VARIANCE ESTIMATION IN COMPLEX SURVEYS

Benjamin J. Tepping, Bureau of the Census

1. Introduction

Clearly, the variance of an estimator based on a sample survey depends upon the design of the sample as well as upon the form of the estimator. For many of the sample designs in common use, estimators of the variance of estimated totals, means, ratios and differences are readily available in the literature. For simple random samples, estimators of the variance of more complex statistics such as variances, correlation coefficients and regression coefficients are also readily available. Thus, the problem with which this session is concerned is that of estimating the variance of such statistics as compound ratios, regression coefficients, or other complicated functions of the sample observations, when the sample design is other than simple random sampling. The sample design may be multistage, the sampling units stratified at one or more levels, with probabilities of selection varying from unit to unit.

For a given statistic based on a given sample design, there will usually exist alternative estimators of the variance. The proper choice among these alternatives will be made on the basis of consideration of characteristics of their sampling distributions, such as variance, mean square error or bias, as well as on the basis of cost considerations.

In the simplest cases the variance of the statistics, for a given sample design, is a known function of certain population parameters. Those parameters may themselves be estimated from the sample, and the estimates substituted for the parameters in the variance function to obtain an estimate of the variance. For example, with a simple random sample of n units from a population of N units without replacement, the variance of the sample mean is

$$(1) \qquad \sigma_{\bar{x}}^2 = \frac{N-n}{N-1}\frac{\sigma^2}{n},$$

and the population variance σ^2 may be estimated by

$$(2) \qquad s^2 = \frac{1}{n-1}\Sigma(x_i-\bar{x})^2.$$

If, as in the case of this example, the variance function is linear in the parameters and the estimators of those parameters are unbiased, we obtain an unbiased estimator of the variance. If the function is rational, we obtain at least a consistent estimator of the variance. For example, for a simple random sample, the variance of the ratio of sample means $r = \bar{y}/\bar{x}$ is given approximately by

$$(3) \qquad \sigma_r^2 \doteq \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{X}^2}(\sigma_y^2+R^2\sigma_x^2-2R\sigma_{xy})$$

where $R = \bar{Y}/\bar{X}$, and $\bar{X}$, $\bar{Y}$ are the expected values of $\bar{x}$ and $\bar{y}$. One usually estimates σ_r^2 by

$$(4) \qquad s_r^2 = \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{x}^2}(s_y^2+r^2s_x^2-2rs_{xy})$$

where s_x^2, s_y^2 and s_{xy} are the usual estimates of the population variances and covariance.

For other statistics, or for more complex designs, the variance may not be a known function of parameters that are easily estimated for substitution into the variance formula. One may then estimate the variance by dividing the sample into random subgroups in such a way that the variance of the statistic, for a sample the size of a subgroup, can be estimated from the differences, among subgroups, of the desired statistic. If the dependence of the variance on the size and type of the subgroup is known, this leads to an estimate of the desired variance. Deming [2] has long insisted on the utility of designing the sample in such a way that the computation of such variance estimates is particularly simple and easy. While this approach has much to recommend it, it is not always appropriate.

Where it is not convenient or possible to divide the sample into sufficiently many random subgroups, the method of half-samples, termed "pseudo-replication" by McCarthy [8], has been employed by the Bureau of the Census, the Survey Research Center of the University of Michigan, and the National Center for Health Statistics and perhaps others. This method, which may be regarded as a special case of Tukey's "Jackknife" [9], involves defining subgroups which are half the size of the full sample. The subgroups are not independent but, properly constructed, lead to an estimate of the variance.

2. Half-sample estimates of variance

An attractive feature of the half-sample or pseudo-replication estimates of variance is that it is not necessary to know the exact functional form of the variance, but only the dependence of the variance on sample sizes. It should be noted that the latter cannot be taken for granted, for it is not always true that the variance is inversely proportional to sample size, although that is frequently a useful approximation.

