SAMPLE CONFIGURATION AND CONDITIONAL VARIANCE IN POSTSTRATIFICATION

Dhiren Ghosh, Pragmatics, Inc., and
Andrew Vogt, Georgetown University
Dhiren Ghosh, 1714 Rupert Street, McLean, VA 22101

KEY WORDS: balanced configuration, unbalanced configuration, unconditional variance, bias

1. INTRODUCTION. Poststratification of data is sometimes compared with prestratification, and it is noted that the former often gives precision comparable to the latter (see, for example, [3], p. 232, or [1], p. 134). A more pointed comparison is between poststratification and no stratification at all. In this note we make such a comparison between a poststratified sample mean and the regular sample mean. We offer evidence that conditional variance, where the condition is a given sample configuration, is the proper instrument for comparing the estimators. Our discussion is similar in spirit to that of Holt and Smith in [6].

Consider a population of size N on which a variable x is defined, said population being divided into k strata of sizes $N_1, \ldots, N_k$ with $\sum_i N_i = N$. A simple random sample of size n is chosen from the population with mean $\bar{x}$. If the sample consists of $n_1, \ldots, n_k$ elements (with $n_i \geq 1$ for each i) from the respective strata of the population, the sample means of x restricted to each stratum are denoted by $\bar{x}_1, \ldots, \bar{x}_k$. Then $n = \sum_i n_i$ and $n\bar{x} = \sum_i n_i \bar{x}_i$.

We compare two estimators for the population mean $\bar{X}$ of x: the regular sample mean $\bar{x}$ and a poststratified sample mean $\bar{x}_{pst}$, given by

$$\bar{x} = \frac{\sum_i n_i \bar{x}_i}{n} \quad \text{and} \quad \bar{x}_{pst} = \frac{\sum_i N_i \bar{x}_i}{N}. \tag{1}$$

These estimators are unbiased, that is,

$$E(\bar{x}) = \bar{X} \text{ and } E(\bar{x}_{pst}) = \frac{\sum_i N_i \bar{X}_i}{N} = \bar{X},$$

where $\bar{X}_1, \ldots, \bar{X}_k$ are the means of the strata over the entire population.

Conditional means can also be computed, where the condition, denoted by $\{n_i\}$, is that the sample have a given configuration $(n_1, \ldots, n_k)$ with n_1 elements in stratum 1, n_2 in stratum 2, ..., n_k in stratum k. Thus

$$E(\bar{x} \;/\; \{n_i\}) = \frac{\sum_i n_i \bar{X}_i}{n} \tag{2}$$

and

$$E(\bar{x}_{pst} \;/\; \{n_i\}) = \frac{\sum_i N_i \bar{X}_i}{N} = \bar{X}. \tag{3}$$

It is evident that the regular sample mean $\bar{x}$ is conditionally biased whenever the sample configuration fails to satisfy $(\sum_i n_i \bar{X}_i) = n\bar{X}$.

With regard to variances we find:

$$V(\bar{x}) = (1 - \frac{n}{N}) \cdot \frac{S^2}{n} \tag{4}$$

and

$$V(\bar{x}_{pst}) = \sum_i (\frac{N_i}{N})^2 \cdot (S_i)^2 \cdot \{E(\frac{1}{n_i}) - (\frac{1}{N_i})\} \tag{5}$$

where S^2 is the variance of x for a finite population, $(S_i)^2$ is the variance of x in the i-th stratum of the population, and $E(1/n_i)$ is the expected value of the reciprocal of the size of the i-th stratum in the sample.

