The Distribution of the Chi-Squared Statistic Under Clustered Sampling from Contingency Tables

JOEL E. COHEN*

When the entries in a contingency table arise from clustered sampling, the chi-squared statistic conventionally calculated to test simple and complex hypotheses about the parameters of the table may not have the distribution of a χ^2 variate. Assuming a model for positively associated clustering, this article finds the distribution of the conventional chi-squared statistic and shows how to correct it to an asymptotically χ^2 variate. A numerical example from the epidemiology of mental illness is given.

1. INTRODUCTION

Clustered sampling is frequent in both experimental science and sample surveys of populations.

In a large demographic sample survey which motivates the present analysis, households were selected in clusters of size six. In each household, the survey recorded the survival or failure to survive to one year of children born at least a year earlier, along with social, economic, and biological characteristics of the family. To study the interaction of these other characteristics with infant mortality, it is necessary to allow for a possible dependence in infant mortality experienced by members of the same sibship. There may also be dependence in infant mortality among households within a cluster. A substantial correlation in infant mortality within sibships has been demonstrated [1, pp. 87, 93] for the same population in a different survey.

A general strategy for analyzing clustered or matched samples from contingency tables[1] is first to test the clusters for independence among members. Cochran [8] develops methods for assaying whether the variation within a cluster differs from the expected binomial variation (see, also [6, Ch. 8]). If there is no evidence against independence, the clustering may be ignored.

If clustering seems *a priori* likely or is apparent from the data, methods developed by Cochran [9] and Bennett [3, 4] and reviewed in an epidemiological context by Pike and Morrow [13] permit a correct analysis.

Here, performance of the conventional chi-squared test under clustering is compared to that for independence. The conventional statistic may be replaced by one which has asymptotically a χ^2 distribution; this is useful both as a computational shortcut and for studying the effect of neglecting clustering.

2. THE MODEL

Suppose each individual in the sample falls into one of $r \geq 2$ cells. The sample design consists of N independent clusters each of 2 individuals. Each cluster has a first and a second individual. Clusters of size K will be considered and the present results greatly extended by Altham [2].

When observations on each individual in the cluster are discretely categorized, the frequency counts X_{ij} of the number of clusters in which the first individual falls in cell i and the second falls in cell j ($i, j = 1, \ldots, r$) may be arrayed in a two-way $r \times r$ contingency table.

Such clustered sampling may arise [6, p. 281] when each sampled individual is classified twice by the same criteria, either because observations are repeated in time (in panel studies) or because of some intrinsic symmetry of the individual (right eye color versus left) or when the sample consists of couples of individuals each of whom are subject to the same categorization. The couples may be paired by some preexisting relationship (husbands and wives, mothers and daughters, sibs) or by experimentally imposed case-control matching on some nuisance variables. In these cases, there is a natural way of identifying, within each cluster, which observation or individual is first (the earlier observation in the panel study, the right eye, the wife) and which second.

Clustered sampling also arises when there is no natural ordering, e.g., in clusters of households or adjacent counties. In such cases, ordering within a cluster is assigned by randomization.

The $r \times r$ table just described has a margin X_{i+} of row sums and a margin X_{+j} of column sums. These two margins can be rewritten into another $2 \times r$ table in which the first row gives X_{i+} and the second row gives X_{+i}, $i = 1, \ldots, r$. The row sums of this new table are each necessarily equal to N. The column sums Y_i of this

* Joel E. Cohen is Professor, The Rockefeller University, New York, NY 10021. The Institute of Population Studies, Hacettepe University, Ankara; King's College Research Centre, Cambridge CB2 1ST, England; and the U.S. National Science Foundation provided partial support. Yvonne M.M. Bishop and P.M.E. Altham provided helpful references and discussion and a referee made useful suggestions. The University of Cambridge Computing Service provided computational facilities.

[1] Suggested by Yvonne M.M. Bishop in private communication.

© Journal of the American Statistical Association
September 1976, Volume 71, Number 355
Theory and Methods Section

new table give the distribution of individuals among the r categories, irrespective of order within the cluster.

In the following model of clustering, the categorization of an individual as first or second is assumed *not* to enter the hypotheses regarding the r proportions $Y_i/(2N)$.

