CONFIDENCE INTERVAL FOR THE DIFFERENCE IN BINOMIAL PROPORTIONS FROM STRATIFIED 2X2 SAMPLES

Peng-Liang Zhao, John K. Troxell, Hui Quan, Michael Lee, and James A. Bolognese
Merck Research Laboratories, P.O. Box 2000, RY 33-404, Rahway, NJ 07065

Key Words: Weighted Average Approaches, Likelihood Approaches, Coverage Probability

Abstract
Confidence intervals for the difference of binomial proportions between two treatments in stratified 2x2 samples are usually obtained using the normal approximation for a weighted average of the differences over all strata. For this approach, different weights such as Cochran-Mantel-Haenszel (CMH), inverse of variance (INVAR), and minimum risk (MR) can be used. Alternatively, under the assumption of a common difference in binomial proportions across strata, confidence intervals can be constructed using likelihood approaches including the profile likelihood (PL) method, simple asymptotic (SA) method, and Wilson score (WS) method. In this paper, we first compare the coverage probabilities of these 6 different methods, without and with a continuity correction. A simulation study shows that for the case of event rates less than 10% in one or both treatment groups, the CMH weighted average method with the continuity correction has the best coverage probability; while all other methods can have an inadequate coverage probability in some parameter settings even for large sample sizes. For the case of event rates between 10% and 90%, the CMH method and the PL method have similar performances and are the best for small and mid-large sample sizes, and all methods have similar coverage probabilities for large sample sizes. Two other weighted average methods, the sample size (SS) weight and equal (EQ) weight, were also compared with the CMH method in a different sample size configuration, and the CMH method is better than or similar to the SS and EQ methods.

1. Introduction
In clinical trials, the parameter of interest in many situations is the difference of the binomial response rates between two treatments. For example, the rates of clinical adverse experiences between the test therapy and placebo are usually compared to evaluate the safety of the test therapy, and in this case the difference of the rates can be the focal point since it is easy to interpret and meaningful for most clinicians. When combining data from several studies or from patient subgroups such as age groups, gender, prior therapies or disease status, stratification for adjusting for these factors is frequently used. This article considers the interval estimation for the difference of binomial proportions based on stratified 2x2 samples.

Confidence intervals for the difference of binomial proportions between two treatments from stratified 2x2 samples are commonly obtained using the normal approximation for a weighted average of the differences over all strata. For this weighted average approach, the two most popular used weights are the Cochran-Mantel-Haenszel weight (Cochran 1954, Mantel and Haenszel 1959) and Inverse of Variance weight (Radhakrishna 1965). Recently, the Minimum Risk weight was proposed by Mehrotra and Raikar (2000). These three weighting methods estimate the same parameter if the true underlying differences of binomial rates between two treatments are constant across strata. If this homogeneity assumption does not hold, however, they estimate different parameters and some are more appropriate and meaningful than others. For this reason, the performances of these three weighting methods can only be compared under the homogeneity condition. The testing for homogeneity of rate differences across strata is not the topic of this paper, but can be conducted using the method given by Lui and Kelly (2000). Under the homogeneity assumption, confidence intervals for the common difference of binomial proportions can be constructed alternatively using likelihood approaches including the profile likelihood method, simple asymptotic method, and Wilson score method. The performances of these 6 methods mentioned above have not been well studied yet. In this paper, we use simulation to compare the coverage probabilities of these 6 different methods. Two other weighted average methods, the sample size weight and equal weight, can have the same weight as the Cochran-Mantel-Haenszel method for certain sample size settings. Here we also compare the sample size weight and equal weight with the other weighted average methods in a different sample size configuration.

2. Weighted Average Approaches
Throughout this paper, we assume that patients are randomly assigned into two treatment groups \( i = 0,1 \) and there are \( K \) strata \( (K \geq 2) \). Let \( n_j \) be the number of patients and \( x_{ij} \) be the number of responders in stratum \( j \) and treatment \( i \). Let \( p_j \) be the true
responder rate and $\hat{p}_{ij} = x_{ij} / n_{ij}$ be the observed responder rate in stratum $j$ and treatment $i$. Let $N_j = \sum_{i=1}^{K} n_{ij}$ denote the total number of patients in treatment $i$. Let $\delta_j = p_{1ij} - p_{0ij}$ be the true rate difference and $\hat{\delta}_j = \hat{p}_{1ij} - \hat{p}_{0ij}$ be the observed rate difference between the two treatments in stratum $j$. We are interested in the overall treatment difference, which does not necessarily need the homogeneity of the $\delta_j$, but does require that there be no qualitative treatment-by-stratum interaction.

