SIMULTANEOUS CONFIDENCE INTERVALS: A COMPARISON UNDER COMPLEX SAMPLING

Richard W. Andrews and William C. Birdsall, University of Michigan
Richard W. Andrews, Michigan Business School, Ann Arbor, MI 48109-1234

KEY WORDS: Bayesian, simulation, earnings

1. Introduction

Procedures have been developed to validate the output of microeconomic simulation models by comparing such output to survey data. The properties of the validation procedures are not well understood and are not easy to determine analytically, particularly when the survey data has been sampled with a complex design. This paper reports on a simulation experiment which compares the properties of three simultaneous confidence interval procedures. These procedures are compared within the context of simulation validation.

Consider two finite populations, designated I and II. Defined on each population there is a single categorical variable which has one of K values. Let p_k be the proportion of population I units in category k, $(k = 1, \ldots, K)$. Likewise, define q_k on population II.

The purpose of this paper is to report on a simulation study of three methods for developing simultaneous confidence intervals for $(p_k - q_k), (k = 1, \ldots, K)$. The sampling design for population I is simple random sampling (SRS), and for population II a stratified two stage (S2S) design is used. The choice of these two designs is consistent with the motivating application which is discussed in section 3.

Three simultaneous confidence interval (SCI) procedures are described in the next section and referred to as: (1) ordinary-χ^2 , (2) full design, and (3) Bayesian. The simulation experiment as described in section 3, uses a survey sample of individual earnings to generate the finite populations. The purpose of the application is to compare the generated output from a microeconomic simulation model with survey sample data.

The simulation results indicate that, when $p_k = q_k, (k = 1, \ldots, K)$, the full design method yields acceptable intervals. However, when $p_k \neq q_k$, the full design SCI's frequently cause one to falsely conclude that the two populations have the same values for p and q. The ordinary-χ^2 and Bayesian SCI's do not falsely conclude equivalence of the two populations as frequently. In addition, the interval lengths of the ordinary-χ^2 and Bayesian methods show better properties than the full design.

There is little evidence of difference between Bayesian and ordinary-χ^2 methods; however, where differences do occur the Bayesian procedure points toward the correct conclusion more often and its length is smaller with less variation.

The full design methodology has been recommended [1] as the procedure to use when full design information is available. Because of the high frequency with which the full design method falsely indicates that two populations have no difference, it is not recommended.

2. Confidence Interval Procedures

The three CIP's considered are (1) ordinary-χ^2 , (2) full design, and (3) a Bayesian posterior interval. The ordinary-χ^2 is derived under the assumption of simple random sampling. The full design method takes into account that population II is sampled using a S2S design. The Bayesian method ignores the design and states the posterior distribution of $\{p_k - q_k\}_{k=1}^K$, given the data. The presentation of these intervals follows.

Ordinary χ^2

From population I, a SRS of size n is selected. Define $\hat{p}_k$ = proportion of population I observations from category k; $\sum_{k=1}^K p_k = 1$. From population II, a S2S sample of total size n is selected. In like manner, define $\hat{q}_k$ = proportion of population II observations from category k; $\sum_{k=1}^K q_k = 1$. Using Scheffe type simultaneous confidence intervals, the $100(1-\alpha)\%$ simultaneous confidence intervals for $(p_k - q_k)$; $k = 1, \ldots, K$ are:

$$(\hat{p}_k - \hat{q}_k) \pm S_{kc}\sqrt{\chi^2_\alpha(K-1)}; \quad (k = 1, \ldots, K).$$

where,

$$S^2_{ko} = \frac{\hat{p}_k(1-\hat{p}_k) + \hat{q}_k(1-\hat{q}_k)}{n}; \quad (k = 1, \ldots, K).$$

and, $\chi^2_\alpha(K-1)$ = the $100(1-\alpha)$ percentile of the chi-squared distribution with $K-1$ degrees of freedom.

Full Design

Population II is stratified into H strata. Stratum h has M_h $(h = 1, \ldots, H)$ clusters; and the j^{th} cluster from the h^{th} stratum has L_{hj} individuals.

