# THE EFFECT OF A TWO-STAGE SAMPLE DESIGN ON TESTS OF INDEPENDENCE IN A 2 by 2 TABLE

J. Cowan and D.A. Binder[1]

When a two-stage sample is used to collect data, the correlations between the sampled units make the $\chi^2$ test of independence invalid. Use of the ordinary $\chi^2$ tables generally results in a test which is greater than the desired level of significance. The effect of the sample design comes from two main areas: the sample size within PSU's and the degree to which the characteristics are independent within each PSU. The effect of the sample size within PSU's is greatest when there is no independence within each PSU, and diminishes as the degree of independence increases.

## 1. INTRODUCTION

Classical statistical inference has been developed through the years under the assumption of independent observations. In recent years, attention has turned to attempts to develop data analysis techniques for complex sampling procedures, especially in the area of social surveys (Kish & Frankel [5]). The Canada Health Survey has set up a data analysis group to monitor new developments in this field, to attempt to adapt existing techniques for its own use, and to look into new areas. This paper deals with one of these new areas.

One of the basic statistical tools currently in popular use is contingency table analysis for testing the independence of two or more characteristics in a population. One of the key assumptions in developing the distributional theory is the independence of the observations. This leads to the multinomial distribution and, asymptotically, the chi-squared test. If the assumption of independence is violated, as in a complex sample survey, this theory loses its validity. Several studies have been done on the analysis of contingency tables

[1] J. Cowan and D.A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

when correlations are present between observations, notably by Cohen [4], Altham [1], and Nathan [6]. Shuster and Downing [7] have derived a test statistic applicable for any sampling scheme.

The question has been raised as to how much the sampling scheme affects the inference. For example, is it possible to assume independent observations without distorting the true situation too much. The answer will depend, of course, on the sampling scheme and what the analysts consider "too much". A study is currently being done which empirically investigates the effect that a two-stage sample has on inferences about independence of two characteristics in a population, although the theory developed can be applied to any self-weighting sampling scheme.

## 2. DISTRIBUTION OF TEST STATISTICS

Denote the usual Pearson chi-square statistics by P:

$$P = \sum_{i,j} \frac{(n\hat{\Pi}_{ij} - n\hat{\Pi}_{i.}\hat{\Pi}_{.j})^2}{n\hat{\Pi}_{i.}\hat{\Pi}_{.j}}\text{, where the } \hat{\Pi} \text{ are the MLE's of } \Pi \text{ in the case of i.i.d. sampling.}$$

Next, approximate P by taking its Taylor series expansion around the point $\{\Pi\colon\ \Pi_{ij} = \Pi_{i.}\Pi_{.j}\}$, where the null hypothesis of independence is true. It can easily be shown that if $\hat{\Pi}$ is unbiased for $\Pi$, the constant and linear terms are zero. To evaluate the quadratic term, we find that

$$\left.\frac{\delta P}{\delta\hat{\Pi}_{pq}\ \delta\hat{\Pi}_{rs}}\right|_{\hat{\Pi}=\Pi} = \begin{cases} 2n\ (2 - \frac{1}{\Pi_{p.}} - \frac{1}{\Pi_{.q}} + \frac{1}{\Pi_{pq}}) & \text{if } p=r,\ q=s \\ 2n\ (2 - \frac{1}{\Pi_{p.}}) & \text{if } p=r,\ q\neq s \\ 2n\ (2 - \frac{1}{\Pi_{.q}}) & \text{if } p\neq r,\ q=s \\ 2n\ (2) & \text{if } p\neq r,\ q\neq s \end{cases}$$

Let $A = \frac{1}{2n}\left(\frac{\delta P}{\delta\hat{\Pi}_{pq}\ \delta\hat{\Pi}_{rs}}\bigg|_{\hat{\Pi}=\Pi}\right)$;

then P is approximated by:

$$P \doteq n(\hat{\Pi}-\Pi)'\ A(\hat{\Pi}-\Pi) = Q\ .$$

Since $\hat{\Pi}$ is assumed to be asymptotically normal, the distribution of Q is the weighted sum of independent central chi-squared random variables, each on one degree of freedom, where the weights $(\lambda_1, \lambda_2, \ldots)$ are the eigenvalues of the matrix AV (see Box [3], Theorem 2.1). Since the asymptotic distribution of Q is known, it is possible to perform an empirical study of the effect that the sample design has on inference by generating values of Q and calculating the proportion of these values which are greater than the $\alpha$-level critical point of the usual $\chi^2$ - test. This should give some idea how the probability of a Type 1 error is changed by the sampling scheme.