In the weighted average approaches, the overall treatment difference is usually defined by $\delta_w = \sum_{j=1}^{K} w_j \delta_j$ and estimated by $\hat{\delta}_w = \sum_{j=1}^{K} w_j \hat{\delta}_j$, where $w_j$ is the weight for stratum $j$ with $\sum_{j=1}^{K} w_j = 1$. The $(1-\alpha)$ 100% confidence interval for $\delta_w$ is given by

$$\hat{\delta}_w \pm z_{\alpha/2} \sqrt{\sum_{j=1}^{K} w_j^2 \hat{V}(\hat{\delta}_j)},$$

where $z_{\alpha}$ is the $(1-c)$ percentile of the standard normal distribution, and

$$\hat{V}(\hat{\delta}_j) = (\hat{p}_{0ij}(1-\hat{p}_{0ij})/n_{0ij}) + (\hat{p}_{1ij}(1-\hat{p}_{1ij})/n_{1ij}).$$

Continuity corrections are commonly used to deal with small sample sizes for binomial data. We use the continuity correction factor suggested by Mehrotra and Railkar (2000). This continuity correction factor is approximately $3/8$ times that used in Mantel and Haenszel (1959) since the latter was found to be overly conservative. The $(1-\alpha)$ 100% confidence interval for $\delta_w$ with this continuity correction is given by

$$\hat{\delta}_w \pm z_{\alpha/2} \sqrt{\sum_{j=1}^{K} w_j^2 \hat{V}(\hat{\delta}_j) + \frac{3}{16} \left( \sum_{j=1}^{K} n_{0ij} n_{1ij} \right)^{-1}}.$$

(3)

The two most popular weights are the Cochran-Mantel-Haenszel (CMH) weight and Inverse of Variance (INVAR) weight. The CMH weight, first discussed by Cochran (1954) and then by Mantel and Haenszel (1959), is proportional to the harmonic mean of the sample sizes in each stratum:

$$w_j = \left[ n_{0ij} + n_{1ij} \right]^{-1} \left[ \sum_{j=1}^{K} \left( n_{0ij} + n_{1ij} \right) \right]^{-1}.$$

The INVAR weight, studied by Radhakrishna (1965), is proportional to the inverse of the variance of the observed rate difference in each stratum:

$$w_j = \hat{V}_{j}^{-1} / \left( \sum_{j=1}^{K} \hat{V}_{j}^{-1} \right),$$

(5)

where $\hat{V}_{j} = \hat{V}(\hat{\delta}_j)$ is given by (2). The CMH method gives more weight to strata that have more patients. The INVAR method puts more weight on strata that have greater precision to estimate treatment differences.

Recently Mehrotra and Railkar (2000) proposed the Minimum Risk (MR) weight. The targeted overall treatment difference for MR method is

$$\delta_j = \sum_{j=1}^{K} f_j (p_{1ij} - p_{0ij})$$

with

$$f_j = \left[ n_{0ij} + n_{1ij} \right] \left[ \sum_{j=1}^{K} \left( n_{0ij} + n_{1ij} \right) \right]^{-1}$$

for prognostic stratification factors (such as sex and age), the SS weight $w_j$ represents the fraction of patients that entered stratum $j$ in the target patient population. The SS method also gives more weight to strata that have more patients, but in a different manner from the CMH method. When $n_{0ij} = c * n_{ij}$ with a constant $c$ for all $j$, the SS and CMH methods have the same weight. The EQ method treats all strata equally. When the number of patients across strata are the same in each treatment group, the EQ and CMH methods have the same weight.

