Journal of Official Statistics, Vol. 12, No. 1, 1996, pp. 95–100

Comment

Alan M.Zaslavsky[1]

1. Introduction

Sample design for the coverage measurement survey for the year 2000 census in the United States is a problem of tremendous practical importance. The cost of the survey will be of the order of $250 million, but under the Integrated Coverage Measurement plans, the entire $3.9 billion census (cost estimates based on U.S. Bureau of the Census 1995) will depend on whether the survey can produce sufficiently precise and credible estimates of net coverage for relevant domains. Very often statisticians can make their greatest contributions when designing data collection, so I am pleased that Joseph Kadane has added his thoughts on design to his many contributions to analysis and evaluation of previous census coverage measurement efforts.

Kadane's conclusions, under some simplifying assumptions about the structure of the population, imply that sample sizes should be equal in every state. I first explain my (non-technical) disagreements with his argument, and then offer some principles that support what I consider a more plausible sample allocation, roughly proportional to a power of the state population.

2. Kadane's argument

The argument of Kadane's article can be summarized as follows:

1. The Hill algorithm, used for converting non-integer "deserved" seats into integral seat apportionments, minimizes a particular loss function similar to the chi-square statistic.
2. This function implicitly measures loss due to the different state population shares and shares of legislative representation.
3. Posterior risk for this loss equals loss evaluated at the value of population shares given by the squared root of the posterior expectation of the squared population shares, $\phi_i^* = \hat{\phi}_i\sqrt{1 + CV_i^2}$.
4. Therefore, a Bayesian who accepts the chi-square-like loss must apply the Hill algorithm to $\{\phi_i^*\}$ instead of to $\{\hat{\phi}_i\}$.
5. If the ratio $\phi_i^*/\hat{\phi}_i$ is equal for every state, the Hill algorithm produces the same results whether applied to $\{\phi_i^*\}$ or to $\{\hat{\phi}_i\}$.
6. The ratio $\phi_i^*/\hat{\phi}_i$ is equal for every state if and only if the coefficient of variation CV_i is the same in every state.

[1]Alan Zaslavsky is Associate Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A. The author thanks Joseph Kadane for his detailed comments which have helped to clarify the arguments.

7. If $\phi^*/\hat{\phi}$ is the same in every state, a strong argument can be made that uncertainty about $\{\phi_i\}$ does not have an inequitable effect on apportionment when $\{\hat{\phi}_i\}$ is used.
8. These arguments support a design that makes the coefficient of variation of estimated population equal across states.

Points (1), (3), (5) and (6) are theorems, and (4) follows if we accept its premises. (The converse of (5) is not necessarily true, because small changes in the proportions will not necessarily be reflected in changed apportionments.) Furthermore, I consider loss function (2) very useful, although not primarily because of its relationship to apportionment rules. Therefore, I next focus on points (7) and (8) and on whether they constitute an adequate argument for the design that Kadane proposes.

3. Is This Argument Compelling?

The very restrictive design constraint in (8) follows because the Bayesian argument of (4) is combined with the external (non-Bayesian) constraint of (7). To make the Bayesian's allocation based on loss function (1) agree with the procedures established in law (using $\{\hat{\phi}_i\}$), Kadane's procedure requires the survey designer to control the information available to the Bayesian so that the Bayesian will calculate a posterior distribution that satisfies constraints (7), even without being told what those constraints are! This proposal implies that the survey designer's loss is not a function of the errors in the population estimates, but only of the disagreement between the apportionment calculated by the Bayesian and the legislated apportionment.

I would find (7) more compelling if I were convinced that the terms of the argument over apportionment following the 2000 census will be set mainly by the arguments of the Bayesians described in (4) against the equity of apportionments using $\{\hat{\phi}_i\}$ as mandated by current legislation. I believe it is more likely that a challenge will be based either on an attack on the accuracy of the population estimates, or on some theory about the equity of the entire apportionment procedure. If some notion of equity is an important criterion, then loss (2) should be minimized within the class of decision rules (apportionment schemes) that are equitable in the desired sense (or the loss function should be modified). This imposes no special constraints on the sample design. If we attempt to carry through this program by making the notion of equity expressed in (7) precise, we immediately confront the fact that the equity of the Hill algorithm between small and large states is hotly disputed, even when there is no sampling error. I believe that it is highly unlikely that Congress will revise its apportionment procedures to replace $\hat{\phi}_i$ with ϕ_i^*, or that a demand for such a change would be a potent basis for a challenge to the apportionment. This is because the current consensus is built around a particular procedure rather than a loss function, even though the two are mathematically equivalent for known ϕ_i.

