On the Comparison of Two Correlated Proportions in the Analysis of Clustered Binary Data from a Cross-Sectional Study

Krishna Saha*

Abstract

Correlated binary data from a cross-sectional study frequently arise in the health sciences. In a randomized trial, patients with multiple myeloma from the same institution were randomly assigned to one of the two chemotherapy treatment groups, and it is often of interest in comparing overall success rates between the two chemotherapy treatment groups. Due to a cross-sectional design, the success rates between the two chemotherapy treatment groups are no longer independent. Moreover, due to cluster effect, post-treatment responses from each cluster (institution) for each treatment group can be highly correlated. By taking both correlation structures into account, we develop three efficient methods for the above inference problem. An extensive simulation study is conducted for the purpose of evaluating and comparing the performance of the proposed methods. An application to a chemotherapy study is used to illustrate the proposed methods.

Key Words: correlated binary data, confidence interval, coverage probability, difference between the proportions, independent binary data

1. Introduction

Correlated binary data from a cross-sectional study frequently arise in many biomedical, toxicological, clinical medicine, and epidemiological applications when the treatment conditions are available within each cluster. For instance, in a cancer and leukemia group B randomized trial (Cooper et al., 1993), patients with multiple myeloma from the same institution were randomly assigned to one of the two chemotherapy treatment groups, where each institution was considered as the randomization unit or cluster. In each group there were 21 institutions with the number of patients ranging from 2 to 12 in each treatment. A total of 72 eligible patients for treatment I and a total of 84 eligible patients for treatment II were accrued. In this study, it is of interest to compare two chemotherapy treatments with respect to success rates of the patients with multiple myeloma who survived at the end of this study. Note that posttreatment responses from the same institution for each treatment group can be highly correlated, which leads to inflated variances of the posttreatment response rates. Furthermore, the success rates between the two chemotherapy treatment groups from the same institute or cluster are no longer independent. For independent binary data, there are numerous binomial interval procedures available in literature for the estimation of the difference between the response rates in two treatment groups.

Let P_1 and P_2 be the the success rates between the two chemotherapy treatment groups, respectively. Then estimating the difference $P_1 - P_2$ will determine whether there is any difference between the two chemotherapy treatment groups. It is worthwhile to note that due to a cross sectional design the estimates of P_1 and P_2 are no longer independent. We, therefore, need to develop some efficient methods for estimating the difference $P_1 - P_2$ for such a design. This inference problem is commonly addressed by computing the confidence interval (CI) for the difference $P_1 - P_2$ by taking both correlations into account: (i) correlation among patients from the same institute for each treatment group and (ii) correlation between the two treatment groups for each institution. For a non-cross sectional

*Department of Mathematical Sciences, Central Connecticut State University, 1615 Stanley Street, New Britain, CT 06050, USA

design, Saha and Wang (2018) introduced several methods for estimating the difference $P_1 - P_2$. Of these methods they proposes two methods which are remarkably simple and have good properties. However, such an analysis would bias the inference procedures regarding $P_1 - P_2$ for a cross sectional design. Furthermore, inference methods concerning $P_1 - P_2$ that does not incorporate correlation between the two treatment groups in a cross sectional design may significantly inflate the Type I error rate. To address the issue, we construct several explicit asymptotic two-sided confidence intervals (CIs) for the difference $P_1 - P_2$ using the method of variance of estimates recovery (MOVER). The basic idea is to recover variance estimates required for the proportion difference from the confidence limits for single proportions. The CI estimators for a single proportion, which are incorporated with the MOVER, will include the CIs proposed by Saha et al. (2016).

2. Proposed Method

First, we present the data layout in the cross-sectional designs shown in Table 1, where $X_{ijl};\ l = 1, 2, \ldots, m_{ij};\ j = 1, 2, \ldots, k;\ i = 1, 2$ be the binary response (success or failure) for lth patient from the jth institution (or cluster) in the ith treatment. The parameter of interest in this article is $\Delta = P_1 - P_2$. In particular, we would like to construct simple but reliable CIs for Δ based on the method of variance estimates recovery (MOVER) as outlined by Tang et al. (2010). Here we briefly review this method for our case.

