APPROXIMATE TESTS OF INDEPENDENCE AND GOODNESS OF FIT BASED ON STRATIFIED MULTI-STAGE SAMPLES

I.P. Fellegi[1]

The impact on linear statistics of the sample design used in obtaining survey data is the subject of much of sampling literature. Recently, more attention has been paid to the design's impact on non-linear statistics; the major factor inhibiting these investigations has been the problem of estimating at least the first two moments of such statistics. The present article examines the problem of estimating the variances of non-linear statistics from complex samples, in the light of existing literature. The behaviour of the chi-square statistic computed from a complex sample to test hypotheses of goodness of fit or independence is studied. Alternative tests are developed and their properties studied in simulation experiments.

1. INTRODUCTION

The impact of the actual sample design used in obtaining data from a given survey has been recognized and studied by a number of authors. Its impact on linear statistics (e.g. population means and totals) has, of course, been the main subject of a large part of sampling literature. In the last ten years, or so, increasing attention has been paid to this impact as it affects non-linear statistics -- regression coefficients, correlations, multiple and partial correlations, etc. Among numerous related papers, the landmark contribution of Kish and Frankel [4] must be mentioned.

The major limiting factor inhibiting the investigation of the impact of the actual sample design on non-linear statistics has been the problem of estimating at least the first two moments of such statistics. It is well known that even from complex stratified multi-stage cluster samples, if the design is self-weighting (i.e. the inclusion probability of each unit in the population is the same), approximately unbiased

---

[1] I.P. Fellegi, Assistant Chief Statistician, Statistical Services Field, Statistics Canada.

estimates can be obtained of the population variances and covariances of the variables which were collected. Moreover, these estimates are formally identical to those derived under the assumption of simple random sampling. Therefore, consistent estimates are available for those non-linear statistics which can be constructed as functions of the estimated population variances and covariances. The estimation of the variance of such non-linear statistics has, however, been a major obstacle -- certainly variance estimators assuming simple random sampling can be quite misleading, as Kish and Frankel [4] have shown.

For purposes of the present paper the most important development enabling the estimation of variances of non-linear statistics from complex samples is the paper by McCarthy [5]. McCarthy's method of variance estimation, known as balanced repeated replication (BRR), is predicated only on the availability of two primary sampling units being selected in each stratum independently (or at least with a correlation between them which is negligible) and a within primary sampling unit sample design which is independent (although not necessarily identical) as between the two psu's of each stratum. The procedure boils down to forming two overall half-samples by combining one of the two psu's from each stratum. Any overall statistic which can be estimated from the complete stratum can also be estimated from each of the two half-samples. If there are L strata, there are 2^L different ways of forming half-samples and each of these is called a replicate. The following points have been made by McCarthy or subsequent authors:

a) If $\hat{T}^{(k)}$ and $\tilde{T}^{(k)}$ are statistics based on the two half-samples of the k-th replicate, $\bar{T}$ is the corresponding estimate made from the full sample (we will use throughout the paper the symbols $\hat{}$ and $\sim$ to refer to estimates derived from half-samples and - to estimates derived from the complete sample), then

$$(\hat{T}^{(k)} + \tilde{T}^{(k)}) \, /2 = \bar{T} \qquad \text{for all } k \tag{1}$$

if $\bar{T}$ is a linear statistic. However, even for non-linear statistics (1) is found to apply approximately.

b) The variance of $\bar{T}$ can be estimated by each of the following expressions:

$$v_1^k = (\hat{T}^{(k)} - \bar{T})^2$$
$$v_2^k = (\tilde{T}^{(k)} - \bar{T})^2 \qquad \text{for all } k.$$
$$v_3^k = (\hat{T}^{(k)} - \tilde{T}^{(k)})^2/4$$

In the case of linear statistics the three expressions are identical and provide an unbiased estimate of $\mathrm{Var}(\bar{T})$. In the case of non-linear statistics the three estimates are observed to be very close.

c) In the case of linear statistics, there is a way of creating K replicates ($L \leq K \leq L+3$), called balanced repeated replication (BRR), in such a fashion that these K out of the 2^L possible replicates capture all the available L degrees of freedom for estimating the variance of $\bar{T}$. In this case

$$\frac{1}{K} \sum_{k=1}^{K} v_i^{(k)} \qquad i=1, 2, 3$$

provides an unbiased variance estimate of $\bar{T}$ with L degrees of freedom. The same phenomenon is conjectured to hold approximately for non-linear statistics.

The main contribution of the present paper is to call attention to the fact that the chi square statistic, when computed from a complex sample to test a hypothesis of goodness of fit or of independence in a contingency table, behaves in a way which is fundamentally different from that of the many common descriptive statistics investigated by Kish and Frankel (e.g. regression coefficients, multiple and partial correlations) -- not only the dispersion of the statistic is altered (with the

mean being more or less unchanged) but the distribution is both shifted and the dispersion affected in a more or less predictable way. Alternative tests are also developed and their behaviour is studied in simulation experiments.

Consider, first of all, the chi square statistic as a test of goodness of fit. First, assume a simple random sample. Let there be m categories and denote the number of observations in the sample of n units which falls into the i-th category by

$$\bar{T}_i \qquad i=1, \ldots, m$$

where $\sum_i \bar{T}_i = n.$

To test the null hypothesis H_o, that

$$E(\bar{T}_i/n) = P_i \qquad i=1, \ldots, m,$$

the statistic s is computed

$$s = \sum_{i=1}^{m} \frac{(\bar{T}_i - nP_i)^2}{nP_i} \tag{2}$$

and, as is well known, it is distributed asymptotically under H_o as chi square with m-1 degrees of freedom.

