SOME THEOREMS RELATING POSTSTRATIFICATION AND SAMPLE CONFIGURATION

Dhiren Ghosh, Statistical Consultant, and Andrew Vogt, Georgetown University

Dhiren Ghosh, 1714 Rupert Street, McLean, VA 22101

KEY WORDS: conditional variance and bias, proportional and optimum allocations

1. INTRODUCTION. In an earlier paper [1] the authors observed that a poststratified mean is often superior to the regular mean as an estimator of population mean if conditional variance or conditional mean square error is used for comparing estimators and the condition is the sample configuration actually obtained.

The conditional variance has a computational advantage since one need not estimate the expected value of $\frac{1}{n_i}$ where n_i is the number of units in the i-th stratum of the sample, but it also has theoretical advantages noted in [1] and elsewhere (see [2] and [3]). Not only do the conditional measures give satisfactory confidence intervals and error bars, but they permit one to distinguish which estimator - the regular mean or the poststratified mean - is preferable for a particular configuration. The information extracted from the sample is thereby enhanced.

Indeed, the following rules of thumb, based on the sample configuration actually achieved, can be applied:

i) if large variance strata are somewhat overrepresented in the sample and small variance strata somewhat underrepresented, the poststratified mean is preferable;

ii) if the configuration is near optimum, for example, intermediate between proportional and pseudoproportional ($n_i \propto N_i S_i^k$, $0 < k < 2$), the poststratified mean is preferable;

iii) if the sample size is sufficiently small and the stratum with the largest variance has sufficiently large variance and all other strata are overrepresented in the sample, the regular mean is preferable.

In this note we state and prove theorems that support these assertions.

2. NOTATION. A finite population is divided into k strata of known sizes N_1, N_2, ..., N_k, with $N = \sum_i N_i$. A variable X is defined on the units of the population. Its mean and variance over the whole population are $\overline{X}$ and S^2, and over the i-th stratum $\overline{X}_i$ and S_i^2 for $i = 1, .., k$. A random sample of size n is drawn from the population, yielding n_1, n_2, ..., n_k units in the respective strata, with $\sum_i n_i = n$, and the values of X are determined on the units of the sample. To estimate the population mean $\overline{X}$, two estimators are available:

the regular sample mean $\overline{x}$

and

the poststratified sample mean

$$\overline{x}_{pst} = \sum_i \frac{N_i}{N} \overline{x}_i,$$

where $\overline{x}_i$ is the mean of X on the units of the sample that belong to the i-th stratum.

All samples are assumed to be of known size n. In addition, we consider two conditions:

$\{n_i \geq 1\}$: the sample size in each stratum is ≥ 1

and

$\{n_i\}$: the sample size in each stratum is known.

When means, variances, etc. are calculated with respect to arbitrary samples of size n, one commonly calls them "unconditional" means, variances, etc. The condition $\{n_i \geq 1\}$ is imposed in order that $\overline{x}_{pst}$ be defined. We assume that $n \geq k$ so that this condition can be achieved. The second condition $\{n_i\}$ asserts

that the sample configuration, i. e., the k-tuple $(n_1, n_2, ..., n_k)$, is known. When the configuration is known, we assume that it satisfies $n_i \geq 1$ for each i.

The estimator $\overline{x}$ satisfies:

$$E(\overline{x}) = \overline{X} \text{ and } V(\overline{x}) = (1 - \frac{n}{N}) \cdot \frac{S^2}{n},$$

while the other estimator $\overline{x}_{pst}$ satisfies:

$$E(\overline{x}_{pst}/\{n_i \geq 1\}) = \overline{X}$$

and

$$V(\overline{x}_{pst}/\{n_i \geq 1\}) = \sum_i (\frac{N_i}{N})^2 \cdot S_i^2 \cdot (E(\frac{1}{n_i}/\{n_i \geq 1\}) - \frac{1}{N_i}).$$

Both estimators are unbiased estimators of the true mean $\overline{X}$, but with regard to different sets of samples. In general, $\overline{x}$ is biased when the condition $\{n_i \geq 1\}$ is imposed. Likewise, if the estimator $\overline{x}_{pst}$ is extended to all samples of size n (e.g. by setting $\overline{x}_i = 0$ whenever $n_i = 0$), this estimator in general is also biased. These biases were overlooked in [1] as well as in other studies.