For example, Tsuang [14, p. 288] determined the birth order (elder or younger), sex, and diagnosis of each individual hospitalized for mental disorder in pairs of siblings. In testing the hypothesis that sex and diagnosis are independent, he wanted to regard birth order as irrelevant. But the possibility of clustering within sibships must be considered (and Tsuang did so in a reasonable, but ad hoc, way). His data are reanalyzed in Section 6. Another example of data and hypotheses with the same formal structure is [12].

In these examples, the hypotheses tested concern marginal tables with entries Y_i obtained by collapsing contingency tables (of dimension one higher than the margin) across the dimension which specifies the ordering in the cluster. For example, in [14], let r be the number of cells in the array Sex × Diagnosis; then the $2 \times r$ table Birth Order × (Sex × Diagnosis) would be collapsed across Birth Order to test a (compound) hypothesis regarding the Y_i, namely, that Sex and Diagnosis are independent. Section 5 sets out this model more explicitly.

When each observation is independent of all others, i.e., in the absence of clustering, Bishop [5] has determined the conditions under which collapsing a contingency table across a dimension will preserve invariant the relations of interest between the remaining dimensions.

When clustering is not excluded by the sample design, we proceed as follows. Let X_{ij} be the number of clusters in which the first individual falls in cell i and the second falls in cell j; $i, j = 1, \ldots, r$. Define

$$X_{i+} = \sum_{j=1}^{r} X_{ij} , \quad X_{+j} = \sum_{i=1}^{r} X_{ij} ,$$
$$Y_i = X_{i+} + X_{+i} , \quad i, j = 1, \ldots, r . \tag{2.1}$$

Then,

$$\sum_{i=1}^{r} Y_i = 2N , \quad \sum_{i=1}^{r} X_{i+} = \sum_{j=1}^{r} X_{+j} = N . \tag{2.2}$$

To formalize the notion that there is independence between clusters, suppose that the set of random variables $\{X_{ij}\}$ is jointly distributed in a multinomial distribution with parameters N and $\{P_{ij}\}$ where $\sum_{i,j} P_{ij} = 1$. To formalize a notion of clustering, meaning positive association within clusters, suppose there are r constants $p_i > 0$, $i = 1, \ldots, r$, $\sum_i p_i = 1$ and a constant a, $0 \leq a \leq 1$, such that

$$P_{ij} = p_i(a\delta_{ij} + (1 - a)p_j) , \quad i, j = 1, \ldots, r . \tag{2.3}$$

Here, $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$. As would be expected if the ordering of individuals in a cluster is irrelevant to the analysis, $P_{ij} = P_{ji}$. Thus, (2.3) is a special case of the model of symmetry in [6, p. 282]. Define P_{i+} and P_{+j} by analogy to (2.1). Then,

$$P_{i+} = P_{+i} = p_i , \quad i = 1, \ldots, r . \tag{2.4}$$

The model (2.3) may be interpreted as follows. p_i is the (marginal) probability that the first individual falls in cell i. Then the conditional probability that the second individual falls in the same cell i, given that the first is in cell i, is a linear interpolation between one and p_i. When the weight a is one, the second member of a cluster is slave to the first. Then the sample really contains only N independent observations on the p_i. When the weight a is zero, the second member is independent of the first. Then the sample contains $2N$ independent observations on the p_i. When a is greater than zero, the increased probability that the second member will fall in the same cell as the first is removed in constant proportions from the probabilities of the remaining cells.

If (2.3) is rewritten as $P_{ij} = ap_i\delta_{ij} + (1 - a)p_ip_j$, then P_{ij} is seen to be the mixture of $p_i\delta_{ij}$ and p_ip_j with mixing probabilities a and $1 - a$, respectively. Altham has pointed out privately that (2.3) may be viewed as a specially symmetric mover-stayer model for social mobility tables.

This model of clustering allows only for positive association or independence between members of a cluster, as is the case for many demographic, social and economic characteristics. The possibility of negative association is not modeled here.