Now assume that the estimates $\bar{T}_i$ arise from a complex self-weighting design. The variance of $\bar{T}_i$ is modified. Kish calls the quantity below the design effect

$$Deff_i = Var(\bar{T}_i)/nP_i(1-P_i) \quad .$$

The value of Deff depends on the nature of the sample design and the variables being measured, and in well-designed surveys it ranges typically between 1 and 3, although values as high as 6 have been reported and the most common values appear to be between 1.4 and 2. If all the values $Deff_i$ above are equal, then s does not have the chi square distribution. It is conjectured (and supported by empirical

investigations of the last section) that the distribution might be close to Deff times the χ^2 distribution. At least the mean of s can easily be shown to be multiplied by Deff. More generally, if $Deff_i$ are not equal, then s is conjectured to have the distribution of chi square multiplied by a weighted average of the $Deff_i$. The impact of Deff (or its applicable weighted average) is profound on the achieved significance levels (under the null hypothesis) as the table below indicates.

Table 1

Achieved significance levels (as opposed to a nominal significance level of 0.05) for different values of Deff and different degrees of freedom

df \ Deff	1.1	1.2	1.4	1.6	1.8	2.0	2.5	3	4	5	6
1	0.062	0.074	0.098	0.121	0.144	0.166	0.215	0.258	0.327	0.381	0.424
2	0.066	0.082	0.118	0.154	0.189	0.224	0.302	0.368	0.473	0.549	0.607
3	0.069	0.089	0.134	0.181	0.227	0.272	0.373	0.457	0.582	0.668	0.729
4	0.071	0.095	0.148	0.204	0.261	0.315	0.434	0.531	0.668	0.755	0.812
5	0.073	0.100	0.161	0.227	0.292	0.354	.491	0.595	0.736	0.819	0.870
6	0.076	.105	.174	.248	.321	.391	.539	.650	.790	.866	.910
7	.077	.110	.186	.268	.349	.425	.584	.698	.833	.902	.938
8	.079	.115	.197	.287	.376	.458	.625	.739	.868	.928	.958
9	.081	.119	.209	.306	.401	.489	.661	.775	.896	.947	.971
10	.083	.123	.219	.324	.426	.518	.695	.807	.918	.961	.980
15	.091	.142	.270	.408	.534	.641	.820	.910	.975	.992	.997
20	.097	.160	.317	.481	.624	.735	.874	.959	.993	.999	1.000
25	.103	.177	.361	.546	.697	.805	.940	.981	.998	1.000	1.000
30	.109	.193	.402	.604	.757	.858	.955	.992	.999	1.000	1.000

A quick perusal of the table should indicate that the use of the standard chi square as a test of goodness of fit can be drastically misleading. It will be seen in subsequent sections of this paper (particularly the simulation results) that much the same holds for chi square tests of independence in contingency tables. In fact, the achieved significance levels are particularly misleading for higher degrees of freedom and, not surprisingly, for higher values of Deff. But even with the degrees of freedom only as large as 4 (or more) and with Deff as large as or larger than 1.6, the significance tests are practically useless: the null hypothesis would be rejected with a probability of 0.2, rapidly rising to .5 or more with larger values of Deff and/or larger degrees of freedom. It should be kept in mind that large degrees of freedom arise quite commonly in the case of contingency tables: e.g. a 4x5 table gives rise to 12 degrees of freedom, a 5x6 table to 20.

One final introductory note on the chi square test is necessary. Conceptually, one can at least contemplate using the test in the case of self-weighting designs (although, as seen above, the results can be most misleading). However, in a not negligible proportion of sample designs actually used in practice the inclusion probability of all units in the population is not equal. In this case the unweighted sample frequencies will not provide unbiased estimates of the corresponding population statistics, thus even if H_o holds in the population as a whole, one would expect that a chi square test based on the unweighted sample frequencies would be rejected far more often than the nominal significance level (whether used as a test of independence or of goodness of fit). However, the chi square statistic (2) does not lend itself to weighting at all, because the numerator increases with the square of any weight whereas the denominator increases linearly.

2. OBSERVATIONS ON RELATED WORK IN THE LITERATURE

Cohen [3] investigates a very special case of the general problem of testing goodness of fit from complex samples. He assumes a simple random sample of n clusters consisting of two units each. This sample of 2n units is classified into r cells. In the model studied by Cohen, if p_i is the probability of unit 1 of a cluster being in cell i, then the probability of the two units being in cells i and j respectively is

$$p_{ij} = (1-a)\, p_i p_j \qquad \text{if } i \neq j$$

$$p_{ii} = p_i\, [a + (1-a)p_i]$$

for values of a between 0 and 1.

Under this model it can easily be shown that

$$\mathrm{Var}(\bar{T}_i) = (1+a)\, 2n\, p_i\, (1-p_i) \qquad i=1, \ldots, m\ .$$

So the design effect Deff is equal to 1+a for all i. Cohen shows that the statistic

$$s/1+a$$

is, in fact, distributed as chi square (under H_o) with m-1 degrees of freedom -- as is conjectured in the introduction more generally.

The most sustained work on tests of independence from complex samples has been carried out by Nathan ([6], [7], [8], [9], [10]). He also reviews the work of several other authors, such as Bhapkar and Koch [1] and Chapman [2].