## 3. SIMPLE TWO STAGE DESIGNS

The distribution derived for the usual test of independence is used to examine the effect that a two-stage sample has on Type I errors in testing the independence of two characteristics, each with two categories. The population is divided into M primary sampling units (PSU), from which m are drawn. Each PSU contains H units from which h secondary sampling units (SSU) are chosen. Each sampled unit is then classified according to the two characteristics, so a 2 by 2 table can be constructed displaying the proportions of the sample population falling into each of the four categories, and P can be calculated. If the observations had been independent, this statistic asymptotically would have a chi-squared distribution with one degree of freedom. The null hypothesis of independence of the characteristics in the population could be tested by comparing the observed value of the statistic with the $\chi^2$ tables. Since the observations are not independent, use of the $\alpha$ - level critical point from the $\chi^2$ tables results in a test which is generally not at the desired level. Since the asymptotic distribution of P can be calculated for large m, the actual level of the test can be empirically investigated by generating random variables with this distribution and calculating the proportion which are greater than the critical value. With independent observations, this proportion will be $\alpha$ on the average.

## 3.1 Method

The first step in calculating the distribution of P is to calculate A, the matrix of the quadratic form which approximates the statistic. This is done by specifying the parameter $\Pi$, the population proportions falling into each category. The next step in calculating the distribution of P is to derive V, the matrix of variances and covariances of $\sqrt{mh}\ \hat{\Pi}$ for this particular sample design. It was decided to take M as infinite, as this would be a fairly accurate approximation to the Canada Health Survey, and the calculations are simplified. The derivation of V is shown in Appendix A. It is necessary to specify the distribution of ${}_i\Pi = ({}_i\Pi_1, {}_i\Pi_2, {}_i\Pi_3, {}_i\Pi_4)$ over all PSU's in order to calculate V. For simplicity the Dirichlet distribution was used. Appendix B gives the properties of this distribution.

The calculation of the eigenvalues of AV was done by a routine due to Sparks and Todd [8], and generation of normal random variables was done by a routine due to Marsaglia and Bray (see [2]). For fixed values of the parameters, the approximation to P was generated 10,000 times and the proportions which were greater than the .10, .05, .01, and .001 level critical values were recorded. Many combinations of the parameters were tried and the results are given in the next section.

## 3.2 Results

In all cases, three of the four eigenvalues are negligible and can be taken as zero, so the distribution of the Taylor series approximation to P reduces to that of a chi-squared random variable with one degree of freedom multiplied by the largest eigenvalue ($\lambda_1$) of the matrix AV. This eigenvalue determines the level of the test. If it is greater than 1, the null hypothesis will incorrectly be rejected more often than 100 $\alpha$ percent of the time, and if less than 1, the level drops below $\alpha$. The effects of the parameters are described below.

Appendix C gives the results of some of the simulations. Although this is far from a complete listing of the possible combinations of parameters, it still gives some idea how the level of significance is changed by the two-stage design.

a. Effect of $\Pi$ (population proportions)

If all other parameters are fixed, a change in the population proportions falling into each of the four categories results in a negligible change in $\lambda_1$. This means that the level of the test does not depend on $\Pi$. Individually, the A, V and AV matrices are changed as $\Pi$ changes, but the matrix AV will have the same eigenvalues.

b. Effect of h, H (within-PSU sample and population sizes)

For fixed h, variation in H gives limited variation in $\lambda_1$ but for fixed H, variation in h results in a great deal of variation in $\lambda_1$ (see Appendix C). Table 1 displays the largest eigenvalues for various values of H and h. It can be seen that an increase in the PSU population size for fixed number of SSU's within each PSU results in a slight increase in the significance level. This is caused by an increase in the covariance between the elements of $\hat{\Pi}$ as H increases. The table also shows that the effect of the sample size within PSU's is far greater than that of the PSU population size, and that the sampling fraction within PSU's by itself is not informative. This is due to the clustering effect of taking more SSU's within PSU's.

