Using an Inverse Sampling Algorithm for Tests of Independence Based on Stratified Samples

Susan Hinkins and H. Lock Oh, Internal Revenue Service
Fritz Scheuren, George Washington University

For some users, a complex sample design can limit the accessibility of the data and can even result in the incorrect use of a data base. By using an inverse sampling algorithm, complex survey data can be converted into a resampling setting, where each sample is a simple random sample (srs). An inverse sampling algorithm to draw a simple random sample from a stratified sample was described in Hinkins, Oh, and Scheuren (1994). By resampling from a stratified sample, one can produce a simple random sample as if it had been selected from the original population. This approach is customer-driven; it makes complex databases more accessible. There are, however, two serious drawbacks to this methodology. First, the resampling algorithm may be computer intensive. This difficulty becomes less of an obstacle every day, as fast, cheap computing becomes more and more available.

A more serious problem is that the power or the precision of the resulting statistic may be seriously reduced, because of sample size limitations. In the case of a cluster sample of k clusters each of size M, the size of the largest srs that can be selected is k. In the case of a stratified sample with sample sizes n_h, the largest simple random sample that may be selected is of size $\min\{n_h\}$. By subsampling from the stratified sample, power is lost both by decreasing the sample size and by losing whatever increase in precision was due to stratification. This difficulty may be overcome by resampling multiple times from the original design. For example, it can be shown that for many linear estimators, an estimator using multiple simple random samples can be made almost as precise as the original estimator.

In particular, let d be a sample from any invertible sample design, such as a stratified design. Let T_d denote an estimator calculated from the sample d, and let T_i denote the equivalent estimate calculated from one simple random sample obtained from the sample d using the inverse algorithm. Then, if

$$E\ (T_i \mid \text{sample d}) = T_d\ ,$$

and if k **independent** simple random samples are chosen from the sample d, then the estimator

$$\overline{T}_* = \frac{1}{k}\sum_1^k T_i$$

has variance

$$= Var(T_d) + \frac{Var(T_i) - Var(T_d)}{k}\ .$$

One can make the variance of $\overline{T}_*$ arbitrarily close to the variance of the original estimator, T_d, by making k large -- i.e., by selecting enough simple random samples. More generally, one can show that this result holds for vectors of estimates and their associated covariance matrices.

The simple random samples are independent conditional on the original sample d, but they are not unconditionally independent. This makes the estimation of the variance of $\overline{T}_*$ a problem. A method for calculating unbiased estimates of the variance of $\overline{T}_*$ was also given in Hinkins, Oh, and Scheuren (1994).

Many important statistical techniques, such as regression and contingency table analysis, were developed largely in an Independent and identically distributed (IID) world. To use these techniques in the setting of sample surveys, it is often assumed that the sampling process makes the observed random variables independent and identically distributed. Further adjustments are needed in complex survey settings and much attention has been paid to the sample design's impact on linear and nonlinear

statistics. In this paper, we consider the chi-square test of independence based on a stratified sample and compare our method with the approach suggested by Fellegi (1980) and Scheuren (1972).

■ The Problem

Consider a 2x2 contingency table. Let P_{ij} denote the proportion of the population in cell ij. We want to test whether the rows and columns are independent. If a simple random sample of size m is taken, then the usual chi-square statistic can be calculated from the estimates, and

$$t = \sum_{i,j} \frac{m\,(\hat{P}_{ij} - \hat{P}_{i.}\hat{P}_{.j})^2}{\hat{P}_{i.}\hat{P}_{.j}}$$

has an asymptotic chi-squared distribution under the null hypothesis.

In stratified samples and other complex surveys, the equivalent statistic is not necessarily distributed asymptotically as a chi-square under H_0. If the chi-square statistic is computed from a complex sample as if the sample design were srs, then entirely misleading results may occur.

Alternative test statistics have been proposed for use when the data come from a sample design other than srs. In this paper we consider the approach suggested by Fellegi (1980) and Scheuren (1972), based on the idea of using balanced repeated replication and having the form:

$$t'' = \sum_{i,j} \frac{1}{b} \frac{(\hat{P}_{ij} - \hat{P}_{i.}\hat{P}_{.j})^2}{\hat{P}_{i.}\hat{P}_{.j}} ,$$

where 1/b is an estimate of the "effective sample size." When

$$b = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\hat{Var}(\hat{P}_{ij})}{\hat{P}_{ij}(1-\hat{P}_{ij})} ,$$

then t" is approximately distributed as a chi-square with one degree of freedom.

In this paper we consider an alternative approach of resampling from the stratified design to get a sequence of simple random samples. These samples are conditionally independent, but unconditionally dependent samples. It needs to be determined how to combine these samples to test the null hypothesis, i.e., how to calibrate the test to achieve the desired level.

