The Korean Journal of Applied Statistics (2009)
22(1), 129–140

Constructing Simultaneous Confidence Intervals for the Difference of Proportions from Multivariate Binomial Distributions

Hyeong Chul Jeong[1] · Daehak Kim[2]

[1]Dept. of Applied Statistics, University of Suwon;
[2]School of Computer & Information Communications, Catholic University of Daegu

(Received October 2008; accepted November 2008)

Abstract

In this paper, we consider simultaneous confidence intervals for the difference of proportions between two groups taken from multivariate binomial distributions in a nonparametric way. We briefly discuss the construction of simultaneous confidence intervals using the method of adjusting the p-values in multiple tests. The features of bootstrap simultaneous confidence intervals using non-pooled samples are presented. We also compute confidence intervals from the adjusted p-values of multiple tests in the Westfall (1985) style based on a pooled sample. The average coverage probabilities of the bootstrap simultaneous confidence intervals are compared with those of the Bonferroni simultaneous confidence intervals and the Šidák simultaneous confidence intervals. Finally, we give an example that shows how the proposed bootstrap simultaneous confidence intervals can be utilized through data analysis.

Keywords: Multivariate binomial distribution, simultaneous confidence intervals, bootstrap, multiple test, pooled sample.

1. Introduction

Binary data is very common in social science or clinical settings; examples include success or failure, survival or death, and occurrence or non-occurrence of cancer. An observed value follows a multivariate binomial distribution if it contains multiple categories and each of the categories involves a binary decision. As an example, given two groups and k tissues that exhibit a binary response to a certain treatment, we are interested in the difference in the proportions of the two groups' tissues that respond to that treatment. We may want to determine whether the cancer rates in tissues or organs differ when drug A is injected in two different species of animals or whether they differ between different tissues if two different treatments, A and B, are injected in the same species. We have to consider that many tissues intermingle with one another and form relationships.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund)(KRF-2006-521-C00031).
[1]Corresponding author: Assistant Professor, Dept. of Applied Statistics, University of Suwon, Wau-ri, Bongdam, Hwasung, Kyunggi-do 445-743, Korea. E-mail:jhc@suwon.ac.kr

In other words, the occurrence of cancer in one tissue is likely to increase the chances of cancer elsewhere. Therefore, we must observe the difference in the cancer ratios as a whole instead of looking at the cancer ratios of each tissue separately. Thus, a researcher must use simultaneous reasoning through multiple comparisons if one is interested in the *familywise error rate* of the entire set rather than the *experimentwise error rate.*

When thinking about error control in simultaneous confidence intervals, we inevitably have to consider both familywise and eachwise errors. Jeong *et al.* (2007) discuss how to estimate the familywise error rate when the eachwise error rate of each variable is given for two multivariate binomial distributions. In this study, as opposed to the approach in Jeong *et al.*(2007), we construct the confidence intervals of a marginal variable by maintaining its given familywise error rate for the difference of proportions between two multivariate binomial distributions.

Defining the simultaneous confidence intervals for the difference of proportions $d_j = p_{1j} - p_{2j}$ $(j = 1, \ldots, k)$ between two groups with k variables is equivalent to adjusting the p-values computed in multiple testing $H_j : p_{1j} = p_{2j}$ $(j = 1, \ldots, k)$. Westfall and Young (1989, 1993) dealt with methods to adjust the p-values for the difference of proportions. Westfall and Young (1989) emphasized the advantages of nonparametric methods by comparing the properties of bootstrap and permutation p-value adjustments with those of Bonferroni and Šidák p-value adjustments. Further, they recommended using pooled samples when calculating test statistics because tests depend on the null hypothesis. However, in this paper, we do not intend to discuss p-value adjustments but rather define the simultaneous confidence intervals using the bootstrap method. The definition of a confidence interval does not depend on the given null hypothesis. Therefore, the samples used to calculate the standard deviation of d_j for a confidence interval are different from those used to calculate the standard deviation of d_j for the hypothesis test. The simultaneous confidence interval that uses a non-pooled sample is much more advantageous than the simultaneous confidence interval that uses a pooled sample. In addition, through simulation study we know that the bootstrap simultaneous confidence intervals are much more advantageous than the Bonferroni or Šidák intervals in terms of average coverage probabilities (Woodroofe and Jhun, 1988).

In fact, in the case of binary, Poisson, or other discrete data, the bootstrap method has no advantage over the normal approximation because of the discreteness (Singh, 1981). However, because Woodroofe and Jhun (1988) proved that bootstraps have enough advantages in terms of average coverage probability for discrete data on average, we intend to evaluate the average execution capability of bootstraps in computing a simultaneous confidence interval.

Jhun *et al.* (2007) defined a simultaneous confidence interval for a multivariate Poisson distribution. Jhun *et al.* (2007) proposed an asymmetric simultaneous confidence interval that considered the skewness of a Poisson distribution. However, similar to the results from Jhun *et al.* (2007), this study will show that the bootstrap simultaneous confidence interval is well applied even for a difference in group ratios. We will also address the possibility that the method of defining the confidence interval may vary depending on the distribution.

Section 2 introduces bootstrap simultaneous confidence intervals for differences of proportions between two groups. In particular, we perform simulations to compare pooled and non-pooled samples and determine average coverage probabilities, which is possibly a more relevant description of confidence interval performance. Section 3 compares simultaneous confidence intervals with p-value adjustments using data analyses. Section 4 concludes the paper.