The full design Scheffe type SCI's for $(p_k - q_k)$ are:

$$(\hat{p}_k - \hat{q}_k) \pm S_{kf}\sqrt{\chi^2_\alpha(K-1)}; \quad (k = 1, \ldots, K).$$

where,

$$S^2_{kf} = \frac{\hat{p}_k(1-\hat{p}_k)}{n} + \hat{V}_q(kk); \quad (k = 1, \ldots, K).$$

and,

$$\hat{V}_q(kk) = \sum_{h=1}^{H} \frac{1-f_1}{m} s^2_{kh1} + \frac{f_1(1-f_2)}{ml} s^2_{kh2};$$

and,

$$f_1 = \frac{m}{M} \quad f_2 = \frac{l}{L}.$$

The between cluster variance is given by:

$$s^2_{kh1} = \frac{\sum_{i=1}^m (\hat{p}_{khi} - \hat{p}_{kh})^2}{(n-1)}.$$

The within cluster variance is given by:

$$s^2_{kh2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{l}(I_{khij} - \hat{p}_{khi})^2}{m(l-1)}.$$

The indicator variable, I_{khij} is one if the j^{th} observation from the i^{th} cluster of the h^{th} stratum belongs to the k^{th} category; and zero otherwise. The estimates of the proportions are as follows:

$$\hat{p}_{khi} = l^{-1}\sum_{j=1}^{l} I_{khij};$$

$$\hat{p}_{kh} = (ml)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{l} I_{khij}.$$

Bayesian

The Bayesian approach uses a multinomial distribution with a conjugate Dirichlet prior. For the sample of size n from population I define $\tilde{u} = (u_1, u_2, \ldots, u_K)'$; such that u_k = the number of observations in the k^{th} category ($k = 1, 2, \ldots, K$). Likewise, for the sample of size n from population II, define $\tilde{v} = (v_1, v_2, \ldots, v_K)'$.

Let $\tilde{p} = (p_1, p_2, \ldots, p_K)'$ and $\tilde{q} = (q_1, q_2, \ldots, q_K)'$. The conditional distribution of $\tilde{u}$ given $\tilde{p}$ is assumed multinomial with parameters $(\tilde{p}, n)$; that is, $\tilde{u} \mid \tilde{p} \sim MN(\tilde{p}, n)$. So,

$$f(\tilde{u} \mid \tilde{p}) = n!\prod_{i=1}^{K}(\frac{p_i^{u_i}}{u_i!}), \text{ for which,}$$

$$\sum_{i=1}^{K} u_i = n, \text{ and,}$$

$$\sum_{i=1}^{K} p_i = 1.$$

Also assume, $\tilde{v} \mid \tilde{q} \sim MN(\tilde{q}, n)$.

The conjugate prior distribution on $\tilde{p}$ is Dirichlet with parameters $\tilde{\alpha}$; $\tilde{p} \sim D(\tilde{\alpha})$. So,

$$f(\tilde{p}) = \Gamma(\alpha)\prod_{i=1}^{K}(\frac{p_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}), \text{ for which,}$$

$$\sum_{i=1}^{K} p_i = 1, \text{ and,}$$

$$\sum_{i=1}^{K} \alpha_i = \alpha.$$

Throughout this analysis we assume $\tilde{p} \sim D(\tilde{1})$ and independently, assume $\tilde{q} \sim D(\tilde{1})$; so, $\alpha_1 = \alpha_2 = \ldots = \alpha_K = 1$. This prior can be thought of as a uniform distribution over the simplex of the $K-1$ dimension space spanned by the vectors $\tilde{p}$ and $\tilde{q}$; under the constraint that $\sum_{i=1}^{K} p_i = 1$ and $\sum_{i=1}^{K} q_i = 1$.

By conjugate prior distribution theory we have that the posterior distribution of $\tilde{p}$ given $\tilde{u}$ is Dirichlet with parameter $\tilde{\alpha} + \tilde{u}$; where $\tilde{\alpha} + \tilde{u} = (\alpha_1 + u_1, \alpha_2 + u_2, \ldots, \alpha_K + u_K)$. So,

$$f(\tilde{p} \mid \tilde{u}) = \Gamma(\sum_{i=1}^{K}(\alpha_i + u_i))\prod_{i=1}^{K}(\frac{p_i^{\alpha_i + u_i - 1}}{\Gamma(\alpha_i + u_i)});$$

and,

$$f(\tilde{q} \mid \tilde{v}) = \Gamma(\sum_{i=1}^{K}(\alpha_i + v_i))\prod_{i=1}^{K}(\frac{q_i^{\alpha_i + v_i - 1}}{\Gamma(\alpha_i + v_i)}).$$

Define, $\tilde{\Delta} = \tilde{p} - \tilde{q} = (p_1 - q_1, p_2 - q_2, \ldots, p_{K-1} - q_{K-1})'$, the difference in the first $K-1$ components of $\tilde{p}$ and $\tilde{q}$. As an approximation to the posterior distribution of $\tilde{\Delta}$ given $\tilde{u}$ and $\tilde{v}$, we let $\tilde{\Delta} \mid \tilde{u}, \tilde{v}$ be multivariate $(K-1)$ normal. The mean vector is

$$\tilde{\mu} = E[\tilde{p} \mid \tilde{u}] - E[\tilde{q} \mid \tilde{v}];$$

where the expectation is taken on the first $K-1$ components of $\tilde{p}$ and $\tilde{q}$.