Consider the usual contingency table: there are r rows, c columns, overall sample size n, P_{ij} are the estimated proportions of frequencies in the (i,j)-th cell (i=1, ..., r; j=1, ..., c). The statistic

$$t = \sum_{i,j} \frac{(n\bar{P}_{ij} - n\bar{P}_{i.}\,\bar{P}_{.j})^2}{n\bar{P}_{i.}\,\bar{P}_{.j}} \tag{3}$$

has approximately the chi square distribution under the null hypothesis if n is based on a simple random sample and is sufficiently large. The quantities $\bar{P}_{.j}$ and $\bar{P}_{i.}$ are obtained by summation over the missing subscripts. Under the null hypothesis the expression

$$\bar{P}_{ij} - \bar{P}_{i.}\,\bar{P}_{.j} \tag{4}$$

has zero expected value but generally an expected value different from zero if the null hypothesis does not hold. The zero expected value of (4) is the result of a number of variance and covariance terms in $E(\bar{P}_{i.}\,\bar{P}_{.j})$ cancelling. In effect, under simple random sampling and the null hypothesis

$$\begin{aligned} E(\bar{P}_{i.}\,\bar{P}_{.j}) &= P_{i.}\,P_{.j} + \mathrm{Var}\,(\bar{P}_{ij}) \\ &\quad + \sum_{(k,k')\neq(i,j)} \mathrm{Cov}(\bar{P}_{kj},\bar{P}_{ik'}) \\ &= P_{ij} + \frac{1}{n}P_{ij}(1-P_{ij}) - \frac{1}{n}\sum_{(k,k')\ (i,j)} P_{kj}\,P_{ik'} \\ &= E(\bar{P}_{ij}). \end{aligned} \tag{5}$$

In complex surveys the variances and covariances in the penultimate line of (5) would each have to be multiplied by their respective design effect multipliers and therefore may not cancel out -- thus the expected value of $\bar{P}_{ij}-\bar{P}_{i.}\,\bar{P}_{.j}$ may not be equal to zero even under H_o.

The work of both Nathan and Bhapkar and Koch starts out with the construction of an expression involving estimates of P_{ij}, $P_{i.}$ and $P_{.j}$ which has zero expected value under the null hypothesis, even in the case of complex samples. For this purpose they both resort to balanced repeated replication and make use of the fact that the two half-samples of any replicate are uncorrelated under the assumptions outlined in the introduction. Thus if $\hat{P}^{(k)}_{ij}$, $\hat{P}^{(k)}_{i.}$, $\hat{P}^{(k)}_{.j}$ are estimated, under a complex sample design, from the first half-sample of the k-th replicate and $\tilde{P}^{(k)}_{ij}$, $\tilde{P}^{(k)}_{i.}$, $\tilde{P}^{(k)}_{.j}$ are the corresponding quantities estimated from the

second half-sample, Nathan's test is based on the expression

$$\bar{U}_{ij}^{(k)}(N) = \hat{P}_{ij}^{(k)} + \tilde{P}_{ij}^{(k)} - \hat{P}_{i.}^{(k)}\tilde{P}_{.j}^{(k)} - \tilde{P}_{i.}^{(k)}\hat{P}_{.j}^{(k)} \tag{6}$$

and Bhapkar and Koch's is based on

$$\bar{U}_{ij}^{(k)}(B) = \hat{P}_{ij}^{(k)}\tilde{P}_{rc}^{(k)} - \hat{P}_{ic}^{(k)}\tilde{P}_{rj}^{(k)} \tag{7}$$

Both (6) and (7) have zero expected values under the null hypothesis. Chapman's test is based on

$$\bar{U}_{ij}^{(k)}(C) = \hat{P}_{ij}^{(k)} - \tilde{P}_{i.}^{(k)}\tilde{P}_{.j}^{(k)} \tag{8}$$

and it does not necessarily have a zero expected value even under H_o.

Now if an estimate $\hat{V}$ can be constructed for the covariance matrix for the $(r-1)x(c-1)$ linearly independent quantities among the rxc U_{ij} values, and if U is the corresponding vector of these values, then

$$\bar{U}' \, (\hat{V})^{-1} \, \bar{U} \tag{9}$$

would, for large enough n, and apart from a suitable constant multiplier, be either distributed approximately as F or a χ^2 -- depending on whether $\hat{V}$ is estimated from a large enough number of degrees of freedom. Note that (9) overcomes the problem of weighting in the case of disproportionate sampling -- its effect would, so to speak, automatically be reflected in $\hat{V}$. The problem, however, is to estimate $\hat{V}$.

In the case of simple random sampling each cell of the covariance matrix of (4) is readily estimated approximately as

$$\hat{v}_{ij,fg} = \frac{1}{n}\bar{P}_{i.}\bar{P}_{.j}(1-\bar{P}_{i.})(1-\bar{P}_{.j}); \qquad (i,j)=(f,g)$$

$$= -\frac{1}{n}\bar{P}_{i.}\bar{P}_{.j}\bar{P}_{f.}\bar{P}_{.g} \qquad ; \qquad (i,j)\neq(f,g)$$

and the resulting estimate of $\hat{V}$ is based on a large number of degrees of freedom so (9) would be distributed as chi square. In the case of

complex samples the analogous estimation cannot be carried out without some very strong simplifying assumptions.