When we pass to a given sample configuration, the two estimators behave as follows:

$$E(\overline{x}/\{n_i\}) = \sum_i \frac{n_i}{n} \overline{X}_i$$

and

$$E(\overline{x}_{pst}/\{n_i\}) = \overline{X} = \sum_i \frac{N_i}{N} \overline{X}_i,$$

so that the first estimator is in general conditionally biased while the second is not. The conditional mean square error of the first estimator $\overline{x}$ is:

$$MSE(\overline{x}/\{n_i\}) =$$

$$\sum_i (\frac{n_i}{n})^2 (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i} + \left[\sum_i (\frac{n_i}{n} - \frac{N_i}{N}) \overline{X}_i\right]^2, \quad (2.1)$$

where the first term is the conditional variance and the second the conditional bias squared. The conditional variance of the second estimator $\overline{x}_{pst}$ is:

$$V(\overline{x}_{pst}/\{n_i\}) = \sum_i (\frac{N_i}{N})^2 (1 - \frac{n_i}{N_i}) \frac{S_i^2}{n_i}. \quad (2.2)$$

3. THEOREMS. Given the sample configuration $\{n_i\}$, a natural way to choose between $\overline{x}$ amd $\overline{x}_{pst}$ is to estimate the mean square error and variance in equations (2.1) and (2.2), and - provided one is comfortable with one's estimates of these quantities - use the estimator that gives the lower value. The theorems below indicate some situations that can arise. In all cases the measure of precision is the conditional mean square error (MSE) or conditional variance, relative to the condition that the sample configuration is given.

Theorem 3.1: Suppose that the strata and the configuration $\{n_i\}$ satisfy, for some real numbers M and d with $M > d > 0$:

i) $\left(\frac{M+d}{M}\right)\left(\frac{N_i}{N} - \frac{n_i}{N} - \frac{n_i^2}{nN_i}\right) \geq \frac{n_i}{n} \geq \frac{N_i}{N}$ for all i such that $S_i^2 \geq M + d$;

ii) $\frac{n_i}{n} = \frac{N_i}{N}$ for all i such that $|S_i^2 - M| < d$; and

iii) $\frac{N_i}{N} \geq \frac{n_i}{n} \geq \left(\frac{M-d}{M}\right)\left(\frac{N_i}{N} - \frac{n_i}{N} - \frac{n_i^2}{nN_i}\right)$ for all i such that $S_i^2 \leq M - d$.

Then $V(\overline{x}_{pst}/\{n_i\}) \leq V(\overline{x}/\{n_i\})$.

Before proving Theorem 3.1, let us interpret it. Roughly speaking, it is statement i) of the Introduction: if large variance strata are somewhat overrepresented in the sample and small variance strata somewhat underrepresented, then the variance of the poststratified mean is smaller than that of the regular mean relative to the given sample configuration. Since the regular mean may also be biased, this conclusion is more than sufficient to guarantee that the poststratified mean is as precise as or more precise than the regular mean.

Theorem 3.1 hypothesizes that strata with variances somewhat larger than M are oversampled while strata with variances somewhat smaller than M are undersampled. Since $\sum_i \frac{n_i}{n}$

$= 1$, not all strata can be oversampled nor can all be undersampled. Thus M must be a number intermediate between the smallest and largest of the stratum variances. M can be regarded as a kind of average value for S_i^2, $i = 1, ..., k$.

If all strata have the same variance, then M is this common value and the theorem requires that the configuration be proportional ($n_i \propto N_i$). In the general case, by hypothesis ii) strata having variances within d units from M are sampled proportionally. This restriction can be rendered vacuous by choosing M and d so that no stratum is eligible. If the strata are indexed in order of increasing variance: $S_1^2 \leq S_2^2 \leq ... \leq S_k^2$, and if M is chosen intermediate between two successive variances: $S_i^2 < M < S_{i+1}^2$, with $d = \min(S_{i+1}^2 - M, M - S_i^2)$, then Theorem 3.1 requires that strata $i+1, ..., k$ be oversampled while strata $1, ..., i$ are undersampled.

The remaining restriction is that the oversampling and undersampling not be excessive. If we ignore the terms $\frac{n_i}{N}$ and $\frac{n_i^2}{nN_i}$ by considering a population large in each stratum compared to the sample size, the theorem requires that $1 + \frac{d}{M} \geq (\frac{n_i}{n})/(\frac{N_i}{N}) \geq 1$ in any oversampled stratum, and $1 - \frac{d}{M} \leq (\frac{n_i}{n})/(\frac{N_i}{N}) \leq 1$ in any undersampled stratum. To provide maximum leeway, one might choose $\frac{d}{M}$ as large as possible - say, $M = \frac{1}{2} \cdot (S_i^2 + S_{i+1}^2)$ and $d = \frac{1}{2} \cdot (S_{i+1}^2 - S_i^2)$ where i is an index giving the largest value for $\frac{d}{M} = (S_{i+1}^2 - S_i^2)/(S_i^2 + S_{i+1}^2)$.

