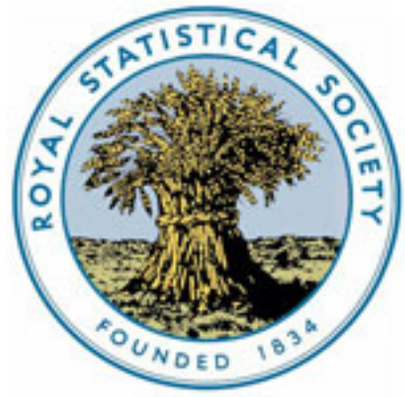

J. R. Statist. Soc. A (1980)
143, *Part* 3, *pp.* 303–320

Chi-squared Tests with Survey Data

By D. HOLT, A. J. SCOTT and P. D. EWINGS

University of Southampton and University of Auckland

SUMMARY

It is usually unrealistic to regard observations from a complex survey as independent, but practitioners often use standard chi-squared tests with survey data. We look at the effect of correlations among elements on the performance of the tests. Empirical results are given for two national surveys which suggest that the effect is severe for tests of goodness-of-fit or homogeneity but less severe for tests of independence.

Keywords: CHI-SQUARED TEST; DESIGN EFFECT; COMPLEX SAMPLES

1. INTRODUCTION

KISH AND FRANKEL (1974) have drawn attention to the problems that arise when standard statistical methods, based on the assumption that observations are independent, are applied to survey data. In most surveys, the assumption of independence is far from realistic. Any large-scale survey will involve stratified multi-stage sampling and correlations between units in the same cluster (or stratum) can have a substantial impact. In this paper we look at the effect of this structure on the behaviour of ordinary chi-squared tests for goodness-of-fit, homogeneity and independence.

Examples of the use of standard chi-squared tests in survey data are plentiful. An investigator may wish to compare the achieved sample proportions for the categories of variables such as age, sex or marital status with known population proportions, for example, as a check on the quality of the sampling. Brackstone and Gosselin (1973) give such an example for the underenumeration in the 1971 Canadian Census. More generally, we might make comparisons between several different surveys from the same population or, between different regions of the country or between similar surveys from different countries as in the World Fertility Survey comparisons. An ordinary chi-squared test of homogeneity would be natural here except for the complexity of the sampled populations. Examples of use of the standard chi-squared test for independence in a two-way contingency table with survey data can be found in almost any issue of most methodological journals in the Social Sciences. Casperis and Vaz (1975) or Chase (1975) are typical examples.

In all these situations, it is common practice to ignore the complexity of the survey and proceed as if the ordinary chi-squared tests will behave much the same as under multinomial sampling. In Section 2, we examine the behaviour of the goodness-of-fit test. A brief sketch of the large sample theory is given and we look at the effect in practice for variables from two large national surveys; the General Household Survey (GHS) of 1971 and the British Election Study (BES) of 1974. The problem of homogeneity is tackled in Section 3. The results are very similar in both cases; the chi-squared test can be very seriously affected by the lack of independence in the observations. However a simple multiplying factor, which can be calculated from information which should be available in many surveys, improves the performance dramatically.

In Section 4 we look at the important problem of testing for independence in a two-way table. The theoretical results are much more complicated here and an appropriate modifying factor is more difficult to compute. In contrast to the results of homogeneity and goodness-of-fit the impact of dependence among the observations seems much less severe in practice, and the ordinary chi-squared test often needs very little modification. These conclusions can be regarded as a natural extension of established results for the difference of sub-class means, of

which the 2×2 table is a special case. It is well known (see Kish, Groves and Krotki, 1976, for example) that lack of independence has a much smaller effect on the variance of the difference between the means of two domains which cut across strata and clusters than on the variance of the individual means. However, if the domains can be put in separate strata, so that the estimated means are independent, the effect on the variance of the difference is about the same as the effect on the individual means. The former case corresponds to the results for independence and the latter to those for homogeneity.

The performance of chi-squared tests when applied to complex samples has attracted considerable attention in recent years and several authors have proposed modified procedures. Armitage (1966) suggests a simple correction for stratified random sampling and similar corrections for more complex schemes have been proposed by Altham (1976), Cohen (1976), Fellegi (1978), Brier (1979), Fienberg (1979) and Rao and Scott (1979). Our results suggest that the corrections work very well for tests of goodness-of-fit and homogeneity but tend to be very conservative in the independence case, at least for the type of variable measured in the GHS and BES surveys. If enough information about the covariance matrix of estimated proportions is available, Bhapker and Koch (1968), Nathan (1973) and Schuster and Downing (1976) have discussed procedures which are asymptotically valid under very general conditions. These procedures require considerably more information than the proposed modifications to the simple chi-squared tests and may not always be possible, particularly for investigations involved in secondary analyses from published tables.

2. The Chi-squared Goodness of Fit Test

2.1. *Asymptotic Theory*

We begin with a brief review of the results in Rao and Scott (1979). The population is split into k categories, according to the value of the variable of interest, with population proportions $p_1, p_2, \ldots, p_k$ ($\Sigma p_j = 1$, summing over j from 1 to k). It is convenient to define $\mathbf{p} = (p_1, \ldots, p_{k-1})^{\mathrm{T}}$, omitting the last category. The null hypothesis to be tested is

$$H_0\colon \mathbf{p} = \mathbf{p}_0 = (p_{01}, \ldots, p_{0,k-1})^{\mathrm{T}}.$$

The survey data produce unbiased estimates $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ of the proportions and the conventional chi-squared statistic is given by

$$\bar{X}^2 = n \sum_{1}^{k} \frac{(\hat{p}_j - p_{0j})^2}{p_{0j}} \tag{2.1}$$

Note that, if $\mathbf{P}_0 = \mathrm{diag}(\mathbf{p}_0) - \mathbf{p}_0 \mathbf{p}_0^{\mathrm{T}}$ denotes the covariance matrix of $\hat{\mathbf{p}}$ for independent sampling when H_0 is true, then we can write $\bar{X}^2$ in the form

$$\bar{X}^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)^{\mathrm{T}} \mathbf{P}_0^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0) \tag{2.2}$$

In most large surveys, the survey design will involve stratification and multistage sampling and it is not realistic to assume that the observations are independent. Instead we suppose that

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}) \tag{2.3}$$

for some positive-definite covariance matrix $\mathbf{V}$ as the sample size $n \to \infty$. If we have a consistent estimator, $\hat{\mathbf{V}}$ say, of $\mathbf{V}$ we can test H_0 using the Wald statistic

$$\bar{X}_w^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)^{\mathrm{T}} \hat{\mathbf{V}}^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0) \tag{2.4}$$

which is asymptotically χ^2_{k-1} under the null hypothesis (see Schuster and Downing, 1978). A good example of this approach is given by Koch, Freeman and Freeman (1975). The disadvantage is that we need to be able to calculate $\hat{\mathbf{V}}$ and this can be very difficult if the design is particularly complex. Unfortunately, it is fair to say that the calculation of variances for

complex surveys is the exception rather than the rule in current survey practice, and the calculation of the full covariance matrix is even rarer.

In the absence of a reasonable estimate of $\mathbf{V}$, it is very common to ignore the lack of independence and use the standard chi-squared test. Obviously it is important to understand how much the properties of this are affected in practice. The formal theory is relatively straightforward. Under the assumption in (2.3), we can write

$$\bar{X}^2 \sim \sum_{1}^{k-1} d_i Z_i^2, \tag{2.5}$$

where $Z_1, Z_2, \ldots$ are asymptotically independent $N(0,1)$ random variables and $d_1, d_2, \ldots, d_{k-1}$ are the eigenvalues of $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V}$. Thus, the asymptotic distribution of $\bar{X}^2$ is that of a linear combination of χ_1^2 random variables, and is only exactly χ_{k-1}^2 in the multinomial case when all the d_i's are equal to one. We look at the effect on the nominal significance levels for two well-known surveys in the next section.

