Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Hypotheses Testing from Categorical Survey Data Using Bootstrap Weights

Jae-kwang Kim and J.N.K. Rao[1]

Abstract

Standard statistical methods that do not take proper account of the complexity of survey design can lead to erroneous inferences when applied to survey data. In particular, the actual type I error rates of tests of hypotheses based on standard tests can be much bigger than the nominal level. Methods that take account of survey design features in testing hypotheses have been proposed, including Wald tests and quasi-score tests (Rao, Scott and Skinner 1998) that involve the estimated covariance matrices of parameter estimates. The bootstrap method of Rao and Wu (1983) is often applied at Statistics Canada to estimate the covariance matrices, using the data file containing columns of bootstrap weights. Standard statistical packages often permit the use of survey weighted test statistics and it is attractive to approximate their distributions under the null hypothesis by their bootstrap analogues computed from the bootstrap weights supplied in the data file. Beaumont and Bocci (2009) applied this bootstrap method to testing hypotheses on regression parameters under a linear regression model, using weighted F statistics. In this paper, we present a unified approach to the above method by constructing bootstrap approximations to weighted likelihood ratio statistics and weighted quasi-score statistics. We report the results of a simulation study on testing independence in a two way table of categorical survey data. We studied the relative performance of the proposed method to alternative methods, including Rao-Scott corrected chi-squared statistic for categorical survey data.

Key Words: Bootstrap weight; Test of independence; Weighted likelihood ratio; Weighted quasi-score.

1. Introduction

Testing statistical hypothesis is one of the fundamental problem of statistics. In the parametric model approach, testing statistical hypothesis can be implemented using Wald test, likelihood ratio test, or score test. In each test, a test statistic is computed and then is compared with the $100\alpha\%$-quantile of the reference distribution which is the limiting distribution of the test statistic under the null hypothesis. The limiting distribution is often a chi-squared distribution due to the central limit theorem of the point estimators.

In survey sampling, however, the samples are selected with complex sampling methods involving clustering, stratification, or unequal probability selection. If the design features are ignored in the statistical analysis, the standard errors are usually underestimated. As a result, the associated coverage rates are underestimated and the test level are inflated. In fact, the limiting distribution of the test statistic is not necessarily a chi-squared distribution. Rather, it can be expressed as a weighted sum of p independent random variables from $\chi^2(1)$ distribution and the weights depend on unknown parameters which depend on the sampling design. To handle such problem, one may consider some correction of the test statistics to obtain a chi-square limiting distribution. Such an approach usually involves computing the design effect (Rao and Scott, 1984) to the test statistics. Rao, Scott and Skinner (1998) used this approach to obtain quasi-score tests in survey data.

In this paper, we use a different approach of computing the limiting distribution using parametric bootstrap. Use of bootstrap to compute the limiting distribution of test statistics under complex sampling has been discussed by Beaumont and Bocci (2009), and did not discuss extensions to likelihood ratio test not to score test. We present a unified approach of using the bootstrap method to obtain the limiting distribution of test statistics under complex sampling. The sampling design is allowed to be informative. The proposed method is presented in the context of the simple goodness-of-fit and testing independence in a two-way table for categorical survey data.

[1] Jae-kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa 50014, U.S.A ; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

In Section 2, the basic setup is introduced in the context of simple goodness-of-fit test. In section 3, the proposed method is introduced and its asymptotic properties are discussed. In section 4, the proposed method is applied to test independence in a two-way table of cell counts of proportions. Results from a limited simulation study are presented in Section 5. Concluding remarks are made in Section 6.