Assuming the same prior for $\tilde{p}$ and $\tilde{q}$,

$$\tilde{\mu} = \begin{pmatrix} \frac{u_1 - v_1}{\alpha + n} \\ \frac{u_2 - v_2}{\alpha + n} \\ \vdots \\ \frac{u_{K-1} - v_{K-1}}{\alpha + n} \end{pmatrix}$$

The variance–covariance of $\tilde{\Delta}$ is

$$\mathbf{V} = V(\tilde{p} \mid \tilde{u}) + V(\tilde{q} \mid \tilde{v}).$$

The diagonal terms are

$$V_{kk} = \frac{(\alpha_k + u_k)(\alpha + n - \alpha_k - u_k) + (\alpha_k + v_k)(\alpha + n - \alpha_k - v_k)}{(\alpha + n)^2(\alpha + n + 1)};$$

for $k = 1, 2, \ldots, K-1$.

The off diagonal terms are

$$V_{kl} = -\frac{(\alpha_k + u_k)(\alpha_l + u_l) + (\alpha_k + v_k)(\alpha_l + v_l)}{(\alpha + n)^2(\alpha + n + 1)};$$

for $k = 1, 2, \ldots, K-1; l \neq k$.

From Berger[2; p.143] the $100(1-\alpha)\%$ credible set for $\tilde{\Delta}$ will be

$$\{\tilde{\Delta} : (\tilde{\Delta} - \tilde{\mu})'\mathbf{V}^{-1}(\tilde{\Delta} - \tilde{\mu}) \leq \chi^2_\alpha(K-1)\}.$$

For each component of $\tilde{\Delta}$, $\Delta_k = p_k - q_k$, the SCI is given by the projection which is the interval

$$\mu_k \pm \sqrt{\frac{\chi^2_\alpha(K-1)}{A_{kk}}};$$

where A_{kk} is the k^{th} diagonal element of the inverse of $\mathbf{V}$.

3. Simulation Experiment

The purpose of the experiment was to investigate the statistical properties of simultaneous confidence interval procedures applied to finite populations. The motivating example for this experiment is the output validation of a microeconomic simulation model

(MSM). Specficially, we want to compare the earnings distribution from the MASSII [3] MSM with the earnings distribution from the Panel Study of Income Dynamics (PSID) [4].

We consider the output from the MSM as a simple random sample. The design of the PSID is stratified two stage. To make the conclusions from the simulation experiment as applicable as possible we used the 1980 results of the PSID to generate the finite experimental population from which the samples were drawn.

The purpose of the experiment is to investigate the statistical behavior of SCI procedures on the difference of proportions, when one population is sampled using a SRS and the other population is sampled using a S2S design. Using the 1980 PSID as a basis, a finite population with a stratified–cluster structure is constructed. The PSID is a stratified cluster sample with 32 strata with 2 clusters per stratum.

The PSID is a combination of two samples. One is an equal probability sample (EPSEM) and the other is a sample which over sampled individuals at the lower end of the earnings scale. As our basis we used only the EPSEM sample. Furthermore, we used only the subset of this EPSEM sample which were prime age (35-50) white males, with reported earnings in the interval between zero (we excluded those with no earnings) and \$99,999. Any individual with earnings above \$99,999. was recorded as having earnings of \$99,999; therefore, individuals with that response were not included in our basis sample. This resulted in 2089 individuals. On each individual we had the recorded earnings, and the strata and cluster designation. Let,

$$Y_{hij} = \text{the earnings of the } j^{th} \text{ individual from}$$

$$\text{the } i^{th} \text{ cluster of the } h^{th} \text{ stratum.}$$

To generate the individual earnings, we used the components of variance model, which is now described. Since the variable of interest is earnings, which has positive skewness we used the transformation $X_{hij} = ln(Y_{hij})$. The components of variance model on the transformed variable is

$$X_{hij} = \mu_h + \alpha_{hi} + \epsilon_{hij}$$

$$h = 1, 2, \ldots, 32$$

$$i = 1, 2$$

$$j = 1, 2, \ldots, J_{hi}.$$

The assumptions on the components are: for any h, α_{h1}, α_{h2} are independent indentically distributed (iid) normal with a mean of zero and a variance designated by $\sigma^2_{\alpha(h)}$; and $\epsilon_{hi1}, \epsilon_{hi2}, \ldots, \epsilon_{hiJ_{hi}}$ are iid normal with variance designated σ^2_h. The $\alpha's$ and $\epsilon's$ are mutually independent.