Alternatively, one may observe that, even in the case of complex samples, the vectors $\bar{U}^{(k)}$ in (6)-(8) are identically distributed and hope to derive estimates of variances and covariances from

$$\sum_{k=1}^{K} (\bar{U}_{ij}^{(k)} - \bar{U}_{ij})(\bar{U}_{fg}^{(k)} - \bar{U}_{fg}) \qquad (10)$$

Now if the K replicate values of $\bar{U}_{ij}^{(k)}$ were independent, the expression (10) above, divided by K-1, would provide an unbiased estimate of $v_{ij,fg}$. However, far from being independent, they are very highly correlated. In the case of $\bar{U}^{(k)}(N)$ the correlations are very close to one -- not too surprisingly, since Kish and Frankel noted that for all replicates the sum of two analogous non-linear statistics, computed respectively from the two half-samples, is very nearly the same for all K replicates and is identically the same in the case of linear statistics. In order to correct (10) for the correlations involved, one would have to estimate these correlations and that, in turn, again requires strong simplifying assumptions. Moreover, when the correlations are close to one the numerical behaviour of the estimates is very bad.

Thus whichever of the two methods of estimating the covariance matrix is attempted, strong simplifying assumptions are needed in the case of complex samples. Nathan [9] is forced to make the assumption, among others, that for each stratum h there is a number n_h which depends only on the number of final units selected in stratum h in each of the two primary sampling units (psu), and if $\hat{P}_{ijha}$ is the estimate of the proportions in cell (i,j) derived from psu a (a=1,2) of stratum h, then $n_h \hat{P}_{ijha}$ has approximately the multinomial distribution with parameters n_h and P_{ijh}. However, this assumption implies that the expected value of an estimate $\hat{P}_{ijha}$ derived from any selected psu, conditional on that psu being in the sample, depends on the stratum only and <u>not</u> on

the selected psu. Thus the total between-psu component of variance is assumed away. But, for example, in the case of two stage stratified sampling, with simple random sampling within each of the psu's, this assumes away all the within-stratum design effects. Other assumptions of Nathan, less important to his development, assume away the effects of stratification and disproportionate sampling in different strata as well.

In light of the comments above, it is not too surprising to find that the test statistic proposed by Nathan behaves very badly with respect to its achieved significance levels. The simulation results reported in his paper [9] are flawed, as pointed out by the author in his own subsequent paper, Nathan [10]. The results reported in [10] refer to stratified cluster sampling with a self-weighting design, so the traditional chi square test can be applied and serves as a measure of comparison. The achieved significance levels of his test statistic under H_o are .038, .144 and .190 for the nominal significance levels of .01, .05 and .10 respectively. However, these are almost identical to the achieved significance levels of the traditional chi square test: the latter differ from those of Nathan's test by at most 0.002. It may be noted for interest that if one assumed that the statistic t of (3) is distributed as chi square multiplied by a factor of 1.4, the achieved significance level of this hypothetical variable at the nominal levels of .01, .05 and .10 would be .037, .118 and .193 respectively -- quite remarkably close to the values reported by Nathan. One might conjecture that Deff in his example was about 1.4.

Finally, a few comments will be made relating to the special case of stratified simple random sampling with proportional allocation.

Nathan in [10] claims* that in this case the traditional chi square statistic is asymptotically distributed under H_o as chi square with

---

* In a private communication with the author, Nathan identified the error in his proof.

(r-1)x(c-1) degrees of freedom. This is not generally true, as a simple counter-example proves. Indeed, suppose that there are exactly rxc strata, each of equal size. Suppose also that corresponding to each cell of the contingency table there is one and only one stratum which contributes to that cell and, conversely, every unit in a stratum is classified to exactly one cell of the contingency table. Now, proportional allocation is equivalent to selecting the same number of units in each stratum, say d, the total sample size being n=rcd. Then every cell of the contingency table will contain exactly d entries with probability equal to one. Thus the statistic t in (3) will be equal to zero with probability equal to one -- no matter how large n becomes. Parenthetically, one may observe that the design effect in this case is equal to zero.

In the relatively simple case of stratified simple random sampling with proportional allocation to strata, it is easy to prove under H_o for the test of goodness of fit, s

$$E(s) = m-1 - \sum_{i=1}^{m} \frac{1}{nP_i} \sum_h n_h (P_{ih}-P_i)^2$$

where P_{ih} is the proportion of units in stratum h belonging to category i and n_h is the sample size in stratum h. Since the expected value of s under simple random sampling is m-1, a <u>reduction</u> is observed in the expected value of s -- as indeed in such cases Deff is known to be less than or equal to 1, the extent of the reduction increasing, roughly speaking, with the between stratum differences in the proportions P_{ih}. In the case of the example of the previous paragraph E(s) = 0 which, given that $s \geq 0$, is only possible if s=0 with probability equal to one.

3. TWO ALTERNATIVE TESTS

The desirable feature of the two half-samples of a replicate is that they have the same distribution and are uncorrelated. This can either

be used to construct, in the case of complex samples, a quantity like Nathan's or Bhapkar and Koch's U(N) or U(B) in (6) or (7) whose expectation is zero under H_o, or it can be used to estimate the variance of linear or non-linear statistic -- but two half-samples cannot be used for both purposes. A simple way out would be available if more than two psu's were selected per stratum, but this is so rarely the case that it is hardly worth considering.

Given the difficulties of variance estimation when the half-samples are used for another purpose, it is a natural motivation to go back to the question: how far away from zero is actually the expected value under H_o of the quantity

$$\bar{U}_{ij} = \bar{P}_{ij} - \bar{P}_{i.}\ \bar{P}_{.j}$$

when it is based on the whole sample or, in fact, on one of the two half-samples of a replicate?