2. Basic setup

Suppose that a finite population U of size N is partitioned into K categories with $U = U_1 \cup \cdots \cup U_K$ being such a partition. Let $p_k = N_k / N$ be the population proportion of category k, where $N_k = |U_k|$. From the finite population U, a probability sample s of size n is selected, and w_i is the sampling weight associated with unit $i \in s$. Let $s_k = s \cap U_k$ be the sample partition of $s = s_1 \cup \cdots \cup s_K$. From the sample, we compute $\hat{p}_k = \hat{N}_k / \hat{N}$ as an estimator of p_k, where $\hat{N}_k = \sum_{j \in s_k} w_j$ is design-unbiased estimator of N_k and $\hat{N} = \sum_{k=1}^{K} \hat{N}_k = \sum_{i \in s} w_i$.

From the sample data $(\hat{p}_k, k = 1, \cdots, K)$, suppose that we are interested in testing $H_0 : p_k = p_{k,0}, k = 1, \cdots, K$, for specified $(p_{1,0}, \cdots, p_{K,0})$ satisfying $\sum_{k=1}^{K} p_{k,0} = 1$. The Pearson Chi-squared goodness-of-fit test statistic for this hypothesis is computed by

$$X^2 = n \sum_{k=1}^{K} \left(\hat{p}_k - p_{k,0} \right)^2 / p_{k,0}.$$

Also, we can compute the LRT (likelihood-ratio test) statistic (assuming multinomial distribution) as

$$G^2 = 2n \sum_{k=1}^{K} \hat{p}_k \log\left(\frac{\hat{p}_k}{p_{k,0}} \right).$$

Writing $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_{K-1})'$ and $\mathbf{p}_0 = (p_{1,0}, \cdots, p_{K-1,0})'$, we have, under H_0,

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}_0) \xrightarrow{\mathcal{L}} N(0, V),$$

for sufficiently large n, where $V = nV(\hat{\mathbf{p}})$. Under simple random sampling (SRS) with replacement, V is equal to $P_0 = diag(\mathbf{p}_0) - \mathbf{p}_0 \mathbf{p}_0'$. For other sampling design, V is more complicated. Under some regularity conditions, according to Rao and Scott (1981),

$$X^2, G^2 \xrightarrow{\mathcal{L}} \sum_{i=1}^{K-1} \lambda_i Z_i^2, \tag{1}$$

under H_0, where $\lambda_1 \le \cdots \le \lambda_{K-1}$ are the eigenvalues of the design effect matrix $\mathbf{D} = P_0^{-1} V$, $Z_1, \cdots, Z_{K-1} \overset{iid}{\sim} N(0,1)$, and $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution. Under SRS, since $V = P_0$, the limiting distribution in (1) reduces to a χ^2 distribution with $K-1$ degrees of freedom.

In a two-stage cluster sampling design with $\lambda_i = \lambda (> 1)$, type 1 error rate is approximately equal to $Pr\{X_{k-1}^2 > \lambda^{-1} \chi_{k-1}^2(\alpha)\}$ which increases with λ and thus can be made arbitrarily large by increasing λ. To overcome this problem, Rao and Scott (1981) proposed a first-order correction which treats $X^2(\hat{\lambda}_+) = X^2 / \hat{\lambda}_+$ as χ_{k-1}^2 under H_0, where

$$\hat{\lambda}_+ = \frac{1}{k-1} \sum_{i=1}^{k} \frac{\hat{p}_i}{p_{0i}} (1 - \hat{p}_i) \hat{d}_i$$

and $\hat{d}_i$ = estimated design effect of $\hat{p}_i$. The second-order Rao-Scott correction (Rao and Scott, 1981) requires the knowledge of the full estimated covariance matrix of the estimated proportions, but inversion of the covariance matrix

is not involved unlike in the case of Wald statistics. STATA and other survey software use Rao-Scott corrections as default option.

3. Proposed bootstrap method

We now propose a new method based on the bootstrap procedure in survey sampling. Bootstrap method in survey sampling has been mainly discussed in the context of replication variance estimation (Rao and Wu, 1988; Rao, Wu, and Yue 1992). In the bootstrap, we use data file consisting of response variables, final survey weights w_i and final bootstrap replication weights $w_i^*(b), b = 1, \cdots, B$. Typically, $B = 500$ columns of bootstrap weights are reported.