2. Simultaneous Confidence Intervals

2.1. Motivation

Assume that the observable data vectors $Y' = (Y_1, \ldots, Y_k)$ with binomial marginal distributions $Y_j \sim B(n, p_j)$ may be modeled as having multivariate binomial distributions. Define the notation

$$Y \sim \text{MVB}_k(P, n, D),$$

where $P = (p_1, \ldots, p_k)$ denotes the population proportions vector for a k-component and the dependence structure is specified by D.

When $n = 1$, the distribution $\text{MVB}_k(P, 1, D)$ is a multivariate Bernoulli distribution. If there are two treatment groups with n_1, n_2 experimental units, respectively, the available data vectors are $X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$, where $X'_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijk})$, $(i = 1, 2)$ has an independent and identically distribution(iid) as $\text{MVB}_k(P_i, 1, D_i)$ and X_{ijl} is the response falling in l^{th} variable of the j^{th} observation for the i^{th} group. Let the i^{th} group proportion vector $P_i = (p_{i1}, p_{i2}, \ldots, p_{ik})$. Then $E(X_{ij}) = P_i$. Let the i^{th} group frequency vector $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik})$. Then $Y_i = \sum_j^{n_i} X_{ij}$ is distributed as $\text{MVB}_k(P_i, n_i, D_i)$. The maximum likelihood estimator of the probability p_{ij} is $\hat{p}_{ij} = Y_{ij}/n_i$ (Jeong *et al.*, 2007).

Let us assume that we are sampling from two different multivariate binomial distributions, where Y_1 is from $\text{MVB}_k(P_1, n, D_1)$ and Y_2 is from $\text{MVB}_k(P_2, n, D_2)$. We are interested in constructing a simultaneous confidence region for $d_j = p_{1j} - p_{2j}$, $j = 1, \ldots, k$. The two sample $Z_j = \{\hat{d}_j - d_j\}/sd(\hat{d}_j)$ statistics are used to construct the marginal confidence intervals, so that the marginal confidence intervals are $\hat{d}_j \pm z(\alpha/2)sd(\hat{d}_j)$, $j = 1, \ldots, k$. Assuming that the two population proportions are different, the detailed formula for $sd(\hat{d}_j)$ is $\sqrt{\hat{p}_{1j}(1-\hat{p}_{1j})/n_1 + \hat{p}_{2j}(1-\hat{p}_{2j})/n_2}$, whereas if the population proportions are the same, $sd(\hat{d}_j)$ is $\sqrt{\hat{p}_j(1-\hat{p}_j)(1/n_1 + 1/n_2)}$, where $\hat{p}_j$ is the sample proportion computed from the pooled sample.

Looking at the experimentwise marginal confidence level, false simultaneous confidence regions are most likely to occur when one uses a marginal confidence level. Therefore, the marginal confidence level $1 - \alpha$ should be adjusted. The adjusted confidence levels will depend on the unknown parameters P and D. These simultaneous confidence regions and adjusted confidence levels may be estimated using the data and bootstrap resampling. To conduct a multiple test, one can obtain the adjusted p values using the technique of Westfall and Young (1989). Let $\{pv_j,\ j = 1, \ldots, k\}$ denote the marginal p values relative to the null hypotheses. Let $X^*_{11}, \ldots, X^*_{1n_1}, X^*_{21}, \ldots, X^*_{2n_2}$ be *iid* according to a multivariate probability distribution that assigns a mass $1/n$ $(n = n_1 + n_2)$ to each of the observed vectors X_{ij} $(i = 1, 2;\ j = 1, \ldots, n)$. Thus the X^*_{ij} are distributed as *iid* $\text{MVB}_k(\hat{P}, 1, \hat{D})$, where $\hat{P} = (1/n)\sum Y_i$ and $\hat{D}$ is the empirical probability measure of the observed k tuples in the combined sample. Then the bootstrap adjusted p-values are given by

$$\text{ad } pv^*_j = \Pr\left[\min\{PV^*_j;\ j = 1, \ldots, k\} \leq pv_j\right],$$

where the pv^*_j are the random p-values from the bootstrap sample (see *e.g* Westfall and Young, 1989). If we invert the two-sided multiple tests, we obtain the simultaneous confidence regions. Jhun and Jeong (2000) proposed the bootstrap simultaneous confidence intervals to contrast several multinomial populations by using the maximum quantities of each pivotal statistics.

2.2. Bootstrap simultaneous confidence regions

We apply the bootstrap method to construct simultaneous confidence intervals for the difference of proportions. The maximum order statistics of the random variable $Z_1, Z_2, \ldots, Z_k$ are given by

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

and to estimate the sampling distribution of $Z_{(k)}$ we use the bootstrap distribution of

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

Let $G(z) = P(Z_{(k)} \leq z_{(k)})$, $G^*(z)$ be the bootstrap *cdf* of $Z^*_{(k)}$ and $G^{*B}(z)$ be its Monte Carlo estimate. The conditional distribution of $Z^*_{(k)}$ converges weakly to the distribution of the maximum order statistics of $G(z)$ because the real-valued maximum functions are continuous functions of $Z_1, Z_2, \ldots, Z_k$. Therefore, we have that $G^*(z)$ converges in distribution to $G(z)$. If B bootstrap replicates have been obtained, a $100(1-\alpha)\%$ simultaneous confidence region for $d_j = p_{1j} - p_{2j}, j = 1, \ldots, k$ is given as

$$\left[d_j \in \hat{d}_j \pm Q^{*B}(1-\alpha) sd(\hat{d}_j), \text{ for all } j = 1, \ldots, k\right],$$

where $Q^{*B}(1-\alpha)$ is $(G^{*B})^{-1}(1-\alpha)$.