Suppose k simple random samples, each of size m, are independently selected from the given stratified sample. Let I and J denote the two categorical variables of interest. Then, we have k 2x2 tables showing the interaction of interest, IxJ. Or, in other words, we have a 2x2xk table. Let K denote the dimension corresponding to the k simple random samples. Because each simple random sample is selected independently from the given stratified sample, each sample must have the same expected value for the IxJ cell estimates. Therefore, the IxJxK interaction must be zero. Similarly, the IxK and the JxK interactions must be zero. This means that the test of independence between I and J can be done by collapsing the tables over the variable K, i.e., by combining the k simple random samples into one table (Bishop, Fienberg, and Holland, 1975). If the overall sample size is large enough, the Pearson test statistic will be approximately chi-square with one degree of freedom. (If the minimum discrimination test statistic were used rather than the Pearson chi-square, then, by hypothesis, independent test statistics are created that partition the information in the sample into additive pieces.)

As k increases, the probability of rejecting the null hypothesis also increases. Therefore, the test must be calibrated so that the desired level (.05 in our case) is achieved. In the next section, two methods are discussed for calibrating the test -- i.e., for determining k, the number of srs' to be selected.

■ A Brief Outline of the Simulation

The simulation compares the Fellegi statistic for testing independence using stratified sample

data to the results using chi-square statistics from multiple srs' drawn from the same stratified sample. The comparisons are made over four populations, with varying patterns of "dependence" between the two variables. The measure of the dependence is the cross-product ratio (cpr).

Each population is defined with four strata, with population sizes 1,377, 553, 678, and 436. All population 2x2 tables are generated with marginal probabilities fixed at .5. The first population considered is the population satisfying the null hypothesis, namely independence between the two variables.

The other three populations are generated with dependence between the variables, represented by a cpr of 1.69. In defining each population, the properties within each stratum are considered as well as over the entire population. In the first case, the cpr is constant over all the strata. This is referred to as the homogeneous case; the stratification has no effect. In the other two cases, referred to as "nonhomogeneous," the rows and columns are independent (cpr=1) in each of three strata, and in the remaining stratum the cpr is as large as necessary, in order to make the overall cpr equal to 1.69. The four populations are summarized in Table 1.

For example, the "homogeneous" (but not independent) population is generated with cell probabilities $P_{11}=P_{22}=.28$ and $P_{12}=P_{21}=.22$ in each stratum. In the "nonhomogeneous population #1," the first three strata are generated with $P_{ij}=.25$ in all four cells and the last stratum is generated with cell probabilities $P_{11}=P_{22}=.48$ and $P_{12}=P_{21}=.02$

Two stratified sample designs are considered, each with a total sample size of n=156. The sample sizes and the sampling rates, by strata, for each design are:

Strata	Design A		Design B	
	n_h	p_h	n_h	p_h
1	39	.028	10	.007
2	22	.040	10	.018
3	40	.059	20	.029
4	55	.125	116	.266

Therefore the largest simple random samples that can be selected are m=22 for Design A, and m=10 for Design B.

From each population generated, 1,000 stratified samples are drawn for each design. The Fellegi test of independence is made for each stratified sample and these results provide estimates of the level of the test when the null hypothesis is true and estimates of the power of the test under the three alternatives.

Table 1.--Population Cross-product Ratios

Populations	Overall cpr	Stratum 1	Stratum 2	Stratum 3	Stratum 4
Independent	1.0	1.0	1.0	1.0	1.0
Homogeneous	1.69	1.69	1.69	1.69	1.69
Nonhomog-1	1.69	1.0	1.0	1.0	432.6
Nonhomog-2	1.69	1.0	38.2	1.0	1.0

For each population and sample design, 5 stratified samples are selected from the 1,000, in order to investigate the properties of the chi-square test statistic calculated from resampled simple random samples. (The intention is to draw 50 stratified samples from the 1,000, but, to begin with, only 5 are chosen.) For each stratified sample, one can calculate an estimated cpr:

$$f = \frac{\hat{P}_{11}\,\hat{P}_{22}}{\hat{P}_{12}\,\hat{P}_{21}} .$$

The 1,000 stratified samples are ordered by their estimated cpr's and approximately every 200th sample is selected. (The very extreme values on both ends are not included.)

From each of the 5 stratified samples selected in this manner, simple random samples are drawn. For each stratified sample in Design A, 700 simple random samples, each of size 22, are selected and for each stratified sample in Design B, 1,500 srs' of size 10 are resampled.