A serious limitation to be considered is that the precision of the variance estimate depends upon the number of replications, and a sufficient number of replications may be quite costly. The Current Population Survey conducted by the Bureau of the Census is illustrative.

To simplify matters somewhat, this description* applies to the sample design and estimating procedure as used before January 1967. The sample design is multistage. The primary sampling units are large, and are classified into 357 strata, of which 112 contain only a single primary sampling unit. One primary unit is selected from each stratum with probability proportionate to its population in 1960, and a subsample of dwellings is selected in several stages in such a manner that the overall probability of selection of a dwelling is constant over the whole population of the United States.

The estimation procedure can be thought of as consisting of two successive stages of ratio estimation followed by the construction of a composite estimate. The first-stage ratio estimate applies only to those primary units in strata containing more than a single primary unit, and consists of computing an inflation factor for each of 24 groups defined by geographic region, urban-rural residence, and race, based on the characteristics of the sample primary units in 1960. The second-stage ratio estimate makes use of independent estimates of the current population by race, age, and sex to modify the inflation factors produced by the first stage.

The composite estimate reduces the variance further for many statistics by taking into account the rotation of the sample from month to month. At any given time, the sample consisted of eight subsamples, called rotation groups, of about 4,000 households each. Each rotation group itself constitutes a national sample. The rotation of the sample is such that each rotation group is retained in the sample for four successive calendar months, dropped from the sample for the next eight calendar months, and then included in the sample again for four calendar months. The rotation pattern is such that in any given month six of the rotation groups were also in the sample the preceding month. The composite estimate for month j is of the form

$$(5) \qquad x''_j = \omega[x''_{j-1} + x'_{Ij} - x'_{I,j-1}] + (1-\omega)x'_j$$

where ω is a weight between 0 and 1, x'_j is the two-stage ratio estimate of the number of people having a particular characteristic based on the whole sample surveyed at time j, x'_{Ij} is the same kind of two-stage ratio estimate for month j based only on the six rotation groups that are in the sample in both months j and j-1, $x'_{I,j-1}$ is the same kind of two-stage ratio estimate for month j-1 based on the same six rotation groups, and x''_{j-1} is the composite estimate obtained for month j-1.

It can be seen that the nature of the estimator is such that a large amount of data needs to be available and processed for each replication, including data for previous months.

* For a more complete description, see [10] and [11].

McCarthy [8] has suggested a way of controlling the selection of half-samples so as to reduce the variability of the variance estimator. He has shown that, for linear statistics based on a sample of L strata, half-sample selections balanced in a certain way provide variance estimates as precise as if all possible 2^L half-samples had been used. He has also suggested the use of partially balanced replicates. From results obtained by him and by Margaret Gurney [3], [4] of the Bureau of the Census, it appears that the variability can be reduced significantly by using balanced or partially balanced replicates rather than a purely random selection of replicates. For a sample design like the Current Population Survey, the reduction in variance seems to be of the order of 1/3 when 40 partially balanced replicates are used, or about 1/6 when 20 are used. On the other hand, the variance for 40 partially balanced replicates seems to be about twice that for a completely balanced set of replicates, and the variance for 20 about four times the variance for a completely balanced set (see [4]). If these degrees of variability are not tolerable, and if larger numbers of replicates are costly, alternative ways of estimating variances are attractive.

3. Direct methods of estimating variances

The distinction between "direct" methods of estimating variances and replication methods is not always a real one. For example, in estimating the variance of the mean of a simple random sample, one may calculate the variance among the means of random subgroups (which may be the elementary units themselves) and then make use of the fact that the desired variance is a known function of the expected values of the calculated variance.