In the special case $k = 2$, $\mathbf{P}_0$ and $\mathbf{V}$ are scalars and d_1 can be interpreted in familiar sampling terms as the ratio of the variance of the estimated proportion in the complex situation to its variance under random sampling; in other words, d_1 is the design effect of the estimated proportion. In the more general case, we might call $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V}$ the design effect matrix. The Z_i's are special linear combinations of the estimated cell proportions and d_i is the design effect of Z_i. In particular if $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_{k-1}$, then d_1 represents the largest possible design effect for any linear combination of the cell proportions, d_2 the largest design effect among any linear combination orthogonal to Z_1, and so on. There is one important case where the result in (2.5) is of immediate use; all design effects are less than one with proportionally allocated stratified sampling (provided the strata are reasonably large) so $\bar{X}^2 \leqslant \Sigma_1^{k-1} Z_i^2$ and the ordinary chi-squared test is conservative.

More generally, the effect of the design on the nominal significance level will depend on the size of the d_i's and on the degrees of freedom. Table 1 illustrates the general effect for the special case in which $d_i = d$ $(i = 1, \ldots, k-1)$. (It follows from (2.5) that $\bar{X}^2 \sim d\chi_{k-1}^2$ in this case.)

The results show very clearly that the effect can be very serious, especially when the degrees of freedom are large. Fellegi (1978) gives further examples of the impact of design effects of the size often encountered in practice.

The effect is more complex in the general case with unequal d_i's. There is a large literature on the distribution of linear combinations of chi-squared random variables and tables of such distributions have been produced by Grad and Solomon (1955), Johnson and Kotz (1968) and Solomon and Stevens (1977). Table 2 illustrates the effect of varying d_i's.

More extensive results are given in Ewings (1979). We see that by far the most important factor is the value of $\bar{d}$. Changes in the distribution of the d_i's about $\bar{d}$ produce only a small change in the significance level, increasing variability making things worse when $\bar{d} = 1$ but actually improving the performance when $\bar{d} = 2$.

The results in Tables 1 and 2 show clearly that the naive chi-squared test can be very misleading in general, but they also suggest that a simple modified test based on

$$\bar{X}_m^2 = \bar{X}^2/\bar{d} \tag{2.6}$$

might work well. There are other modifying factors that might give more accurate results (the geometric mean for example) but $\bar{d}$ has one very important advantage. Any estimate of the d_i's in general will require an estimate of $\mathbf{V}$ and our starting point was that such an estimate is not normally available. It can be shown, however, that $\bar{d}$ can be expressed in the form

$$(k-1)\bar{d} = \sum_{j=1}^{k} V_{jj}/p_{0j} \tag{2.7}$$

TABLE 1

Actual size of $\bar{X}$ test with nominal 5 per cent significance level ($d_i = d$ ($i = 1,...,k-1$))

Design effect d	*Degrees of freedom*		
	2	5	10
1	5	5	5
1·2	8	10	12
1·5	14	19	27
3·0	37	59	81

TABLE 2

Size of $\bar{X}^2$ test with nominal 5 per cent significance level ($k = 6$)

$d_i/\sum d_i$					$\bar{d}$		
					1·0	1·5	2·0
0·20	0·20	0·20	0·20	0·20	5	19	35
0·22	0·22	0·22	0·18	0·16	5	20	36
0·36	0·36	0·12	0·08	0·08	7	21	35
0·50	0·24	0·10	0·08	0·08	8	21	34
0·60	0·10	0·10	0·10	0·10	9	21	32

and hence can be calculated from the cell variances alone. We need no information on the covariance terms at all. Information about the variances is available in many surveys (as in Kish *et al.*, 1976, for example) so the modification is often feasible. It is also the case that survey practitioners have more experience about the inflation of the variance of a proportion rather than the covariance of two estimated proportions and can often put a reasonable upper bound on the value of $\bar{d}$. This would lead to a conservative test procedure. We look at the performance of $\bar{X}^2_m$ in practice in the next section. If no information about the V_{ij}'s is available, the unmodified chi-squared test is potentially so misleading that there seems little point in carrying it out.

2.2. *Empirical Results*

To investigate the performance of $\bar{X}^2$ and $\bar{X}^2_m$ in practice, data from two large-scale national surveys were analysed. The first was the General Household Survey (GHS), involving responses from more than 13 000 households and the second was the British Election Survey (BES) in which data were collected from almost 2500 individuals. Both surveys are stratified, three stage designs sampling polling districts within constituencies and are typical of large-scale national surveys. Sampling with probability proportional to size was used in both surveys to ensure approximately equal selection methods. Full details of the GHS design are given in the *Annual Report* (1973) and of the BES design in Crewe, Sarlvic and Alt (1974).

Our aim was to evaluate the actual size of the tests based on $\bar{X}^2$ and $\bar{X}^2_m$ under the null hypothesis. This was done very simply by estimating $\hat{\mathbf{p}}$ and $\hat{\mathbf{V}}$ for each variable using standard survey sampling methods and then taking $\mathbf{p}_0 = \hat{\mathbf{p}}$ to ensure that $\hat{\mathbf{D}} = \mathbf{P}_0^{-1}\hat{\mathbf{V}}$ gave a reasonable estimate of the null behaviour. The eigenvalues $d_1,...,d_{k-1}$ were obtained from $\hat{D}$, and the distribution of $\bar{X}^2$ and $\bar{X}^2_m$ could then be found using standard methods for linear combinations of χ^2_1 random variables (Solomon and Stephens, 1977). Although the d_i's are only estimates they are based on large samples and the probabilities are very insensitive to small changes in the d_i's so the estimated sizes should be a reasonably good approximation to the actual values. We wanted to gain insight into the direction and magnitude of effects occurring in practice so a variety of variables were studied.

Table 3 shows the estimated size of nominal 5 per cent tests based on $\bar{X}^2$ and $\bar{X}^2_m$ for 13 variables measured on the GHS. A list of the variables analysed from both surveys is given in Appendix 1 together with the maximum number of categories for each variable. For some parts of the empirical work categories may have been collapsed.

As expected the actual size is much larger than the nominal 5 per cent levels for variables like age of building (G1) which have high design effects. It is disturbing to note how high the level can be for rather low design effects when the degrees of freedom are large. The modification is very satisfactory in all cases and seems to be an adequate substitute for the Wald test if cell variances are available.

TABLE 3

Estimated significance levels (%) for nominal 5 per cent goodness-of-fit tests for GHS data

Variable	*d.f.*	$\bar{m}$	$\bar{d}$	*Size* ($\bar{X}^2$)	*Size* ($\bar{X}^2_m$)
G1	3	34·7	3·27	48	6
G2	2	33·1	3·42	41	5
G3	3	33·4	2·54	37	6
G4	3	27·7	2·17	30	6
G5	9	31·3	1·23	16	7
G6	3	34·6	1·48	16	6
G7	6	34·7	1·29	15	6
G8	9	34·6	1·19	14	6
G9	5	26·6	1·14	10	7
G10	2	34·6	1·26	10	5
G11	4	34·5	1·01	6	6
G12	1	34·7	1·13	6	5
G13	3	34·7	1·02	6	5

Note: $\bar{m}$ is the average sample size per cluster. Apart from G9: Gross Household Weekly income and G4: Type of Accommodation, the sample sizes are approximately equal. Variations which do occur are due to non-response and categories such as "don't know" which have been excluded.