3. TESTING A SIMPLE HYPOTHESIS

The conventional chi-squared statistic calculated to test the goodness of fit of the Y_i to a multinomial model of r cells with probabilities $p_i > 0$, $i = 1, \ldots, r$, is

$$X^2 = \sum_{i=1}^{r} (Y_i - 2Np_i)^2/(2Np_i) . \tag{3.1}$$

Let $\sim$ mean "has the same distribution in large samples $(N \to \infty)$." Then when $a = 0$, $X^2 \sim \chi^2_{r-1}$ where the subscript shows the degrees of freedom (df). When $a = 1$, $X^2/2 \sim \chi^2_{r-1}$, since the observed sample size $(2N)$ is twice the number (N) of independent observations. We now show that, generally, under the model of Section 2,

$$W^2 = X^2/(1 + a) \sim \chi^2_{r-1} , \quad 0 \leq a \leq 1 . \tag{3.2}$$

Distributions of this form have arisen previously [7, 11].

To prove (3.2), recall that, according to the multinomial model for X_{ij},

$$E(X_{ij}) = NP_{ij} ,$$
$$\text{Cov}\,(X_{ij}, X_{k\ell}) = N(P_{ij}\delta_{ik}\delta_{j\ell} - P_{ij}P_{k\ell}) , \quad i, j, k, \ell = 1, \ldots, r . \tag{3.3}$$

Combining (2.1) and (2.4) with (3.3) gives

$$E(Y_i) = 2Np_i , \quad i = 1, \ldots, r . \tag{3.4}$$

Let Z_i be a random variable which counts the number of individuals in cell i in a given cluster. Since each Y_i is the sum of N independent, identically distributed copies

of Z_i, which has the marginal distribution

$$Z_i = 2 \text{ with probability } P_{ii} ,$$
$$= 1 \text{ with probability } \sum_{j \neq i} P_{ij} + \sum_{k \neq i} P_{ki} = 2(p_i - P_{ii}) ,$$
$$= 0 \text{ otherwise} ,$$

the variance of Y_i is N times the variance of Z_i, and for $i \neq j$, Cov $(Y_i, Y_j) = N$ Cov (Z_i, Z_j), where Z_i and Z_j are jointly distributed random variables referring to the same cluster. Hence,

$$\text{Var } (Y_i) = (1 + a)2Np_i(1 - p_i) , \quad i = 1, \ldots, r , \tag{3.5a}$$

and for $i \neq j$, $i, j = 1, \ldots, r$,

$$\begin{aligned} \text{Cov } (Y_i, Y_j) &= N[E(Z_iZ_j) - 4p_ip_j] \\ &= N \Pr[Z_i = 1 \text{ and } Z_j = 1] - 4Np_ip_j \\ &= N(P_{ij} + P_{ji}) - 4Np_ip_j \\ &= -(1 + a)2Np_ip_j . \end{aligned} \tag{3.5b}$$

Hence, the covariance matrix $\mathbf{L} = (L_{ij})$ of $\{Y_i\}$ is just $(1 + a)\mathbf{S}$ where $\mathbf{S} = (S_{ij})$ is the covariance matrix of the cell counts in a simple random sample of size $2N$ from a multinomial distribution on $\{p_i\}$. Hence, $\mathbf{L}^{-1} = \mathbf{S}^{-1}/(1 + a)$. Then following exactly Wilks' Theorem 9.3.2 [15, p. 261] on the asymptotic distribution of quadratic forms and its application in Theorem 9.3.2a to the multinomial distribution, $(1 + a)^{-1}$ factors out to give (3.2).

4. ESTIMATION AND TESTING OF COMPLEX HYPOTHESES

Crude estimates of parameters are first obtained by assuming that random variables are near their expected values. We then obtain maximum likelihood estimates.

From (3.4), an obvious estimator of p_i is

$$p_i^* = Y_i/(2N) , \quad i = 1, \ldots, r , \tag{4.1}$$

with variance obtainable from (3.5a). This estimator is unbiased and converges in probability to p_i.