Proof of Theorem 3.1: Let $\sum_1$ respectively $\sum_2$ denote sums over indices i satisfying i) respectively iii). From (2.1) and (2.2) we obtain:

$$V(\overline{x}/\{n_i\}) - V(\overline{x}_{pst}/\{n_i\}) =$$

$$\left(\sum_1 + \sum_2\right)\left((\frac{n_i}{n})^2 - (\frac{N_i}{N})^2\right)(1 - \frac{n_i}{N_i})\frac{S_i^2}{n_i} \geq$$

$$\sum_1(\frac{n_i}{n} - \frac{N_i}{N})(\frac{n_i}{n} + \frac{N_i}{N})(1 - \frac{n_i}{N_i})(\frac{M+d}{n_i})$$

$$-\sum_2(\frac{N_i}{N} - \frac{n_i}{n})(\frac{n_i}{n} + \frac{N_i}{N})(1 - \frac{n_i}{N_i})(\frac{M-d}{n_i})$$

$$= \frac{2d}{n}\sum_1(\frac{n_i}{n} - \frac{N_i}{N})$$

$$+(M+d)\sum_1 \frac{1}{n_i}(\frac{n_i}{n} - \frac{N_i}{N})(\frac{N_i}{N} - \frac{n_i}{N} - \frac{n_i^2}{nN_i})$$

$$-(M-d)\sum_2 \frac{1}{n_i}(\frac{N_i}{N} - \frac{n_i}{n})(\frac{N_i}{N} - \frac{n_i}{N} - \frac{n_i^2}{nN_i})$$

$$\geq \frac{2d}{n}\sum_1(\frac{n_i}{n} - \frac{N_i}{N})$$

$$+\frac{M}{n}\sum_1(\frac{n_i}{n} - \frac{N_i}{N}) - \frac{M}{n}\sum_2(\frac{N_i}{N} - \frac{n_i}{n})$$

$$= \frac{2d}{n}\sum_1(\frac{n_i}{n} - \frac{N_i}{N}) \geq 0.\square$$

The next theorem indicates what happens when the configuration is of the form $n_i \propto N_i S_i^k$. This includes several familiar configurations - namely proportional ($k = 0$), optimum ($k = 1$), and pseudoproportional ($k = 2$). The optimum configuration is the configuration that gives the smallest conditional variance for the poststratified mean, the condition being the given sample configuration. However, the optimum configuration may not be the configuration where the poststratified mean gives the greatest improvement in precision over the regular mean. The pseudoproportional configuration is dual to the proportional, in the sense that both configurations give the same conditional variance for the poststratified mean. More generally, $n_i \propto N_i S_i^k$ is dual to $n_i \propto N_i S_i^{2-k}$ in this sense, and the optimum configuration is self-dual.

Theorem 3.2: Let k be a real number in the interval $[0, 2]$, and let $n_i \propto N_i S_i^k$ for $i = 1, ..., k$. Then $V(\overline{x}_{pst}/\{n_i\}) \leq V(\overline{x}/\{n_i\})$ provided the sampling fraction $\frac{n}{N}$ can be ignored.

Proof: As in the proof of (3.1)

$$V(\overline{x}/\{n_i\}) - V(\overline{x}_{pst}/\{n_i\}) =$$

$$\sum_i\left((\frac{n_i}{n})^2 - (\frac{N_i}{N})^2\right)(1 - \frac{n_i}{N_i})\frac{S_i^2}{n_i} =$$

$$\sum_i\left(C_k^2\frac{S_i^{2k}}{n^2} - \frac{1}{N^2}\right)N_i\left(1 - \frac{C_k N_i S_i^k}{N_i}\right)\frac{S_i^{2-k}}{C_k} =$$

$$\frac{1}{n}\left(\frac{u_{k+2}}{u_k} - u_k u_{2-k}\right) - \frac{1}{N}\left(\frac{u_{2k+2}}{u_k^2} - u_2\right),$$

where $u_j = \sum_i \frac{N_i}{N} S_i^j$ and $C_k = \frac{n}{N u_k}$. Note that since $n_i \propto N_i S_i^k$ for all i, $\frac{n_i}{N_i} = C_k S_i^k$.

If we ignore the second term on the right on the assumption that $\frac{n}{N}$ is very small, it suffices to establish that

$$u_{k+2} \geq (u_k)^2 u_{2-k},$$

or after the application of the logarithm function that

$$g(k+2) \geq 2g(k) + g(2-k), \qquad (*)$$

where $g(x) = \log u_x = \log \sum_i \frac{N_i}{N} S_i^x$ for a real number x. But g satisfies $g(0) = 0$ and $g''(x) \geq 0$ for $x \geq 0$ by the Cauchy-Schwarz inequality (we omit details). Hence g is subadditive on the nonnegative reals, and $(*)$ holds for $0 \leq k \leq 2$. □

Now we give a result in the opposite direction.