Several observations will be made on this question.

a) The expected value of $\bar{U}_{ij}$ is zero under simple random sampling or, slightly more generally, if the Deff of all the variances and covariances in (5) is equal to one. Actually, an even more general sufficient condition is that all the Deff's are equal. While this cannot be assumed to be the case generally, it is often the case that Deff's from the same survey for a wide variety of variables are within a quite narrow range of one another.

b) Under the most general conditions,

$$\sum_i \bar{U}_{ij} = 0 \qquad j=1, 2, \ldots, c$$

and

$$\sum_j \bar{U}_{ij} = 0 \qquad \text{for all } i=1, 2, \ldots, r$$

so that the expectations of $\bar{U}_{ij}$ are subject to r+c-1 linear constraints. Whereas this does not exclude the possibility that $E(\bar{U}_{ij})$ may be quite large in absolute value, it most certainly ensures that they are not all of the same sign.

c) Most important of all, it can easily be shown that so long as all Deff values are bounded for all variances and covariances considered (i.e. |maximum Deff| $\leq$ B for some B in whatever way $n\to\infty$),

$$E(\bar{U}_{ij}^2) = Var(\bar{U}_{ij}) + 0\ (\frac{1}{n^2}) \tag{11}$$

where the left hand side is of 0 $(\frac{1}{n})$.

Indeed,

$$E(\bar{U}_{ij}) = -\ Cov\ (\bar{P}_{i.}, \bar{P}_{.j}). \tag{12}$$

The right hand side of (12) is obviously 0 $(\frac{1}{n})$ in the case of simple random sampling, but in the case of complex designs it will only be multiplied by the appropriate Deff. Now there are no reported values in the survey literature of Deff exceeding 10, in fact values above 3 or 4 are exceedingly rare -- for the simple reason that the survey would never be allowed if it was that inefficiently designed. At any rate, so long as maximum|Deff| $\leq$ B as $n\to\infty$ over any class of designs subject only to the independent selection of two psu's per stratum and the availability of unbiased stratum-level estimates of linear statistics from each of the two psu's,

$$|E(\bar{U}_{ij})| \leq \frac{B}{n} P_{i.} P_{.j}\ . \tag{13}$$

Further,

$$E(\bar{U}_{ij}^2) = Var\ \bar{U}_{ij} + [E(\bar{U}_{ij})]^2\ , \tag{14}$$

and (11) follows immediately from (13) and (14).

Since, however, under the same conditions

$$\text{Var } \bar{U}_{ij} = 0\ (\tfrac{1}{n})$$

it follows that for sufficiently large n one may well be able to treat $\bar{U}_{ij}$ as if it did have a zero expected value. Typically, in complex surveys n is quite large: at least of the order 1000-2000 and very often, in the case of large national surveys, tens of thousands.

That being the case, one can construct U_{ij} from each of the two half-samples of the k-th replicate. Consider the vector of U_{ij} values corresponding to the (r-1)x(c-1) upper left hand corner of a contingency table, constructed from one of the two half samples of the k-th replicate

$$\hat{U}^{(k)} = \hat{U}^{(k)}_{1,1},\ \hat{U}^{(k)}_{1,2,\ldots},\ \hat{U}^{(k)}_{1,c-1},\ \hat{U}^{(k)}_{2,1,\ldots},\ \hat{U}^{(k)}_{r-1,c-1}$$

and similarly $\tilde{U}^{k}$ constructed analogously from the second half sample of the same replicate

$$\hat{V} = \frac{1}{4K} \sum_{k=1}^{K} \hat{U}^{(k)}\ (\tilde{U}^{(k)})' \tag{15}$$

provides an approximately unbiased estimate of the variance of

$$\bar{U} = \sum_{k=1}^{K} (\hat{U}^{(k)} + \tilde{U}^{(k)})/2K \tag{16}$$

where K is the number of orthogonal replicates.

Therefore, under the null hypothesis, the statistic

$$\bar{U}'\ (\hat{V})^{-1}\ \bar{U} \tag{17}$$

is approximately distributed as Hotelling's T^2, or multiplied by an appropriate constant, as F. From standard textbooks this constant is easily seen as (L-m)/m(L-1) where L is the number of strata and m=(r-1)x(c-1). It follows that

$$t' = \frac{L-m}{m(L-1)}\ \bar{U}'\ (\hat{V})^{-1}\ \bar{U} \tag{18}$$

is approximately distributed as F(m,L-m). It is easy to see that (18) also provides a test of goodness of fit: the vector U has to be replaced with the vector $\bar{P}-P^{O}$, where $\bar{P}$ is the vector of observed proportions, P^{O} the proportions under H_o, $\hat{V}$ is the covariance matrix of $\bar{P}$ estimated through BRR and m=rc-1.

The second test is more heuristically constructed than the first. Consider first the test of goodness of fit

$$\sum_i \frac{(n\bar{P}_i - n\ P_i^O)^2}{nP_i^O} = \sum_i \frac{n(\bar{P}_i - P_i^O)^2}{P_i^O} \ . \qquad (19)$$

As discussed above, when H_o holds the <u>expected</u> value of the numerator of each term in (19) under the given design is the appropriate Deff times its expected value under simple random sampling. Assume that not only the expected value of the numerator but its <u>distribution</u> was also equal to that obtained by multiplying by Deff the corresponding statistic under simple random sampling. This would then suggest that dividing each term by the estimated Deff of the numerator would restore the distribution (under H_o) to chi square.