Two estimators a_1^* and a_2^* of a follow from (2.3) by considering the case $i = j$ and replacing P_{ii} by X_{ii}/N, p_i by p_i^*. If one sums over i and then solves for a, one obtains

$$a_1^* = (\sum_{i=1}^{r} X_{ii}/N - \sum_{i=1}^{r} p_i^{*2})/(1 - \sum_{i=1}^{r} p_i^{*2}) . \tag{4.2}$$

If one first solves (2.3), $i = j$, for a and then averages the r estimates obtained, one finds

$$a_2^* = r^{-1} \sum_{i=1}^{r} (X_{ii}/N - p_i^{*2})/(p_i^*(1 - p_i^*)) . \tag{4.3}$$

Asymptotic variances of these estimators may be derived by the usual truncated Taylor series approximation, but they are not attractive. Appropriate weighting of the terms averaged in (4.3) might improve its variance.

The estimators a_1^* and a_2^* depend only on $\{X_{ii}\}$ and $\{Y_i\}$. The same should remain true of the maximum likelihood estimators because the model (2.3) is a linear combination of a model (when $a = 1$) for which the diagonal frequencies $\{X_{ii}\}$ are sufficient and a model (when $a = 0$) for which the margins $\{Y_i\}$ are sufficient. Equations (4.4) and (4.5) confirm the sufficiency of $\{X_{ii}\}$ and $\{Y_i\}$. The full array $\{X_{ij}\}$ is required only to test fit to the model (2.3).

As the observations are assumed to be intrinsically discrete, and not the result of grouping some quantitative variable, the maximum likelihood and modified minimum χ^2 methods [10, p. 426] coincide. Then using (2.3) and $p_r = 1 - \sum_{j=1}^{r-1} p_j$ in

$$L = N! \prod_{i,j=1}^{r} (P_{ij}^{X_{ij}}/X_{ij}!) , \tag{4.4}$$

the maximum likelihood equations

$$\frac{\partial \ln L}{\partial a} = 0 , \quad \frac{\partial \ln L}{\partial p_i} = 0 , \quad i = 1, \ldots, r - 1 , \tag{4.5}$$

reduce to

$$\sum_{i=1}^{r} X_{ii}(1 - p_i)/[a/(1 - a) + p_i] = N - \sum_{i=1}^{r} X_{ii} \tag{4.6}$$

and

$$p_i = [Y_i - X_{ii}/(1 + (a^{-1} - 1)p_i)]/ [2N - \sum_{j=1}^{r} X_{jj}/(1 + (a^{-1} - 1)p_j)] . \tag{4.7}$$

The maximum likelihood estimator $\hat{a}$ of a is the larger of 0 and the solution of (4.6). It is easily checked that (4.6) and (4.7) are consistent with results known to be true in the limiting cases $a = 0$ and $a = 1$.

If a is in (0, 1) and if $\{p_i\}$ are assumed known, the variance of $\hat{a}$ may be estimated by replacing a by $\hat{a}$ in

$$\begin{aligned} -(E(\partial^2 \ln L/\partial a^2))^{-1} &= N^{-1}(\sum_{i=1}^{r} p_i(1 - p_i)^2/(a + (1 - a)p_i) \\ &\quad + (1 - \sum_{i=1}^{r} p_i^2)/(1 - a))^{-1} . \end{aligned} \tag{4.8}$$

When a solution $(\hat{a}, \hat{p}_1, \ldots, \hat{p}_r)$ to (4.6) and (4.7) exists, a suggested procedure for finding it is to start with $a^{(0)} = a_1^*$ or a_2^*, $p_i^{(0)} = p_i^*$. With $k = 0$,

1. find $a^{(k+1)}$ by numerical solution of (4.6), using $p_i^{(k)}$ for p_i;
2. find $p_i^{(k+1)}$ from (4.7) using $a^{(k+1)}$ for a;
3. increment k by 1 and go to step 1. Stop when the solutions quasiconverge.

With the r linearly independent parameter estimates $\hat{a}$ and $\hat{p}_i$, the adequacy of the model of Section 2 may be tested with the statistic

$$\sum_{i,j=1}^{r} (X_{ij} - N\hat{P}_{ij})^2/(N\hat{P}_{ij}) \sim \chi^2_{r^2-r-1} , \tag{4.9}$$

where $\hat{P}_{ij} = \hat{p}_i(\hat{a}\delta_{ij} + (1 - \hat{a})\hat{p}_j)$. When some of the r^2 cells must be pooled to obtain expected frequencies of

reasonable size, the term r^2 in the expression for df must be reduced appropriately.