TABLE 4

Estimated significance levels (%) for a nominal 5 per cent goodness-of-fit test based on $\bar{X}^2$ and $\bar{X}^2_m$ (BES data)

Variable	*d.f.*	$\bar{m}$	$\bar{d}$	*Size* ($\bar{X}^2$)	*Size* ($\bar{X}^2_m$)
B1	3	12·23	3·38	50	6
B2	3	11·87	2·97	44	6
B3	3	12·17	1·93	26	5
B4	12	12·14	1·26	19	6
B5	4	10·53	1·43	17	6
B6	4	11·67	1·43	17	5
B7	3	11·75	1·55	17	5
B8	2	9·39	1·57	16	5
B9	2	11·64	1·34	12	5
B10	2	11·10	1·32	11	5
B11	3	12·31	1·19	10	5
B12	2	12·15	1·25	10	5
B13	10	11·19	1·13	10	6
B14	5	12·20	1·13	8	5
B15	1	11·48	1·29	8	5
B16	1	12·31	1·06	6	5

Results for 16 variables from the BES are given in Table 4. The pattern of results and general conclusions are very similar to those for the GHS data. The naive chi-squared test can perform very badly but the modified test seems to work well in all cases.

It is clear that the number of degrees of freedom makes a considerable difference to the performance of the ordinary $\bar{X}^2$ test, and we obtained some tables with a larger number of degrees of freedom by cross-classifying pairs of variables from the same survey. Results are given in Table 5. The main observation is that, although the value of $\bar{d}$ for the cross-classification tends to be smaller than for the individual variables, this effect is swamped by the increase in the degrees of freedom so that the chi-squared test performs very badly indeed. Although not shown, the test based on $\bar{X}^2_m$ again performed well in all cases.

TABLE 5

Estimated significance level (%) for a nominal 5 per cent goodness-of-fit test based on $\bar{X}^2$ for cross-classified variables

Variables		*Variable 1*			*Variable 2*			*Cross-classified variable*		
1	2	$\bar{d}$	*d.f.*	*Size* ($\bar{X}^2$)	$\bar{d}$	*d.f.*	*Size* ($\bar{X}^2$)	$\bar{d}$	*d.f.*	*Size* ($\bar{X}^2$)
B9	B2	1·33	2	11	2·94	3	42	1·61	11	37
B9	B11	1·34	2	12	1·22	3	10	1·11	11	10
B2	B13	2·96	3	43	1·16	3	9	1·50	15	32
B13	B11	1·15	3	9	1·19	3	9	1·08	15	12
G2	G6	3·41	2	41	1·76	2	18	1·85	8	39
G6	G7	1·81	2	19	1·62	3	18	1·28	11	18
G6	G11	1·81	2	19	1·01	4	6	1·01	14	6
G6	G8	1·81	2	19	1·51	4	19	1·21	14	15
G6	G5	1·78	2	19	1·56	3	17	1·16	11	17
G6	G3	1·82	2	19	2·53	3	37	1·53	11	31
G3	G7	2·55	3	38	1·62	3	19	1·44	15	28
G3	G8	2·55	3	38	1·51	4	19	1·33	19	26

Notes: (i) Where a variable is used in different cross-classifications the individual $\bar{d}$ may vary since the records used will depend upon valid responses to both variables in the particular cross-classification.
(ii) Some degrees of freedom and design effects are different to those presented in Tables 3 and 4 since categories have been collapsed if they contained very small proportions.

TABLE 6

Effect of collapsing categories on the size of a nominal 5 per cent test

Variable	*d.f.*	$\bar{d}$	*Size* ($\bar{X}^2$)
B6	2	1·61	16
	4	1·43	17
B13	2	1·14	8
	10	1·13	10
B5	2	1·59	15
	4	1·43	17

Finally, we looked briefly at the effect of collapsing categories of a variable and thus reducing degrees of freedom. Results are shown in Table 6. The average design effect tends to increase when the categories are collapsed (we have no explanation as to why this should be so) partially, but not completely, offsetting the improvement that results from fewer degrees of freedom.

3. TESTING HOMOGENEITY

3.1. *General Theory*

We now consider the more general situation with independent samples, of sizes n_1 and n_2 from two populations. The null hypothesis is that the category proportions (for some variable measured in both samples) is the same for both populations. Let $\mathbf{p}_j = (p_{j1}, \ldots, p_{j,k-1})^T$ represent the population proportions for the jth population ($j = 1, 2$) so the hypothesis is

$$H_0: \mathbf{p}_1 = \mathbf{p}_2 \quad (= \mathbf{p} \text{ say}). \tag{3.1}$$

As before, we suppose that estimates $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ are calculated from the sample data and that

$$\sqrt{n_j}(\hat{\mathbf{p}}_j - \mathbf{p}_j) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_j) \quad (j = 1, 2) \tag{3.2}$$

as $n_j \to \infty$. If we have consistent estimators, $\hat{\mathbf{V}}_1$ and $\hat{\mathbf{V}}_2$, of $\mathbf{V}_1$ and $\mathbf{V}_2$ then we can base a Wald test on

$$\bar{X}^2_{WH} = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)^T \left[\frac{\hat{\mathbf{V}}_1}{n_1} + \frac{\hat{\mathbf{V}}_2}{n_2}\right]^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)$$

which is asymptotically χ^2_{k-1} under the null hypothesis, or on a modified statistic with $\mathbf{V}_1$ and $\mathbf{V}_2$ estimated under H_0.

As in the goodness-of-fit case, if no estimates of $\mathbf{V}_1$ and $\mathbf{V}_2$ are available, practitioners often use the ordinary chi-squared statistic

$$\bar{X}^2_H = \sum_{i=1}^{2} \sum_{j=1}^{k} n_i \frac{(\hat{p}_{ij} - \hat{p}_j)^2}{\hat{p}_j}, \tag{3.3}$$

where $\hat{p}_j = (n_1 \hat{p}_{1j} + n_2 \hat{p}_{2j})/(n_1 + n_2)$. Note that $\bar{X}^2_H$ can be written in the form

$$\bar{X}^2_H = \frac{n_1 n_2}{n_1 + n_2} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)^T \hat{\mathbf{P}}^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \tag{3.4}$$

where $\hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^T$, so that $\bar{X}^2_H$ is equivalent to a modified Wald statistic for multinomial sampling since $\mathbf{V}_1$ and $\mathbf{V}_2$ are both equal to $\mathbf{P}$ under H_0 in this case. Again the asymptotic distribution of $\bar{X}^2_H$ under the general variance structure is straightforward to derive. It follows from standard results on quadratic forms (see Johnson and Kotz, 1970, p. 150) that

$$\bar{X}^2_H = \sum_{i=1}^{k-1} d_i Z_i^2, \tag{3.5}$$

where $Z_1, Z_2, \ldots, Z_{k-1}$ are asymptotically independent $N(0, 1)$ and the d_i's are eigenvalues of

$$\begin{aligned} D_H &= \mathbf{P}^{-1} \left(\frac{n_2 \mathbf{V}_1 + n_1 \mathbf{V}_2}{n_1 + n_2}\right) \\ &= \frac{n_2 \mathbf{D}_1 + n_1 \mathbf{D}_2}{n_1 + n_2}, \quad \text{say} \end{aligned} \tag{3.6}$$

where $\mathbf{D}_i = \mathbf{P}^{-1} \mathbf{V}_i$ is the design effect matrix for the ith population. The d's can be regarded as design effects of the Z_i's again and we note that

$$d_{r+s-1} \leqslant \frac{n_2 d_{1r} + n_1 d_{2s}}{n_1 + n_2} \quad (r + s \leqslant k), \tag{3.7}$$

where $d_{i1}, d_{i2}, \ldots$ are the ordered eigenvalues of $\mathbf{D}_i$ ($i = 1, 2$). In particular, with proportionally allocated stratified sampling $d_{ij} \leqslant 1$ for all i, j so that $d_j \leqslant 1$ and the ordinary chi-squared test is conservative.

The comments about approximations to the distribution in the goodness-of-fit case carry over directly here. The simple device of dividing $\bar{X}^2_H$ by $\bar{d}$ and treating the resulting statistic, $\bar{X}^2_{Hm}$ say, as χ^2_{k-1} under H_0 gives a natrual approximation (although there are even more accurate modifying factors if enough is known about the d_i's). All of the empirical results presented in the next section indicate that $\bar{X}^2_{Hm}$ is a very satisfactory approximation.