The rationale and theoretical merits of these weighting methods have been discussed in Mehrotra and Railkar (2000) and Mehrotra (2001). The INVAR and MR weights involve the estimated variance $\hat{V}_j$. If the observed variance $\hat{V}_j$ is zero in a stratum, the INVAR and MR weights can not be computed. To overcome this problem, the procedure of adding 0.5 has been commonly used (Mehrotra 2001). That is, for the estimation of the variance $\hat{V}_j$ in (2), we replace $\hat{p}_y = 0$ with $\hat{p}_y + (0.5/n_y)$ and $\hat{p}_y = 1$ with $\hat{p}_y - (0.5/n_y)$. This “adding 0.5 correction” is applied only to the estimation of the variances $\hat{V}_j$, but not to the observed rate difference $\hat{\delta}_j = \hat{p}_{1ij} - \hat{p}_{0ij}$. So we refer this procedure as “variance correction” throughout this paper. Although this variance correction is not required for the CMH, SS, and EQ weights, it can improve the coverage probability when the sample sizes are small.

The performances of these weighting methods can only be compared under the condition that the $\delta_j$ are constant across strata. The comparisons were made without was the continuity correction, as given in
The variance correction was used for the INVAR and MR methods regardless of the sample sizes, and for the CMH, SS, and EQ methods when the sample size $N_1$ or $N_2$ was less than 100.

### 3. Likelihood Approaches

The likelihood approaches are derived under the assumption of a common difference across strata. That is, we assume that $\hat{\delta}_j = p_{ij} - p_0j = \delta$ for $j = 1, 2, ..., K$. Under this assumption, the parameters are reduced to $\theta = (p_{o1}, p_{o2}, ..., p_{ok}, \delta)'$, and the likelihood function is

$$G(\theta) = \prod_{j=1}^{K} F(x_{oj}, n_{oj} | p_{oj}) F(x_{ij}, n_{ij} | p_{oj} + \delta),$$

where $F(x, n | p) = \binom{n}{x} p^x (1-p)^{n-x}$. For the purpose of computation, it is necessary to set $0^0 = 1$. It can be verified that there exists an unique maximum likelihood estimator (MLE) $\hat{\theta} = (\hat{p}_{o1}, \hat{p}_{o2}, ..., \hat{p}_{ok}, \hat{\delta})'$ for (7). This MLE can be computed using the Newton-Raphson iteration method or a search algorithm. We used the search algorithm for our simulation study, since it is more reliable for the low event rate case. Let $(\hat{p}_{o1}(\delta), \hat{p}_{o2}(\delta), ..., \hat{p}_{ok}(\delta), \hat{\delta})'$ be the MLE for (7) for a given value $\delta$. Then the maximum likelihood equation for $\hat{p}_{oj}(\delta)$ becomes a cubic equation and has an unique closed-form solution (see Miettinen and Nurminen, 1985). The solutions $\delta_L$ and $\delta_H$ of (13) can be computed numerically using the bisection method.

The first likelihood approach discussed here is the profile likelihood (PL) method (also called the likelihood ratio method in the setting of the hypothesis test). Let $\hat{\theta}(\delta) = (\hat{p}_{o1}(\delta), \hat{p}_{o2}(\delta), ..., \hat{p}_{ok}(\delta), \delta)'$. The $(1 - \alpha)$ 100% confidence limits $(\delta_L, \delta_H)$ based on the PL method are the values of $\delta$ such that

$$2[\log(G(\hat{\theta}^+)) - \log(G(\hat{\theta}(\delta)))] = \chi^2_{\alpha/2},$$

where $\hat{\theta}^+ = \hat{p}_{oj}^* + \hat{\delta}^*$. So the $(1 - \alpha)$ 100% confidence interval for $\delta$ based on the SA method is

$$\delta^* = \hat{\delta}^* \pm z_{\alpha/2} \sqrt{\hat{V}(\delta^*)}.$$  

The third likelihood approach is the Wilson Score (WS) method. Wilson (1927) proposed the score method for constructing the confidence interval for the single proportion of one binomial sample. To apply the WS method to our case, the idea is to construct the confidence interval using $\hat{\delta}^*$ and its variance under the true value $\delta$ assuming that $p_{ij} - p_0j = \delta$ for all $j$. Let $\tilde{p}_{ij}(\delta) = \hat{p}_{oj}(\delta) + \delta$ and

$$\tilde{V}(\delta) = \left( \sum_{j=0}^{K} \frac{\tilde{p}_{ij}(\delta)(1-\tilde{p}_j(\delta))}{n_j} \right) \left( \frac{n_{oj} + n_{ij} - 1}{n_{oj} + n_{ij}} \right).$$