5. TESTING COMPLEX HYPOTHESES USING THE Y_i

In the demographic survey which provoked the development of these methods, only the values of $\{Y_i\}$ are routinely available. It is desired to use these values to test hypotheses concerning the $\{p_i\}$. We consider the case where the $\{Y_i\}$ are entries in a two-way contingency table with R rows and C columns, $R, C \geq 2$. We wish to test for independence of rows and columns.

Suppose that p_i and Y_i have been reindexed as p_{uv} and Y_{uv}, $u = 1, \ldots, R$, $v = 1, \ldots, C$, $r = RC$. Within the model of Section 2, we wish to test the null hypothesis that there exist constants $g_u > 0$, $u = 1, \ldots, R$, $h_v > 0$, $v = 1, \ldots, C$ such that $\sum g_u = \sum h_v = 1$ and $p_{uv} = g_u h_v$. The row and column margins of the table of Y_{uv} are

$$Y_{u+} = \sum_{v=1}^{C} Y_{uv} , \quad Y_{+v} = \sum_{u=1}^{R} Y_{uv} . \tag{5.1}$$

The test statistic conventionally calculated is

$$(X')^2 = \sum_{u=1}^{R} \sum_{v=1}^{C} (Y_{uv} - Y_{u+}Y_{+v}/(2N))^2/ [Y_{u+}Y_{+v}/(2N)]. \tag{5.2}$$

We shall show that, under the model of Section 2 and the preceding null hypothesis,

$$(W')^2 = (X')^2/(1 + \hat{a}) \sim \chi^2_{(R-1)(C-1)} , \tag{5.3}$$

whether $\hat{a}$ is based on the same sample as that on which the Y_{uv} are based or $\hat{a}$ is based on any other large sample from a population with the same value of a.

From (3.4) and (3.5) we have, under the null hypothesis

$$E(Y_{uv}) = 2Ng_u h_v ,$$

$$\text{Cov}\,(Y_{uv}, Y_{wz}) = (1 + a)2N(g_u h_v \delta_{uw}\delta_{vz} - g_u h_v g_w h_z) , \tag{5.4}$$

$$u, w = 1, \ldots, R, \quad v, z = 1, \ldots, C .$$

Then,

$$\begin{aligned} E(Y_{u+}) &= 2Ng_u , \quad u = 1, \ldots, R , \\ E(Y_{+v}) &= 2Nh_v , \quad v = 1, \ldots, C . \end{aligned} \tag{5.5}$$

Define

$$\begin{aligned} g_u^* &= Y_{u+}/(2N) , \quad u = 1, \ldots, R , \\ h_v^* &= Y_{+v}/(2N) , \quad v = 1, \ldots, C . \end{aligned} \tag{5.6}$$

Then, g_u^* is an unbiased estimator of g_u that converges in probability to g_u, and similarly for h_v^*. Altogether these constitute $R + C - 2$ linearly independent estimators. The distribution of the sequence (indexed on N) of random variables $\{Y_{uv} - 2Ng_u^* h_v^*\}$ approaches a distribution of rank

$$RC - 1 - (R + C - 2) = (R - 1)(C - 1)$$

whose covariance matrix is given by the limit as $N \to \infty$ of

$$(1 + a)2N(g_u^* h_v^* \delta_{uw}\delta_{vz} - g_u^* h_v^* g_w^* h_z^*) \tag{5.7}$$

The combination of the previously used theorems on quadratic forms and multinomial covariance matrices [15, pp. 261–2] with theorems on the limiting distribution of sums of squares from singular distributions [10, pp. 298–300, 313–14] permits the conclusion

$$(X')^2/(1 + a) \sim \chi^2_{(R-1)(C-1)} . \tag{5.8}$$

Since $\hat{a}$ converges in probability to a, the ratio $(1 + a)/(1 + \hat{a})$ converges in probability to 1; hence, by another convergence theorem [10, p. 254], the product, which is $(W')^2$, of $(1 + a)/(1 + \hat{a})$ and the left side of (5.8) has a distribution which converges to $\chi^2_{(R-1)(C-1)}$. As Cramer points out, this convergence holds whether or not $\hat{a}$ is independent of the Y_{uv}.