TABLE 1:

| $H$ | $h$ | $\lambda_1$ |
|---|---|---|
| 1000 | 50 | 9.13 |
| 500 | 50 | 9.08 |
| 300 | 50 | 9.03 |
| 100 | 50 | 8.75 |
| 1000 | 40 | 7.47 |
| 800 | 40 | 7.46 |
| 100 | 40 | 7.17 |
| 1000 | 30 | 5.81 |
| 400 | 30 | 5.77 |
| 100 | 30 | 5.59 |
| 1000 | 20 | 4.15 |
| 750 | 20 | 4.15 |
| 500 | 20 | 4.13 |
| 100 | 20 | 4.01 |
| 700 | 15 | 3.32 |
| 400 | 15 | 3.30 |
| 100 | 15 | 3.22 |
| 1000 | 10 | 2.49 |
| 500 | 10 | 2.49 |
| 250 | 10 | 2.47 |
| 100 | 10 | 2.42 |

$\lambda_1$ = largest eigenvalue of AV when $\theta. = \sum_{i=1}^{4} \theta_i = 5$

c. Effect of $\theta.$ (parameter of the distribution of $_i\Pi$)

Small values of $\theta.(= \sum_{i=1}^{4} \theta_i)$ give large values of $\lambda_1$, and as $\theta.$ increases to infinity, $\lambda_1$ decreases to $(1 - \frac{h}{H})$. A typical example is for the case H = 300, h = 50 (Table C4): $1 - \frac{h}{H} = 0.833$.

| $\theta.$ | $\lambda_1$ |
|---|---|
| 5 | 9.030 |
| 25 | 2.727 |
| 125 | 1.226 |
| 625 | 0.915 |
| 3125 | 0.852 |
| 15625 | 0.839 |
| 78125 | 0.837 |

This result shows that the more independent the characteristics are within strata, the less the size of the test. Notice that if all units within strata are sampled (h = H), then independence within strata gives a value of 0 for the approximation to P, as it should. Another result is that for small $\theta.$ , variation in h causes wide variation in $\lambda_1$, but for large $\theta.$ , the variation in $\lambda_1$ becomes a great deal smaller.

## 4. FUTURE DIRECTIONS

The results here are very restrictive because the model is simple. The two-stage model was chosen to be as close to multinomial sampling as possible. However, because the usual chi-squared behaves so poorly here, we would expect things would get worse in a more realistic setting.

One interesting point is that all the above results yielded only one dominant eigenvalue. Is this true for more realistic settings (e.g. unequal sized PSU's, unequal probability samples)? If so then we may be able to estimate the largest eigenvalue and derive a statistic whose distribution is much closer to chi-squared. Also the use of the Dirichlet distribution is only one of many possible distributions that could be considered.

## APPENDIX A: Derivation of V

Let $u_i = 1$ if PSU i, is in the sample $i = 1, \ldots, \infty$
$= 0$ otherwise .

Let $x_{ijk} = 1$ if the $j^{th}$ unit in the $i^{th}$ PSU belongs to category k
$= 0$ otherwise.

Then $\bar{X}_{i.k}$ is the proportion of units in PSU i that belong to category k,

and $\hat{\bar{X}}_{i.k}$ is the proportion of sampled SSU's in PSU i that belong to category k.

Let $S_{ik\ell} = \frac{1}{H-1} \sum_{j=1}^{H} (x_{ijk} - \bar{X}_{i.k})(x_{ij\ell} - \bar{X}_{i.\ell})$.

Now, $y_k = \sum_{i=1}^{\infty} u_i \, h \, \hat{\bar{X}}_{i.k}$ = total number of sampled units that belong to category k

and $E(y_k) = E_{\underset{\sim}{u}}\left(E(y_k|u)\right) = h \, E_{\underset{\sim}{u}}\left(\sum_{i=1}^{\infty} u_i \, \bar{X}_{i.k}\right) = hm \, E(\bar{X}_{i.k})$ .

$$\text{Also, } \operatorname{Cov}(y_k, y_\ell) = E_{\underset{\sim}{u}}\left(\operatorname{Cov}(y_k, y_\ell|\underset{\sim}{u})\right) + \operatorname{Cov}_{\underset{\sim}{u}}\left[E(y_k|\underset{\sim}{u}),\ E(y_\ell|\underset{\sim}{u})\right]$$

$$= mh(1 - \tfrac{h}{H})\ E(S_{ik\ell}) + h^2 m\left[E(\bar{X}_{i.k}\ \bar{X}_{i.\ell}) - E(\bar{X}_{i.k})\ E(\bar{X}_{i.\ell})\right],$$

and since $E(S_{ik\ell}) = \frac{H}{H-1} E\{\delta_{k\ell}\ \bar{X}_{i.k} - \bar{X}_{i.k}\ \bar{X}_{i.\ell}\}$ ($\delta_{k\ell} = 1$ if $k = \ell$
$= 0$ otherwise)

we have: $\operatorname{Cov}\left(\frac{y_k}{\sqrt{mh}}, \frac{y_\ell}{\sqrt{mh}}\right) = \frac{H-h}{H-1} E\{\delta_{k\ell}\ \bar{X}_{i.k} - \bar{X}_{i.k}\ \bar{X}_{i.\ell}\}$

$$+ h\left[E(\bar{X}_{i.k}\ \bar{X}_{i.\ell}) - E(\bar{X}_{i.k})\ E(\bar{X}_{i.\ell})\right].$$