However, in the context of a complex sample design, for example a multistage sample design with two or more primary sampling units selected from each of several strata, an alternative to the use of pseudo-replication of half-samples is the following, based on an approximate linearization of the statistic involved. We may refer to equation (3), from which it is clear that

$$(6) \qquad \sigma_r^2 \doteq \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{X}^2}\sigma_z^2$$

if the variable z is defined by

$$(7) \qquad z = y - Rx.$$

The estimation of σ_z^2 may well be easier and simpler than the estimation of σ_x^2, σ_y^2 and σ_{xy}.

This illustration is a special case of a long-known* attack on the problem of estimating a variance. The Bureau of the Census has abandoned the use of half-samples for the Current

* See for example Deming ([1], Chapter III), Kendall and Stuart ([6], Sec. 10.6), and Keyfitz [7].

Population Survey and is now estimating the variances of the quite complex composite estimator described earlier by a direct procedure. For estimating the variance of seasonally adjusted statistics, we continue to employ the replication estimator.

The direct procedure essentially amounts to calculating a linear combination of sample totals for each primary sampling unit, and then estimating the variance of the sum of those linear combinations. Thus the problem has been reduced to the simple problem of estimating the variance of a total.

The procedure may be described in the following way. Let $u = (u_1, u_2, \ldots, u_k)$ be a vector of statistics whose expected value is a vector of population parameters $U=(U_1,U_2,\ldots,U_k)$. Suppose that the population parameter of interest is a function $f(U)$, and is to be estimated by $f(u)$. To terms of the first degree in (u_i-U_i), the Taylor's series approximation for $f(u)$ is given by

$$(8) \qquad f(u) \doteq f(U) + \sum_{i=1}^{k} (u_i-U_i)\frac{\partial f(U)}{\partial U_i}$$

and hence the variance and the mean square error of $f(u)$ are, to this approximation, the same as those of the linear function

$$(9) \qquad \ell(u) = \sum_{i=1}^{k} u_i \frac{\partial f(U)}{\partial U_i}$$

where the partial derivatives are to be evaluated at $u = U$.

This "linearization" of the estimator $f(u)$ is frequently useful for the estimation of the variance of a complex estimator. The ratio estimator exhibited earlier is one example. A more complex example, the estimation of the variance of a regression coefficient based on a stratified multistage sample, can be given.

Let x_{hi}, y_{hi} denote values of the variables x, y associated with the i-th elementary sampling unit selected for the sample in stratum h. We consider the statistic

$$(10) \qquad b = \frac{\frac{1}{n}\sum_h\sum_i x_{hi}y_{hi} - (\frac{1}{n}\sum_h\sum_i x_{hi})(\frac{1}{n}\sum_h\sum_i y_{hi})}{\frac{1}{n}\sum_h\sum_i x_{hi}^2 - (\frac{1}{n}\sum_h\sum_i x_{hi})^2}$$

where $n = \sum_h n_h$ denotes the number of elementary units in the sample. This statistic b is sometimes taken to estimate the regression coefficient of y on x in the population, and we may be concerned in estimating its variance. Let us introduce new variables by means of the notation:

$$(11) \quad \begin{cases} w_h = \sum_i x_{hi}y_{hi} & w = \sum_h w_h \\ u_h = \sum_i x_{hi}^2 & u = \sum_h u_h \\ x_h = \sum_i x_{hi} & x = \sum_h x_h \\ y_h = \sum_i y_{hi} & y = \sum_h y_h. \end{cases}$$

Then b may be written

$$(12) \qquad b = \frac{nw - xy}{nu - x^2}.$$

If N, W, U, X, Y denote the expectations of n, w, u, x, y respectively, the derivatives required are

$$(13) \quad \begin{cases} \frac{\partial b}{\partial N} = \frac{X(UY-WX)}{(NU-X^2)^2} \\ \frac{\partial b}{\partial W} = \frac{N}{NU-X^2} \\ \frac{\partial b}{\partial U} = -\frac{N(NW-XY)}{(NU-X^2)^2} \\ \frac{\partial b}{\partial X} = \frac{-NUY+2NWX-X^2Y}{(NU-X^2)^2} \\ \frac{\partial b}{\partial Y} = -\frac{X}{NU-X^2}. \end{cases}$$