Note that

$$\bar{d} = \frac{\text{tr}(\mathbf{D}_H)}{(k-1)} = \frac{n_2 \,\text{tr}(\mathbf{D}_1) + n_1 \,\text{tr}(\mathbf{D}_2)}{(n_1+n_2)(k-1)}$$
$$= \frac{n_2 \bar{d}_{1.} + n_1 \bar{d}_{2.}}{n_1+n_2} \tag{3.8}$$

Now $\bar{d}_{1.}$ and $\bar{d}_{2.}$ can be calculated from the diagonal elements of $\mathbf{V}_1$ and $\mathbf{V}_2$ respectively so $\bar{d}$ only requires a knowledge of the cell variances for the two populations. We look at values of $\bar{d}$ for several alternative designs with the GHS data in the next section. Notice that if one sample is much smaller than the other then $\bar{d}$ is essentially equal to the design effect of the smaller sample (this reduces to the one sample goodness-of-fit problem in the limit).

We can extend these results directly to the problem of testing the homogeneity of r populations given independent samples from each population. Let $\bar{\mathbf{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{i,k-1})^{\mathrm{T}}$ denote the vector of estimated proportions for the ith sample and suppose that

$$\sqrt{n_i}(\mathbf{p}_i - \mathbf{p}_i) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_i) \tag{3.9}$$

as $n_i \to \infty$. We want to test the hypothesis

$$H_0\colon \mathbf{p}_i = \mathbf{p} \quad (i = 1, \ldots, r), \tag{3.10}$$

The usual chi-squared test of homogeneity is

$$\bar{X}^2_H = \sum_{i=1}^{r} \sum_{j=1}^{k} n_i \frac{(p_{ij} - \hat{p}_j)^2}{\hat{p}_j}, \tag{3.11}$$

where $\hat{p}_j = \Sigma_i n_i \hat{p}_{ij} / \Sigma n_i$. This has a $\chi^2_{(r-1)(k-1)}$ distribution under H_0 with independent multinominal sampling in each population. As in previous sections we want to study the behaviour of $\bar{X}^2_H$ under more general sampling schemes. It is shown in Appendix 2 that

$$\bar{X}^2_H = \sum_{i=1}^{(r-1)(k-1)} d_i Z_i^2, \tag{3.12}$$

where $Z_1, Z_2, \ldots$ are asymptotically independent $N(0, 1)$ and the d_i's are the non-zero eigenvalues of

$$\mathbf{A} = \begin{pmatrix} (1-f_1)\,\mathbf{D}_1 & -f_1\,\mathbf{D}_2 & \ldots & -f_1\,\mathbf{D}_r \\ -f_2\,\mathbf{D}_1 & (1-f_2)\,\mathbf{D}_2 & \ldots & -f_2\,\mathbf{D}_r \\ \vdots & \vdots & & \vdots \\ -f_r\,\mathbf{D}_1 & \ldots & \ldots & (1-f_r)\,\mathbf{D}_r \end{pmatrix}, \tag{3.13}$$

where $f_i = n_i/n$ and $\mathbf{D}_i = \mathbf{P}^{-1}\mathbf{V}_i$ is the design effect matrix for the ith population ($i = 1, \ldots, r$). As usual the ordinary chi-squared test is conservative with proportionally allocated stratified sampling.

For more complex designs, we can fall back on the modified test statistic $\bar{X}^2_H/\bar{d}$. Now,

$$(r-1)(k-1)\bar{d} = \text{tr}(\mathbf{A}) = \sum_{i=1}^{r} (1-f_i)\,\text{tr}(\mathbf{D}_i) \tag{3.14}$$

so that

$$\bar{d} = \sum_{1}^{r} \frac{(1-f_i)}{r-1} \bar{d}_{i\cdot}$$

where $\bar{d}_{i\cdot}$ is the average design effect for the ith population. Thus $\bar{d}$ can still be calculated simply from information about cell variances for each population. Notice that as r becomes large $\bar{d}$ will tend towards the unweighted average of the $\bar{d}_{i\cdot}$'s provided no single n_i dominates the others. Notice also that $\bar{d}$ is simply a weighted average of population design effects and should stay relatively stable as r increases. The number of degrees of freedom, $(r-1)(k-1)$, increases with r, however, so that the numerical results of the previous section would lead us to expect the performance of the ordinary chi-squared test to get much worse as r increases.

3.2. *Empirical Results for Tests of Homogeneity*

To investigate the effect of different survey designs on the usual $\bar{X}^2$ test, and on the modified procedure, we treated the GHS data set as a clustered population with 345 clusters and looked at 5 different schemes for drawing a sample of 1000 households from it. Sampling from a common population ensured that the null hypothesis was true, and the underlying vector **p** (and hence **P**) is known exactly for each variable. The 5 designs considered were

A: Two stage pps sampling with 50 psu's and 20 observations per psu.
B: Two stage pps sampling with 100 psu's and 10 observations per psu.
C: Two stage pps sampling with 200 psu's and 5 observations per psu.
D: Simple random sampling with 1000 observations.
E: Proportionally allocated stratified sampling using the original 141 strata of the GHS.

The population defined by the GHS data set has somewhat different structure to that of the original population of the United Kingdom from which it was drawn, since it is the result of a multistage pps survey. Nevertheless, it is felt that the results give sufficient indication of the effects which might be obtained in practice. Survey designs D and E would not usually be used since cost considerations would call for some form of multi-stage sampling but they are included for comparative purposes. The three clustered designs are by no means extreme and are typical of common survey designs. The design effects may be slightly inflated since no account has been taken of the stratification which would usually be used in conjunction with such designs. This is not likely to invalidate the results. For variable G3 (Home ownership), for example, the average design effect computed in Section 2 taking the GHS design into account was 2·54 which is comparable with the average design effects for this variable under different designs presented in Table 7. In all, 7 variables were considered and the corresponding design effects are set out in Table 7.

TABLE 7
Average design effects for the five designs

		Design				
Variable	*d.f.*	*A*	*B*	*C*	*D*	*E*
Type of accommodation (G4)	3	3·64	2·25	1·56	1·00	0·91
Age of building (G2)	2	4·09	2·46	1·65	1·00	0·91
No. of bedrooms (G7)	4	2·33	1·63	1·28	1·00	0·96
Home ownership (G3)	3	3·39	2·13	1·50	1·00	0·92
Bedroom standard (G11)	4	1·96	1·45	1·20	1·00	0·97
Number of rooms (G8)	3	2·66	1·79	1·35	1·00	0·94
Number of cars (G6)	2	2·55	1·73	1·33	1·00	0·95

TABLE 8

Significance level for a nominal 5 per cent test based on $\bar{X}^2$

	Designs														
Variable	*AA*	*AB*	*AC*	*AD*	*AE*	*BB*	*BC*	*BD*	*BE*	*CC*	*CD*	*CE*	*DD*	*DE*	*EE*
G4	51	43	38	34	32	32	26	19	18	17	11	10	5	5	4
G2	48	40	35	30	29	29	23	18	17	15	11	10	5	4	4
G7	38	32	27	23	22	22	17	13	13	12	8	8	5	5	4
G3	49	40	36	30	29	29	23	18	16	15	11	10	5	4	4
G11	30	23	21	18	17	17	14	10	10	10	7	7	5	5	5
G8	39	33	28	23	22	22	17	13	12	12	9	9	5	5	5
G6	30	24	22	18	18	17	14	12	11	11	8	8	5	5	4

TABLE 9

Significance level for a nominal 5 per cent test for design B with several populations

			No. of populations				
	d.f.	$\bar{d}$	2	3	4	5	10
G2	2	2·46	29	42	56	59	85
G3	3	2·13	29	45	52	61	86
G11	4	1·45	17	22	27	31	79

It turned out that the size of the modified test based on $\bar{X}^2_H/\bar{d}$ was always between 5 and 6 per cent so we omit further details. The actual size of the ordinary chi-squared test at a nominal 5 per cent level is given in Table 8 for all possible pairs of designs.