## APPENDIX B: Distribution of ${}_i\Pi$

Within PSU i, we know that the sum of the proportions in each category is 1, or $\sum_{j=1}^{4} {}_i\Pi_j = 1$ for i = 1, 2, ... ∞. We also know that the average value over all strata of the proportion in category j is $\Pi_j$, or $\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} {}_i\Pi_j = \Pi_j$ for j = 1, 2, 3, 4. One simple way to accomplish this is to let ${}_i\Pi$ follow a multivariate analogue of the Beta distribution, called the Dirichlet distribution.

If the random vector $(y_1, \ldots, y_k)$ follows a Dirichlet distribution with parameters $\theta_1, \ldots, \theta_k$, then: $\sum_{i=1}^{k} y_i = 1$; $y_1 \geq 0$, i = 1, ..., k,

the density is given by

$$\frac{\Gamma\left(\sum_{i=1}^{k} \theta_i\right)}{\prod_{i=1}^{k} \Gamma(\theta_i)} \prod_{i=1}^{k} y_i^{\theta_i - 1} ,$$

$$E(y_i) = \frac{\theta_i}{\Sigma\theta_j}, \quad E(y_i^2) = \frac{\theta_i(\theta_i + 1)}{(\Sigma\theta_j)(\Sigma\theta_j + 1)},$$

$$E(y_i y_k) = \frac{\theta_i \, \theta_k}{(\Sigma\theta_j)(\Sigma\theta_j + 1)} .$$

To apply this distribution to the calculation of V, we specify $\Pi_k$, k = 1, 2, 3, 4 and equate $\frac{\theta_k}{\Sigma\theta_j}$ to $\Pi_k$. This determines $(\theta_1, \theta_2, \theta_3, \theta_4)$ to within a constant multiple. Various values for $\theta. = \sum_{i=1}^{4} \theta_i$ will then determine various degrees of independence of the two characteristics within PSU's. As $\theta.$ increases, $\mathrm{Var}({}_i\Pi_j)$ decreases to zero for all j so that the distribution of the proportions of units falling into each of the categories is identical from PSU to PSU. Since the average value over all PSU's is $\Pi_k$, the within PSU distribution must be identical to the population proportion, and since independence holds in the population, it necessarily holds within PSU.

## APPENDIX C: Proportion of times the approximation to P is greater than the critical point at the given significance level

C1. # UNITS/PSU = 400
# SSU /PSU = 30

| max eigenvalue : | 5.77277 | 2.04550 | 1.15805 | .97376 | .93662 | .92918 | .92769 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .491 | .250 | .130 | .094 | .089 | .088 | .088 |
| .05 | .412 | .168 | .070 | .046 | .042 | .043 | .042 |
| .01 | .282 | .071 | .017 | .008 | .007 | .008 | .006 |
| .001 | .171 | .021 | .002 | .0008 | .0007 | .0006 | .0006 |

C2. # UNITS/PSU = 400
# SSU /PSU = 15

| max eigenvalue : | 3.30409 | 1.50472 | 1.07630 | .98733 | .96940 | .96581 | .96509 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .364 | .175 | .110 | .098 | .093 | .093 | .093 |
| .05 | .283 | .108 | .057 | .049 | .047 | .044 | .045 |
| .01 | .157 | .035 | .012 | .009 | .009 | .009 | .008 |
| .001 | .068 | .007 | .002 | .0009 | .0008 | .0007 | .0009 |

C3. # UNITS/PSU = 500
# SSU /PSU = 10

| max eigenvalue : | 2.48497 | 1.32881 | 1.05353 | .99637 | .98485 | .98254 | .98208 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significane level | | | | | | | |
| .10 | .300 | .155 | .108 | .097 | .095 | .095 | .094 |
| .05 | .215 | .088 | .055 | .049 | .048 | .048 | .047 |
| .01 | .102 | .025 | .012 | .009 | .009 | .009 | .009 |
| .001 | .037 | .004 | .0012 | .0009 | .0009 | .0009 | .0008 |

