Simultaneous Confidence Intervals for the Difference of Proportions from Multivariate Binomial Distribution

Hyeong Chul Jeong
Department of Computer Science and Statistics, Pyongtaek University
111 Yongyi-dong, Pyongtaek, Korea
jhc@ptuniv.ac.kr

Myoungshic Jhun, Jae Won Lee
Department of Statistics, Korea University
5-1 Anam-dong, Sungbuk-gu, Seoul, Korea
jhun@korea.ac.kr ; jael@korea.ac.kr

1. Introduction

For the two-group data from multivariate binomial distribution, we consider two problems. First, the simultaneous confidence level and its standard error of a collection of the dependent confidence intervals for the difference of proportions with a *experimentwise error rate* at the α level are presented. Secondly, given a *familywise error rate* at the level α, obtaining simultaneous confidence intervals for the difference of proportions between two groups are considered. The bootstrap method is used to construct the simultaneous confidence intervals for the difference of proportions. This methodology is compared to the usual Bonferroni-Sidak method and it is demonstrated both Bonferroni and Sidak simultaneous confidence intervals are grossly conservative in certain instance because of their failure to account for dependence between confidence intervals. The comparison of the procedures are made in terms of the *average coverage probability*(Woodroofe and Jhun; 1989), which is possibly a more relevant description of the performance.

2. Simultaneous confidence levels

Let us assume that we are sampling from two different multivariate binomial distributions, where Y_1 are from $MVB_k(P_1, n, D_1)$, and Y_2 are from $MVB_k(P_2, n, D_2)$, where $P_i = (p_{i1}, \cdots, p_{ik})$ denotes population proportions vector for a k-component of i-th group and dependence structure specified by D_i. We are interested in simultaneous confidence level of a collection of confidence intervals for each of the parameter $p_{1j} - p_{2j}$, $j = 1, \cdots, k$. The exact simultaneous probability is determined by the underlying multivariate binomial distribution. In this case the goal is to estimate

$$\pi_\alpha = \Pr[p_{1j} - p_{2j} \in I_\alpha(Y_{1j} - Y_{2j}) \quad \text{for all} \quad j = 1, \cdots, k]$$

where $I_\alpha(Y_{1j}-Y_{2j})$ is a marginal confidence interval for $p_{1j}-p_{2j}$. The estimate of π_α is computed as

$$\pi_\alpha^* = \Pr[\hat{p}_{1j} - \hat{p}_{2j} \in I_\alpha(Y_{1j}^* - Y_{2j}^*) \quad \text{for all} \quad j = 1, \cdots, k]$$

by bootstrap. This analysis is an extension of the Westfall(1985)'s bootstrap simultaneous confidence levels with one-group multivariate Bernoulli data to the two-group proportions. In addition, standard errors of estimates of simultaneous coverage level are estimated by double bootstrap method.

3. Simultaneous Confidence intervals

We are interested in constructing a simultaneous confidence region for $d_j = p_{1j} - p_{2j}, j = 1, \cdots, k$. We apply the bootstrap method to construct simultaneous confidence intervals for the difference of proportions. Consider a quantity

$$B_{(k)} = \max\left[\frac{|\hat{d}_1 - d_1|}{sd(\hat{d}_1)}, \frac{|\hat{d}_2 - d_2|}{sd(\hat{d}_2)}, \cdots, \frac{|\hat{d}_k - d_k|}{sd(\hat{d}_k)}\right],$$

and as an estimator of the sampling distibution of $B_{(k)}$, use the bootstrap distribution of

$$B_{(k)}^* = \max\left[\frac{|\hat{d}_1^* - \hat{d}_1|}{sd(\hat{d}_1^*)}, \frac{|\hat{d}_2^* - \hat{d}_2|}{sd(\hat{d}_2^*)}, \cdots, \frac{|\hat{d}_k^* - \hat{d}_k|}{sd(\hat{d}_k^*)}\right]$$

where, $sd(\hat{d}_j^*) = [n_1^{-1}\hat{p}_{1j}^*(1-\hat{p}_{1j}^*) + n_2^{-1}\hat{p}_{2j}^*(1-\hat{p}_{2j}^*)]^{1/2}$.

Using the bootstrap estimator of the sampling distributions of $B_{(k)}$, a $100(1-\alpha)\%$ simultaneous confidence region for $d_j, j = 1, \cdots, k$ is obtained as

$$\left[d_j \in \hat{d}_j \pm B_{(k)}^*(1-\alpha)Sd(\hat{d}_j), \text{ for all } j = 1, \cdots, k\right].$$

We adjust the Westfall and Young(1989)'s multiple tests to the simultaneous confidence intervals with two groups multivariate Bernoulli data. It is demonstrated by using a simulation study that the bootstrap method has some advantages in constructing simultaneous confidence regions for the difference of proportions(Jhun and Jeong; 2000).

REFERENCES

Jhun, M. and Jeong, H.C. (2000). Application of bootstrap methods for categorical data analysis. *Comp. Statist. & Data Ana.*, **35**, 83-91.

Westfall, P.H. (1985). Simultaneous small-sample multivariate Bernoulli confidence intervals. *Biometrics*, **41**, 1001-1003.

Westfall, P.H. and Young, S.S. (1989). P-value adjustments for multiple tests in multivariate binomial models. *JASA*, **84**, 780-786.

Woodroofe, M. and Jhun, M. (1989). Singh's theorem in the lattice case. *Statist. Prob. Lett.*, **7**, 201-205.