In the proposed bootstrap testing procedure, we use the bootstrap sample to approximate the limiting distribution in (1), without having to compute the design effect matrix. To describe the proposed method, let $\hat{\mathbf{p}}^*$ be the estimator of $\mathbf{p}$ based on the bootstrap weights w_i^* and $\hat{V}$ be the design-consistent estimator of V. For stratified multistage sampling, we assume that the PSUs within strata are drawn with replacement or the PSU sampling fraction is negligible. The proposed bootstrap statistics for goodness-of-fit statistics X^2 and G^2 are

$$X^{2*} = n\sum_{i=1}^{K}(\hat{p}_i^* - \hat{p}_i)^2 / \hat{p}_i$$
$$G^{2*} = 2n\sum_{i}\hat{p}_i^* \log\left(\hat{p}_i^* / \hat{p}_i\right),$$

respectively. Note that we do not use p_{0i} in place of $\hat{p}_i$ in the bootstrap test statistics.

The following theorem present the asymptotic properties of the proposed bootstrap test statistics.
Theorem 1 *Under* H_0,

$$X^{2*}, G^{2*} \xrightarrow{\mathcal{L}^*} \sum_{i=1}^{K-1}\hat{\lambda}_i Z_i^2 \tag{2}$$

where $\hat{\lambda}_1 \le \cdots \le \hat{\lambda}_{K-1}$ *are the eigenvalues of estimated design effect matrix* $\hat{P}^{-1}\hat{V}$, $Z_1, \cdots, Z_{K-1} \overset{iid}{\sim} N(0,1)$, *and* $\xrightarrow{\mathcal{L}^*}$ *denotes convergence in bootstrap distribution*.
A sketched proof of Theorem 1 is presented in Appendix A.

Note that the limiting distribution in (2) is asymptotically the same as the limiting distribution in (1). Thus, we can use the bootstrap samples to approximate the sampling distribution of the test statistics. That is, from the histogram of B bootstrap statistics $X^{*2}(1), \cdots, X^{*2}(B)$, find the upper α value and reject H_0 if the observed X^2 exceeds that value. Similarly, likelihood ratio statistic G^2 can be used by computing the corresponding bootstrap statistics $G^{*2}(1), \cdots, G^{*2}(B)$.

4. Test of independence

We now discuss test of independence in two-way tables. Let $p_{ij} = N_{ij} / N$ be the population proportion for cell (i, j) with margins p_{i+} and p_{+j}, where $\{N_{ij}; i = 1, \cdots, R, j = 1, \cdots, C\}$ is the set of population counts with margins N_{i+} and N_{+j}. Let $\hat{N}_{ij}$ be a design unbiased estimator of N_{ij} and $\hat{p}_{ij} = \hat{N}_{ij} / \hat{N}$. The X^2 and G^2 test statistics for testing $H_0 : p_{ij} = p_{i+}p_{+j}$ for all i and j are given by

$$X_I^2 = n\sum_i\sum_j \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} \tag{3}$$

$$G_I^2 = 2n\sum_i\sum_j \hat{p}_{ij}\log\left\{\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right\}.$$

Rao and Scott (1981) have shown that, under H_0, writing $d = (R-1)(C-1)$,

$$X_I^2, G_I^2 \xrightarrow{\mathcal{L}} \sum_{l=1}^{d}\delta_l Z_l^2 \tag{4}$$

where $\delta_1 \le \cdots \le \delta_d$ are the d eigenvalues of a design effect matrix (see Appendix B) and $Z_1, \cdots, Z_d \overset{iid}{\sim} N(0,1)$.