We used the standard classical methods from Graybill [5, Sec. 16.5] to estimate the parameters. Only 8 of the 32 strata resulted in postive estimates of the variance of the cluster effect. Therefore, in generating the finite population we used 8 strata. The parameter values for the generation are the estimates in the 8 strata. This gives the finite population on which the simulation was built a realistic basis. For each stratum, we generated 50 clusters, and within each cluster we generated 100 individual earnings.

The simulation experiment was executed using a FORTRAN program. The steps of the experiment are as follows:

1. Using the components of variance model with the estimated parameters, Population I earnings were generated. There were 8 strata, 50 clusters per strata, and 100 individuals per cluster, resulting in a total population size of 40,000.

2. For each individual earnings in Population I, a corresponding earnings was generated in Population II by multiplying the Population I earnings by δ. The values of δ investigated were $\delta = 1.00, 1.05, 1.10, 1.20, 1.50,$ and 2.00. These values are seen across the top of Table 1, which give the simulation results.

3. The individuals were classified into one of three categories $k = 1, 2, 3 = K$.

   Category 1 = Earnings less than \$10,000.

   Category 2 = Earnings between \$10,000. and \$20,000.

   Category 3 = Earnings greater than \$20,000.

4. From Population I a SRS of size 400 was selected and the earnings categories were noted.

5. From Population II a S2S sample was selected. In each stratum, 5 clusters were sampled; and within each cluster, 10 individuals were selected. Again the earnings categories were noted.

6. Using the three methods described in section 2, SCI's were calculated for $(p_k - q_k)$; $k = 1, 2, 3$; as defined in section 1.

7. For each value of δ, steps 4, 5, and 6 were replicated 500 times.

8. Measures of effectiveness (MOE) of the SCI's were calculated and are reported in Table 1.

4. Simulation Results

The simulation results are reported in Table 1 by giving 5 measures of effectiveness. Also, for the 3 categories of earnings, the finite population proportions are at the bottom of the table (p_1, p_2, p_3 for population I; q_1, q_2, q_3 for population II).

The MOE's follow the guidelines given in Schriber and Andrews[6]. The MOE given in the first row measures the coverage function introduced and discussed by Schruben[7]. Define the random variable η^* as follows:

$$\eta^* = inf\{\eta : \forall k, (p_k - q_k) \in C(\eta)\};$$

where, $C(\eta)$ is an $100\eta\%$ confidence interval employing the procedure under investigation. Hence, η^* is the confidence level that just succeeds in simultaneously covering all components of $\tilde{p} - \tilde{q}$.

It is shown in Schruben[7] that if $C(\eta)$ is an appropriate confidence interval procedure then the random variable η^* will be uniformly distributed on the $[0, 1]$ interval. Therefore, the value of η^* was determined for each of the SCI procedures on every replication. A chi-squared statistic of goodness-of-fit, using 10 equal sized intervals, was calculated to see if the η^*s followed a uniform distribution. These chi-squared values are reported in the first row of Table 1.

The small chi-squared values indicate a good fit. The 95^{th} percentile of the chi-squared with the appropriate 9 degrees of freedom is 16.92. Therefore, at this testing level, the ordinary-χ^2 and Bayesian procedures behave appropriately for all values of δ. However, for all values of δ the full design methodology fails to meet this criteria.

The next four rows of Table 3 pertain to the simultaneous intervals calculated with 95% confidence. The second row gives the percent coverage of the actual finite population difference. This should be approximately .95 for the 500 replications. For all three procedures, at all δ values, the coverage is close to .95. However, the coverage using the full design methodology is consistently larger than either of the other two methodologies. This indicates that the full design intervals are larger than required.

The third row gives the coverage of the point $p_k - q_k = 0$; $k = 1, 2, 3$. Coverage of this point could lead one to the conclusion that the proportions in the two populations are the same. This is only true when $\delta = 1.0$. For all three methods this coverage remains high until $\delta = 1.20$. This coverage is essentially zero for all three methods when $\delta = 1.50$ or 2.00. However, for $\delta's$ smaller than 1.50 the full design method consistently had a false coverage of zero at a larger percent than the other two methodologies.