Two methods are used to calibrate how many samples should be combined in order to give the correct level of the test. The first method estimates the effective sample size, n_{eff}, as in the Fellegi statistic, under the null hypothesis of independence. Population values of the P_{ij} are used and assumptions about the population marginal distribution have to be made. This is a very easy calculation to make, especially with the simulation assumption that the population marginals are all equal to .5,

$$\frac{1}{n_{eff}} = \frac{1}{N^2}\sum_h \frac{N_h(N_h-n_h)}{n_h-1} .$$

The second method uses the independent population to determine which value of k results in an approximately .05 level test. That is, using the simulated independent population, the power of the test under the null hypothesis is estimated for each different value of k. For example, for each stratified sample under Design A, there are 700 srs' drawn, each of size 22. For each stratified sample, combining the srs' 7 at a time (k=7) results in 100 2x2 tables each with sample size k*m=154. For each table, the Pearson chi-square test is made and the null hypothesis of independence is either rejected or not. There are 5*100=500 such tests. The estimated level of the test for Design A with k=7 is, therefore, the number of tests that reject the null hypothesis divided by 500. For each design, this was done for several values of k.

In this example, the two methods give similar results. So, for each design, the number of samples to combine, k, is set. For the samples from the alternative population, k simple random samples are combined, the chi-square statistic is calculated, and the test of independence is made. This results in estimates of power for this technique to be compared to the power estimates for the Fellegi statistics.

■ Simulation Results

The Fellegi Test

For each population and sample design, we have 1,000 results of the test of independence using the Fellegi statistic, with level equal to .05. Table 2 shows the resulting estimated probability of rejecting the null hypothesis, for each simulated population and for each sample design. The last row shows the estimated level of the test. The Fellegi test statistic is achieving approximately the right test level; it rejects the null hypothesis slightly more often than expected.

Table 2.--Estimated Power of the Fellegi Test

Population	Design A	Design B
Homogeneous	.312	.133
NonH - 1	.293	.114
NonH - 2	.291	.151
Independent	.058	.067

The first three rows give estimates of the power of the test under the three alternatives and the two sample designs. The two nonhomogeneous populations (NonH-1 and NonH-2) were selected as populations for which the Fellegi statistic might not be as effective. While there is some loss in power between the homogeneous alternative and the nonhomogeneous alternatives, it is only a relatively small reduction. However, the sample design has a large effect on the power. In Design B, the test shows much less difference in the probability of rejecting the null hypothesis when it is false compared to when it is true. The power under the alternative is very low.

Combining Simple Random Samples

Recall that for samples using Design A, each srs is of size 22; for samples using Design B, each srs is of size 10. As described above, two methods are used for determining the number of simple random samples that should be combined in order to have a .05 level test. Estimating the effective sample sizes, as in the Fellegi statistic, gives $n_{eff} = 121$ for Design A and $n_{eff} = 35$ for Design B. This would mean k = 5 or 6 for design A and k = 3 or 4 for Design B.

Table 3 shows the results using the second method of determining k, based on the simulated independent population.

Table 3.--Independent Population Combining srs'

Design A		Design B	
k	level	k	level
6	.079	4	.085
5	.069	3	.064
4	.054	2	.048

This indicates one should choose k = 4 or 5 for Design A and k = 2 or 3 for Design B. These results overlap the results from the first method.

We chose k = 5 for Design A and k = 3 for Design B. Table 4 gives the estimated power of the test under the three alternative populations, using the Pearson chi-square test on the combined simple samples.

Table 4.--Estimated Power Combining k Simple Random Samples

Population	A (k = 5)	B (k = 3)
Homogeneous	.30	.13
NonH - 1	.30	.11
NonH - 2	.26	.12

The estimates of power using the inverse algorithm to get simple random samples are based on only 5 of the 1,000 stratified samples used to estimate the power of the Fellegi statistic. However, even in this preliminary work, the estimates of power for the two methods are very close.

■ Conclusion and Future Work

The simulation needs to be completed, by looking at the results for the simple random samples using 50 rather than only 5 stratified samples. But preliminary results indicate that this method of redrawing simple random samples from a stratified sample and using the simple Pearson chi-square holds much promise. Once the (conditionally) independent simple random samples are provided to the user, it is an easy, well known procedure for the user. The additional complexity required is determining the correct number of srs' to combine. It appears that it will not be difficult to calibrate the test, either using a fairly simple calculation or by simulating populations under the null hypothesis.

The proposed method appears to give power equivalent to the Fellegi methodology, while being much more user friendly. There are other means to be investigated for improving the power of the combined simple random samples. In this paper, we

assumed that the effective sample size, in terms of k, had to match the number of samples combined to estimate the P_{ij}'s. We will be looking at the possibility of using more srs' for estimating the P_{ij}'s and then estimating the effective sample size. Further work and simulations are needed to refine the methodology.

■ References

Bishop, Yvonne M. M.; Fienberg, Stephen E.; and Holland, Paul W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge, Mass.

Fellegi, Ivan (1980). Approximate Tests of Independence and Goodness of Fit Based on Multistage Samples, *Journal of the American Statistical Association*, 75, 261-268.

Hinkins, Susan; Oh, H. Lock; and Scheuren, Fritz (1994). Inverse Sampling Design Algorithms, *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Scheuren, Fritz (1972). *Topics in Multivariate Finite Population Sampling and Data Analysis*, George Washington University Doctoral Dissertation. ■