The Rao-Scott first order correction to X_I^2 can be written as $X_I^2(\hat{\delta}_+) = X_I^2 / \hat{\delta}_+$ treated as χ^2 with $(R-1)(C-1)$ degrees of freedom under H_0, where $(R-1)(C-1)\hat{\delta}_+ = \sum_l \hat{\delta}_l$ requires only cell deffs and row and column marginal deffs (Rao and Scott, 1984). Two way tables should report those deffs in addition to estimated cell counts or proportions and their marginals. Rao and Scott (1984) provided unified theory for log linear models to cover multi-way tables and other extensions.

We now consider bootstrap tests of H_0 in this case. Let $\hat{p}_{ij}^*$ be the bootstrap cell proportions computed using the bootstrap weights. Define $\hat{p}_{i+}^* = \sum_j \hat{p}_{ij}^*$ and $\hat{p}_{+j}^* = \sum_i \hat{p}_{ij}^*$. The proposed bootstrap version of $X_I^2 = n\sum_i\sum_j (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2 / (\hat{p}_{i+}\hat{p}_{+j})$ is given by

$$X_I^{2*} = n\sum_i\sum_j \frac{\{(\hat{p}_{ij}^* - \hat{p}_{i+}^*\hat{p}_{+j}^*) - (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})\}^2}{(\hat{p}_{i+}\hat{p}_{+j})}$$

Note that, under H_0, terms in the numerator of sample X^2 are identical to $\{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}) - (p_{ij} - p_{i+}p_{+j})\}^2$. That is, the bootstrap test statistic is computed by simply replacing $\{\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{+j}\}$ and $\{p_{ij}, p_{i+}, p_{+j}\}$ by $\{\hat{p}_{ij}^*, \hat{p}_{i+}^*, \hat{p}_{+j}^*\}$ and $\{\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{+j}\}$, respectively.

Let $\Delta_{ij} = p_{ij} / (p_{i+}p_{+j})$, then under H_0 $\Delta_{ij} = 1$ and we can express

$$G_I^2 = 2n\sum_i\sum_j \left[\hat{p}_{ij}\log\left\{\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}\Delta_{ij}}\right\} - (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}\Delta_{ij})\right].$$

Bootstrap version G^{2*} is now obtained by replacing $\{\hat{p}_{ij}, \hat{p}_{i+}, \hat{p}_{+j}\}$ by $\{\hat{p}_{ij}^*, \hat{p}_{i+}^*, \hat{p}_{+j}^*\}$ and Δ_{ij} by $\hat{\Delta}_{ij}$. That is, the proposed bootstrap version of G_I^2 is given by

$$G_I^{2*} = 2n\sum_i\sum_j \left[\hat{p}_{ij}^*\log\left\{\frac{\hat{p}_{ij}^*}{\hat{p}_{i+}^*\hat{p}_{+j}^*\hat{\Delta}_{ij}}\right\} - (\hat{p}_{ij}^* - \hat{p}_{i+}^*\hat{p}_{+j}^*\hat{\Delta}_{ij})\right]$$

where $\hat{\Delta}_{ij} = \hat{p}_{ij} / (\hat{p}_{i+}\hat{p}_{+j})$. Note that G_I^{2*} is always nonnegative.

The following theorem presents some asymptotic properties of the proposed bootstrap test statistics.

Theorem 2 *Under* H_0,

$$X_I^{2*}, G_I^{2*} \xrightarrow{\mathcal{L}^*} \sum_{i=1}^{d} \hat{\delta}_i Z_i^2 \tag{5}$$

where $\hat{\delta}_1 \le \cdots \le \hat{\delta}_d$ *are the eigenvalues of estimated design effect matrix which converges in probability to the design effect matrix corresponding to the eigenvalues* $\delta_i, i = 1, \cdots, (R-1)(C-1)$.

A sketched proof of Theorem 2 is presented in Appendix B.