The corresponding conditional variances are:

$$V(\bar{x} \;/\; \{n_i\}) = \sum_i (\frac{n_i}{n})^2 (1 - \frac{n_i}{N_i}) \frac{(S_i)^2}{n_i} \tag{6}$$

and

$$V(\bar{x}_{pst} \;/\; \{n_i\}) = \sum_i (\frac{N_i}{N})^2 (1 - \frac{n_i}{N_i}) \frac{(S_i)^2}{n_i}. \tag{7}$$

2. AN UNCONDITIONAL COMPARISON. If we use unconditional variance to compare the precisions of the two estimators, we find from (4) and (5) that the poststratified sample mean $\bar{x}_{pst}$ is more precise than the regular sample mean $\bar{x}$ when

$$S^2 - \sum_i (\frac{N_i}{N}) (S_i)^2 \tag{8}$$

is larger than

$$\sum_i (\frac{N_i}{N}) (S_i)^2 (\frac{n(N_i/N)}{1 - (n/N)}) \{E(\frac{1}{n_i}) - \frac{1}{n(N_i/N)}\} \tag{9}$$

An approximation for $E(1/n_i)$ (see [4], pages 139 and 116) is usually made that differs from the second term inside the braces in (9) by a factor proportional to $1/n^2$. If the approximation is made, the poststratified mean is preferable provided that the average of the within-stratum variances multiplied by a factor of the form $(1 + C/n)$ is smaller than S^2. Poststratification, like stratification, is thus appropriate when the strata have small variances and the sample size is reasonably large. If the expression in (8) is negative, the regular mean is preferable. Since (8) is identically equal to

$$\frac{N}{N - 1} \sum_i (\frac{N_i}{N}) (\bar{X}_i - \bar{X})^2 - \frac{1}{N} \sum_i (\frac{N - N_i}{N - 1}) (S_i)^2,$$

poststratification is inappropriate when all strata have approximately the same mean.

3. A CONDITIONAL COMPARISON. The qualitative arguments just given leave something to be desired. It would be nice to avoid approximations and associated implicit assumptions (or explicit technicalities). It would also be nice to obtain a more detailed case-by-case understanding of the structure of the two estimators.

The use of conditional variances is a method for achieving these goals. Indeed, we have already assumed - at least for the poststratified mean - that all configurations have $n_i \geq 1$.

The original mean and variance of $\bar{x}_{pst}$ are both conditional: the condition is that $n_i \geq 1$ for all i. $E(1/n_i)$ is a conditional mean based on this same condition. (In fact, the assumption that the sample size is a fixed number n is itself a condition: all variances are conditional variances.) The method proposed here and in [6] is to study the estimators relative to the condition that the configuration is given.

Let us compare the two estimators configuration by configuration with conditional variances. Because the regular mean is in general conditionally biased, the appropriate comparison is between the conditional variance of $\bar{x}_{pst}$ and the conditional mean square error of $\bar{x}$. The poststratified mean will be preferable when its conditional variance is smaller than the conditional MSE of the regular mean. This happens if and only if $V(\bar{x}_{pst} / \{n_i\})$ in (7) is less than

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

$$\textstyle\sum_i (\frac{n_i}{n})^2 (1 - \frac{n_i}{N_i}) \frac{(S_i)^2}{n_i} + [\sum_i (\frac{n_i}{n} - \frac{N_i}{N}) \bar{X}_i]^2 \qquad (10)$$

The last term in (10) is the square of the conditional bias.

For a balanced configuration, that is, one with $n_i/n = N_i/N$ for all i, a configuration that may be only approximately achievable, the estimators $\bar{x}$ and $\bar{x}_{pst}$ have the same value on the sample, their conditional variances (6) and (7) are equal, and they are both conditionally unbiased. However, the unconditional variances (4) and (5) may be quite different. The two estimators agree on all samples with the balanced configuration, and this is why their conditional variances are equal. The unconditional variances take into account how they behave on samples with other configurations, samples on which the two estimators may have distinct values.

In general, for unbalanced configurations the estimators have distinct values, $\bar{x}$ is conditionally biased, and the conditional variances of $\bar{x}$ and $\bar{x}_{pst}$ are distinct from each other and vary from configuration to configuration.

An indication of the structure of these variations can be obtained by consideration of a simple example.

Imagine a population with two strata of sizes N_1 and $N_2 = N - N_1$, having means $\bar{X}_1$ and $\bar{X}_2$ and variances $(S_1)^2$ and $(S_2)^2$ where $0 < S_1 < S_2$. We shall assume that N is quite large. A sample of size n is drawn, and we wish to compare various configurations (n_1,n_2) with $n_2 = n - n_1$.