Table 1: Typical data layout in the cross-sectional designs

	Institution (or Cluster)						
	1		2		...	k	
Patients↓	Treatment 1	Treatment 2	Treatment 1	Treatment 2	...	Treatment 1	Treatment 2
1	X_{111}	X_{211}	X_{121}	X_{221}	...	X_{1k1}	X_{2k1}
2	X_{112}	X_{212}	X_{122}	X_{222}	...	X_{1k2}	X_{2k2}
⋮	⋮	⋮	⋮	⋮		⋮	⋮
l	X_{11l}	X_{21l}	X_{12l}	X_{22l}	...	X_{1kl}	X_{2kl}
⋮	⋮	⋮	⋮	⋮		⋮	⋮
m_{ij}	$X_{11m_{11}}$	$X_{21m_{21}}$	$X_{12m_{12}}$	$X_{22m_{22}}$	...	$X_{1km_{1k}}$	$X_{2km_{2k}}$
Total	X_{11}	X_{21}	X_{12}	X_{22}	...	X_{1k}	X_{2k}

We first consider that $\hat{P}_1$ and $\hat{P}_2$ are uncorrelated. By central limit theorem, a $100(1-\alpha)\%$ CI for $\Delta = P_1 - P_2$ is given by (L, U), where

$$L = \hat{P}_1 - \hat{P}_2 - z_{\alpha/2}\sqrt{\text{Var}(\hat{P}_1) + \text{Var}(\hat{P}_2)} \text{ and } U = \hat{P}_1 - \hat{P}_2 + z_{\alpha/2}\sqrt{\text{Var}(\hat{P}_1) + \text{Var}(\hat{P}_2)}.$$

The above confidence limits (L, U) are not yet available without the appropriate variance estimates, $\widehat{\text{Var}}(\hat{P}_i), i = 1, 2$. Suppose a $100(1-\alpha)\%$ CI for P_i is $(l_i, u_i), i = 1, 2$, where $l_i = \hat{P}_i - z_{\alpha/2}\sqrt{\text{Var}(\hat{P}_i)}$ implies $\widehat{\text{Var}}(\hat{P}_i = (\hat{P}_i - l_i)^2/z^2_{\alpha/2}$ under $P_i \approx l_i$. Similarly, $u_i = \hat{P}_i + z_{\alpha/2}\sqrt{\text{Var}(\hat{P}_i)}$ implies $\widehat{\text{Var}}(\hat{P}_i = (u_i - \hat{P}_i)^2/z^2_{\alpha/2}$ under $P_i \approx u_i$. Based on the possible values (l_1, u_1) of P_1 and (l_2, u_2) of P_2, the values closest to the minimum L and maximum U are $l_2 - u_1$ and $u_2 - l_1$, respectively. As a result, for setting L with $P_1 \approx l_1$ and $P_2 \approx u_2$, we have $\text{Var}(\hat{P}_1) + \text{Var}(\hat{P}_2) = (\hat{P}_1 - l_1)^2/z^2_{\alpha/2} + (u_2 - \hat{P}_2)^2/z^2_{\alpha/2}$, which gives

$$L = \hat{P}_1 - \hat{P}_2 - \sqrt{(\hat{P}_1 - l_1)^2 + (u_2 - \hat{P}_2)^2}.$$

Similarly, we have

$$U = \hat{P}_1 - \hat{P}_2 + \sqrt{(u_1 - \hat{P}_1)^2 + (\hat{P}_2 - l_2)^2}.$$

Obviously, $\hat{P}_1$ and $\hat{P}_2$ are correlated in the present setting and the covariance between $\hat{P}_1$ and $\hat{P}_2$ can be obtained as

$$\text{Cov}(\hat{P}_1, \hat{P}_2) = \text{Corr}(\hat{P}_1, \hat{P}_2)\sqrt{\text{Var}(\hat{P}_1)\text{Var}(\hat{P}_2)},$$

which can be used to extend obtaining the confidence limits to the case of correlated proportions. Thus, a $100(1-\alpha)\%$ CI based on MOVER for $\Delta = P_1 - P_2$ is given by (Δ_l^M, Δ_u^M), where

$$\Delta_l^M = \hat{P}_1 - \hat{P}_2 - \sqrt{(\hat{P}_1 - l_1)^2 + (u_2 - \hat{P}_2)^2 - 2\widehat{\text{Corr}}(\hat{P}_1, \hat{P}_2)(\hat{P}_1 - l_1)(u_2 - \hat{P}_2)}$$

and

$$\Delta_u^M = \hat{P}_1 - \hat{P}_2 + \sqrt{(u_1 - \hat{P}_1)^2 + (\hat{P}_2 - l_2)^2 - 2\widehat{\text{Corr}}(\hat{P}_1, \hat{P}_2)(u_1 - \hat{P}_1)(\hat{P}_2 - l_2)},$$

where $\widehat{\text{Corr}}(\hat{P}_1, \hat{P}_2)$ is the estimated correlation between $\hat{P}_1$ and $\hat{P}_2$. We obtain this estimate based on the ANOVA estimate discussed by Donner and Klar (2000) after ignoring the cross-sectional data structure.