In practice, a special analysis which recognizes the clustering in the design is necessary to obtain an estimate $\hat{a}$. Then the correction factor $(1 + \hat{a})^{-1}$ can be applied to the values of $(X')^2$.

The argument leading from (5.7) to (5.8) is heuristic rather than rigorous, but it suggests that in complete multidimensional contingency tables, more elaborate hierarchical loglinear hypotheses regarding the (variously subscripted) parameters p_i can similarly be tested using the Y_i and marginal configurations of the Y_i to calculate a chi-squared statistic. An asymptotically χ^2 variable should be obtained by multiplying the chi-squared statistic by the correction factor $(1 + \hat{a})^{-1}$.

When clusters are of size $K \geq 2$, and a chi-squared statistic X^2 analogous to (3.1) or (5.2) is calculated without regard to the clustering, the null hypothesis regarding the p_i may be accepted or rejected. If it is accepted, i.e., if the value of X^2 is not large enough to be significant, then recognition of clustering with positive association would not alter the conclusion. If the null hypothesis is rejected and if X^2/K is also significantly large, then recognition of clustering with positive association would again not alter the conclusion, since the worst situation is that all K members of a cluster always fall in a single cell. If X^2 is significantly large but X^2/K is not, a more detailed study of the dependencies within a cluster is required to estimate an "effective" sample size.

When clusters are of unequal size within a single sample, as in surveys of individuals in households or sibships, the preceding qualitative analysis applies if K is now interpreted as the size of the largest cluster which occurs in the sample. The crude bounds established by such analysis of extreme cases leave much to be desired.

6. NUMERICAL EXAMPLE

The methods in the preceding sections were designed for data formally similar to those in the following

example. However, I was unaware of the existence of the particular data analyzed here when I developed the methods. The satisfactory fit of the model of clustering (2.3) to these data gives hope that the same model may apply usefully to other data.

In this example, $r = 4 = 2 \times 2$. It is desired to test a complex hypothesis (independence) using the marginal frequencies Y_i arranged in a 2×2 table.

Table 1 gives the $r \times r$ distribution by sex and diagnosis (schizophrenic versus not schizophrenic) of pairs of siblings, both hospitalized, according to rank in age (elder versus younger). These data, labeled "Ob.," are the X_{ij}. Here, each of the two subscripts is, in fact, an ordered couple (Diagnosis, Sex). It is desired to test whether there is interaction between diagnosis and sex.

1. Diagnosis and Sex of Hospitalized Sibling Pairs

Elder sibling	Item	Younger sibling			
		SM	SF	NM	NF
SM	Ob.	13	5	1	3
	In.	6.51	4.85	2.27	7.87
	Cl.	10.48	3.38	1.61	5.28
SF	Ob.	4	6	1	1
	In.	4.85	3.61	1.69	5.86
	Cl.	3.38	7.67	1.29	4.21
NM	Ob.	1	1	2	4
	In.	2.27	1.69	0.79	2.75
	Cl.	1.61	1.29	2.99	2.01
NF	Ob.	3	8	3	15
	In.	7.87	5.86	2.75	9.52
	Cl.	5.28	4.21	2.01	14.33

NOTE: S = Schizophrenia, N = Not schizophrenia, M = Male, F = Female, Ob. = Observed numbers of pairs, In. = Expected given independence, Cl. = Expected given clustering.
Source: Tsuang [13].

2. Estimates of Parameters Assuming Independence and Clustering

Parameter	Independence	Clustering
a	0	0.3006
p_{SM}	0.3028	0.2923
p_{SF}	0.2254	0.2330
p_{NM}	0.1056	0.1112
p_{NF}	0.3662	0.3636
X^2 (4.9)	26.631	13.109
df	12	11
P	$0.001 < P < 0.01$	$0.2 < P < 0.3$

First, it is necessary to check whether there is independence within clusters. The values "In." under the observations give the numbers of each cluster expected assuming that elder and younger siblings are identically and independently distributed. Expected values are calculated from (2.3) and (4.1), fixing $a = 0$. Independence within clusters is rejected at the one-percent level.