By Theorem 2, we can use the bootstrap distribution to approximate the sampling distribution of the test statistic X_I^2 or G_I^2. That is, from the histogram of B bootstrap statistics $X_I^{*2}(1), \cdots, X_I^{*2}(B)$, find the upper α value and reject H_0 if the observed X_I^2 exceeds this value. Similarly, likelihood ratio statistic G_I^2 can be used by computing the corresponding bootstrap statistics $G_I^{*2}(1), \cdots, G_I^{*2}(B)$.

5. Simulation Study

We conducted a limited simulation study of test of independence to check the performance of the bootstrap method under cluster sampling. In the simulation study, we generated $n = 50$ clusters each of size $m = 20$. We considered $R = 3$ rows and $C = 3$ columns. Given p_{ij}, the samples are drawn in two steps:

1. For each cluster i, given $\mathbf{p} = (p_{11}, \cdots, p_{33})'$, generate $\mathbf{p}_i$ from a Dirichlet distribution with parameter $C\mathbf{p}$, where C is to be determined by a common design effect δ.
2. Using $\mathbf{p}_i$ generated from Step 1, the cell counts for cluster i are generated from a multinomial distribution with sample size m and probability $\mathbf{p}_i$.

Under this procedure, we have $\delta = 1 + (m-1)/(C+1)$. We set $\delta = 1, 2$ and 3.

For the parameter $\mathbf{p}$, we considered four scenarios:

1. Case 1: $p_{11} = 1/4, p_{12} = p_{13} = p_{21} = p_{31} = 1/8, p_{22} = p_{23} = p_{32} = p_{33} = 1/16$.
2. Case 2: $p_{11} = 1/4, p_{12} = p_{13} = (1.2)/8$, $p_{21} = p_{31} = (0.8)/8$, $p_{22} = p_{33} = 1/16$, $p_{23} = (1.2)/16, p_{32} = (0.8)/16$.
3. Case 3: $p_{11} = 1/4, p_{12} = p_{13} = (1.4)/8$, $p_{21} = p_{31} = (0.6)/8$, $p_{22} = p_{33} = 1/16$, $p_{23} = (1.4)/16, p_{32} = (0.6)/16$.
4. Case 4: $p_{11} = 1/4, p_{12} = p_{13} = (1.5)/8$, $p_{21} = p_{31} = (0.5)/8$, $p_{22} = p_{33} = 1/16$, $p_{23} = (1.5)/16, p_{32} = (0.5)/16$.

In Case 1, the two way table satisfies independence. Cases 2-4 do not satisfy independence and the level of non-independence can be expressed using a noncentrality parameter γ, given by

$$\gamma = mn \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}}.$$

The values of γ are $0.0, 2.6, 11.7, 19.9$ for Case 1, Case 2, Case 3, Case 4, respectively. From each sample, we considered 4 test procedures:

1. Naive Pearson: Reject H_0 if $X_I^2 > \chi_4^2(0.95)$, where $\chi_4^2(0.95)$ is the upper 5% point of χ^2

distribution with 4 degrees of freedom.

2. Naive LR: Reject H_0 if $G_I^2 > \chi_4^2(0.95)$.
3. Bootstrap Pearson: Reject H_0 if $X_I^2 > q_1^*(0.95)$, where $q_1^*(0.95)$ is the 0.95-th quantile of the bootstrap distribution of X_I^{2*}.
4. Bootstrap LR: Reject H_0 if $G_I^2 > q_2^*(0.95)$, where $q_2^*(0.95)$ is the 0.95-th quantile of the bootstrap distribution of G_I^{2*}.

In computing the bootstrap quantiles, we used $B = 5{,}000$ bootstrap samples. Simulation results are presented in Table 1.

Table 1 reports the size $(\gamma = 0)$ and the power of the four test procedures, using $R = 1{,}000$ Monte Carlo simulated samples. Under H_0, the size of the four tests are similar and close to the nominal size, $\alpha = 0.05$ level when the design effect is not present $(\delta = 1)$. On the other hand, for $\delta = 2$ and 3, the naive X^2 and G^2 lead to inflated sizes which increase with δ, as expected. The proposed bootstrap tests show valid test size under H_0 even when the design effect is greater than 1. The test power of the proposed bootstrap tests increases with γ.