The large-sample validity of the proposed intervals depends on the limiting behavior of the distribution $G^*(\cdot)$. Because the bootstrap *cdf* of $Z^*_{(k)}$ converges weakly to the true conditional distribution $Q^*(p) \to Q(p)$ (pointwise for $0 < p < 1$), where $Q^* = G^{*-1}$ is theoretical bootstrap percentile function. In practice, the endpoints of the interval are estimates of $Q^*(1-\alpha)$, since a finite value of B is used. From the Glivenko-Cantelli lemma, we know that $Q^{*B}(p)$ converges to $Q^*(p)$ in probability as $B \to \infty$ (Freedman, 1984; Thombs and Schucany, 1990).

To test the null hypotheses $H_{0j} : p_{1j} = p_{2j}$, the two-sample Z statistics are used as follows.

$$\hat{Z}_j = \frac{\hat{p}_{1j} - \hat{p}_{2j}}{\sqrt{\hat{p}_j(1-\hat{p}_j)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $\hat{p}_j = (n_1 \hat{p}_{1j} + n_2 \hat{p}_{2j})/(n_1 + n_2)$ with pooled samples.

Constructing a confidence interval using the maximum pivots in $Z(i)$, $i = 1, \ldots, k$ is equivalent to the method of adjusting the p-values in multiple tests. The difference between the confidence intervals and testing procedures, however, is in the resampling methods for bootstrap samples. In multiple tests, in order to adjust p-values, a pooled sample should be used since the pivotal statistic depends on the null hypothesis. When constructing confidence intervals, however, there is no need to pool the samples. It is therefore reasonable to use non-pooled samples when constructing confidence intervals. Moreover, constructing simultaneous confidence intervals provides more information than adjusting the p-values. In a real computational problem, adjusting p-values and constructing simultaneous confidence intervals can be done regardless of pooling the sample. We expect that the use of a pooled or non-pooled sample will impact statistical robustness. We will therefore examine which of the two methods is more robust.

2.3. Bonferroni-style simultaneous confidence regions

Bonferroni's inequality plays an important role in the construction of simultaneous confidence intervals for the difference of proportions. Let k denote a component of a multivariate binomial random vector, then $100(1-\alpha)\%$ Bonferroni simultaneous confidence intervals are

$$\hat{d}_j \pm z\left(\frac{\alpha}{2k}\right) sd(\hat{d}_j), \quad \text{for all } j = 1, \ldots, k,$$

where $z(\alpha/(2k))$ is the $100(1-\alpha/2k)^{th}$ percentile of the standard normal distribution. If positive dependence can be assumed, the Šidák method may be less conservative than the Bonferroni method (see *e.g.* Holland and Copenhaver, 1987). Šidák simultaneous confidence intervals take the form

$$\hat{d}_j \pm z\left(1-\left(1-\frac{\alpha}{2}\right)^{\frac{1}{k}}\right) sd(\hat{d}_j), \quad \text{for all } j = 1, \ldots, k.$$

These techniques are simple functions of the experimentwise marginal confidence level. They are computationally quick and easy, but these methods lack utility in that they fail to account for the discreteness and dependence structures of multivariate binomial data. For example, with perfectly correlated multivariate binomial distributions, these simultaneous confidence regions are overly conservative. Therefore, these procedures may be wasteful, in the sense that the probability error rate is less than, rather than equal to, α.

2.4. Pooled sample and non-pooled sample

Although it is reasonable to use the pooled bootstrap sample since only the test statistics rely on the null hypothesis for multiple testing to control the significance level, we can anticipate there will be discrepancies between using the pooled sample and the non-pooled sample when constructing the confidence interval. The choice of sample will also be affected by the correlation between the groups and the sample proportion vector. The upper and lower bounds for defining the simultaneous confidence interval of d_j are

$$L = (\hat{p}_{j1} - \hat{p}_{j2}) - Q^{*B}(1-\alpha)\sqrt{n_1^{-1}\hat{p}_{j1}(1-\hat{p}_{j1}) + n_2^{-1}\hat{p}_{j2}(1-\hat{p}_{j2})},$$

$$U = (\hat{p}_{j1} - \hat{p}_{j2}) + Q^{*B}(1-\alpha)\sqrt{n_1^{-1}\hat{p}_{j1}(1-\hat{p}_{j1}) + n_2^{-1}\hat{p}_{j2}(1-\hat{p}_{j2})},$$

where non-pooled sample statistics are

$$Z_{(k)}^{*NP} = \max\left[\frac{|(\hat{p}_{1j}^* - \hat{p}_{2j}^*) - (\hat{p}_1 - \hat{p}_2)|}{\sqrt{\hat{p}_{1j}^*(1-\hat{p}_{1j}^*)/n_1 + \hat{p}_{2j}^*(1-\hat{p}_{2j}^*)/n_2}},\ j = 1, \ldots, k\right] \tag{2.1}$$

although the following two statistics can be used for pooled samples

$$Z_{(k)}^{*P1} = \max\left[\frac{|(\hat{p}_{1j}^* - \hat{p}_{2j}^*)|}{\sqrt{\hat{p}_{1j}^*(1-\hat{p}_{1j}^*)/n_1 + \hat{p}_{2j}^*(1-\hat{p}_{2j}^*)/n_2}},\ j = 1, \ldots, k\right] \tag{2.2}$$

$$Z_{(k)}^{*P2} = \max\left[\frac{|(\hat{p}_{1j}^* - \hat{p}_{2j}^*)|}{\sqrt{\hat{p}_j^*(1-\hat{p}_j^*)(1/n_1 + 1/n_2)}},\ j = 1, \ldots, k\right] \tag{2.3}$$

to obtain $Q^{*B}(1-\alpha)$.