Furthermore, the correction in (4), and therefore the discrepancy addressed by (7), will be very small. Suppose for illustrative purposes that the coefficient of variation of the population estimate for a state is .02 (2%). This quantity is large compared to typical differentials by state, and therefore probably at least as large as typical target coefficients of variation in the census. Then $\phi_i^*/\hat{\phi}_i$ is 1.0004 and the relative discrepancy between $\hat{\phi}$ and ϕ^* is .0004 (ignoring the relatively small uncertainty in the denominator $\Sigma\phi_i$), compared to another state with zero coefficient of variation. This is negligible relative to the coefficient of variation. It is also much smaller than biases due to measurement error (e.g., see Mulry and Spencer 1993) that will remain after applying estimated corrections for differential coverage. Therefore, neither Congress nor the courts would be very wrong to ignore it.

My objections to design constraint (8) stem primarily from its implications for census design. Assume, for purposes of discussion, that the sampling variance of the estimated coverage rate is inversely proportional to the corresponding sample size, $\text{Var}(\hat{\phi}_i - \phi_i)/\phi_i = CV_i^2 \propto 1/n_i$, with the same constant for all states. This "homogeneous variance assumption" (HVA) allows us to draw some broad conclusions about the consequences of the differing population of states by ignoring differences in conditions between regions. HVA is unlikely to be true, but if an argument leads to a sample allocation that is undesirable under HVA, there is no reason to count on HVA being violated in just the way that fixes up the problem.

Under HVA, (8) requires that the sample size in every state should be the same, despite tremendous disparities in population. This counterintuitive conclusion should give us a pause.

4. Alternative Approaches to Integrated Coverage Measurement Design

I now turn to some alternative principles for sample allocation that do not require us to use the same sample size in California and Wyoming (ratio of populations ≈ 66), expanding on the discussion in a National Academy of Sciences report (Steffey and Bradburn 1994, pp. 125–126). Given the loss function defined by Kadane, we would naturally seek a design that minimizes risk, i.e., that minimizes the expectation of the chi-squared loss of apportionment under the Hill algorithm. This is difficult because the number of seats apportioned to each state is a discontinuous function of the estimated shares. However, if we assume that loss from sampling error $L_S = \Sigma(\hat{\phi}_i - \phi_i)^2/\phi_i$ is approximately independent of loss due to integral apportionment $L_A = \Sigma(h\hat{\phi}_i - a_i)^2/a_i$, then the optimal allocation minimizes the expectation of L_S. Under HVA, optimal sample size for this risk is $n_i \propto \phi_i^{1/2}$. (As in most design problems, the parameters that control the design, in this case $\{\phi_i\}$, are unknown before the census, but we have good estimates from the previous census, or better yet, from postcensal estimates.)

Another popular and intuitively appealing loss function is the number of seats that are misapportioned. This is related to the size of the absolute errors, because the cutoffs for states to receive an extra seat (beyond what each deserves) are approximately uniformly distributed over the interval $[\phi_i, \phi_i + 1/435]$ and the cutoffs for losing a deserved seat are approximately uniformly distributed over $[\phi_i - 1/435, \phi_i]$. (See the Appendix.) Minimization (under the HVA) of the expected sum of absolute errors $\Sigma|\hat{\phi}_i - \phi_i|$ implies $n_i \propto \phi_i^{2/3}$. This allocation puts slightly more sample into larger states than the last, because chi-squared loss gives smaller weight to an error of the same absolute size in a larger state than in a smaller one.

Considerations other than accuracy of apportionment support unequal sample sizes. Zaslavsky (1993, sec. 2.1) argues for the two loss functions considered above (chi-squared and absolute error) as general-purpose measures of the cost of misallocation of continuous benefits such as monetary block grants. Another consideration is that a design with equal sample size for every state (and therefore very different sampling *rates*) would almost certainly be highly inefficient for estimation of total national population or population of nationally-distributed domains such as racial groups. Estimates for Blacks would be particularly bad because of the concentration of Blacks in the larger states.

Estimation of relative populations of substate domains is also important, and perhaps more so for parts of a large state than for corresponding fractions of a smaller state. For example, differences between New York City and upstate New York, or between northern and southern California, are of immense importance for distribution of representation and funds; direct estimates of these estimates may be needed. On the other hand, we may be willing to accept model-based (e.g., synthetic) estimates (possibly constrained to agree with direct estimates for states, as suggested in Steffey and Bradburn 1994, p. 126) for the difference between northern and southern Idaho. In particular, recent decisions on Congressional districting have established very stringent standards for equality of district populations within each state. Assuming (unrealistically) that direct estimates for every district must be precise enough to support these standards, state sample size would have to be proportional to the number of districts (and therefore approximately proportional to population). Similar arguments apply to distribution of funds to localities and districting for state legislatures.

Thibaudeau and Navarro (1995) consider what we might call "equity of accuracy" in sample allocation for sampled follow-up of nonrespondents to the mailed census questionnaire. This follow-up involves a much larger sample than the coverage measurement survey, so direct estimates will be obtained for relatively small substate domains. Thibaudeau and Navarro propose that target levels of accuracy be defined for domains of equal population, thereby avoiding the issues that are the main focus of Kadane's article and this discussion. Instead, they focus on equitable sample allocation between domains with different rates of mail nonresponse. If we ignore state boundaries, however, this argument leads to a sample allocation of this operation that is roughly proportional to state populations, modified for differing mail response rates in different states.