The main feature of the results is the severe distortion which can occur with any of the clustered designs. Since none of those designs is particularly extreme it follows that any application of a standard test of homogeneity to survey data should be viewed with great suspicion. The results become even worse with more than two populations. Taking design B as being fairly typical of many real surveys, we show in Table 9 what happens to the significance level of the ordinary chi-squared test when the same design is used for r different populations whose structures are similar enough to regard the design effects as constant. As expected the level rises steadily as the number of populations increases.

Finally for variable G5 (Head of household gross weekly income) we have investigated the effect of collapsing for the two sample case for the possible combinations of designs A–E. The results are presented in Tables 10 and 11 and as for the goodness of fit case we see that collapsing categories leads to an increases in $\bar{d}$ but this is more than compensated for by the reduction in

TABLE 10

Average design effects for G5 (Head of household gross weekly income) for the five designs for various degrees of freedom

	Design				
d.f.	*A*	*B*	*C*	*D*	*E*
9	1·86	1·41	1·18	1·00	0·98
7	1·92	1·43	1·19	1·00	0·98
5	2·00	1·47	1·21	1·00	0·98
3	2·25	1·59	1·26	1·00	0·97
2	2·47	1·70	1·31	1·00	0·96

TABLE 11

Significance level for a nominal 5 per cent test based on $\bar{X}^2$ for G5 (Head of household gross weekly income) for various degrees of freedom

d.f.	*AA*	*AB*	*AC*	*AD*	*AE*	*BB*	*BC*	*BD*	*BE*	*CC*	*CD*	*CE*	*DD*	*DE*	*EE*
9	41	34	25	22	22	21	18	12	11	11	9	9	5	5	5
7	39	31	25	22	21	21	17	12	11	11	8	8	5	5	5
5	38	27	24	21	20	20	16	11	11	10	8	7	5	5	5
3	33	26	22	19	19	19	15	11	11	10	8	7	5	5	5
2	30	24	20	18	18	18	14	11	11	10	8	7	5	5	5

degrees of freedom so that the distortion of the significance level for $\bar{X}^2$ is reduced for fewer degrees of freedom.

4. TESTING INDEPENDENCE IN A TWO-WAY CONTINGENCY TABLE

4.1. *Asymptotic Theory*

We now turn to the most important (certainly the most common in practice) of the three problems examined in the paper, that of testing for independence in a two-way table. The general theory has been outlined by Rao and Scott (1979) and we give a brief review of their results here. Suppose the table has r rows and c columns, and let $\mathbf{p} = (p_{11}, p_{12}, \ldots, p_{rc})^{\mathrm{T}}$ denote the vector of cell probabilities ($\Sigma_1^r \Sigma_1^c p_{ij} = 1$). For reasons of symmetry it is convenient to include all rc cell probabilities in the definition of $\mathbf{p}$ in this section. As usual we suppose that we have estimated probabilities $\hat{\mathbf{p}} = (\hat{p}_{11}, \ldots, \hat{p}_{rc})^{\mathrm{T}}$ and that

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}). \tag{4.1}$$

Notice that since $\mathbf{p}$ has dimension rc and $\sum^{rc} p_{ij} = 1$, the covariance matrix $\mathbf{V}$ will be singular throughout this section. We also define the vectors of marginal probabilities $\mathbf{p}_r = (p_{1+}, \ldots, p_{(r-1)+})^{\mathrm{T}}$, where $p_{i+} = \Sigma_{j=1}^c p_{ij}$, and $\mathbf{p}_c = (p_{+1}, \ldots, p_{+(c-1)})^{\mathrm{T}}$, where $p_{+j} = \Sigma_{i=1}^r p_{ij}$.

The null hypothesis of interest is that of independence between rows and columns, i.e.

$$H_0\colon p_{ij} = p_{i+}\, p_{+j} \quad (i = 1, \ldots, r;\ j = 1, \ldots, c).$$

It is natural to consider tests based on the quantities

$$h_{ij}(\hat{\mathbf{p}}) = \hat{p}_{ij} - \hat{p}_{i+}\, \hat{p}_{+j} \quad (i = 1, \ldots, r-1;\ j = 1, \ldots, c-1) \tag{4.2}$$

Let $\mathbf{h}(\mathbf{p}) = (h_{11}(\mathbf{p}), h_{12}(\mathbf{p}), \ldots, h_{(r-1)(c-1)}(\mathbf{p}))^{\mathrm{T}}$ and let $H(\mathbf{p})$ denote the $(r-1)(c-1) \times rc$ matrix of partial derivatives $H(\mathbf{p}) = \partial \mathbf{h}(\mathbf{p})/\partial \mathbf{p}$. Then under the assumptions above, $\sqrt{n}(h(\hat{\mathbf{p}}) - (\mathbf{p} - \mathbf{p}_0))$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{HVH}^{\mathrm{T}}$, where $\mathbf{p}_0 = (p_{1+}\, p_{+1}, \ldots, p_{r+}\, p_{+c})^{\mathrm{T}}$. If a consistent estimator of $\mathbf{V}$ is available, we can base a test on the Wald statistic

$$\bar{X}_W^2 = n\mathbf{h}^{\mathrm{T}}(\hat{\mathbf{p}})(\hat{\mathbf{H}}\hat{\mathbf{V}}\hat{\mathbf{H}}^{\mathrm{T}})^{-1}\mathbf{h}(\hat{\mathbf{p}}), \tag{4.3}$$

where $\hat{\mathbf{H}} = \mathbf{H}(\hat{\mathbf{p}})$, which is asymptotically $\chi^2_{(r-1)(c-1)}$ under H_0, or on a modified version with $\mathbf{V}$ (and $\mathbf{H}$) estimated under H_0. With random sampling, when the distribution is multinomial, $\mathbf{HVH}^{\mathrm{T}}$ reduces to $\mathbf{P}_r \otimes \mathbf{P}_c$ under the null hypothesis $\mathbf{p} = \mathbf{p}_0$, where $\mathbf{P}_r = \mathrm{diag}(\mathbf{p}_r) - \mathbf{p}_r \mathbf{p}_r^{\mathrm{T}}$ and $\mathbf{P}_c = \mathrm{diag}(\mathbf{p}_c) - \mathbf{p}_c \mathbf{p}_c^{\mathrm{T}}$.

When no estimate of $\mathbf{V}$ is available it is very common to ignore the sample structure and use a test based on the ordinary chi-squared statistic:

$$\bar{X}_I^2 = n\sum_1^r\sum_1^c \frac{(\hat{p}_{ij}-\hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}. \tag{4.4}$$

We note that $\bar{X}_I^2$ can be written in the form

$$\bar{X}_I^2 = n\mathbf{h}^{\mathrm{T}}(\hat{\mathbf{p}})(\hat{\mathbf{H}}\hat{\mathbf{P}}_0\hat{\mathbf{H}}^{\mathrm{T}})^{-1}\mathbf{h}(\hat{\mathbf{p}}) \tag{4.5}$$

so the ordinary statistic is just a special case of a modified Wald statistic. As in previous sections, the asymptotic behaviour of $\bar{X}_I^2$ under a general sampling scheme is relatively simple to obtain; $\bar{X}_I^2$ can be represented as

$$\bar{X}_I^2 = \sum_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2, \tag{4.6}$$

where the Z_i's are asymptotically independent $N(0, 1)$ under H_0 and the δ_i's are the eigenvalues of

$$\mathbf{D}_I = (\mathbf{H}\mathbf{P}_0\mathbf{H}^{\mathrm{T}})^{-1}(\mathbf{H}\mathbf{V}\mathbf{H}^{\mathrm{T}}). \tag{4.7}$$

Again the δ_i's can be interpreted as design effects (in this case, of the components of $\mathbf{H}\hat{\mathbf{p}}$). As usual this means that the ordinary chi-squared test is conservative for proportionally allocated stratified sampling. Since the gains from proportional allocation are generally rather modest we would expect the ordinary chi-squared test to perform very well in this special case. This approximate validity has been noted in empirical studies by Kish and Frankel (1974) and Nathan (1975) and the result above gives a formal explanation of their results.