As we have discussed earlier, one needs two separate CIs for P_1 and P_2 to construct CI for $\Delta = P_1 - P_2$ based on the above method, MOVER. It can be seen that $X_{ij} = \sum_{l=1}^{m_{ij}} X_{ijl}$ $(i = 1, 2; j = 1, 2, \ldots, k)$ follows an over-dispersed binomial distribution such as beta-binomial distribution with parameters P_i and ϕ_i $(i = 1, 2)$. Saha et al. (2016) investigated 16 asymptotic CIs for a single proportion for the over-dispersed binary data and compared the performance of those methods through an extensive simulation for a variety of parameter combinations. From their results, it shows that Wald CI based on beta-binomial may tend to be liberal , particularly for the small number of clusters, but the Wilson score and the profile likelihood-based CIs outperform the other CIs considered. We use these three methods to obtain two separate CIs (l_i, u_i) for $P_i,\ i = 1, 2$. We briefly review these CIs for P_i as follows:

The Wilson score CIs: Using the central limit theorem, it can be shown that $M_i^{1/2}(\hat{P}_i - P)/\sqrt{P_i(1-P_i)\hat{\omega}_i}$ converges in distribution to the standard normal distribution as $k \to \infty$, where $\hat{\omega}_i = \sum_{j=1}^{k}[m_{ij}\{1 + (m_{ij} - 1)\hat{\phi}_i]/M_i$ with $\hat{\phi}_i$ being the ANOVA-type estimate of ϕ_i. Then, the approximate $100(1-\alpha)\%$ Wilson CIs (l_i, u_i) for $P_i,\ i = 1, 2$ are the roots of the quadratic equation:

$$P\left(\frac{M_i(\hat{P}_i - P)^2}{P_i(1-P_i)\hat{\omega}_i} \le z_{\alpha/2}^2\right) = 1 - \alpha.$$

After some straightforward algebra, it can be obtained as

$$l_i = \tilde{P}_i - \frac{z_{\alpha/2}}{\tilde{M}_i}\sqrt{M_i\hat{P}_i(1-\hat{P}_i)\hat{\omega}_i + \frac{\hat{\omega}_i^2 z_{\alpha/2}^2}{4}}$$

and

$$u_i = \tilde{P}_i + \frac{z_{\alpha/2}}{\tilde{M}_i}\sqrt{M_i\hat{P}_i(1-\hat{P}_i)\hat{\omega}_i + \frac{\hat{\omega}_i^2 z_{\alpha/2}^2}{4}},$$

where

$$\tilde{P}_i = \frac{M_i\hat{P}_i + 0.5\hat{\omega}_i z^2_{\alpha/2}}{M_i + \hat{\omega}_i z^2_{\alpha/2}} \text{ and } \tilde{M}_i = M_i + \hat{\omega}_i z^2_{\alpha/2}.$$

The Profile Likelihood CIs: The log-likelihood of the beta-binomial model ($X_{ij} \sim BB(m_{ij}, P_i, \phi_i)$), apart from a constant, can be written as

$$l_i(P_i,\phi_i) = \sum_{j=1}^{k}\left[\sum_{r=0}^{x_{ij}-1}\ln\{(1-\phi_i)P_i + r\phi_i\} + \sum_{r=0}^{m_{ij}-x_{ij}-1}\ln\{(1-P_i)(1-\phi_i) + r\phi_i\} - \sum_{r=0}^{m_{ij}-1}\ln\{(1-\phi_i) + r\phi_i\}\right].$$

The ML estimator $\hat{P}_i$ of P_i can be obtained by maximizing $l_i(P_i, \phi_i)$ while the ML estimator $\hat{\phi}_i$ of ϕ_i can be obtained by maximizing $\sum_i l_i(P_i, \phi_i)$. Let $l^p_i(P_i) = l_i(P_i, \hat{\phi}(P_i))$ be the profile likelihood for P_i, where $\hat{\phi}(P_i)$ is obtained from the reduced model with respect to ϕ_i keeping P_i fixed. Then the approximate $100(1-\alpha)\%$ profile likelihood (PL) based CI (l_i, u_i) for P_i is given by

$$\{P_i : l^p_i(P_i) \geq l_i(\hat{P}_i, \hat{\phi}_i) - \frac{1}{2}\chi^2_{1,\alpha}\},$$

where $\chi^2_{1,\alpha}$ is the $100(1-\alpha)$ percentile of a chi-squared distribution with one degree of freedom. The endpoints (l_i, u_i) of the CI can be obtained by solving the system of nonlinear equations following the methodology introduced by Venzon and Moolgavkar (1988).