In Figures 1 and 2 we have graphed PST = $V(\bar{x}_{pst} / \{n_i\})$ and REG = $MSE(\bar{x} / \{n_i\})$ as functions of n_1 for n_1 between 0 and n. The curves intersect when $n_1 = n_{prop} = (N_1/N)n$, a point corresponding to proportional allocation. So long as $S_1 \neq S_2$, the curves will cross each other at this point. The PST curve has its minimum to the left of the crossing point at

$$n_1 = n_{opt} = (\frac{N_1 S_1}{N_1 S_1 + N_2 S_2}) \cdot n ,$$

the point of optimum allocation under prestratification; while the REG curve has its minimum to the right of the crossing point. There is thus an interval to the left of proportional allocation where the poststratified mean is preferable, and one to the right where the regular mean is preferable.

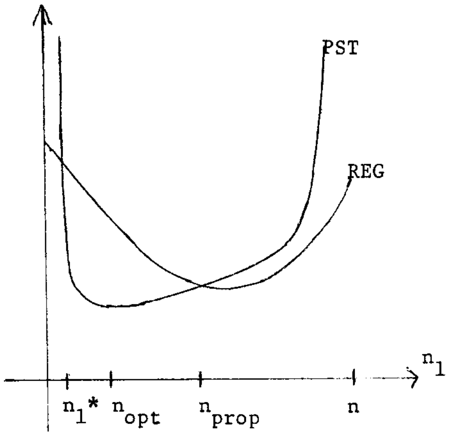

Figure 1

PST = $V(\bar{x}_{pst} / n_1,n_2)$

REG = $MSE(\bar{x} / n_1,n_2)$

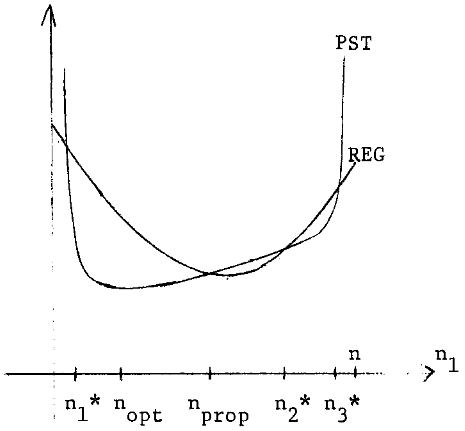

Figure 2

The PST curve and REG curve intersect in from two to four points as n_1 varies from 0 to n. There will be exactly one intersection point n_1* to the left of n_{prop} provided that the stratum weight $W = N_1/N$ satisfies

$W \leq 1/2$; or else

$W > 1/2$ and

$$nW(1 - W)D^2 + (3W - 1)((S_2)^2 - (S_1)^2) > (S_2)^2 .$$

Here $D = |\bar{X}_1 - \bar{X}_2|$ is the absolute difference of the stratum means.

If the quantity D is sufficiently small, n_1* and n_{prop} are the only intersection points, as shown in Figure 1. But if D is sufficiently large, two further intersection points n_2* and n_3* can be found to the right of n_{prop} (see Figure 2) and there is a second interval where the poststratified mean is preferable.

Passing to the general case with the example in mind, we make a few qualitative observations. At proportional allocation the estimators are equivalent. In the special case when proportional and optimum allocation coincide ($S_1 = S_2$ in our example), the poststratified mean is preferable when the within-stratum variances are small compared to the between-strata variance and the sample size is reasonably large. The regular mean should be preferred otherwise. If proportional allocation and optimum do not coincide, the poststratified mean is to be preferred for configurations tending away from proportional toward the optimum and beyond, while the regular mean should be preferred if the configuration deviates from proportional away from the optimum.

In other words, the poststratified mean should be used when the configuration overrepresents strata with large variances (if S_2 and n_2 are large, then n_1 is small and the PST curve is the lower curve). If the configuration underrepresents strata with large variances, the regular mean should be used (if S_2 is large but n_2 is small, then n_1 is large and in general the REG curve will be lower).

A direct comparison of the estimators in (1) suggests the same conclusions. It is better in general to use the estimator that gives lesser weights to strata with unreliable means. Which estimator that is depends on the configuration.