With more general sampling schemes, we get a good approximation by treating $\bar{X}_I^2/\bar{\delta}$ as $\chi^2_{(r-1)(c-1)}$ under H_0. This requires an estimate of $\bar{\delta}$, however, and here $\bar{\delta}$ is not a function just of the diagonal elements of $\mathbf{V}$. We can calculate $\bar{\delta}$ from the variances of the $h_{ij}(\hat{\mathbf{p}})$'s but as far as we know this information is not available at present for any major survey. If independence is of primary importance there may be a case for calculating these variances directly. Even the diagonal elements of $\mathbf{V}$ in a large cross-classification are rarely estimated so the average design effect may not be available but, more seriously, the relationship between this average and $\bar{\delta}$ appears very complicated for general $\mathbf{V}$. At best, the only information available in practice seems to be the variances of the row and column marginals $\hat{p}_{i+}$ and $\hat{p}_{+i}$. Altham (1976), Cohen (1976) and Brier (1979) have proposed models in which all cells and all linear combinations of the cells have a common design effect. In this case $\bar{\delta}$ and all marginal design effects are equal to this common value and we could estimate $\bar{\delta}$ by some average of the marginal effects. Fellegi (1978) has suggested using the average cell design effect as a divisor and in the case of a common design effect this too would be appropriate. The results in Section 3.2 indicate that Fellegi's suggestion will give good results for testing homogeneity and we might anticipate that this would carry across to independence in view of the close relationship between the two problems under the multinomial distribution. Unfortunately the assumption of a common design effect does not seem very realistic in many cross-classifications. For example Kish *et al.* (1976) report average design effects for socio-economic characteristics in the range 4–8 whereas demographic variables range from 1·0 to 1·6. Thus a two-way table formed from one socio-economic and one demographic variable would not yield a single common design effect. More importantly perhaps, the limited results available on values of $\bar{\delta}$ suggest that this tends to be smaller than the design effects for individual cells. Kish and Hess (1959) explore this point in some detail for 2×2 tables and Kish and Frankel (1974) and Nathan (1975) have some further results. Additional results of empirical studies presented in the next section confirm this tendency.

4.2. *Empirical Results for Testing Independence*

To get some feeling for the size of $\bar{\delta}$, and of its relationship to the cell and marginal design effects, we examined a large number of cross-classifications of variables from both the BES and

TABLE 12
Observed and estimated proportions under independence for G1 vs G2

		G2		
		Rented	*Owned*	
G1:	Post 1919	0·36 (0·35)	0·33 (0·35)	0·69
	Pre 1919	0·14 (0·15)	0·17 (0·15)	0·31
		0·50	0·50	

GHS data sets. In all cases studied the spread of the δ_i's was small and treating $\bar{X}^2/\bar{\delta}$ as approximately $\chi^2_{(r-1)(c-1)}$ under H_0 gave a very good approximation so we have reported only the average value, $\bar{\delta}$, here. The marginal design effects for the individual variables are contained in Tables 3 and 4. We show values of $\bar{d}$, calculated from the individual cell variances $\hat{V}_{ii}$, and also values of the Wald statistic to give some feeling for the degree of dependence between the variables. It must be noted that the sample sizes are large in both surveys so that very small deviations of $\hat{p}_{ij}$ from $\hat{p}_{i+}\,\hat{p}_{+j}$ can produce quite large values of $\bar{X}^2_w$. For example, the cell and marginal proportions for the variables G1 and G2 (shown in Table 12) give a value 22·87 for $\bar{X}^2_w$ with one degree of freedom although the observed proportions are very close to those expected under independence. Table 13 show values of $\bar{\delta}$, $\bar{d}$ and $\bar{X}^2_w$ for cross-classified BES variables. The table also contains the estimated sizes of tests based on $\bar{X}^2$, $\bar{X}^2/\bar{d}$ and $\bar{X}^2/\bar{d}_m$ where $\bar{d}_m$ is the smaller of the two average marginal design effects. The outstanding feature of the table is obviously the small values for $\bar{\delta}$. The $\bar{X}^2$ test needs no modification at all for these variables.

The main reason for considering $\bar{X}^2/d$ and $\bar{X}^2/\bar{d}_m$ is that $\bar{\delta}$ is rarely available at present. We feel that there is a greater possibility of $\bar{d}$ having been calculated although more frequently it will be only the marginal design effects which are available. The best that can be said about the modified tests is that they are both conservative, with the test based on $\bar{d}$ slightly better than the one based on the marginal effects. Notice that the pattern does not seem to change with different values of $\bar{X}^2_w$; $\bar{\delta}$ is close to one whether the variables involved are approximately independent or not.

The corresponding results for the GHS data are shown in Table 14. Here the relationships are a little more complicated; the value of $\bar{\delta}$ is still much smaller than $\bar{d}$ or $\bar{d}_m$ and still very close to one in most cases. Thus $\bar{X}^2/\bar{d}$ and $\bar{X}^2/\bar{d}_m$ give very conservative tests in general (and hence must sacrifice a fair amount of power). However, some cross-classifications give a high value of $\bar{\delta}$ and the naive chi-squared test is not adequate in these cases. Some warning is given by the marginal design effects, since the high $\bar{\delta}$ values occur with the cross-classifications of the subset (G1, G2, G9, G12) which are the variables with the highest individual effects, although high values of the marginal design effects need not indicate a high value of $\bar{\delta}$. Moreover, high values of $\bar{\delta}$ tend to be associated with high values of $\bar{X}^2_w$ and hence may not reflect behaviour under the null hypothesis, which is our main concern. Certainly, $\bar{\delta}$ is close to one in all cases in which $\bar{X}^2_w \leqslant \chi^2_{(0\cdot99)}$. Thus the unmodified $\bar{X}^2$ test works well in every case for which the null hypothesis of independence seems reasonable.

Of course, the overall pattern of results will not seem particularly surprising to many practitioners. The 2×2 table is a special case of the problem of comparing two sub-class means which has been discussed extensively (see Kish and Frankel, 1974; Kish *et al.*, 1976). It is well known that when the sub-classes cut across strata and clusters, the design effect for the difference between two sub-class means tends to be much smaller than that for the individual means. This ties in exactly with the conclusion that $\bar{\delta}$ is smaller than $\bar{d}_m$ in our results for independence. If the sub-classes can be placed in separate strata, however, the estimates of the two means are independent and there is no attenuation at all in the design effect of the difference. This

TABLE 13

Design effects for BES variables and significance levels for various nominal 5 per cent test procedures