The last two MOE's deal with the width of the intervals. We define W^* to be the width of the widest interval among the simultaneous intervals being calculated for the three components. $E[W^*]$ is the average over the replications of these widest intervals; and $SD[W^*]/E[W^*]$ is the corresponding coefficient of variation.

Ideally a confidence interval procedure needs to have the correct coverage with a small average width and furthermore, the width should have a small variation over the replications. Of the three confidence intervals investigated, the Bayesian procedure had the best properties according to these ideal criteria. Its value of $E[W^*]$ was always the smallest, even though it was only slightly smaller than the ordinary-χ^2 method. The full design average widths were always the largest which explains some of the coverage problems.

In addition the coefficient of variation of the full design widths were larger than the other two by a factor of 10. This indicates that the size of the intervals reported by the full design will vary considerably and in some instances give a false sense of precision by reporting very small intervals.

This simulation study casts some serious doubts on the recommended full design methodology. It is especially troublesome on two accounts. First, there is a large probability that the method will lead to falsely concluding that two populations have the same proportions. Secondly, the widths of the full design intervals have excessive variance.

5. Acknowledgment

This research was supported by the Department of Health and Human Services, Social Security Administration, under grant No. 10-P-98285-5-01.

REFERENCES

[1] Holt, D., Scott, A. J. and Ewings, P. D., (1980). Chi-squared Tests with Survey Data. *Journal of the Royal Statistical Society, A* **143**, Part 3, 303-320.

[2] Berger, J. O., (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

[3] United States Department of Health and Human Services (1985). Report on Earnings Sharing Information Study, SSA Publ. No. 12-004 ICN443150, Washington, D.C.

[4] Institute of Social Research (1984). User Guide to the Panel Study of Income Dynamics. Survey Research Center, University of Michigan, Ann Arbor.

[5] Graybill, F. A. (1961) *An Introduction to Linear Statistical Models, Vol. 1*, McGraw-Hill, New York.

[6] Schriber, T. J. and Andrews, R. W. (1981) A Conceptual Framework for Research in the Analysis of Simulation Output. *Communications of the ACM*, Vol. 24, No.4, pp. 218-232.

[7] Schruben, L. W. (1980) Coverage function of interval estimators of simulation response. *Management Science*, 26, 1, pp 18-27.

TABLE 1
Simulation Results (SRS vs. Two Stage)
(500 replications, Sample Size = 400)
(Seeds = 4543, 8771, 811) Prior = (1,1,1)

		∂ =1.0			∂ =1.05			∂ =1.10			∂ =1.20			∂ =1.50			∂ =2.0		
		Ord	FD	Bayes	Ord	FD	Bayes	Ord	FD	Bayes	Ord	FD	Bayes	Ord	FD	Bayes	Ord	FD	Bayes
Chi-squared Test of Coverage of Diff.		10.44	33.48	6.12	9.64	28.88	11.68	14.32	44.76	12.52	9.68	34.80	9.60	11.68	21.32	16.52	3.92	19.16	1.92
95% Interval	% Coverage of Population Difference	0.956	0.982	0.942	0.958	0.980	0.946	0.956	0.976	0.944	0.964	0.988	0.950	0.950	0.972	0.940	0.952	0.960	0.932
	% Coverage of Zero	0.956	0.982	0.942	0.912	0.946	0.898	0.784	0.860	0.746	0.300	0.420	0.258	0.000	0.000	0.000	0.000	0.000	0.000
	$\hat{E}[W^*]$	.1696	.1932	.1687	.1704	.1940	.1695	.1709	.1942	.1700	.1712	.1935	.1704	.1686	.1879	.1678	.1608	.1759	.1602
	$\frac{\hat{SD}[W^*]}{E[W^*]}$	.0074	.0663	.0074	.0064	.0682	.0064	.0055	.0690	.0055	.0049	.0664	.0049	.0081	.0633	.0079	.0137	.0645	.0134
P_1, P_2, P_3		0.258,0.339,0.404			0.258,0.339,0.404			0.258,0.339,0.404			0.258,0.339,0.404			0.258,0.339,0.404			0.258,0.339,0.404		
Q_1, Q_2, Q_3		0.258,0.339,0.404			0.241,0.328,0.431			0.226,0.318,0.456			0.199,0.297,0.504			0.141,0.235,0.625			0.087,0.170,0.742		