To analyze the performance of the statistics using pooled and non-pooled samples, we performed a Monte Carlo investigation and computed the familywise error rates. We limited our attention to the cases where the proportions of the two groups are the same and when they are different and when the correlations are same and when they are different. For convenience, equal proportions and an equivalent correlation matrix were used for the correlation, where P_1, P_2 and ρ are as follows when $k = 5$.

- Case 1 : $P_1 = \{0.50, \ldots, 0.50\}$, $\rho = 0.50$; $P_2 = \{0.50, \ldots, 0.50\}$, $\rho = 0.25$.
- Case 2 : $P_1 = \{0.50, \ldots, 0.50\}$, $\rho = 0.75$; $P_2 = \{0.50, \ldots, 0.50\}$, $\rho = 0.50$.
- Case 3 : $P_1 = \{0.25, \ldots, 0.25\}$, $\rho = 0.50$; $P_2 = \{0.75, \ldots, 0.75\}$, $\rho = 0.50$.
- Case 4 : $P_1 = \{0.25, \ldots, 0.25\}$, $\rho = 0.25$; $P_2 = \{0.75, \ldots, 0.75\}$, $\rho = 0.75$.

We consider cases where the sample sizes $n_1(n_2)$ are the same and different and cases where the nominal coverage $1 - \alpha = 0.80$, 0.90, 0.95 and 0.99. The algorithm of Park *et al.* (1996) was used to generate the pseudo-random samples of the multivariate Bernoulli distribution with predefined proportions and dependence structure. Table 2.1 shows the $(1-\alpha)\%$ bootstrap simultaneous confidence levels computed using statistics from (2.1), (2.2), (2.3), the Bonferroni simultaneous confidence levels, and the Šidák simultaneous confidence levels for the four cases. The Monte Carlo simulation was repeated 1000 times.

Table 2.1 demonstrates many facts. First, there is a difference between statistics from (2.2) (BP1) and statistics from (2.3) (BP2). The estimated simultaneous confidence levels from (2.3) have lower nominal probabilities than the statistics from (2.2). In case 1, in fact, the BP2 method appears optimal, but the results come from the fact that the confidence levels of Wald confidence intervals have lower estimated confidence probabilities (Jeong *et al.*, 2007). Secondly, in case 1, the bootstrap method using a pooled sample (BP1) is slightly closer to the nominal probabilities than the bootstrap method using a non-pooled sample (BNP). As the sample size increases, however, the estimated coverages for the two methods converge to true the nominal coverage. Thirdly, in cases 2, 3 and 4, the non-pooled bootstrap method is closer to the nominal probability than pooled sample. This observation becomes even clearer when the difference between the sample sizes of the two groups is large. In case 4, the non-pooled bootstrap method is much better than the pooled bootstrap method. Fourth, the Bonferroni simultaneous confidence interval and the Šidák simultaneous confidence interval are over-estimated relative to the nominal levels when the sample size is larger. The results are not always over-estimated because discrete statistics do not converge to asymptotic normality. For smaller samples, the Bonferroni and Šidák methods are underestimated relative to the given nominal level, indicating that the estimated confidence levels of each of the marginal variables is lower than the nominal level.

In general, the bootstrap simultaneous confidence level using a non-pooled sample yields more feasible results. This trend is more apparent when the sample sizes of the two groups are different.

2.5. Average coverage probability

In this section, we investigate the performance of the bootstrap method when obtaining regions for the difference of proportions, which is the lattice case. Some related theoretical background for bootstrap methods is as follows. Let T_n be a studentized sum of n *iid* random variables with a

Table 2.1. Simulation results(BNP is the bootstrap method with non-pooled sample using Equation (2.1), BP1 is the bootstrap method with pooled sample using Equation (2.2) and BP2 is the bootstrap method with pooled sample using Equation (2.3). BON is Bonferroni method and SID is Šidák method)