As before, we approximate

$$(14) \quad \mathrm{Var}(b) \doteq \mathrm{Var}(n\frac{\partial b}{\partial N} + w\frac{\partial b}{\partial W} + u\frac{\partial b}{\partial U} + x\frac{\partial b}{\partial X} + y\frac{\partial b}{\partial Y}).$$

The right-hand member may be written as a sum over the strata, so that

$$(15) \quad \mathrm{Var}(b) \doteq \mathrm{Var}\sum_h(\frac{\partial b}{\partial N}n_h + \frac{\partial b}{\partial W}w_h + \frac{\partial b}{\partial U}u_h + \frac{\partial b}{\partial X}x_h + \frac{\partial b}{\partial Y}y_h)$$

$$\doteq \sum_h \mathrm{Var}(\frac{\partial b}{\partial N}n_h + \frac{\partial b}{\partial W}w_h + \frac{\partial b}{\partial U}u_h + \frac{\partial b}{\partial X}x_h + \frac{\partial b}{\partial Y}y_h)$$

since sampling is independent in the several strata. Thus the estimation of the variance of the estimator b has been reduced to the problem of estimating the variance of a linear combination of sample sums for each stratum.

The manner in which the variance of that linear combination

$$(16) \quad \ell_h = \frac{\partial b}{\partial N}n_h + \frac{\partial b}{\partial W}w_h + \frac{\partial b}{\partial U}u_h + \frac{\partial b}{\partial X}x_h + \frac{\partial b}{\partial Y}y_h$$

may be estimated will, of course, depend upon the sample design. If, for example, the sample within each stratum is a single-stage, simple random sample of elementary units, the sample may be subdivided into, say, m equal, random subgroups. The linear expression

$$(17)\quad \ell_{hj} = \frac{\partial b}{\partial N} n_{hj} + \frac{\partial b}{\partial W} w_{hj} + \frac{\partial b}{\partial U} u_{hj} + \frac{\partial b}{\partial X} x_{hj} + \frac{\partial b}{\partial Y} y_{hj}$$

is then formed for each of the subgroups, so that

$$(18)\quad \ell_h = \sum_j \ell_{hj}$$

and the variance estimate is based on the variance among the ℓ_{hj}. On the other hand, if two or more primary sampling units were selected from each stratum and then subsampled, a quantity ℓ_{hj} may be formed for each primary unit and the variance of ℓ_h estimated from the differences among the ℓ_{hj} in precisely the same manner as the variance x_h is estimated from the differences among the x_{hj}.

One difficulty that cannot be ignored is that the coefficients in the linear form ℓ_h are unknown population parameters. The usual practice is to substitute sample estimates in the expressions for the derivatives, just as in the case of the ratio estimates one substitutes the sample ratio r = y/x for the population ratio R = Y/X in the Taylor's approximation to the variance of r. For large samples, this procedure yields satisfactory estimates. In the case of estimating the variance of the regression coefficient b, one would take

$$(19)\quad \begin{cases} \frac{\partial b}{\partial N} \doteq \frac{\bar{x}(\bar{y}-b\bar{x})}{ns_x^2} \\ \frac{\partial b}{\partial W} \doteq \frac{1}{ns_x^2} \\ \frac{\partial b}{\partial U} \doteq \frac{b}{ns_x^2} \\ \frac{\partial b}{\partial X} \doteq -\frac{\bar{y}-2b\bar{x}}{ns_x^2} \\ \frac{\partial b}{\partial Y} \doteq -\frac{\bar{x}}{ns_x^2} \end{cases}$$

where $\bar{x} = x/n$, $\bar{y} = y/n$, $\bar{u} = u/n$, $s_x^2 = \bar{u}-\bar{x}^2$.