Next, since the full data are available in this situation, it is necessary to check whether the model of clustering describes the data (Table 2). Four iterations of the numerical procedure proposed in Section 4, starting with $a_1^* = 0.3079$, gave parameter estimates that quasi-converged to 10^{-4}. Within each of these iterations, the Newton-Raphson method never required more than two iterations to solve (4.6) for a. The model of clustering describes the data well.

A lower bound on the variance of $\hat{a}$ is obtained by treating $\{p_i\}$ in (4.8) as if they were known in advance to be $\{\hat{p}_i\}$ (instead of themselves being estimated from the data). The lower bound on the standard deviation of $\hat{a}$ thus obtained is 0.0818, which substantially exceeds the difference 0.0073 between a_1^* and $\hat{a} = 0.3006$. Thus, a_1^* may be viewed as a good first guess in this example.

The conventional 2×2 table which would be formed from Y_{SM}, Y_{SF}, Y_{NM}, Y_{NF} is shown in the tabulation,

Diagnosis	Sex	
	M	F
S	43	32
N	15	52

where $(X')^2 = 17.885$, $(W')^2 = 13.751$, df $= 1$, $P > 0.001$, and the abbreviations and data are given in Table 1. The conventional test statistic $(X')^2$ from (5.2) to test for independence between sex and diagnosis is significant at the 0.1 percent level. The corrected test statistic $(W')^2$ from (5.3) remains significant at the 0.1 percent level.

Hence, in this case, there is strong evidence of positive association within clusters. But adjusting the measure of association between sex and diagnosis to allow for this clustering does not eliminate its statistical significance.

[*Received September 1974. Revised April 1975.*]

REFERENCES

[1] Adlakha, Arjun Lal, "A Study of Infant Mortality in Turkey," Doctoral dissertation, University Microfilms 71-4553, University of Michigan, Ann Arbor, 1970.

[2] Altham, Patricia M.E., "Discrete Data Analysis for Individuals Grouped into Families," *Biometrika*, 63, No. 2 (1976).

[3] Bennett, B.M., "Test of Hypotheses Concerning Matched Samples," *Journal of the Royal Statistical Society, Ser. B*, 29, No. 3 (1967), 468–74.

[4] ———, "Note on χ^2 Tests for Large Samples," *Journal of the Royal Statistical Society, Ser. B*, 30, No. 2 (1968), 368–70.

[5] Bishop, Yvonne M.M., "Effects of Collapsing Multidimensional Contingency Tables," *Biometrics*, 27 (September 1971), 545–62.

[6] ———, Fienberg, Stephen E., and Holland, Paul, *Discrete Multivariate Analysis*, Cambridge, Mass.: M.I.T. Press, 1974.

[7] Chase, Gerald R., "On the Chi-Square Test When the Parameters Are Estimated Independently of the Sample," *Journal of the American Statistical Association*, 67 (September 1972), 609–11.

[8] Cochran, William G., "Analysis of Variance for Percentages Based on Unequal Numbers," *Journal of the American Statistical Association*, 38 (September 1943), 287–301.

[9] ———, "The Comparison of Percentages in Matched Samples," *Biometrika*, 37, Parts 3 and 4 (1950), 256–66.

[10] Cramer, H., *Mathematical Methods of Statistics*, Princeton, N.J.: Princeton University Press, 1946.

[11] Murthy, V.K. and Gafarian, A.V., "Limiting Distributions of Some Variations of the Chi-Squared Statistic," *Annals of Mathematical Statistics*, 41 (February 1970), 188–94.

[12] Nielsen, Johannes, "Mental Disorders in Married Couples (Assortative Mating)," *British Journal of Psychiatry*, 110 (August 1964), 683–97.

[13] Pike, Malcolm and Morrow, Richard, "Statistical Analysis of Patient-Control Studies in Epidemiology; Factor Under Investigation an All-or-None Variable," *British Journal of Preventive and Social Medicine*, 24 (February 1970), 42–4.

[14] Tsuang, Ming-Tso, "A Study of Pairs of Sibs Both Hospitalized for Mental Disorder," *British Journal of Psychiatry*, 113 (March 1967), 283–300.

[15] Wilks, S.S., *Mathematical Statistics*, New York: John Wiley & Sons, Inc., 1962.