The Wald CIs: From the above, we see that the sample proportion $\hat{P}_i = X_i/M_i$, where $X_i = \sum_{j=1}^{k} X_{ij}$ $(i = 1, 2)$ is an unbiased estimator of P_i with the variance of $\hat{P}_i$ given by Var($\hat{P}_i$) = $P_i(1-P_i)\lambda_i/M_i$, where $\lambda_i = \sum_{j=1}^{k}[m_{ij}\{1 + (m_{ij}-1)\phi_i]/M_i$.

Then, as $k \to \infty$, $\hat{P}_i$ follows the normal distribution with mean P_i and variance $P_i(1-P_i)\lambda_i/M_i$. The resulting approximate $100(1-\alpha)\%$ Wald CI (l_i, u_i) for P_i is given by

$$l_i = \hat{P}_i - z_{\alpha/2}\sqrt{\hat{P}_i(1-\hat{P}_i)\hat{\lambda}_i/M_i} \text{ and } u_i = \hat{P}_i + z_{\alpha/2}\sqrt{\hat{P}_i(1-\hat{P}_i)\hat{\lambda}_i/M_i}.$$

Similar to the Wilson score CI, we also obtain the Wald CI (l_i, u_i) for P_i using the ANOVA-type estimate of ϕ_i in the above equation for $\hat{\lambda}_i$.

3. Simulations

In this section, we investigate the performance of the small and moderate sample behavior of the proposed methods in terms of observed coverage probability and average interval length using the pre-assigned confidence level of 95%. We considered the number of clus-

Table 2: Median coverage probability (CP) and median expected length (EL) of the 95% confidence intervals for $P_1 - P_2$ based on all parameter combinations considered here.

Method	Median CP	Median EL	Length Comparison individual/WA
WI	0.951	0.294	1.024
WA	0.947	0.287	1.000
PL	0.943	0.279	0.972

ters $k = 15, 25, 50$ and the response probabilities $P_1 = 0.1, 0.3, 0.5$ and $P_2 = P_1 + 0.2$.

Based on the historical data in biomedical applications, the common intraclass correlation coefficient was set at $\phi = \phi_1 = \phi_2 = 0.0, 0.1, 0.3, 0.5$ and the common correlation between two diagnostic tests was set at $\omega_{12} = \phi/2$. We also considered # of units per cluster (i.e. cluster size = m_{ij}) to be either fixed or variable. For the fixed cluster size case, m_{ij} was taken from the cross-sectional study example discussed in the introduction. For the variable cluster size case, m_{ij} was generated from the empirical distribution (ED) of 523 litter sizes where the cluster sizes range from 1 to 19 with a mean of 12 and standard deviation 2.98 (Kupper et al., 1986). We generated data X_{ijl} based on the bivariate beta-binomial distribution and generated 10,000 data set for each assessment.

The observed coverage probability (CP) and the expected interval length (EL) for two-sided confidence intervals (l_j, u_j) for $\delta = P_1 - P_2$ were obtained by

$$\text{CP} = \frac{\sum_{t=1}^{10000} I(l_t \leq \delta \leq u_t)}{10000} \quad \text{and} \quad \text{EL} = \frac{\sum_{t=1}^{10000}(u_t - l_t)}{10000},$$

where $I = 1$ if $l_t \leq \delta \leq u_t$, and $I = 0$, otherwise. The results are reported in Table 2 from which we make the following observations:

- The CPs for all three methods are virtually the same and reasonability close to the nominal level.
- As expected, the WA and PL methods show somewhat conservative coverage; however, the CP for the PL method shows a bit more conservative than that of the WA method.
- The WI method produces better coverage compared to the other two methods and maintain the coverage very close to the nominal level.
- All three methods tend to have similar ELs; however, the WI and WA methods tend to have larger ELs compared to the PL method.

4. Example: A Chemotherapy Study

We revisit the example of estimating the difference between the success rates of the two chemotherapy treatments that we have considered in Section 1. In this clinical trial example, patients enrolled in the trial are randomly assigned to one of the two chemotherapy treatment groups, and the success rates on the treatments are compared to determine which treatment is more efficacious. Table 3 provides summary statistics of this study.