The illustration can be extended in a straightforward way to multiple regression coefficients. If x_{ij} is taken to be the value of the i-th regressor variable for the j-th element of the sample, the estimates b_i of the regression coefficients are computed as the solution of the system of linear equations

$$(20)\quad \sum_{i=0}^{p} \sum_{j=1}^{n} x_{kj}x_{ij}b_i = \sum_{j=1}^{n} x_{kj}y_j \qquad k = 0, 1, \ldots, p$$

where $x_{0j} = 1$. We introduce new variables u_{ki} and v_k, defined by

$$(21)\quad \begin{cases} u_{ki} = \sum_{j=1}^{n} x_{kj}x_{ij} \\ v_k = \sum_{j=1}^{n} x_{kj}y_j. \end{cases}$$

The system (20) can then be written

$$(22)\quad \sum_{i=0}^{p} u_{ki}b_i = v_k \qquad k = 0, 1, \ldots, p.$$

Differentiation with respect to $u_{h\ell}$ yields

$$(23)\quad \sum_{i=0} u_{ki} \frac{\partial b_i}{\partial u_{h\ell}} = -(1-\delta_{h\ell})\delta_{kh}b_\ell - \delta_{k\ell}b_h \qquad h, k, \ell = 0, 1, \ldots, p$$

and differentiation with respect to v_h yields

$$(24)\quad \sum_{i=0}^{p} u_{ki} \frac{\partial b_i}{\partial v_k} = \delta_{hk} \qquad h, k = 0, 1, \ldots, p$$

where δ_{ij} is the Kronecker delta. The system (23) can be subdivided in $(p+1)^2$ subsystems $(h, \ell = 0, 1, \ldots, p)$, each of p+1 equations in p+1 variables $\left(\frac{\partial b_i}{\partial u_{h\ell}}, i = 0, 1, \ldots, p\right)$. Actually, only $\frac{1}{2}(p+1)(p+2)$ of the subsystems are distinct because of the symmetry $u_{h\ell} = u_{\ell h}$. The system (24) can be subdivided into p+1 systems $(h = 0, 1, \ldots, p)$, each of p+1 equations in p+1 variables $\left(\frac{\partial b_i}{\partial v_h}, i = 0, 1, \ldots, p\right)$. Thus the determination of the coefficients of the linear approximations to the regression coefficients will require the solution of $\frac{1}{2}(p+1)(p+4)$ systems while the half-sample method using L replications will require the solution of L systems of the same size. The amount of computation depends then on the relative values of L and $\frac{1}{2}(p+1)(p+4)$.

The limitations of this approach arise primarily from the use of the Taylor's series, for precautions must be taken to assure that the linear approximation is acceptably good. With sufficiently large sample sizes, this can usually be assured. An illustration is provided by the variance of the estimated variance for a simple random sample of size n. Here the statistic whose variance is desired is

$$(25)\quad s^2 = \frac{1}{n-1}\sum^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\Sigma x_i^2 - \frac{1}{n(n-1)}(\Sigma x_i)^2.$$

If we take the variables in the estimator function f(u) to be the elementary variables x_i, then

the Taylor's series is identical with the function, all of whose terms are quadratic in the x_i. Thus the linear approximation to s^2 is taken to be zero, so that the variance is estimated to be zero. The difficulty here is that the Taylor approximation is not necessarily a good one when the variables in the Taylor's series are the x_i, each of which is based on a single sample observation. But if we define new variables u and v by

$$(26) \qquad \begin{cases} u = \frac{1}{n} \Sigma x_i^2 \\ v = \frac{1}{n} \Sigma x_i \end{cases}$$

then s^2 may be written

$$(27) \qquad s^2 = \frac{n}{n-1} (u-v^2)$$

so that, evaluated at the expected values of u and v,

$$(28) \qquad \begin{cases} \frac{\partial s^2}{\partial u} = \frac{n}{n-1} \\ \frac{\partial s^2}{\partial v} = -\frac{2n\bar{X}}{n-1} . \end{cases}$$