Theorem 3.3: Suppose that $N > 2n$, that the k-th stratum has the largest variance, and with

$$D_k = \max_i \{|\overline{X}_i - \overline{X}_k|\}$$

that

$$S_k^2 \geq \max_{i=1,\ldots,k-1} \left(S_i^2 + \frac{2n}{1 - \frac{2n}{N}}\left(\frac{N_k}{N}\right) D_k |\overline{X}_i - \overline{X}_k|\right). \qquad (3.3)$$

If

$$\frac{n_i}{n} \geq \frac{N_i}{N} \text{ for } i = 1, \ldots, k-1,$$

then $MSE(\overline{x}/\{n_i\}) \leq V(\overline{x}_{pst}/\{n_i\})$.

This theorem says that if some stratum has sufficiently large variance and all other strata are overrepresented in the sample, then the regular mean is more precise than the poststratified. The variance of this stratum must be larger not only than all other within-stratum variances but also larger than each such variance plus the sample size times a measure of between-strata variability. In fact, the second term in inequality (3.3) could be replaced by

$$\frac{2n}{1 - \frac{2n}{N}}\left(\frac{N_k}{N}\right) D_k^2$$

for a somewhat simpler (but less general) theorem. Inequality (3.3) is satisfied if the sample size is relatively small and the stratum means are close together while one stratum has a large variance. But it is also satisfied if the sample size is small, the stratum means differ significantly, and one stratum has extremely large variance.

Proof of Theorem 3.3: Let

$$F(n_1, \ldots, n_{k-1}) = V(\overline{x}_{pst}/\{n_i\}) - MSE(\overline{x}/\{n_i\})$$

be regarded as a function of all stratum sizes but the k-th, with $n_k = n - \sum_{i<k} n_i$ as a dependent variable. At proportional allocation, from (2.1) and (2.2) we obtain:

$$F(n(\frac{N_1}{N}), \ldots, n(\frac{N_{k-1}}{N})) = 0,$$

as well as:

$$\frac{\partial F}{\partial n_i}(n_1, \ldots, n_{k-1}) = \frac{S_k^2 - S_i^2}{n^2} + \left(\frac{N_k}{N}\right)^2 \frac{S_k^2}{n_k^2} - \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i^2} + \frac{2n_i S_i^2}{n^2 N_i} - \frac{2n_k S_k^2}{n^2 N_k} - 2\left(\sum_j \left(\frac{n_j}{n} - \frac{N_j}{N}\right)\overline{X}_j\right)\left(\frac{\overline{X}_i - \overline{X}_k}{n}\right).$$

This expression has six terms in it. If each n_i, $i < k$, satisfies $n_i \geq n(\frac{N_i}{N})$ and accordingly $n_k \leq n(\frac{N_k}{N})$, then the second and third terms together equal the nonnegative quantity:

$$\left(\frac{N_k}{N}\right)^2 \left(\frac{S_k^2 - S_i^2}{n_k^2}\right) + \left(\frac{N_k}{N}\right)^2 \frac{Si^2}{n_i^2}\left(\frac{n_i^2}{n_k^2} - \frac{N_i^2}{N_k^2}\right).$$

Thus the partial derivatives will be nonnegative provided that for each $i < k$ the first, fourth, fifth, and sixth terms combined stay nonnegative, that is:

$$S_k^2\left(1 - 2\frac{n_k}{N_k}\right) \geq$$

$$S_i^2(1-2\frac{n_i}{N_i})+2n\left(\sum_j(\frac{n_j}{n}-\frac{N_j}{N})\overline{X}_j\right)(\overline{X}_i-\overline{X}_k).$$

The sum in the large parentheses is just the bias, and can be rewritten after a bit of algebra as

$$\sum_{j<k}(\frac{n_j}{n}-\frac{N_j}{N})(\overline{X}_j-\overline{X}_k),$$

and thus is dominated by:

$$\sum_{j<k}(\frac{n_j}{n}-\frac{N_j}{N})D_k=(\frac{N_k}{N}-\frac{n_k}{n})D_k.$$

If each n_i, $i<k$, is larger than or equal to its proportional allocation value, then $\frac{n_i}{N_i}\geq\frac{n}{N}\geq\frac{n_k}{N_k}$ and inequality (3.3) suffices for all the partials to be nonnegative. Since a path from proportional allocation to such an allocation can be chosen consisting of line segments along which only one n_i, $i<k$, increases and the others stay constant, F is non-decreasing along the path and its value is nonnegative at the endpoint allocation. □

REFERENCES

[1] Ghosh, D. and Vogt, A.(1988), "Sample Configuration and Conditional Variance in Poststratification," 1988 American Statistical Association Proceedings, Section on Survey Research Methods, 289-292.

[2] Holt, D. and Smith, T. M. F.(1979), "Poststratification," Journal of the Royal Statistical Society, Ser. A, 142, 33-46.

[3] Valliant, R.(1993), "Poststratification and Conditional Variance Estimation," Journal of the American Statistical Association, 88, 89-96.