Table 1: Power of the test procedures for independence based on 1,000 Monte Carlo simulation samples

Design Effect	Method	Case 1 ($\gamma = 0$)	Case 2 ($\gamma = 2.6$)	Case 3 ($\gamma = 11.7$)	Case 4 ($\gamma = 19.9$)
1	Naive Pearson	0.050	0.220	0.799	0.972
	Naive LR	0.051	0.227	0.804	0.971
	Bootstrap Pearson	0.063	0.219	0.800	0.970
	Bootstrap LR	0.060	0.216	0.801	0.967
2	Naive Pearson	0.304	0.449	0.818	0.951
	Naive LR	0.309	0.449	0.819	0.952
	Bootstrap Pearson	0.047	0.113	0.516	0.749
	Bootstrap LR	0.051	0.117	0.520	0.751
3	Naive Pearson	0.543	0.643	0.864	0.948
	Naive LR	0.546	0.646	0.864	0.948
	Bootstrap Pearson	0.062	0.094	0.295	0.536
	Bootstrap LR	0.070	0.105	0.304	0.540

6. Concluding Remarks

We plan to extend our bootstrap method to tests for multi-way tables of counts or proportions, using a loglinear model approach. We also plan to develop bootstrap tests for logistic regression and other models, using pseudo likelihood ratio and quasi-score approaches (Rao, Scott, and Skinner, 1998).

References

Beaumont, J.-F. and Bocci, C. (2009). A practical bootstrap method for testing hypothesis from survey data. *Survey Methodology*, **35**, 25–35.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-Squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, **76**, 221–230.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multistage contingency tables with cell proportions

estimated from survey data. *Annals of Statistics*, **12**, 46–60.

Rao, J. N. K., Scott, A. J., and Skinner, C. J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, **8**, 1059–1070.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83**, 231–241.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209–217.

Appendix A: Proof of Theorem 1

The goodness-of-fit chi-squared statistic X^2 can be expressed as

$$X^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0). \tag{6}$$

Using the matrix representation (6) of X^2 and noting that $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}_0) \xrightarrow{\mathcal{L}} N(0, \mathbf{V})$ under H_0, the result (1) for X^2 follows by appealing to standard results on the distribution of quadratic forms (Rao and Scott, 1981). Result (1) for G^2 can be obtained by noting that X^2 and G^2 are asymptotically equivalent under H_0.

Turning to the bootstrap chi-squared statistics X^{2*}, we can express it as

$$X^{2*} = (\hat{\mathbf{p}}^* - \hat{\mathbf{p}})'\hat{\mathbf{P}}_0^{-1}(\hat{\mathbf{p}}^* - \hat{\mathbf{p}}) \tag{7}$$

where $\hat{\mathbf{P}}_0 = diag(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$. For stratified multistage sampling with PSUs drawn with replacement within strata, we have

$$\sqrt{n}(\hat{\mathbf{p}}^* - \hat{\mathbf{p}}) \xrightarrow{\mathcal{L}^*} N(0, \hat{V}) \tag{8}$$

under the Rao-Wu (1988) bootstrap method. It now follows from (7) and (8) that (2) holds. Results (2) for G^{2*} also holds, noting that X^{*2} and G^{*2} are asymptotically equivalent with respect to the bootstrap distribution.