The linear form ℓ is then

$$(29) \qquad \ell = \frac{n}{n-1} (u-2\bar{X}v)$$

whose variance is easily found to be

$$(30) \qquad \mathrm{Var}(\ell) = \frac{n}{(n-1)^2} (\mu_4 - \mu_2^2)$$

where μ_2, μ_4 denote the second and fourth moments of x. The actual variance of s^2 is

$$(31) \qquad \mathrm{Var}(s^2) = \frac{n}{(n-1)^2} (\mu_4 - \frac{n-3}{n-1} \mu_2^2)$$

which differs only trivially from $\mathrm{Var}(\ell)$ for sufficiently large n. Thus the proper choice of the variables used in this approach can be important.

4. Precision of variance estimates

It is easy to exaggerate the precision of estimates of variance for complex surveys, whether the variance is estimated by replication or by other methods, including those mentioned above. Discussions of the precision of the variance estimate usually assume that the contributions to the variance from the individual primary strata are approximately equal. Experience indicates that this is far from true. In one example (Table 1) in which there were 120 primary strata, a single stratum contributed 40 percent of the total variance between primary sampling units. Four other strata made an additional contribution of 21 percent. As a result (see Table 1), a single stratum contributed more than 80 percent of the variance of the estimated variance, and 5 of the 120 strata contributed about 95 percent of the total.

Table 2 lists the contribution of the top 5 strata to the variance between strata, the coefficient of variation of that variance estimated by the collapsed stratum method, and the percentage contribution of the top 5 strata to the variance of the estimated variance, for a number of estimates of totals and ratios. The table emphasizes the marked inequality of the contributions of the individual strata.

Tables 1 and 2 were concerned with the between-psu component of the variance. The distribution of the total variance among the strata will be somewhat less skewed, but may still be quite marked. Even in the extreme case when the within-psu variance of an estimated total is precisely the same for each of L strata, it can be shown that the proportion of the total variance contributed by stratum h is

$$(32) \qquad (\frac{1}{L} - K_h)P + K_h$$

where P is the ratio of the within-psu component of the variance to the total variance and K_h is the contribution of stratum h to the between-psu variance. Thus, for example, if the within-psu component is half the total variance, a stratum that contributed 40 percent of the between-psu variance would contribute more than 20 percent of the total variance.

Table 1: PERCENTAGE CONTRIBUTIONS TO THE BETWEEN-PSU VARIANCE AND TO THE VARIANCE OF THE ESTIMATED VARIANCE, OF THE ESTIMATED NET INCREASE OF THE U.S. POPULATION FROM MIGRATION, 1955-1960.