Table 3: Summary statistics for the data set in a chemotherapy study

Chemotherapy Treatments	# of subjects	# of clusters	mean cluster size	success rate
Treatment I	72	21	3.43	0.542
Treatment II	84	21	4.00	0.524

The distributions of cluster-level proportions for both treatment groups are shown in Figure 1. The estimated success probability and the estimated intraclass correlation for both treatment groups and the estimated common intraclass correlation are provided in Table 4. In

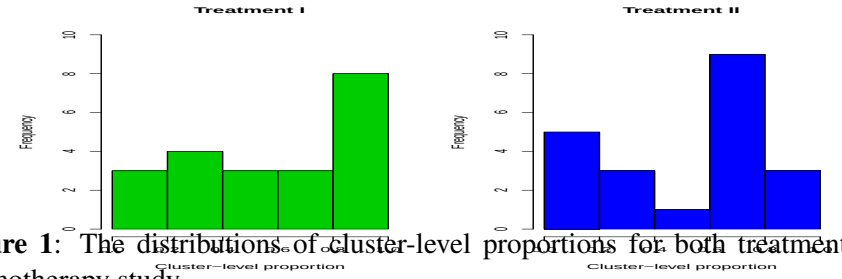

Figure 1: The distributions of cluster-level proportions for both treatment groups in a chemotherapy study.

Table 4: The point estimates of the parameters obtained based on the four different methods for the data set in a solar protection study.

Methods	π_1	π_2	ϕ_1	ϕ_2
ML	0.586	0.521	0.194	0.083
AOV	0.542	0.524	0.226	0.142

this study, it is of interest to compare two chemotherapy treatments with respect to success rates of the patients with multiple myeloma who survived at the end of this study. Then, the 95% confidence intervals for P_1 - P_2 obtained using the proposed methods are given in Table 5. It is seen from Table 5 that all three confidence internals contain 0, indicating that there is no statistical evidence of different success rates of the two chemotherapy treatments. For unpaired case, the 95% confidence interval for P_1 - P_2 based on the MW1 method used by Saha and Wang (2019) is given as $(-0.113, 0.302)$ with the interval width of 0.415 (see Table 7 of Saha and Wang (2019)). As expected due to positive correlation between treatment groups, our proposed method WI for the split-clustered case shows the shorter width compared to the method MW1 for non-split-clustered case.

Table 5: The 95% confidence intervals for P_1 - P_2 obtained using the WI, WA and PL methods.

Method	Lower Limit	Upper Limit	Comparison Length
WI	-0.1634	0.1958	0.3592
WA	-0.1694	0.2051	0.3745
PL	-0.1222	0.2181	0.3403

5. Conclusion

This paper proposed three methods to construct the confidence intervals for $P_1 - P_2$ for a clustered binary data from a cross-sectional study based on the MOVER using the two

separate CIs for a single proportion. The results of a simulation study suggest that the proposed methods generally perform well as their observed CPs are very close to the nominal coverage level. The PL method is preferable compared to the other methods in the sense that they generally possess shorter ELs in almost all data situations considered here. However, due to simplicity in the computation, we recommend the WI CI for $P_1 - P_2$ for a clustered binary data from a cross-sectional study.

Acknowledgements

This paper was supported in part by a CSU-AAUP University research grant.

REFERENCES

Cooper, M. R., Dear, K. B. G., McIntyre, 0. R., Ozer, H., Ellerton, J., Cannellos, G., Duggan, B., and Schiffer, C. (1993), "A Randomized Clinical TrialCcomparing Melphalan/Prednisone with and without a-2b Interferon in Newly-Diagnosed Patients with Multiple Myeloma: A Cancer and Leukemia Group B Study". *Journal of Clinical Oncology*, **11**, 155-160.

Donner, A. and Klar, N. (2000), "Design and Alalysis of Cluster Randomized Trials in Health Research". Arnold, New York, USA.

Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. (1986). "The Impact of Litter Effects on Dose-Response Modeling in Teratology". *Biometrics*, **42**, 8598.

Saha, K. K. and Wang, S., and Miller, D. "A Comparison of Some Approximate Confidence Intervals for a Single Proportion for Clustered Binary Outcome Data. *International Journal of Biostatistics*, bf 12, 124.

Saha, K. K. and Wang, S. "Confidence Intervals for the Difference in the Success Rates of Two Treatments in the Analysis of Correlated Binary Responses". *Biometrical Journal*, **61**, 9831002.

Venzon, D. J. and Moolgavkar, S. H. (1988), "A method for computing profilelikelihood-based confidence intervals". *Applied Statistics*, **37**, 87-94.