Appendix B: Proof of Theorem 2

We present a brief justification of the proposed bootstrap method for testing independence in a two-way table of cell proportions or counts. Using the notation of Rao and Scott (1981), let $\mathbf{h}(\mathbf{p})$ be the $d=(R-1)(C-1)$ dimensional vector with elements $h_{ij}(\mathbf{p})=p_{ij}-p_{i+}p_{+j}$, $i=1,\cdots,R-1; j=1,\cdots,C-1$, where $\mathbf{p}=(p_{11},p_{12},\cdots,p_{RC-1})'$. Then the chi-squared statistic X_I^2, under H_0, may be expressed in a matrix form as

$$X_I^2 = n\{\mathbf{h}(\hat{\mathbf{p}})-\mathbf{h}(\mathbf{p})\}'(\hat{\mathbf{P}}_{R+}^{-1}\otimes\hat{\mathbf{P}}_{+C}^{-1})\{\mathbf{h}(\hat{\mathbf{p}})-\mathbf{h}(\mathbf{p})\},$$

where $\hat{\mathbf{P}}_{R+}=diag(\hat{\mathbf{p}}_{R+})-\hat{\mathbf{p}}_{R+}\hat{\mathbf{p}}_{R+}'$ and $\hat{\mathbf{P}}_{+C}=diag(\hat{\mathbf{p}}_{+C})-\hat{\mathbf{p}}_{+C}\hat{\mathbf{p}}_{+C}'$ with $\hat{\mathbf{p}}_{R+}=(\hat{p}_{1+},\cdots,\hat{p}_{R-1,+})'$ and $\hat{\mathbf{p}}_{+C}=(\hat{p}_{+1},\cdots,\hat{p}_{+,C-1})'$ and $\otimes$ denotes direct product. Now, noting that $\sqrt{n}(\hat{\mathbf{p}}-\mathbf{p})\xrightarrow{\mathcal{L}}N(\mathbf{0},\mathbf{V})$, it follows that

$$\sqrt{n}\{\mathbf{h}(\hat{\mathbf{p}})-\mathbf{h}(\mathbf{p})\}\xrightarrow{\mathcal{L}}N(\mathbf{0},\mathbf{HVH}'),$$

where $\mathbf{H}=\partial\mathbf{h}(\mathbf{p})/\partial\mathbf{p}'$ is the $d\times(RC-1)$ matrix of partial derivatives of $\mathbf{h}(\mathbf{p})$. Using the above result, we get (4) where the δ_l $(l=1,\cdots,d)$ are the eigenvalues of the design effect matrix $\mathbf{D}_h=\left(\mathbf{P}_{R+}^{-1}\otimes\mathbf{P}_{+C}^{-1}\right)(\mathbf{HVH}')$.

Turning to the proposed bootstrap method, we can express the bootstrap version of X_I^2 in a matrix form as

$$X_I^{2*} = n\{\mathbf{h}(\hat{\mathbf{p}}^*)-\mathbf{h}(\hat{\mathbf{p}})\}'(\hat{\mathbf{P}}_{R+}^{*-1}\otimes\hat{\mathbf{P}}_{+C}^{*-1})\{\mathbf{h}(\hat{\mathbf{p}}^*)-\mathbf{h}(\hat{\mathbf{p}})\}.$$

Now, noting that

$$\sqrt{n}\left\{\mathbf{h}(\hat{\mathbf{p}}^*)-\mathbf{h}(\hat{\mathbf{p}})\right\}\xrightarrow{\mathcal{L}^*}N(\mathbf{0},\hat{\mathbf{H}}\hat{\mathbf{V}}\hat{\mathbf{H}}'),$$

the representation (5) of X_I^{2*} holds, where the $\hat{\delta}_l$ are eigenvalues of the estimated design effect matrix $\hat{\mathbf{D}}_h=(\hat{\mathbf{P}}_{R+}^{-1}\otimes\hat{\mathbf{P}}_{+C}^{-1})(\hat{\mathbf{H}}\hat{\mathbf{V}}\hat{\mathbf{H}}')$. Now, using $\hat{\mathbf{D}}_h\rightarrow_p\mathbf{D}_h$, it follows that $\hat{\delta}_l\rightarrow_p\delta_l, l=1,\cdots,d$ and the limiting distribution of X_I^{*2} is the same as the limiting distribution in (4).