It can be proved that the variance of $\hat{\delta}^*$ under the given true value of $\delta$ can be estimated by

$$\hat{V}(\delta) = 1 \left( \sum_{j=1}^{K} \frac{V_{ij}(\delta)}{\hat{V}_{ij}(\delta)} \right)^{-1}.$$  

So the $(1 - \alpha)$ 100% confidence limits $(\delta_L, \delta_H)$ based on the WS method are the values of $\delta$ such that

$$\frac{(\hat{\delta}^* - \delta)^2}{\hat{V}(\delta)} = \chi^2_{\alpha/2}.$$  

The solutions $\delta_L$ and $\delta_H$ of (13) can be computed numerically using the bisection method.

The PL and WS methods proposed by Miettinen and Nurminen (1985) are different from the ones studied in this paper. Their PL method uses the unrestricted MLE $\hat{\theta} = (\hat{p}_{o1}, ..., \hat{p}_{ok}, \hat{\delta}_1, ..., \hat{\delta}_K)'$ instead of $\hat{\theta}$ and replaces $z_{\alpha/2}$ by $\chi^2_{K, \alpha}$ in (8). However, their PL method does not have confidence limits for some data. Their WS method defines the $(1 - \alpha)$ 100% confidence limits $(\delta_L, \delta_H)$ to be the values of $\delta$ such that

$$\frac{\sum_{j=1}^{K} w_j (\hat{p}_{ij} - \hat{p}_{ij} - \hat{\delta}_j)^2}{\sum_{j=1}^{K} w_j \hat{V}_{ij}(\delta)} = \chi^2_{\alpha/2}.$$  

This method requires choice of the weight $w_j$. The weight they suggested depends on the value of $\delta$ and the computation becomes more complex. Therefore, their WS method is not evaluated here. Note that for the unstratified case ($K = 1$), the PL and WS methods of Miettinen and Nurminen and our proposed PL and WS methods become the same, respectively.

The performances of the PL, SA, and WS methods given in (8), (10) and (13) are evaluated, without and with the continuity correction. The same continuity correction factor $[3/16] (1/\sum_{j=1}^{K} (n_{oj} + n_{ij}))/\sum_{j=1}^{K} (n_{oj} + n_{ij})$] as in the previous section is used. The continuity corrected versions of the confidence intervals from (8), (10), and (13) are obtained by subtracting and adding the
continuity correction factor from the lower limit and to the upper limit, respectively.

4. Simulation Results

4.1. Simulation Methods

We compare the performances of different methods discussed in Sections 2 and 3 under the setting of a common difference across strata (i.e., $p_{ij} - p_{0j} \equiv \delta$ for all $j$). The true underlying parameters are $p_{01}, p_{02}, ..., p_{0K}$, and $\delta$. The simulation study was conducted under the following settings for strata, sample sizes and parameters.

1. Number of strata: $K = 4$.
2. Total sample size per treatment: $N = N_1 = N = 40, 100, 200, 500, 1000$.
3. Number of patients per treatment in stratum $j$: $n_{o_j} = n_{1j} = N \times \gamma_j$, where $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.25, 0.25, 0.25, 0.25)$ and $(0.1, 0.2, 0.35, 0.35)$ are the two cases representing the balanced and unbalanced stratum frequencies.
4. The true treatment difference: $\delta = 0, 0.05, 0.1, 0.2$.
5. Three cases are considered regarding the range of the true $(p_{01}, p_{02}, p_{03}, p_{04})$:
   - 4(a). Thirty different $(p_{01}, p_{02}, p_{03}, p_{04})$ randomly chosen from $0$ to $0.1$.
   - 4(b). Twenty different $(p_{01}, p_{02}, p_{03}, p_{04})$ randomly chosen from $0.1$ to $0.5$.
   - 4(c). Twenty different $(p_{01}, p_{02}, p_{03}, p_{04})$ randomly chosen from $0.4$ to $0.7$.