In effect, by dividing each term in the numerator by its corresponding Deff, the multiplier n becomes what is known as the <u>effective sample size</u>

$$n_i' = \frac{n}{deff_i}$$

so that the statistic

$$\sum_i \frac{n_i' (\bar{P}_i - P_i^O)^2}{P_i^O}$$

is distributed as chi square.

Since

$$\frac{1}{2}\left(\hat{P}_i^{(k)} + \tilde{P}_i^{(k)}\right) = \bar{P}_i \qquad \text{for all } k$$

and the variance of the expression above can be estimated as

$$a_i = \frac{1}{4K} \sum_{k=1}^{K} \left(\hat{P}_i^{(k)} - \tilde{P}_i^{(k)}\right)^2 \tag{20}$$

and since the variance of $\bar{P}_i$ under simple random sampling (under H_o) is estimated as

$$\frac{1}{n} \bar{P}_i \left(1-\bar{P}_i\right) \tag{21}$$

it follows that (20) divided by (21) provides an estimator of $Deff_i$. In fact,

$$b_i = a_i / \bar{P}_i \left(1-\bar{P}_i\right)$$

is an estimator of the inverse of the effective sample size, n_i'. Thus

$$\sum_i \frac{\left(\bar{P}_i - P_i^o\right)^2}{b_i P_i^o}$$

might be distributed approximately as chi square. In simulation studies the statistic above tended to be too large. However, by using the average of the b_i values, good results were obtained. So the second statistic proposed as a test of goodness of fit and evaluated through simulation studies is

$$t'' = \frac{1}{\bar{b}} \sum_i \frac{\left(\bar{P}_i - P_i^o\right)^2}{P_i^o} \tag{22}$$

where

$$\bar{b} = \frac{1}{r} \sum_{i=1}^{r} b_i. \tag{23}$$

Note that (22) can be computed whether or not disproportionate sampling among the strata has been used.

The test (22) can readily be generalized to obtain a test of independence. Let a_{ij}, b_{ij} and $\bar{b}$ be defined, respectively, as

$$a_{ij} = \frac{1}{4K} \sum_{k=1}^{K} (\hat{P}_{ij}^{(k)} - \tilde{P}_{ij}^{(k)})^2$$

$$b_{ij} = a_{ij} / \bar{P}_{ij} \, (1-\bar{P}_{ij})$$

$$\bar{b} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} b_{ij} \tag{24}$$

then

$$t'' = \frac{1}{\bar{b}} \sum_{i,j} \frac{(\bar{P}_{ij} - \bar{P}_{i.} \bar{P}_{.j})^2}{\bar{P}_{i.} \bar{P}_{.j}} \tag{25}$$

Note that (25) differs from Pearson's chi square test simply by the replacement of the actual sample size by the average effective sample size under complex designs. Note also that (25) can be computed whether or not proportionate allocation among the strata has been used. Also, since $\bar{P}_{ij}$ are linear statistics, their variances (hence their Deff's) can be estimated through traditional methods, i.e. without BRR. This makes the calculation of t'' quite easy: apart from a package to compute the traditional chi square statistic, only an efficient variance estimation program is needed.

4. EMPIRICAL RESULTS

The results of seven simulated examples are presented in this concluding section. In every instance the simulated sample design is stratified, two primary sampling units are selected with equal probabilities and replacement, and the sampling within the psu's is simple random, also

with replacement. Except for the with replacement sampling of psu's (which may well enough be approximated in practice if there are a large number of psu's, say 20, and no more than two of them are selected), the remaining simplifications in the simulations were imposed by the need to keep the programs simple -- as opposed to theoretical restrictions. All examples, except example 7, are based on 500 simulations, example 7 on 250. In all simulations, except example 6, the contingency tables are based on 2 rows and 3 columns. In examples 1 and 6 proportional allocation to the strata was used, in the others the sampling rates differed in the proportions 1:2:3. The total sample size in all examples was 1200.

The features of the examples are summarized in Table 2 below while the behaviour of the unweighted chi square test values is shown in Table 3.

Table 2

Summary of features of seven examples

	No. of rows	No. of columns	No. of strata	Relative sampling rates	Range of P_{ij} (H_o)	Range of Deff	No. of simulations
Example 1	2	3	6	1:1:1	1/6	1.98	500
Example 2	2	3	6	1:2:3	.161- .172	1.49- 2.07	500
Example 3	2	3	6	1:2:3	.158- .173	1.57- 2.55	500
Example 4	2	3	6	1:2:3	.158- .175	1.62- 3.03	500
Example 5	2	3	6	1:2:3	.156- .178	1.66- 3.51	500
Example 6	3	4	12	1:1:1	1/12	1.89	500
Example 7	2	3	30	1:2:3	.150- .180	2.07- 3.09	250

Table 3

Observed significance level (under H_o) of unweighted sample counts in Pearson's chi square test

Nominal level	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	Avg. test ÷ d. of f.	Avg. Deff
Example 1	.090	.198	.290	.420	.536	.604	.696	.770	.826	.886	.938	1.87	1.98
Example 2	.100	.212	.290	.426	.542	.596	.656	.728	.778	.868	.936	1.91	1.76
Example 3	.150	.294	.388	.502	.566	.636	.702	.772	.846	.914	.962	2.45	2.10
Example 4	.190	.330	.416	.528	.594	.666	.734	.810	.870	.918	.958	2.75	2.29
Example 5	.232	.362	.436	.556	.632	.688	.740	.820	.880	.922	.964	3.08	2.48
Example 6	.174	.350	.452	.606	.706	.782	.856	.896	.938	.960	.978	1.90	1.89
Example 7	.212	.324	.416	.532	.616	.684	.776	.860	.900	.940	.976	2.81	2.60