n_1	n_2	$1-\alpha$	Case 1					Case 2				
			BNP	BP1	BP2	BON	SID	BNP	BP1	BP2	BON	SID
10	10	0.80	.821	.811	.643	.814	.814	.813	.809	.638	.803	.803
		0.90	.885	.881	.842	.842	.842	.890	.875	.840	.843	.843
		0.95	.962	.940	.854	.848	.848	.966	.940	.854	.846	.846
		0.99	1.000	.991	.941	.943	.943	1.000	.988	.944	.949	.949
30	30	0.80	.831	.825	.738	.834	.834	.821	.823	.755	.825	.825
		0.90	.909	.910	.869	.909	.909	.909	.913	.877	.909	.909
		0.95	.956	.954	.934	.952	.952	.951	.953	.933	.949	.949
		0.99	.995	.992	.986	.988	.988	.989	.989	.982	.985	.985
50	50	0.80	.795	.804	.766	.847	.799	.807	.810	.779	.855	.808
		0.90	.905	.901	.871	.918	.918	.912	.911	.880	.924	.924
		0.95	.959	.952	.934	.959	.959	.960	.959	.938	.961	.961
		0.99	.989	.987	.982	.985	.985	.992	.994	.989	.989	.989
100	100	0.80	.805	.805	.779	.846	.846	.791	.797	.781	.841	.841
		0.90	.887	.894	.873	.915	.914	.886	.888	.872	.908	.908
		0.95	.944	.941	.930	.957	.957	.946	.949	.936	.953	.953
		0.99	.986	.984	.982	.985	.985	.991	.991	.984	.991	.991
10	100	0.80	.872	.813	.707	.733	.733	.883	.697	.568	.668	.668
		0.90	.960	.925	.787	.818	.818	.936	.875	.692	.775	.775
		0.95	.986	.950	.853	.894	.885	.972	.922	.801	.867	.843
		0.99	.990	.989	.927	.936	.936	.991	.990	.894	.908	.908
20	100	0.80	.806	.802	.758	.809	.798	.812	.740	.688	.787	.774
		0.90	.906	.899	.859	.880	.880	.918	.877	.815	.870	.870
		0.95	.957	.952	.904	.929	.929	.962	.936	.886	.925	.925
		0.99	.997	.991	.960	.966	.966	.996	.991	.961	.970	.970

n_1	n_2	$1-\alpha$	Case 3					Case 4				
			BNP	BP1	BP2	BON	SID	BNP	BP1	BP2	BON	SID
10	10	0.80	.920	.676	.651	.672	.672	.891	.689	.670	.690	.690
		0.90	.976	.876	.763	.801	.801	.969	.881	.778	.802	.802
		0.95	.983	.914	.819	.911	.911	.978	.919	.817	.908	.908
		0.99	.989	.941	.916	.918	.918	.984	.936	.921	.922	.922
30	30	0.80	.788	.764	.741	.792	.792	.811	.784	.752	.814	.814
		0.90	.900	.870	.824	.896	.896	.903	.879	.842	.901	.901
		0.95	.968	.922	.906	.909	.909	.972	.929	.914	.917	.917
		0.99	.999	.974	.962	.962	.962	1.000	.977	.971	.973	.973
50	50	0.80	.799	.762	.710	.806	.806	.802	.784	.750	.810	.810
		0.90	.881	.867	.866	.886	.884	.890	.873	.871	.891	.890
		0.95	.945	.930	.894	.938	.938	.958	.945	.904	.950	.950
		0.99	.997	.978	.972	.984	.984	.996	.982	.976	.987	.987
100	100	0.80	.796	.782	.777	.852	.852	.788	.771	.767	.828	.828
		0.90	.894	.876	.867	.910	.907	.892	.865	.850	.910	.909
		0.95	.946	.938	.929	.957	.957	.943	.932	.927	.951	.950
		0.99	.987	.979	.975	.992	.992	.990	.983	.979	.991	.991
10	100	0.80	.812	.803	.675	.746	.729	.786	.721	.589	.708	.689
		0.90	.855	.837	.765	.800	.800	.823	.783	.710	.755	.755
		0.95	.895	.929	.805	.818	.817	.890	.840	.755	.776	.775
		0.99	.958	.977	.830	.834	.834	.974	.966	.786	.790	.790
20	100	0.80	.801	.763	.700	.771	.763	.813	.710	.646	.764	.760
		0.90	.911	.869	.807	.851	.848	.924	.836	.774	.838	.835
		0.95	.959	.919	.871	.893	.893	.965	.913	.842	.891	.891
		0.99	.989	.975	.927	.936	.936	.990	.971	.925	.944	.943

Table 2.2. Average coverage probabilities for the simultaneous confidence regions(BON: Bonferroni method, SID: Šidák method, BNP: bootstrap method using non-pooled sample, Bootstrap replication $B = 2000$, simulation replication $M = 5000, k = 5$)

$n_1 = n_2$	$1-\alpha$	$\rho = 0.00$			$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.80$		
		BON	SID	BNP	BON	SID	BNP	BON	SID	BNP	BON	SID	BNP
10	0.80	.7278	.7278	.8426	.7168	.7168	.8758	.7664	.7664	.8334	.8424	.8424	.7886
	0.90	.8640	.8640	.9230	.8542	.8542	.9500	.8590	.8590	.9258	.9054	.9054	.9134
	0.95	.8980	.8980	.9706	.8864	.8864	.9836	.8838	.8838	.9782	.9204	.9204	.9688
	0.99	.9742	.9742	.9994	.9520	.9520	.9998	.9526	.9526	.9998	.9664	.9664	.9998
30	0.80	.7968	.7968	.8136	.7876	.7876	.8084	.8160	.8160	.8050	.8740	.8740	.7992
	0.90	.8924	.8924	.9080	.8894	.8894	.9058	.9018	.9018	.9048	.9266	.9266	.8938
	0.95	.9500	.9460	.9594	.9477	.9475	.9580	.9506	.9496	.9504	.9695	.9590	.9494
	0.99	.9872	.9872	.9920	.9830	.9830	.9936	.9841	.9840	.9912	.9870	.9870	.9908
50	0.80	.8176	.7868	.8014	.8044	.7844	.7960	.8272	.8122	.7980	.8848	.8710	.7946
	0.90	.9030	.8970	.8996	.8972	.8964	.9028	.9106	.9104	.8964	.9362	.9362	.9014
	0.95	.9434	.9434	.9504	.9412	.9412	.9492	.9507	.9504	.9520	.9666	.9666	.9498
	0.99	.9886	.9886	.9904	.9854	.9854	.9902	.9904	.9904	.9902	.9926	.9926	.9928
100	0.80	.8162	.8092	.8088	.8156	.8096	.8018	.8488	.8418	.8064	.9004	.8948	.8038
	0.90	.9048	.9042	.9096	.9024	.9016	.8968	.9174	.9166	.8994	.9450	.9442	.9042
	0.95	.9502	.9502	.9588	.9505	.9502	.9520	.9512	.9512	.9488	.9686	.9686	.9522
	0.99	.9904	.9904	.9926	.9888	.9888	.9914	.9903	.9903	.9910	.9938	.9938	.9922