Variables	*d.f.*	$\bar{\delta}$	$\bar{d}$	$\bar{X}^2_w$	*Size* $\bar{X}^2$	*Size* $\bar{X}^2/\bar{\delta}$	*Size* $X^2/\bar{d}$	*Size* $\bar{X}^2/d_m$
B2 × B3	3	1·30	1·88	41·40	11	5	1	1
B2 × B5	2	1·29	1·76	127·28	10	5	2	3
B2 × B7	3	1·41	1·81	5·56	13	5	2	4
B2 × B10	2	1·22	1·72	32·59	9	5	2	4
B2 × B12	2	1·17	1·73	0·31	8	5	1	4
B2 × B14	5	1·18	1·39	33·76	9	5	2	6
B2 × B15	1	1·16	2·08	1·77	7	5	1	4
B2 × B16	1	0·96	1·98	0·02	5	5	0	4
B3 × B5	6	1·03	1·31	10·81	6	5	2	1
B3 × B7	9	1·13	1·37	6·86	10	6	2	1
B3 × B10	6	1·03	1·30	7·84	6	5	2	2
B3 × B12	6	1·07	1·33	11·53	7	5	2	3
B3 × B14	15	1·01	1·15	49·83	6	6	2	3
B3 × B15	3	1·16	1·47	5·79	9	5	2	4
B3 × B16	3	1·09	1·44	5·23	7	5	2	6
B5 × B7	6	0·97	1·19	20·59	5	5	2	1
B5 × B10	4	1·07	1·24	19·74	7	5	3	2
B5 × B12	4	1·07	1·21	7·02	7	5	3	3
B5 × B14	10	1·09	1·14	29·20	9	5	5	5
B5 × B15	2	0·84	1·21	2·24	3	5	2	1
B5 × B16	2	1·12	1·30	146·75	7	5	3	6
B7 × B10	6	1·03	1·19	17·89	6	5	3	2
B7 × B12	6	0·94	1·16	18·35	4	5	2	1
B7 × B14	15	1·07	1·12	653·96	8	6	4	4
B7 × B15	3	1·00	1·26	0·58	5	5	3	3
B7 × B16	3	1·02	1·25	2·52	6	5	3	5
B10 × B12	4	1·02	1·14	6·89	5	5	3	2
B10 × B14	10	1·05	1·11	40·10	7	5	5	5
B10 × B15	2	0·97	1·13	4·94	5	5	3	2
B10 × B16	2	1·02	1·15	18·68	5	5	4	5
B12 × B14	10	1·01	1·07	39·38	6	5	5	3
B12 × B15	2	1·03	1·14	9·03	6	5	4	3
B12 × B16	2	0·96	1·09	7·93	5	5	4	4
B14 × B15	5	0·99	1·07	7·30	5	5	3	3
B14 × B16	5	0·94	1·04	16·07	4	5	3	3
B15 × B16	1	1·02	1·14	2·39	5	5	4	5

corresponds to our results for homogeneity. Clearly we need more empirical evidence before drawing firm conclusions in general. Both the surveys used for the empirical study have very similar structure and look at similar types of variable. Variables from other subject matter areas may show different patterns. In particular, both these surveys are approximately self-weighting and it is part of the folklore of survey methodology that standard techniques tend to hold up better with self-weighting designs. It would be very interesting to have some results for a non-epsem sample.

In summary, then, the $\bar{X}^2$ test for independence seems to be much less vulnerable than the corresponding tests for goodness-of-fit and homogeneity provided the variables cut across strata and clusters. If both variables have high marginal design effects, it may be a warning that some modification of the test is needed. Dividing by the smaller of the two marginal effects gives a conservative test, but this modification will generally be too severe and some power will be

TABLE 14

Design effects for GHS variables and significance levels for various nominal 5 per cent test procedures

Variable	*d.f.*	$\bar{\delta}$	$\bar{d}$	$\bar{X}^2_w$	*Size* $\bar{X}^2$	*Size* $\bar{X}^2/\bar{\delta}$	*Size* $\bar{X}^2/\bar{d}$	*Size* $\bar{X}^2/\bar{d}_m$
G1 × G2	3	1·96	2·44	878·2	26	6	3	1
G1 × G3	3	2·04	2·60	585·6	28	5	3	1
G1 × G4	6	1·96	2·29	951·8	62	6	3	1
G1 × G8	6	1·57	2·09	895·4	23	6	2	4
G1 × G9	15	0·93	1·12	1341·2	5	7	2	2
G1 × G10	6	0·92	1·61	287·7	4	5	0	1
G1 × G12	3	1·06	2·00	306·1	6	5	0	5
G1 × G13	9	0·87	1·37	137·3	3	5	1	2
G2 × G3	1	1·99	3·18	22·9	16	5	1	1
G2 × G4	2	1·97	2·36	651·3	22	5	3	2
G2 × G8	2	1·24	1·98	309·7	9	5	1	2
G2 × G9	5	0·91	1·23	242·7	4	5	2	2
G2 × G10	2	0·97	1·75	156·4	5	5	1	2
G2 × G12	1	0·99	2·13	178·2	5	5	0	3
G2 × G13	3	0·93	1·41	261·9	4	5	1	5
G3 × G4	2	1·94	2·49	592·4	21	5	3	2
G3 × G8	2	1·41	1·86	883·8	12	5	2	3
G3 × G9	5	1·02	1·18	359·8	6	5	3	4
G3 × G10	2	1·13	1·61	20·2	8	5	2	4
G3 × G12	1	0·98	1·80	5·7	5	5	1	3
G3 × G13	3	1·01	1·34	52·6	5	5	2	5
G4 × G8	4	1·26	1·72	289·1	11	5	1	1
G4 × G9	10	0·93	1·14	242·8	3	5	2	2
G4 × G10	4	0·96	1·51	50·8	5	5	1	2
G4 × G12	2	0·96	1·73	27·5	5	5	1	4
G4 × G13	6	0·96	1·30	169·76	5	5	2	5
G8 × G9	10	0·94	1·05	805·9	5	5	3	2
G8 × G10	4	0·93	1·21	326·0	4	5	2	1
G8 × G12	2	0·86	1·21	1445·1	3	5	2	2
G8 × G13	6	0·94	1·10	298·8	4	5	3	4
G9 × G10	10	0·85	0·94	2934·5	3	5	4	1
G9 × G12	5	0·83	0·96	2241·3	2	5	3	2
G9 × G13	15	0·82	0·91	383·5	2	6	3	1
G10 × G12	2	0·88	1·02	2852·4	4	5	4	3
G10 × G13	6	0·94	0·98	3894·0	4	5	5	4
G12 × G13	3	0·74	0·91	694·8	2	5	3	2

lost. The loss of power may not be very important if the sample size is very large, but it could be crucial in a small research survey and much more work needs to be done on finding a better test that can be calculated without too much information about **V** (or $\mathbf{HVH}^{\mathrm{T}}$).

ACKNOWLEDGEMENTS

The support of the SSRC through Grant No. HR 4973/1 is gratefully acknowledged. The authors would like to thank the Social Science Survey Archive and the Office of Population Censuses and Surveys for making available the data used.

REFERENCES

ALTHAM, P. A. E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika*, **63**, 263–269.

ARMITAGE, P. (1966). The chi-squared test for heterogeneity of proportions after adjustment for stratification. *J. R. Statist. Soc.* B, **28**, 150–163.