Stratum	Variance (%)	Variance of estimated variance (%)	Stratum	Variance (%)	Variance of estimated variance (%)	Stratum	Variance (%)	Variance of estimated variance (%)
1	0	0	41	.260	.010	81	.003	.000
2	.367	.002	42	.210	.152	82	.004	.000
3	.881	.012	43	0	0	83	.805	.130
4	.496	.014	44	.051	.000	84	.380	.007
5	.424	.009	45	.151	.002	85	.087	.000
6	.161	.003	46	0	0	86	.025	.000
7	.014	.000	47	.214	.002	87	.051	.000
8	.002	.000	48	.002	.000	88	.000	.000
9	.325	.001	49	.059	.000	89	.116	.003
10	.402	.003	50	.714	.018	90	.003	.000
11	.215	.000	51	.150	.001	91	0	0
12	.001	.000	52	.025	.000	92	0	0
13	.029	.000	53	.742	.032	93	.369	.003
14	.581	.043	54	.267	.001	94	0	0
15	.001	.000	55	.013	.000	95	.006	.000
16	.177	.000	56	.406	.001	96	1.586	1.755
17	.052	.000	57	38.986	81.883	97	.063	.001
18	.781	.011	58	0	0	98	.280	.008
19	.336	.002	59	.011	.000	99	.001	.000
20	.225	.000	60	.004	.000	100	.274	.007
21	.114	.000	61	.002	.000	101	0	0
22	.014	.000	62	.376	.048	102	2.427	.808
23	.000	.000	63	.004	.000	103	2.476	.268
24	.040	.000	64	0	0	104	.332	.002
25	.181	.001	65	0	0	105	.441	.004
26	.092	.000	66	0	0	106	.573	.004
27	3.155	.423	67	.009	.000	107	2.303	.123
28	.643	.008	68	0	0	108	1.598	.108
29	.021	.000	69	0	0	109	2.495	.144
30	.298	.003	70	0	0	110	.063	.000
31	.004	.000	71	.054	.000	111	1.624	.109
32	.245	.006	72	.702	.037	112	.772	.060
33	.056	.000	73	.050	.000	113	1.203	.030
34	.005	.000	74	.908	.089	114	.997	2.018
35	1.147	.510	75	.005	.000	115	5.731	3.563
36	.012	.000	76	.004	.000	116	.047	.000
37	.560	.023	77	.024	.000	117	.334	.011
38	.138	.002	78	.005	.000	118	.105	.005
39	.002	.000	79	5.160	1.519	119	.764	.104
40	.155	.028	80	6.882	5.665	120	2.860	.162

Each "stratum" listed above consists of a pair or a triple of non-certainty strata of the Census Bureau's Current Population Survey as used in 1966. All but 5 of the 120 groups are pairs. The estimated variance employs the collapsed group method as described in [5], Vol. I, Chapter 9, Sections 15 and 28.

Table 2: CONCENTRATION OF THE BETWEEN-PSU VARIANCE AND OF THE VARIANCE OF THE ESTIMATED VARIANCE, FOR SPECIFIED STATISTICS.

[The entries in columns (1) and (3) are the proportions (of the variances and of the variance of the estimated variance, respectively) contributed by the 5 "strata" that are the largest contributors in each case. For the definition of a "stratum" see footnote to Table 1.]

Statistic	Concentration of variance (1)	Coefficient of variation of estimated variance (2)	Concentration of variance of estimated variance (3)
Estimates of totals:			
Total population, 1960	.38	.41	.90
Rural-farm population, 1960	.37	.16	.40
Non-white population, 1960	.37	.23	.74
Live births, 1960	.31	.27	.75
Marriages, 1960	.78	1.87	.997
Number of families, 1960	.39	.36	.86
Families with 1959 income less than $3,000	.29	.15	.43
Aggregate income in 1959	.39	.29	.82
Elementary school enrollment, 1960	.31	.28	.77
High school enrollment, 1960	.29	.26	.75
Net gain through migration, 1950-1960	.60	.66	.95
Civilian labor force, 1960	.31	.31	.86
Unemployed persons, 1960	.28	.21	.63
Employed persons, 1960	.31	.31	.87
Employed in agriculture, 1960	.38	.16	.65
Employed in manufacturing, 1960	.18	.15	.38
Employed in wholesale or retail trade, 1960	.28	.19	.62
Housing units, 1960	.41	.35	.78
Vacant housing units, 1960	.68	.65	.97
Bank deposits, 1960	.31	.30	.84
Taxable payroll, January-March, 1959	.68	4.33	1.00
Value added by manufacture, 1958	.24	.14	.41
Retail sales, 1958	.23	.23	.61
Retail sales, 1954	.24	.30	.85
Wholesale sales, 1958	.46	.34	.92
Estimates of ratios to total population:			
Rural-farm population, 1960	.36	.15	.60
Non-white population, 1960	.39	.23	.74
Live births, 1960	.22	.21	.57
Marriages, 1960	.78	1.89	.997
Families with 1959 income less than $3,000	.26	.15	.46
Aggregate income in 1959	.44	.22	.85
Elementary school enrollment, 1960	.22	.21	.76
High school enrollment, 1960	.21	.22	.70
Net gain through migration, 1950-1960	.62	.68	.95
Civilian labor force, 1960	.29	.29	.84
Unemployed persons, 1960	.24	.20	.64
Employed persons, 1960	.20	.22	.65
Employed in manufacturing, 1960	.19	.14	.42
Employed in durable goods manufacturing, 1960	.26	.17	.54
Employed in wholesale or retail trade, 1960	.39	.42	.91
Housing units, 1960	.47	.48	.91
Vacant housing units, 1960	.69	.66	.97
Bank deposits, 1960	.26	.32	.82
Taxable payroll, January-March, 1959	.71	4.52	1.00
Value added by manufacture, 1958	.24	.14	.36
Retail sales, 1958	.19	.34	.85