distribution function $F_\omega(\cdot)$, and let $H_n(\omega, t) = P_\omega[T_n \leq t]$. If $H_n(\omega, t)$ is the coverage probability of a confidence set at ω, then $\int_\Omega H_n(\omega, t)\xi(\omega)d\omega$ can be regarded as the long run relative frequency of coverage in many independent replications of the experiment, where ω is drawn from the density ξ; therefore, $\int_\Omega H_n(\omega, t)\xi(\omega)d\omega$ can be called the *average coverage probability* at ξ. In the non-lattice case, Singh (1981) showed that the bootstrap estimator of the sampling distribution of T_n differs from the actual distribution by an order of magnitude smaller than $1/\sqrt{n}$ with probability one as $n \to \infty$. In the lattice case, Woodroofe and Jhun (1988) showed that in terms of average coverage probability, the bootstrap estimator differs from very weak expansions by a term of order $1/\sqrt{n}$. However, it was also shown that the coefficient of the term is very small for any ξ with compact support.

In summary, for a fixed vector of the proportions of a multivariate binomial distribution, the actual coverage probability of a simultaneous confidence region is the probability that the simultaneous confidence region contains that vector. However, one can probably interpret simultaneous confidence coefficients in terms of average performance. In this case, the bootstrap method has an advantage over the normal approximation method. By using the average coverage probability, the performance of the bootstrap simultaneous confidence regions are compared with that of its competitors. We obtained results $\int_\Omega H_n(\omega, t)\xi(\omega)d\omega$ for a uniform distribution defined by $(p_{i1}, \ldots, p_{i5})$ for each group i and four different dependence structures defined by D: (1) Independent structures, (2) D with relatively weak equal correlations (ρ=0.25), (3) D with equal correlations (ρ= 0.50), (4) D with relatively strong equal correlations (ρ=0.80). Though this evaluation may suggest a Bayesian approach to inference, we restrict our attention in this paper to comparing the four correlation structures for the three methods described previously.

Table 2.2 shows the average coverage probabilities for the uniform averages of the parameter values at various correlation structures, for nominal 80%, 90%, 95% and 99%, Bonferroni, Šidák and bootstrap simultaneous confidence levels. This was repeated 5,000 times independently in order to get an estimate of the average coverage probability. Some conclusions can be drawn from the simulations. First, when the sample size is small ($n_1(n_2) = 10$), the Bonferroni and Šidák simultaneous intervals have lower estimated coverage probabilities than the nominal ones. This contradicts the

Table 3.1. Simultaneous confidence intervals and p-values, with multiplicity adjustments for six tests using the Brown and Fears data(BON: Bonferroni methods, SID: Šidák methods, BPO: bootstrap methods using pooled sample, BNP: bootstrap methods using non-pooled sample)

$1-\alpha$	Method	Liver	Lung	Lymph	Cardio.	Pituitary	Ovary
.80	BON	(−.0304 .2501)	(−.1545 .0917)	(−.1954 .0659)	(−.0137 .1329)	(−.0491 .0663)	(−.1137 .0243)
	SID	(−.0293 .2489)	(−.1534 .0907)	(−.1943 .0649)	(−.0131 .1323)	(−.0487 .0658)	(−.1131 .0238)
	BPO	(−.0296 .2492)	(−.1537 .0909)	(−.1946 .0651)	(−.0133 .1324)	(−.0488 .0659)	(−.1132 .0239)
	BNP	(−.0294 .2490)	(−.1535 .0908)	(−.1944 .0650)	(−.0132 .1323)	(−.0487 .0658)	(−.1131 .0238)
.95	BON	(−.0641 .2837)	(−.1840 .1212)	(−.2267 .0973)	(−.0313 .1505)	(−.0630 .0801)	(−.1302 .0409)
	SID	(−.0638 .2835)	(−.1838 .1210)	(−.2265 .0971)	(−.0312 .1503)	(−.0629 .0800)	(−.1301 .0408)
	BPO	(−.0630 .2827)	(−.1830 .1203)	(−.2258 .0963)	(−.0308 .1499)	(−.0625 .0797)	(−.1297 .0404)
	BNP	(−.0657 .2853)	(−.1854 .1226)	(−.2283 .0988)	(−.0322 .1513)	(−.0636 .0808)	(−.1310 .0417)
Raw p-value		.0990	.5871	.2920	.0899	.7530	.1661
Adjusted p-value							
Bonferroni		.5938	1.0000	1.0000	.5392	1.0000	.9965
Šidák		.4649	.9951	.8741	.4317	.9998	.6637
Westfall(BPO)		.4586	.9953	.8833	.4377	.9999	.6673
Westfall(BNP)		.4726	.9954	.8788	.4357	.9997	.6708