BHAPKER, V. P. and KOCH, G. (1968). On the hypothesis of no interaction in contingency tables. *Biometrics*, **24**, 567–594.
BRACKSTONE, G. J. and GOSSELIN, J. F. (1973). 1971 Census Evaluation Programme, Reverse Record Check; Analysis of Missed Persons: Method of Analysis and Preliminary Results. Statistics Canada Report.
BRIER, S. S. (1979). Categorical data models for complex sampling schemes. Ph.D. Thesis, University of Minnesota.
CASPARIS, T. and VAZ, R. W. (1973). Social class and self-reported delinquent acts among Swiss boys. *Int. J. of Comp. Sociol*, **14**, 47–58.
CHASE, I. D. (1975). A comparison of men's and women's intergenerational mobility in the United States. *Amer. Sociol. Rev.*, **40**, 483–505.
COHEN, J. E. (1976). The distribution of the chi-squared statistic under cluster sampling from contingency tables. *J. Amer. Statist. Ass.*, **71**, 665–670.
CREWE, I., SARLVIK, B. and ALT, J. (1974). *The British Election Study at the University of Essex, February Cross-section Sample—Codebook for the Survey.*
EWINGS, P. D. (1979). The effect of complex sample designs on the chi-squared goodness of fit statistic. M.Sc. Dissertation, University of Southampton.
FELLEGI, I. P. (1978). Approximate tests of independence and goodness of fit based upon stratified multi-stage samples. *Survey Methodol.* (Statistics Canada), **4**, No. 1, 29–56.
FIENBERG, S. E. (1979). The use of chi-squared statistics for categorical data problems. *J. R. Statist. Soc.* B, **41**, 54–64.
GRAD, A. and SOLOMON, H. (1955). Distributions of quadratic forms and some applications. *Ann. Math. Statist.*, **26**, 464–477.
JOHNSON, N. L. and KOTZ, S. (1968). Tables of distributions of positive definite quadratic forms in central normal variables. *Sankhyā*, B, **30**, 303–314.
—— (1970). *Distributions in Statistics: Continuous Univariate Distributions—2.* Boston: Houghton Mifflin.
KISH, L. and FRANKEL, M. R. (1974). Inference from complex samples (with Discussion). *J. R. Statist. Soc.* B, **36**, 1–37.
KISH, L., GROVES, R. M., and KROTKI, K. P. (1976). Sampling errors for fertility surveys. Occasional paper no. 17: World Fertility Survey, London.
KISH, L. and HESS, I. (1959). Some sampling techniques for continuing survey operations. *Proc. Soc. Statist. Sec.*, American Statistical Association, pp. 139–143.
KOCH, G. G., FREEMAN, D. H. and FREEMAN, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *Int. Stat. Rev.*, **43**, 59–78.
NATHAN, G. (1973). Approximate tests of independence in contingency tables from complex stratified samples. *N.C.H.S. Vital and Health Statistics, Ser. 2, No. 53*, Washington, D.C.
—— (1975). Tests of independence in contingency tables from stratified proportional samples. *Sankhyā* C, **37**, 77–87.
OFFICE OF POPULATION, CENSUSES AND SURVEYS (1973). *General Household Survey 1973: Annual Report.* London: H.M.S.O.
PERUCCI, C. C. (1978). Income attainment of college graduates: a comparison of employed women and men. *Sociol. Social Res.*, **63**, 3.
RAO, J. N. K. and SCOTT, A. J. (1979). The analysis of categorical data from complex sample surveys. Presented to American Statistical Association Conference, Washington, D.C.
SATTERTHWAITE, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, **2**, 110–114.
SHUSTER, J. J. and DOWNING, D. J. (1976). Two-way contingency from complex sampling schemes. *Biometrika*, **63**, 271–276.
SOLOMON, H., and STEPHENS, M. A. (1977). Distributions of a sum of weighted chi-square variables. *J. Amer. Statist. Ass.*, **72**, 881–885.

APPENDIX 1
Variables used in the Empirical Studies

British Election Study

B1 Residence type (4 categories)
B2 Type of home ownership/rental (4 categories)
B3 Respondent's feeling about neighbourhood (4 categories)
B4 Area grew up in (9 regions, Wales, Scotland, Ireland, other)
B5 Social Grade (5 categories)
B6 Respondent's view of conservatives on 5 point scale from extreme to moderate
B7 Length of residence (4 categories)
B8 Father's party choice (3 categories)
B9 Respondent's view of social services on 3 point scale
B10 Respondent's view of how much Britain can depend on the United States to look at world politics the same way we do, on 3 point scale
B11 Martial status (4 categories)

B12 Respondent's view on whether he/she is better off now than a year or two ago, on a 3 point scale
B13 Marks out of 10 for Liberals (0–10)
B14 Year of birth (6 categories)
B15 Respondent's view of whether the recent election campaign gave people the facts about the problems facing the country (2 categories)
B16 Sex (2 categories)

General Household Survey
G1 Nett rateable value (4 categories)
G2 Age of building (3 categories
G3 Homeownership/rental (4 categories)
G4 Type of accommodation (4 categories)
G5 Head of household gross weekly income (10 categories)
G6 No. of cars (0–3)
G7 No. of bedrooms (0–6)
G8 No. of rooms (1–10)
G9 Household gross weekly income (6 catgories)
G10 Age of head of household (3 categories)
G11 Bedroom standard (5 categories)
G12 No. of persons in household (2 categories)
G13 Head of household length of residence (4 categories)

APPENDIX 2

The Asymptotic Distribution of $\bar{X}^2_H$

Let $\hat{\mathbf{p}}_i$ represent the vector of estimated proportions for the ith population as in Section 3.1, and suppose the sample sizes n_i increase together in such a way the $n_i/\Sigma_1 n_i \to f_i$ with $0<f_1<1$ $(i=1,\ldots,r)$. Suppose also that $\sqrt{n_i}(\hat{\mathbf{p}}_i-\mathbf{p}) \xrightarrow{L} N(\mathbf{0},\mathbf{V}_i)$ as $n_i \to \infty$. Then, if we let $\hat{\mathbf{p}}_0$ be the $r(k-1)$ dimensional vector defined by $\hat{\mathbf{p}}_0^{\mathrm{T}} = (\hat{\mathbf{p}}_1^{\mathrm{T}},\ldots,\hat{\mathbf{p}}_r^{\mathrm{T}})$, it follows that $\sqrt{n}(\hat{\mathbf{p}}_0-\mathbf{p}_0) \xrightarrow{L} N(\mathbf{0},\mathbf{V}_0)$ as $n \to \infty$ where

$$n=\sum_1^r n_i, \quad \mathbf{p}_0^{\mathrm{T}}=(\mathbf{p}^{\mathrm{T}},\mathbf{p}^{\mathrm{T}},\ldots,\mathbf{p}^{\mathrm{T}}) \quad \text{and} \quad \mathbf{V}_0=\bigotimes_1^r(\mathbf{V}_i/f_i).$$

Now consider the ordinary chi-squared statistic for the hypothesis of homogeneity,

$$\bar{X}^2_H=\sum_i\sum_j n_i\frac{(\hat{p}_{ij}-\hat{p}_j)^2}{\hat{p}_j},$$

where $\hat{p}_j=\Sigma_{i=1}^r n_i\hat{p}_{ij}/n$. It is straightforward to show that $\bar{X}^2_H$ has the same asymptotic distribution as

$$\bar{X}^{*2}_H=\sum_i\sum_j n_i\frac{(\hat{p}_{ij}-\hat{p}_j)^2}{p_j}$$

under the assumptions above, and we can re-write $\bar{X}^{*2}_H$ in the form

$$\bar{X}^{*2}_H=n(\hat{\mathbf{p}}_0-\mathbf{p}_0)^{\mathrm{T}}\mathbf{B}(\hat{\mathbf{p}}_0-\mathbf{p}_0),$$

where $\mathbf{B}=\mathbf{F}\otimes\mathbf{P}^{-1}$ with $\mathbf{P}=\mathrm{diag}(\mathbf{p})-\mathbf{p}\mathbf{p}^{\mathrm{T}}$, $\mathbf{F}=\mathrm{diag}(\mathbf{f})-\mathbf{f}\mathbf{f}^{\mathrm{T}}$ and $\mathbf{f}^{\mathrm{T}}=(f_1,\ldots,f_r)$. Note that rank $\mathbf{B}=$ rank $(\mathbf{F})\times$ rank $(\mathbf{P})=(r-1)(k-1)$.

It follows immediately from standard results on quadratic forms (Johnson and Kotz, 1970, p. 150) that

$$\bar{X}_H^{*2} = \sum_{1}^{(r-1)(k-1)} d_i Z_i^2,$$

where $Z_1, Z_2, \ldots$ are asymptotically independent $N(0, 1)$ and d_1, d_2, ... are the non-zero eigenvalues of

$$\mathbf{A} = \mathbf{B}\mathbf{V}_0 = \begin{pmatrix} (1-f_1)\mathbf{D}_1 & -f_1\mathbf{D}_2 & -f_1\mathbf{D}_r \\ -f_2\mathbf{D}_1 & (1-f_2)\mathbf{D}_2 & \\ -f_r\mathbf{D}_1 & & (1-f_r)\mathbf{D}_r \end{pmatrix},$$

where $\mathbf{D}_i = \mathbf{P}^{-1}\mathbf{V}_i (i = 1, \ldots, r)$.