C4. # UNITS/PSU = 300
# SSU /PSU = 50

| max eigenvalue : | 9.03010 | 2.72704 | 1.22631 | .91466 | .85185 | .83927 | .83675 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .587 | .320 | .136 | .084 | .073 | .068 | .069 |
| .05 | .516 | .236 | .077 | .041 | .033 | .030 | .031 |
| .01 | .395 | .119 | .019 | .007 | .005 | .005 | .005 |
| .001 | .274 | .046 | .0025 | .0004 | .0004 | .0002 | .0004 |

C5. # UNITS/PSU = 250
# SSU /PSU = 10

| max eigenvalue : | 2.46988 | 1.31140 | 1.03557 | .97829 | .96674 | .96443 | .96397 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .293 | .150 | .104 | .096 | .095 | .092 | .092 |
| .05 | .212 | .086 | .053 | .047 | .045 | .045 | .045 |
| .01 | .099 | .025 | .011 | .009 | .008 | .008 | .008 |
| .001 | .035 | .004 | .0013 | .0010 | .0007 | .0007 | .0008 |

C6. # UNITS/PSU = 500
# SSU /PSU = 20

| max eigenvalue : | 4.13494 | 1.69416 | 1.11302 | .99234 | .96801 | .96314 | .96217 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .415 | .204 | .117 | .098 | .095 | .092 | .092 |
| .05 | .333 | .131 | .062 | .048 | .045 | .046 | .044 |
| .01 | .204 | .048 | .014 | .009 | .008 | .008 | .008 |
| .001 | .105 | .012 | .0020 | .0011 | .0008 | .0008 | .0009 |

C7. # UNITS/PSU = 750
# SSU /PSU = 20

| max eigenvalue : | 4.14553 | 1.70638 | 1.12563 | 1.00502 | .98072 | .97585 | .97487 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .417 | .211 | .120 | .100 | .093 | .095 | .093 |
| .05 | .337 | .138 | .064 | .051 | .048 | .047 | .047 |
| .01 | .207 | .051 | .017 | .010 | .009 | .008 | .008 |
| .001 | .108 | .012 | .002 | .0012 | .0010 | .0009 | .0007 |

C8.  # UNITS/PSU = 500  
# SSU /PSU = 50

| max eigenvalue : | 9.08484 | 2.79020 | 1.29147 | .98023 | .91751 | .90494 | .90243 |
|---|---|---|---|---|---|---|---|
| θ. sums to → | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 |
| ↓ significance level | | | | | | | |
| .10 | .582 | .325 | .148 | .095 | .083 | .082 | .083 |
| .05 | .513 | .244 | .086 | .047 | .040 | .037 | .038 |
| .01 | .393 | .125 | .023 | .010 | .008 | .007 | .007 |
| .001 | .275 | .050 | .004 | .0009 | .0007 | .0006 | .0005 |

## RESUME

Quand on utilise un échantillon à deux degrés pour rassembler des données, les corrélations entre les unités échantillonnées rendent le test d'indépendance $\chi^2$ invalide. Si on utilise les tables ordinaires de $\chi^2$, on obtient généralement un test qui est plus grand que le seuil significatif voulu. L'effet du plan d'échantillonnage provient de deux facteurs principaux: la taille de l'échantillon dans les UPE et le degré d'indépendance des caractéristiques dans chaque UPE. L'effet de la taille de l'échantillon dans les UPE est à son maximum quand il n'y a pas d'indépendance dans chaque UPE, et diminue à mesure que le degré d'indépendance augmente.

## REFERENCES

[1] Altham, P.M.E. (1976). "Discrete variable analysis for individuals grouped into families", Biometrika, 63, 2, 263-269.

[2] Atkinson, A.C. and Pearce, M.C. (1976). "The computer generation of beta, gamma, and normal random variables", JRSS, A, 139, 431-461.

[3] Box, G.E.P. (1954). "Some theorems on quadratic forms applied in the study of analysis of variance problems I", AMS, 25, 2, 290-302.

[4] Cohen, J.E. (1976). "The distribution of the chi-squared statistic under clustered sampling from contingency tables", JASA 71, 665-670.

[5] Kish, L. and Frankel, M.R. (1974). "Inference from complex samples", JRSS(B), 36, 1, 1-37.

[6] Nathan, G. (1975). "Tests of independence in contingency tables from stratified proportional samples", Sankhya (C), 37, 77-87.

[7] Shuster, J.J. and Downing, D.J. (1976). "Two-way contingency tables for complex sampling schemes", Biometrika, 63, 271-276.

[8] Sparks, D.N. and Todd, A.D. (1973). "Latent roots and vectors of a symmetric matrix", Applied Statistics, 22, 2, 260-265.