Looking at Table 3, one's immediate observation is the rising level of nominal significances from examples 1 to 5 with the rising level of average Deff. However, in examples 2 to 5 another consideration is also important: the unweighted counts do not have an expected value which is consistent with H_o: i.e. H_o is valid over all strata but not in each stratum, hence the unweighted counts are inconsistent with H_o in the case of disproportional allocation. The simulation model used was such that, going from examples 2 to 5, not only the range and average of Deff values but also the within stratum departures from H_o increased. This explains the reason why within these examples the average test value divided by the degrees of freedom (which for a valid chi square test, of course, ought to be 1), rises faster than the average of the Deff values.

It is interesting to note the entries in examples 1 and 6 (for which proportional allocation was used) at the nominal significance level .05: .198 and .350 respectively. Compare these with the corresponding

proportions "predicted" by Table 1: for Deff = 1.87 and 2 degrees of freedom it is .201; for Deff = 1.90 and 6 degrees of freedom it is .356. The agreement is, indeed, very close. Also very close for these two examples is the average value of the test statistic divided by the degrees of freedom: 1.87 and 1.98, and 1.90 and 1.89 respectively.

Finally, it should be emphasized that, as predicted, the classical chi square test provides totally misleading results in the presence of Deff's which are moderately large, even in the case of proportional allocation -- and more so otherwise.

Table 4 shows the observed significance level (under H_o) of the test t', i.e. the F test.

Table 4

Observed significance level (under H_o) of the test t' (F test)

Nominal level	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	Avg. test ÷ Exp. value	Chi square (deciles)
Example 1	.008	.038	.068	.152	.240	.346	.452	.554	.664	.768	.870	0.81	12.60
Example 2	.012	.052	.080	.178	.276	.374	.482	.570	.670	.776	.888	0.96	4.72
Example 3	.016	.054	.102	.188	.280	.384	.474	.578	.688	.804	.906	1.04	3.96
Example 4	.014	.058	.108	.184	.270	.384	.478	.560	.676	.790	.894	0.99	8.38
Example 5	.018	.058	.104	.184	.278	.388	.484	.564	.658	.786	.906	1.03	11.12
Example 6	.018	.072	.124	.222	.336	.430	.500	.604	.708	.786	.872	0.83	16.04
Example 7	.004	.040	.112	.228	.304	.404	.500	.612	.696	.788	.900	1.00	4.00

The one before the last column contains entries obtained by dividing the average of our test statistic by its theoretical expected value (if d is

the degrees of freedom for the denominator of the F test, this value is d ÷ d-2). The last column is the chi square test statistic for goodness of fit applied to the decile values of the table above. The critical value of chi square with 9 degrees of freedom at the 5% level is 16.92, thus at least this test is consistent with the hypothesis that t' is distributed as F. The nominal significance levels, particularly at the .01 and .05 levels, which are usually of greatest interest, behave very well -- their average over the seven examples is .013 and .053 respectively.

Notwithstanding the non-significant values of the chi square goodness of fit test, particularly the first five examples show what appear to be consistent departures from the nominal levels in the range .3 - .8. It appears, however, that this is <u>not</u> due to the approximation whereby (13) is assumed to be zero (i.e. E $(\bar{U}_{ij})$ = 0). Indeed, this approximation holds exactly if all the Deff's in (5) are equal -- as noted before. This is the case in examples 1 and 6. Yet, in many requests, they appear to follow the F test worst (although still well enough at the .01 and .05 levels). Thus it would appear that the problem, if it can be called that, is due to the normal approximation, as opposed to the approximation (13). Note that n = 1200 -- small by standards of survey sampling.

Next the behaviour of the test t'' is shown, still under H_o.

Table 5

Observed significance level (under H_o) of the test t''
(adjusted chi square)

Nominal level	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	Avg. test ÷ d. of f.	Chi square (deciles)
Example 1	.034	.068	.104	.216	.306	.390	.496	.600	.698	.782	.892	1.10	4.96
Example 2	.028	.092	.146	.248	.340	.408	.494	.612	.680	.790	.898	1.16	17.40
Example 3	.040	.102	.164	.250	.324	.416	.510	.600	.696	.790	.906	1.23	27.56
Example 4	.042	.100	.158	.244	.326	.416	.510	.600	.676	.796	.902	1.25	25.68
Example 5	.048	.112	.156	.250	.324	.420	.506	.600	.670	.772	.896	1.28	27.96
Example 6	.032	.072	.126	.238	.326	.412	.502	.610	.696	.816	.912	1.07	9.64
Example 7	.008	.068	.124	.216	.296	.400	.504	.596	.684	.780	.908	1.02	10.88

Clearly, t'' has a distribution which in four of the seven examples is significantly different from chi square at the 5% level of significance. However, in exploring the results in Table 5 somewhat further, the following might be observed.

a) Most notably, if the alternative is between using the unadjusted chi square test (Table 3) or using the simple adjustment which leads to t'', clearly t'' is very much closer to chi square. While the statistics corresponding in Table 3 to the last column of Table 5 were not computed, this much is clear to the naked eye.