In contrast to the discussion based on unconditional variance, we find that, even when the strata have approximately the same means, the poststratified estimator will be appropriate if the configuration favors large variance strata.

4. CONCLUSIONS. Conditional variances can be used in connection with other estimators besides the two treated in this note - e.g., estimators occurring in double sampling or weighted estimators such as those discussed by Fuller [2]. More than one stratification scheme can be applied to a given sample, and different ones can be compared. The proper way to collapse poststrata when a configuration is very unbalanced may also be evaluated by conditional variances.

Holt and Smith [6] indicate that use of the conditional variance is a matter of controversy. However, it is not clear that those authors who propose usage of the so-called unconditional variance intend it to be used exclusively. For example, Hansen et al. in [5], p. 790, appear to recognize the validity of the conditional in context: their quarrel is with unwarranted reliance on model-dependent conditions, rather than with conditions consistent with a probability-sampling design.

Evidently both unconditional and conditional variances are valid indicators of precision- the difference between them is that they refer to different sets of samples. In general there may be no natural ordering of these sets, but in the present case the conditional variance refers to a subset of the set of samples for the unconditional.

The issue is whether to use a poststratified mean or the regular mean to estimate the true mean of a population. If one uses unconditional measures to decide, a definite choice can be made by comparing (4) and (5) (or rather approximations associated with (5)). The assumption underlying this choice is that the same estimator will be used regardless of the sample configuration.

If the configuration is taken into account, as in section 3 above, the decision as to which estimator to use depends on the configuration. If a decision applicable to all possible configurations is sought, an estimator of the following form might be used:

$$\hat{x} = \begin{cases} \bar{x} & \text{for certain configurations} \\ \bar{x}_{pst} & \text{for all others .} \end{cases}$$

The definition of $\hat{x}$ would involve a configuration-by-configuration comparison of the precisions of $\bar{x}$ and $\bar{x}_{pst}$, and the unconditional MSE of $\hat{x}$ would be computed by averaging the conditional MSEs of $\bar{x}$ and $\bar{x}_{pst}$ over the configurations.

Rather than treat the compound estimator $\hat{x}$, it is more natural to focus on the estimator used for a particular configuration and on conditional variances (even if they will ultimately be used to calculate MSE($\hat{x}$)). The statistician, in deciding what estimator to use when faced with a particular configuration, may not wish to bother to specify what estimator will be used for other configurations. His conditional 95% confidence interval for the true mean, based on (7) or (10), will still be sensible. Over many repetitions - where the same configuration occurs - his interval estimate should be correct 95% of the time (if one ignores uncertainties in the estimates of the quantities in (7) or (10)).

In practice, S_i and $\bar{X}_i$ are usually unknown before the sample is chosen. An advocate of unconditional variances estimates S_i from the sample, compares (4) and (5), and announces which estimator he will use (not just for the given sample but in effect for all samples of size n). The advocate of conditional methods estimates S_i and $\bar{X}_i$ from the sample, compares (7) and (10), and announces his estimator (expecting to use the same estimator for all other samples with the same configuration).

Configuration by configuration the conditional advocate will do better most of the time, and so he will do better overall. Even though the estimates of S_i and $\bar{X}_i$ have sampling errors that might affect the two advocates unequally, we believe that this conclusion still stands and could be supported by simulation.

References

[1] Cochran, W. G.(1977), Sampling Techniques (3rd ed.), New York: John Wiley.

[2] Fuller, W. A.(1966), "Estimation Employing Post Strata," Journal of the American Statistical Association, 61, 1172-1183.

[3] Hansen, M. H., Hurwitz, W. N., and Madow, W. G.(1953), Sample Survey Methods and Theory, vol. 1, New York: John Wiley.

[4] Hansen, M. H., Hurwitz, W. N., and Madow, W. G.(1953), Sample Survey Methods and Theory, vol. 2, New York: John Wiley.

[5] Hansen, M. H., Madow, W. G., and Tepping, B. J.(1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," Journal of the American Statistical Association, 78, 776-793.

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