properties of these methods. The low simultaneous coverage may be due in part to the marginal confidence intervals, which can exhibit severe underestimation when the sample sizes are small. Similar phenomena can be found in Beal (1987)'s Table 1 and 7. When sample sizes are large, however, these intervals have greater estimated average coverages than the nominal ones. Increasing the sample size does not help the overestimation problem. Throughout Table 2.2, there is a tiny difference between the Bonferroni and Šidák methods (except for $1-\alpha=0.8$). This is because the Bonferroni marginal confidence levels of 96%, 98%, 99% and 99.8% are the almost the same as the Šidák marginal confidence levels 95.63%, 97.91%, 98.97% and 99.79% when one constructs the simultaneous confidence levels of 80%, 90%, 95% and 99% with $k=5$. Secondly, note that the Bonferroni and Šidák simultaneous confidence levels are severely affected by the correlation structure. For example, when $n_1\,(n_2)=100$ and $1-\alpha=0.8$, the Bonferroni confidence levels increase from 0.8 ($\rho=0.2$) to 0.8488 ($\rho=0.5$) and to 0.9004 ($\rho=0.8$). By contrast, the bootstrap simultaneous confidence levels are surprisingly close to the nominal confidence levels even for the strong concordance correlation. Thirdly, the bootstrap method is more accurate in terms of average coverage probability. In general, when $\rho=0$ (perfect independent structure), there may be no difference between the methods used for simultaneous confidence intervals. The convergence rate of the average coverage probability to the nominal coverage as sample sizes increase is faster under the bootstrap method than using the normal approximation or the Bonferroni and Šidák methods. Note that the bootstrap intervals exhibit stability when $n_1(n_2)\leq 30$. Overall, the bootstrap confidence regions tend to outperform the classical ones in terms of having average coverage probabilities close to nominal confidence levels.

3. Examples

3.1. Weak correlation structure data

We construct simultaneous confidence intervals for the Brown and Fears (1981) data with a weak correlation structure. This example data was also examined by Westfall and Young (1989) for multiple testing (Jeong *et al.*, 2007). We are interested in the difference in the proportions of neoplastic lesions found in tissues between the high-dose group and the low-dose group. The low dose

Table 3.2. Simulated random sample from two different multivariate Bernoulli distribution(Group 1: equal correlation with 0.8, Group 2: equal correlation with 0.9)

Group 1					Group 2				
$X1$	$X2$	$X3$	$X4$	Frequency	$X1$	$X2$	$X3$	$X4$	Frequency
0	0	0	0	344	0	0	0	0	590
0	0	0	1	1	0	0	0	1	0
0	0	1	1	2	0	0	1	1	1
0	1	0	1	5	0	1	0	1	2
0	1	1	0	2	0	1	1	0	1
0	1	1	1	45	0	1	1	1	21
1	0	0	1	4	1	0	0	1	4
1	0	1	0	5	1	0	1	0	1
1	0	1	1	34	1	0	1	1	28
1	1	0	0	3	1	1	0	0	1
1	1	0	1	44	1	1	0	1	19
1	1	1	0	32	1	1	1	0	22
1	1	1	1	479	1	1	1	1	310

group(4ppm and 8ppm) contains $n_1 = 98$, and the high-dose group(16ppm and 50ppm) contains $n_2 = 93$. We assume that the two groups follow a multivariate binomial distribution.

The simultaneous confidence intervals of Bonferroni(BON), Šidák(SID), bootstrap using a pooled sample(BPO) and bootstrap a using non-pooled sample(BNP), raw p-values and adjusted p-values of Bonferroni, Šidák, Westfall and Young's bootstrap method using a pooled sample and the bootstrap method using a non-pooled sample are all given in Table 3.1. The confidence coverages $1 - \alpha = 0.8$ and 0.95 are considered.

The two smallest p-values for these data are cardiovascular, low-dose versus high-dose (p-value = 0.08987) and liver, low-dose versus high-dose (p-value = 0.09896). The adjusted p-values based on 10,000 simulated data sets under the complete null hypothesis show that they are no longer significant. Note that the adjusted p-values for the Bonferroni method are larger than those for other two methods. The adjusted p-values for the Šidák method are similar to those for the bootstrap method. In turn, the simultaneous confidence intervals give the more information than the adjusted p-values. All of the simultaneous confidence intervals contain zero. The results are the same for multiple testing. Note that the simultaneous confidence intervals for the Šidák method are similar or equal to those for the bootstrap method at the 80% confidence level and are similar to those for the Bonferroni methods at the 95% confidence level. For this example, at the 95% confidence level, the bootstrap simultaneous confidence intervals are wider than those of the other methods. These phenomena may be attributed in part to the fact that confidence intervals of Bonferroni and Šidák methods are short primarily because of underestimation stemming from the normal approximation. The similarity of the three methods may be due to the fact that the three sites (cardiovascular, pituitary and ovary) have the low marginal totals (nearly zero) and the fact that correlation structure of six sites is nearly independent. When the data are sparse, the simultaneous bootstrap confidence intervals may be too wide.