b) The extent to which the distribution of t'' departs from chi square seems to follow closely the extent to which the expected value of t'' departs from the expected value of the corresponding chi square distribution (penultimate column of Table 5). However, t'' was constructed in such a manner that its expected value would asymptotically

be equal to that of chi square. The convergence, however, depends on the number of available replications from which to compute $\bar{b}$ of (25) -- not on n. In fact, we are faced with the usual bias of a ratio estimate whose magnitude heavily depends on the variance of the denominator, i.e. of $\bar{b}$, which in turn depends on the number of replicates or, since BRR need not be used, on the number of strata. The good behaviour of t'' in examples 6 and 7, where the number of strata is 12 and 30 respectively, is consistent with this line of reasoning. Thus one might expect t'' to behave acceptably well for the survey designs encountered in practice, where the number of strata is usually quite large.

c) While t' behaves consistently better than t'', one has to note that t' is not applicable if the degrees of freedom is greater than or equal to the number of strata. This may well occur in multi-level consistency tables.

Finally, Table 6 is a tentative attempt to compare the power of t' and t'' under an alternative hypothesis H_1. Since H_1 could not be consistently chosen across the examples, the comparison of the results in Table 6 should be restricted to comparing t' and t'' within the same example.

Table 6

Observed proportion of times t' and t'' exceed their respective significance levels, under H_1

Nominal level		.01	.05	.10
Example 1	t'	.026	.178	.340
	t''	.250	.418	.510
Example 2	t'	.060	.232	.406
	t''	.282	.498	.632
Example 3	t'	.062	.214	.394
	t''	.242	.438	.568
Example 4	t'	.066	.212	.380
	t''	.218	.408	.552
Example 5	t'	.062	.206	.382
	t''	.220	.378	.524
Example 6	t'	.160	.480	.686
	t''	.714	.866	.920
Example 7	t'	.148	.332	.476
	t''	.136	.316	.452

No authoritative conclusions can be drawn from the above, primarily because of the fact that t'' is biased upward under H_o -- at least for the first six examples. In fact, for this reason a comparison of the two test statistics at their respective nominal level is misleading -- it would make t'' appear to have considerably more power than it actually does. Analyzing Table 6 together with Tables 4 and 5 is more realistic. By and large t' at the .10 nominal level would appear to be more comparable with t'' at the .05 nominal level. Even so, t'' would appear to have somewhat more power than t', except for example 7. This is more or less what one could expect, since t', based on the F test, will asymptotically be distributed as chi square if the degrees

of freedom for the denominator is large, i.e. if the number of strata is large compared to the number of subclasses.

In summary, the unadjusted chi square is subject to intolerable biases under complex designs even with moderate Deff's; t' appears to be distributed as expected even for values of n as small as 1200; t'' behaves incomparably better than the unadjusted chi square test but still appears to have higher than nominal significance levels particularly when the number of strata is small; t'' is much easier to calculate than t'; t'' appears to have greater power than t' unless the number of strata is large compared to the number of subclasses.

RESUME

Une grande partie de la littérature sur l'échantillonnage se concentre sur l'effet que le plan d'échantillonnage utilisé pour rassembler des données dans une enquête porte sur les statistiques linéaires. Récemment, on a considéré davantage l'effet du plan d'échantillonnage sur les statistiques non-linéaires. Le facteur le plus important qui empêche ces recherches a été le problème de l'estimation d'au moins les deux premiers moments de ces statistiques. Le présent article étudie le problème de l'estimation des variances des statistiques non-linéaires des échantillons complexes, en considérant la littérature existante. On étudie les attributs de la statistique chi-carré calculée à partir d'un échantillon complexe pour tester des hypothèses de la qualité de l'ajustement ou d'indépendance. On développe des tests alternatifs et on étudie leurs attributs en faisant des expériences simulées.

REFERENCES

[1] BHAPKAR, V.P. and KOCH, G.G., "On the Hypothesis of 'no interaction' in Contingency Tables", Biometrics 24(3), pp. 567-594, September 1968.

[2] CHAPMAN, D.W., "An Approximate Test of Independence Based on Replications of Complex Sample Survey Design", Unpublished master's thesis, Cornell University, 1966.

[3] COHEN, J.E., "The Distribution of the Chi-squared Statistic Under Clustered Sampling from Contingency Tables", Journal of the American Statistical Association, Volume 71, pp. 665-669, September 1976.

[4] KISH, L. and FRANKEL, M.R., "Inference from Complex Samples", Journal of the Royal Statistical Society, Series B, Vol. 36, pp. 1-37, 1974.

[5] McCARTHY, P.J., "Replication - an Approach to the Analysis of Data from Complex Surveys", National Centre for Health Statistics, Vital and Health Statistics, Series 2, No. 14, Washington, D.C., 1966.

[6] NATHAN, G., "Tests of Independence in Contingency Tables from Stratified Samples", In N.L. Johnson and H. Smith, eds., New Developments in Survey Sampling, Wiley, New York, pp. 578-600, 1969.

[7] NATHAN, G., "A Simulation Comparison of Tests for Independence in Stratified Cluster Sampling", Bulletin of the International Statistical Institute, Vol. 44, Book 2, pp. 289-295, 1971.

[8] NATHAN, G., "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", Journal of the American Statistical Association, Vol. 67, pp. 917-920, 1972.

[9] NATHAN, G., "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", National Centre for Health Statistics, Vital and Health Statistics, Series 2, No. 53, Washington, D.C., 1973.

[10] NATHAN, G., "Tests of Independence in Contingency Tables from Stratified Proportional Samples", Sankhya, Vol. 37, Series C, Part 1, pp. 77-87, 1975.