The same chosen $(p_{01}, p_{02}, p_{03}, p_{04})$ was used for different methods, $N$, $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$, and $\delta$.

The case 4(a) is the low event rate case, representing the typical situation in analyses of clinical adverse experience data from clinical trials. For case 4(b) and 4(c), the rates in both treatment groups will be between $0.1$ and $0.9$. The SS and CMH methods have the same weight for the above sample size settings. For $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.25, 0.25, 0.25, 0.25)$, the EQ and CMH methods also have the same weight. Comparisons of the SS and EQ methods with the CMH method are discussed in Section 4.3. Here and in Section 4.2 we focus on the CMH, INVAR, and MR methods as well as the 3 likelihood methods.

For each set of $N$, $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$, $(p_{01}, p_{02}, p_{03}, p_{04})$, and $\delta$, the coverage probability was estimated by simulating 20,000 cases of 4 stratified $2 \times 2$ samples for the weighted average approaches and simulating 1,000 cases for the likelihood approaches, as the latter require much longer computation time. When the true coverage is greater than $0.94$, the standard error of the estimated coverage probability is about $0.0017$ for the weighted average approaches and about $0.0075$ for the likelihood approaches. The true coverage is considered to be less than $0.95$ if the estimated coverage is less than $0.945$ and $0.927$ (3 standard error limit off 0.95) for the weighted average approaches and the likelihood approaches, respectively.

For each $N$, $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$, and $\delta$, the estimated individual coverage probabilities corresponding to different $(p_{01}, p_{02}, p_{03}, p_{04})$ and their first quartile, median, and third quartile are plotted. These plots are similar to the box plot, but do not have whiskers.

4.2. Results

The confidence interval (CI) with the continuity correction has a better coverage probability than that without the continuity correction. Hence we present the results only for the continuity corrected CIs. The results are generally similar between $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.25, 0.25, 0.25, 0.25)$ and $(0.1, 0.2, 0.35, 0.35)$. Because of the limited space, here we only show the results for case 4(a) with $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.25, 0.25, 0.25, 0.25)$ in Figures 1 to 5. The figures for 4(b) and 4(c) can be obtained from authors per request.

For the case 4(a), the CMH method has the best coverage in the sense that the coverage of its 95% CI is closer to the target 0.95 than the others (Figures 1 to 5). The coverage of the 95% CI for the CMH method is at least 0.948. The INVAR and MR methods do not have an adequate coverage even for some large sample sizes (see $N=200$ to 1000 and $\delta = 0.05$, 0.1 for INVAR; see $N=500$, 1000 and $\delta = 0$, 0.05 for MR). For $N=40$ and $\delta = 0$ to 0.1, the coverage for the three weighted average approaches is conservative (well above 0.95) due to the variance correction. As discussed previously, the INVAR and MR methods need this correction to handle zero variance. For the CMH method, without this correction, the coverage for $N=40$ is between 0.951 and 0.997 for $\delta = 0$ and between 0.924 and 0.957 for $\delta = 0.05$ to 0.2. Although a conservative coverage is not good, it is probably better than an inadequate one. Therefore, we recommend using the variance correction for the CMH method when $N$ is less than 100. Among the likelihood approaches, the PL method is the best. Comparing with 0.927 (3 standard error limit off 0.95), the PL method has an adequate coverage for $\delta = 0.1$ and 0.2 when $N \geq 200$. However, the coverage of the PL method can be lower than 0.927 for $\delta = 0$ (see $N=500$ in Figure 4; the coverage is about 0.918). The SA and WS methods have a very poor coverage for $\delta = 0$ with almost all $N$ and for $\delta = 0.05$ with $N \leq 200$. One explanation is that the estimated variances in formulas (9) and (12) are not good for small $N$ and low rates because they involve the inverse of the individual
variances. In summary, the CMH method is the best among these 6 methods for the case 4(a).