3.2. Strong correlation structure data

We construct the simultaneous confidence intervals for the simulated data with a strong correlation structure. The 1,000 random samples of the first group are generated from a multivariate binomial distribution with same proportion of 0.6 and the same correlation value of 0.8 and a second group

Table 3.3. Simultaneous confidence intervals using the simulated data

$1-\alpha$	Method	$X1$	$X2$	$X3$	$X4$
.95	Bonferroni	(.1615 .2705)	(.1797 .2883)	(.1605 .2695)	(.1746 .2834)
	Šidák	(.1616 .2705)	(.1798 .2882)	(.1605 .2695)	(.1747 .2833)
	Bootstrap(pool)	(.1662 .2659)	(.1843 .2837)	(.1652 .2649)	(.1793 .2787)
	Bootstrap(non-pool)	(.1662 .2658)	(.1844 .2836)	(.1652 .2648)	(.1793 .2787)
.99	Bonferroni	(.1500 .2820)	(.1683 .2997)	(.1490 .2810)	(.1632 .2948)
	Šidák	(.1500 .2820)	(.1683 .2997)	(.1490 .2810)	(.1632 .2948)
	Bootstrap(pool)	(.1524 .2796)	(.1707 .2974)	(.1515 .2787)	(.1656 .2925)
	Bootstrap(non-pool)	(.1537 .2783)	(.1719 .2961)	(.1527 .2773)	(.1669 .2912)
Raw p-value		.0000	.0000	.0000	.0000
Adjusted p-value					
Bonferroni-style		.0000	.0000	.0000	.0000
Šidák-style		.0000	.0000	.0000	.0000
Westfall-style (pool)		.0000	.0000	.0000	.0000
Westfall-style (non-pool)		.0000	.0000	.0000	.0000

with same proportion of 0.4 and the same correlation value of 0.9. The simulated data are given in Table 3.2. The simultaneous confidence intervals of the Bonferroni, Šidák and bootstrap methods using a pooled sample and a non-pooled sample are given in Table 3.3. The confidence coverages $1-\alpha = 0.95$ and 0.99 are considered. Bootstrap replication are 10,000 times. The estimates of the proportions with four variables are (.601 .385), (.610 .376), (.599 .384) and (.614 .385). The raw p-values are all zero. In this case, simultaneous confidence intervals are useful for comparing the difference of proportions.

Table 3.3 shows the advantage of the bootstrap method. The adjusted p-values are all computed to be zero for two groups with different proportions when the sample size is large, but the simultaneous confidence intervals are more useful than the adjusted p-values in that they provide more information. Note that the simultaneous confidence intervals for the bootstrap method using a non-pooled sample are shorter than those for the Bonferroni, Šidák and bootstrap methods using a pooled sample. In this example, the Bonferroni and Šidák simultaneous confidence intervals are again grossly conservative because there are strong correlations between the four variables.

4. Conclusion

For two groups with data from a multivariate Bernoulli distribution, techniques for constructing simultaneous confidence regions for differences of proportions have been presented. The adjusted p-values and simultaneous confidence intervals are standing on the same paradigm. The maximum pivot statistics have been employed to construct the simultaneous confidence regions for the entire collection of marginal confidence intervals of d_j, $j = 1, \ldots, k$. Especially for the simultaneous confidence intervals, the proposed bootstrap methods using non-pooled samples are more accurate, at least in terms of average coverage probability, and this result is reasonable since it considers the dependence structure of the multivariate Bernoulli distribution.

References

Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples, *Biometrics*, **43**, 941–950.

Brown, C. C. and Fears, T. R. (1981). Exact significance levels for multiple binomial testing with application to carcinogenicity screens, *Biometrics*, **37**, 763–774.

Freedman, D. A. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models, *The Annals of Statistics*, **12**, 827–842.

Holland, B. S. and Copenhaver, M. D. (1987). An improved sequentially rejective bonferroni test procedure, *Biometrics*, **43**, 417–424.

Jeong, H. C., Jhun, M. and Lee, J. W. (2007). Estimating the simultaneous confidence levels for the difference of proportions from multivariate binomial distributions, *Journal of the Korean Statistical Society*, **36**, 397–410.

Jhun, M. and Jeong, H. C. (2000). Applications of bootstrap methods for categorical data analysis, *Computational Statistics & Data Analysis*, **35**, 83–91.

Jhun, M., Jeong, H. C. and Bahng, J. S. (2007). Simultaneous confidence intervals for the mean of multivariate Poisson distribution: A comparison, *Communications in Statistics-Simulation and Computation*, **36**, 151–164.

Park, C. G., Park, T. P. and Shin, D. W. (1996). A simple method for generating correlated Binary varites, *American Statistician*, **50**, 306–310.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *The Annals of Statistics*, **9**, 1187–1195.

Thombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression, *Journal of the American Statistical Association*, **85**, 486–492.

Westfall, P. H. and Young, S. S. (1989). P-value adjustments for multiple tests in multivariate binomial models, *Journal of the American Statistical Association*, **84**, 780–786.

Westfall P. H. (1985). Simultaneous small-sample multivariate bernoulli confidence intervals, *Biometrics*, **41**, 1001–1013.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, New York.

Woodroofe, M. and Jhun, M. (1988). Singh's theorem in the lattice case, *Statistics & Probability Letters*, **7**, 201–205.