Table 2: CONCENTRATION OF THE BETWEEN-PSU VARIANCE AND OF THE VARIANCE OF THE ESTIMATED VARIANCE, FOR SPECIFIED STATISTICS - continued.

[The entries in columns (1) and (3) are the proportions (of the variances and of the variance of the estimated variance, respectively) contributed by the 5 "strata" that are the largest contributors in each case. For the definition of a "stratum" see footnote to Table 1.]

Statistic	Concentration of variance (1)	Coefficient of variation of estimated variance (2)	Concentration of variance of estimated variance (3)
Estimates of ratios to total population continued:			
Retail sales, 1954.....	.24	.39	.89
Wholesale sales, 1958.....	.43	.32	.90
Estimates of other ratios:			
Employed in agriculture/total employed, 1960.....	.37	.16	.64
Employed in durable goods/total in manufacturing, 1960.....	.17	.15	.39
Vacant/total housing units, 1960..	.70	.66	.97
Retail sales, 1958/1954.....	.47	.35	.81
Wholesale sales, 1959/1958.....	.44	.34	.94

REFERENCES

[1] Deming, W. Edwards. (1943). Statistical Adjustment of Data. New York: John Wiley & Sons, Inc.

[2] Deming, W. Edwards. (1960). Sampling Design in Business Research. New York: John Wiley & Sons, Inc.

[3] Gurney, Margaret. (1964). "McCarthy's Orthogonal Replications for Estimating Variances, with Grouped Strata." Unpublished memorandum, U.S. Bureau of the Census, to be published in Technical Notes, No. 3, (1968).

[4] Gurney, Margaret. (1962). "The Variance of the Replication Method for Estimating Variances for the CPS Sample Design." Unpublished memorandum, U.S. Bureau of the Census, to be published in Technical Notes, No. 3, (1968).

[5] Hansen, M.H., W.N. Hurwitz, and W.G. Madow. (1953). Sample Survey Methods and Theory, Vol. I. New York: John Wiley & Sons, Inc.

[6] Kendall, M.G. and A. Stuart. (1958). The Advanced Theory of Statistics, Vol. I. London: Charles Griffin & Company Limited.

[7] Keyfitz, Nathan. (1957). "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum." Journal of the American Statistical Association: 52, 503-510.

[8] McCarthy, Philip J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. National Center for Health Statistics. Vital and Health Statistics, PHS Pub. No. 1000, Series 2, No. 14. Public Health Service, Washington: U.S. Government Printing Office.

[9] Tukey, J.W. (1958). "Bias and Confidence in Not-Quite Large Samples." Abstracted in Annals of Mathematical Statistics: 29, 614.

[10] U.S. Bureau of the Census. (1963). The Current Population Survey, A Report on Methodology. Technical Paper No. 7. Washington: U.S. Government Printing Office.

[11] U.S. Bureau of the Census. (1967). Concepts and Methods Used in Manpower Statistics from the Current Population Survey. Current Population Reports, Series P-23, No. 22. Washington: U.S. Government Printing Office.