For the cases 4(b) and 4(c), by comparing the results of 20,000 simulations for the 3 weighted average methods and also by comparing the results of 1,000 simulations for all 6 methods, we can summarize the results as follows. When \( N \leq 100 \), the CMH and PL methods (and the WS method for case 4(c)) perform similarly and are the best. When \( N \geq 200 \), all 6 methods have similar coverage probabilities.

4.3. Additional Simulation Results

We also compared the SS and EQ methods with the CMH method under the following artificial sample size setting. The total sample size for the 2 treatments is:

\[
2 \times N_0 = N_1 = N = 40, 100, 200, 500, 1000.
\]

The number of patients in the 4 strata for the 2 treatments is:

\[
(n_{01}, n_{02}, n_{03}, n_{04}) = (0.3, 0.2, 0.1, 0.4) * (N/2),
(n_{11}, n_{12}, n_{13}, n_{14}) = (0.1, 0.2, 0.35, 0.35) * N.
\]

For the low event rate case 4(a), the CMH method is still the best for all \( N \) and the SS and EQ methods also can have an inadequate coverage. For cases 4(b) and 4(c), the CMH method is either better than or similar to the SS and EQ methods. Detailed figures are available from authors.

5. Discussion and Conclusion

We have compared the coverage probabilities of 8 different methods under the condition that the true rate differences are constant across strata. The likelihood approaches require this homogeneity assumption and involve complex computations, yet they do not have a better coverage than the CMH weighted average method. Therefore, the likelihood approaches are not recommended. The weighted average approaches, except the INVAR method, do not require the homogeneity assumption and still define and estimate a meaningful overall treatment difference as long as there is no qualitative treatment-by-stratum interaction. The simulation study has shown that in the case of event rates less than 10% in one or both treatment groups, the CMH method is the best and has a coverage probability very close to or above the target 0.95; while all other methods can have an inadequate coverage. For the case of the event rates between 10% and 90%, the CMH method is still the best among the weighted average approaches when \( N \leq 100 \), and all methods have similar coverage probabilities when \( N \geq 200 \).

In practice, the choice of an appropriate weight for the weighted average approaches can be made based on the above findings and the stratification factor (prognostic versus non-prognostic). For prognostic stratification factors (such as sex and age), a meaningful overall treatment difference is naturally defined by \( \delta_c = \sum_{j=1}^{K} w_j (p_{1j} - p_{0j}) \) with the SS weight

\[ w_j = f_j. \]

The choice is between the SS weight and MR weight since both are targeting to estimate this overall treatment difference. Since patients are randomly assigned into treatment groups equally or in a certain ratio in clinical trials, it is expected that \( n_{0j} = c * n_{1j} \) with a constant \( c \) for all \( j \) would be nearly true and the SS weight is very close to the CMH weight. So, for prognostic stratification factors, in general we recommend using the SS method to obtain the confidence interval. When the event rates are between 10% and 90% and the sample sizes are large, the MR method can be used since it usually has a (slightly) shorter confidence interval width than the SS method. If the SS and CMH weights are very different, simulation evaluations of the SS and MR methods for the real situation are needed.

For non-prognostic stratification factors (such as study center or different studies), the definition of the overall treatment difference is more subjective (unless the homogeneity condition holds) and depends on the choice of the weight. Based on the performance, in general, we recommend using the CMH method to obtain the confidence interval. When the event rates are between 10% and 90% and the sample sizes are large, the INVAR weight can be used if the homogeneity assumption holds, because the INVAR method has the shortest confidence interval width under the homogeneity condition.

Acknowledgment

The authors are grateful to Prof. Scott Zeger for suggesting the PL method and motivating this research. The authors also thank Drs. Devan Mehrotra, Ji Zhang, and Thomas Capizzi for their very helpful suggestions.

References


Figures 1 to 5  Empirical Coverage Probability For 95\% CIs With Continuity Correction 
\((\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.25, 0.25, 0.25, 0.25)\)

4(a) 30 Different \( (p_{01}, p_{02}, p_{03}, p_{04}) \) from 0 to 0.1 
(20,000 Simulations for CMH, INVAR, and MR) 
(1,000 Simulations for PL, SA, and WS)

Note: The coverage of the SA method for \( N=40 \) and \( \delta = 0.05 \) is between 0.51 and 0.746 and hence does not appear in Figure 1.