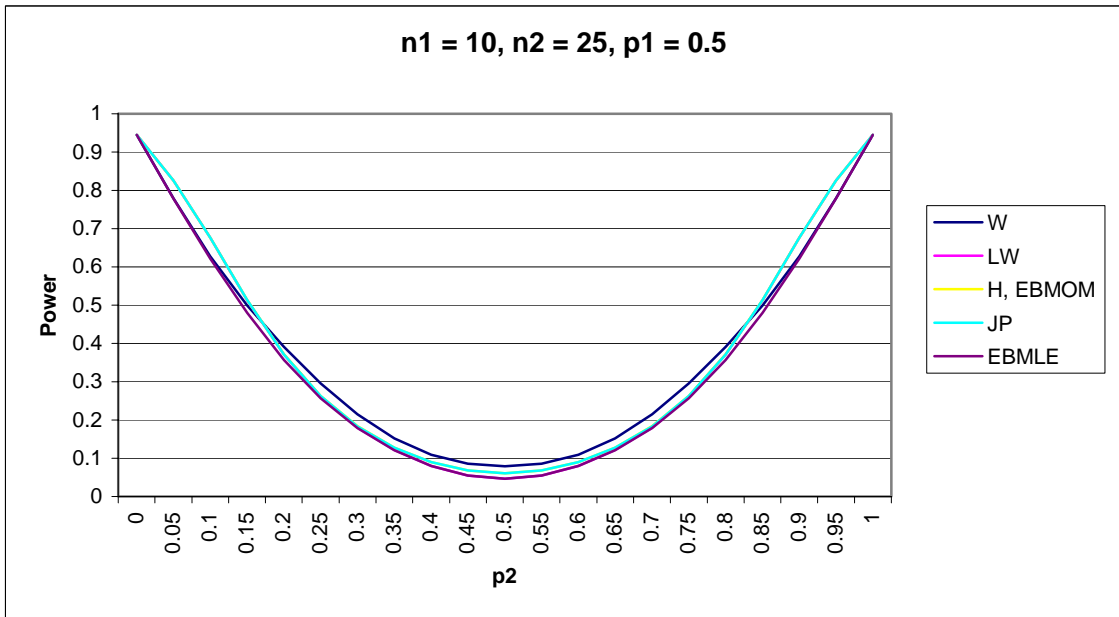
**Figure B-19**



**Figure B-20**

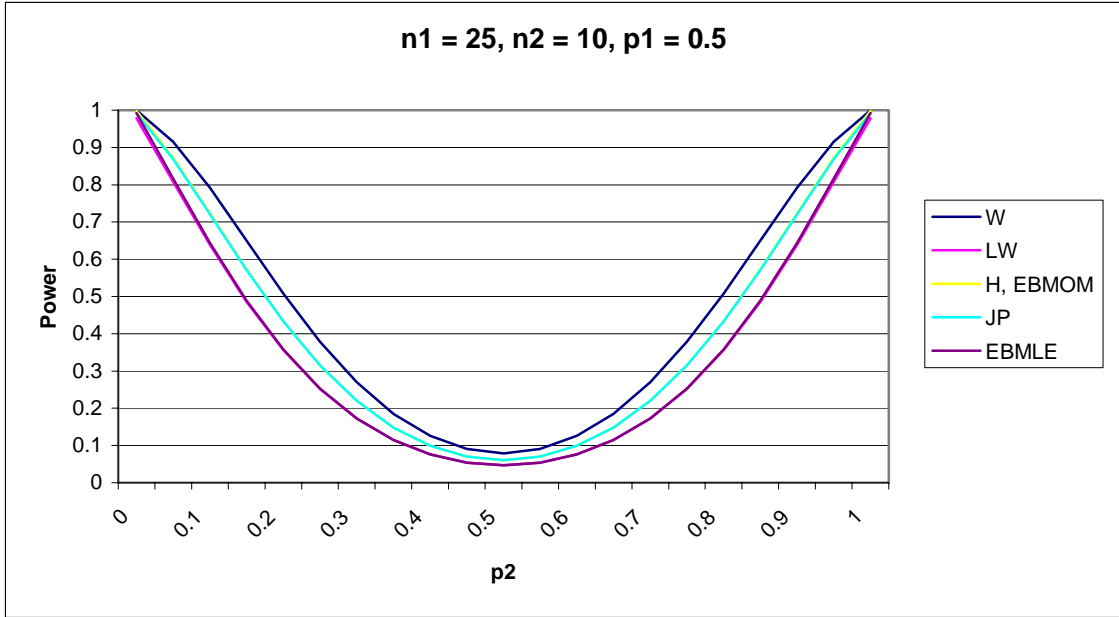


**Figure B-21**

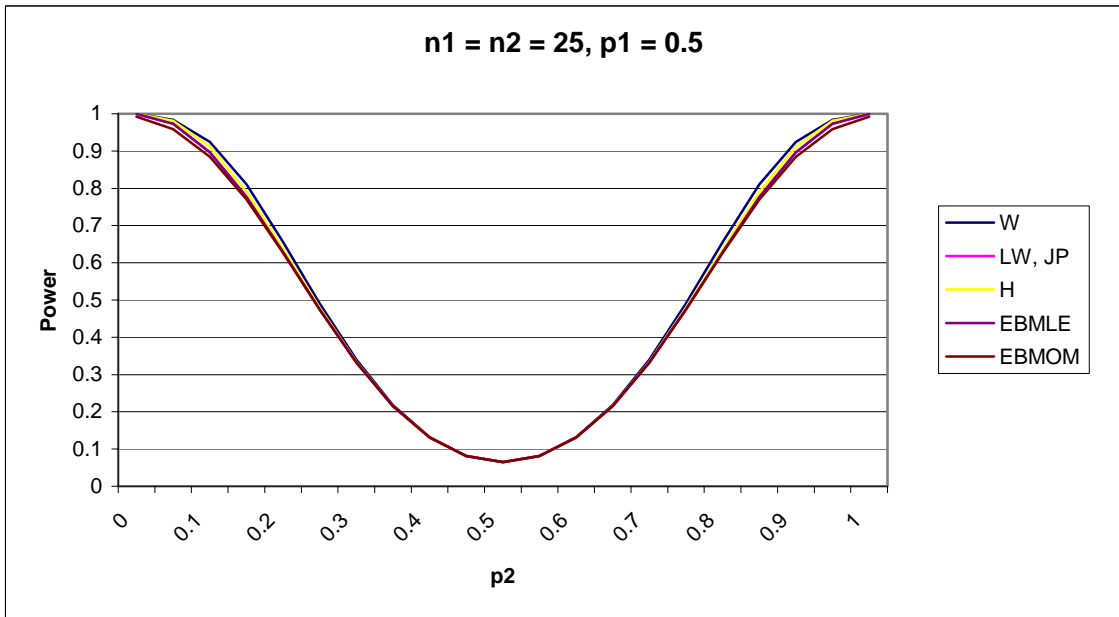


**Figure B-22**





**Figure B-23**



**Figure B-24**