These models and loss functions suggest allocating sample size roughly proportional to some power of state population between 1/2 and 1. In practice, other considerations will be important (see Kadane's Section 4). It may be politically necessary to maintain a ceiling on the CV of direct estimates for any state, implying a floor on any state's sample size. The HVA does not hold, due to varying block sizes and degrees of interblock heterogeneity, which further complicate sample design. Estimates for many domains will be based on models, which have the potential to improve on direct estimates for these domains. States also differ in the heterogeneity of substate domains, which affects the number of such domains into which the state should be divided for estimation purposes. These factors, as well as the experience of the 1995 census tests and the multiplicity of objectives that the census must serve, will all be considered as Census Bureau staff proceed with design for the year 2000.

5. Appendix – Congressional Apportionment and Absolute Error Loss

Regard estimated populations for all states except state i as fixed. If ϕ_i is overestimated by a sufficient amount, $\hat{\phi}_i\sqrt{a_i\,(a_i+1)}$ becomes larger than the corresponding value, before assignment of the last seat, for the state that got that seat; then the seat is erroneously transferred to state i. Conversely, if state i has more than one seat and ϕ_i is underestimated by a sufficient amount, the last seat apportioned to state i is lost to the next state in line for a seat. For each state we may calculate the positive and negative errors in population share estimates that would lead to erroneously gaining or losing a seat, denoted by e_{Gi} and e_{Li}, respectively.

With any fixed set of state populations, the number of seats misallocated is a deterministic and discontinuous function of errors of estimated populations. However, over hypothetical draws of state populations that resemble current true populations but are slightly perturbed for each state, the quantities e_{Gi} and e_{Li} may be regarded as random. (This random population argument is in the spirit of Balinski and Young 1982, sec. A5, and Spencer 1985, sec. 9.1.) The probability that an error e_i will lead to misallocation of a seat for state i is $P(e_i > e_{Gi})$ if $e_i > 0$ and $P(e_i < e_{Li})$ if $e_i < 0$.

I conjectured that the distributions of e_{Gi} and e_{Li} over slightly perturbed populations would be approximately uniform for all but the smallest states. In a simulation study, 5000 pseudo-populations were created by multiplying state populations from the 1990 census by log-normal noise with coefficient of variation 0.1 (representing our uncertainty about how much populations will change between 1990 and 2000). For the 31 states with five or more seats, comprising 91% of the population of the United States, the distributions of e_{Gi} and e_{Li} were close to uniform over the intervals $(0\ ,1/435)$ and $(-1/435,\ 0)$ respectively; dividing the intervals into five equal subintervals, probabilities of falling into each subinterval had standard deviation less than 0.01 across all states and subintervals. Thus, if we regard the fractional parts of state quotas as essentially random noise, the probability of misallocation of a seat is roughly proportional to absolute error for these states.

If very good population projections are available before the census, it may be possible to determine which states are likely to be close to a cutoff, so that a small error in population has a high probability of causing an error in apportionment. Using this information, it would be possible to go beyond the random model used in the simulations and determine for which states it would be critical to get precise population estimates. In principle, it could even be possible to determine which states are more likely to gain (because they are just below a cutoff) or more likely to lose (because they are just above a cutoff) if their population shares are measured with error. This information, however, is not likely to be available at the time at which the coverage measurement survey must be designed. In any case, the objectivity of the census could be called into question if the design were based on anticipation of effects on identifiable states.

At the other extreme, assuming 1990 population shares, the six smallest states (Wyoming, Alaska, Vermont, North Dakota, Delaware, and South Dakota) will receive their single representative despite any conceivable measurement error because to get a second representative would require a combined relative undercount and measurement error of at least 16%. From the standpoint of apportionment, coverage measurement in these states is unnecessary; of course there are other reasons why it would be unacceptable to omit these states from the coverage measurement survey.

6. Additional References

Mulry, M. and Spencer, B.D. (1993). Accuracy of the 1990 Census and Undercount Adjustments. Journal of the American Statistical Association, 88, 1080–1091.

Steffey, D.L. and Bradburn, N.M., eds. (1994). Counting People in the Information Age. Washington, DC, National Academy Press.

Thibaudeau, Y. and Navarro, A. (1995). Optimizing Sample Allocation of the 2000 Non-Response Follow-Up. Proceedings of the Section on Survey Research Methods, American Statistical Association, 736–741.

U.S. Bureau of the Census (1995). The Reengineered 2000 Census. Unpublished draft, May 19, Washington, DC.

Zaslavsky, A.M. (1993). Combining Census, Dual-System, and Evaluation Study Data to Estimate Population Shares. Journal of the American Statistical Association, 88, 1092–1105.

Received January 1996