Journal of Official Statistics, Vol. 12, No. 1, 1996, pp. 85–93

A Bayesian Approach to Designing U.S. Census Sampling for Reapportionment

Joseph B. Kadane[1]

This article proposes a design criterion for sampling in conjuction with the U.S. censuses of 2000 and beyond. Since reapportionment of Congress is the constitutional basis of the census, the loss function used here minimizes apportionment errors in a certain sense. This leads to a stochastic modification of the Hill (equal proportions) method of apportionment now used. If the sampling in the census is designed to achieve minimum constant coefficient of variation of state shares of the national population, the use of the proposed "single-number" census will result in the same apportionment as would have been obtained using the proposed loss function.

Key words: Census adjustment; equity in reapportionment; Hill method of equal proportions; loss functions; small distance asymptotics; Webster method.

1. Introduction: Purposes of the U.S. Census

U.S. Decennial Census data are used in many aspects of American life. The Constitution requires that the census be taken every ten years, and that the results be used to reapportion Congress (although the reapportionment did not happen after the 1920 census). Federal statutes require the use of census data in many formulae dividing Federal funds for various purposes. Private companies use census data to guide marketing by zip code, store location, etc. States and localities use it for drawing legislative and Congressional district boundaries and for planning purposes. Without denying the importance of these manifold uses of the census, this article concentrates on the Constitutional use, the reapportionment of Congress among the states.

Other countries' censuses also have systematic errors, as recent articles about the Australian (Steel 1994), Canadian (Strauss 1993) and United Kingdom (Diamond and Skinner 1994) censuses have pointed out. These censuses are conducted in an institutional environment that may be more amenable to stress the goal of estimation rather than apportionment.

2. The Uses of Sampling in the Census

A recent National Research Council report (1993) endorses the Census Bureau's proposal of a "single number census" for the year 2000. This is a departure from

[1]Department of Statistics, Carnegie Mellon University, 232 Baker Hall, Pittsburgh, PA 15213-3890, U.S.A.
Acknowledgement: This research was supported in part by NSF Grants DMS-9303557 and SES-9123370. I thank Steve Fienberg and Mary Mulry for their helpful comments.

© Statistics Sweden

the practice in 1990, when an enumerative census was reported in December 1990, and an adjusted census was prepared for the following summer. In a much disputed decision, the Secretary of Commerce determined not to adopt the adjustment.

A "single number census" will no doubt involve a traditional enumeration, but may embed sampling in a variety of ways as well. Sampling for follow-up of residences from which no questionnaire has been returned by mail has been advocated both to save money and improve the quality of census data, since by concentrating on reasonably sized samples more careful work can be done in follow-up (Ericksen and Kadane 1985). Additionally, a single-number census will incorporate whatever form the Post-Enumeration Survey takes for the year 2000.

In any sampling design, the issue arises of how to distribute sampling resources: where to sample heavily, and where more sparsely. This article proposes and justifies a criterion for designing such sampling in the census.

3. Uncertainty in State Populations

As traditionally reported, the census gives a single number for the population of each state. For example, in 1990, South Dakota was reported to have 699,999 persons on Census Day, April 1, 1990 (Department of Commerce 1990). The implicit accuracy claimed must of course be treated with some scepticism.

In order to obtain design criteria for sampling in connection with the census, it is necessary to be explicit about what the consequences would be of various different patterns of uncertainty in state populations. Since apportionment is the Constitutional purpose of the census and the focus of this article, I begin by reviewing the currently used method of apportionment, called the Hill (1911), or Equal Proportions, method.

To establish some notation, an apportionment is a choice from the set $A = \{a_i \mid a_i \geq 1,\ a_i \text{ is an integer and } \sum a_i = h\}$, where h is the intended size of the house (currently 435 for the House of Representatives), and the index i ranges over states, now 50 in number. Suppose also that state i has population P_i, presumed for the moment to be a known number. The Hill method may then be stated as an algorithm, as follows:

a. [Initialization]. Set $a_i = 1$ for $i = 1, \ldots, 50$.

b. [Step]. Suppose h' seats have been allocated, i.e., $\sum_{i=1}^{50} a_i = h'$. Choose a state j for which $P_j / \sqrt{a_j(a_j + 1)}$ is maximum. Increase a_j by one, and h' by one.

c. [Stopping]. If $h' < h$, return to (b). Otherwise stop.

Note that P_j appears in the algorithm in a linear, homogeneous way. Thus if $P = (P_1, \ldots, P_{50})$ were replaced by a constant multiple, for example, population shares $\phi = (\phi_1, \ldots, \phi_{50})$, where $\phi_j = Pj/T$ and $T = \sum_{i=1}^{50} P_i$, the Hill algorithm applied to ϕ would yield the same apportionment as would the Hill algorithm applied to populations P.

There are two rather fundamentally different ways to view uncertainty in statistical problems. The classical view, which accepts only distributions on the observations given the parameter, has been used by several authors studying aspects of

reapportionment in a stochastic environment. Gilford, Causey, and Rothwell (1982) and Gilford (1983) used simulation from the twelve Post-Enumeration Program (PEP) series arising from the coverage evaluations of the 1980 census. They show that the PEP series chosen would have an effect on apportionment. Spencer (1985) simulates given a known "true" population, to assess the effects of uncertainty on different algorithms for apportionment, and the extent to which they may systematically favor smaller or larger states. Schirm and Preston (1987) examine the effects of a synthetic estimate of census adjustments using similar methods.

This article uses the other main approach of dealing with uncertainty, the Bayesian method. This presupposes the knowledge of a posterior distribution on the quantities of interest. Although the modeling requirements and agreements needed to produce a 50-dimensional posterior distribution of state populations might seem severe, only certain very simple features of that posterior distribution will turn out below to be relevant. Since the literature concerning how to produce such posterior distributions is voluminous (e.g., Ericksen, Kadane, and Tukey (1989) for the 1980 census, the papers in the special section on the undercount in the *Journal of the American Statistical Association* (JASA) of September 1993 for the 1990 census) and because the methods to be used in the census of 2000 are still in flux, I will not address the question here of how such a posterior distribution might be developed for the census of 2000.

The other fundamental quantity that must be specified to use the Bayesian approach is a loss function. There are two criteria that I believe should apply to the choice of a loss function. First, when the posterior distribution is concentrated at a single point, so that state populations are considered known, the use of the loss function should reproduce the currently used Hill (equal proportions) method. (The Appendix considers an analogous analysis for the main alternative to the Hill method). Second, to be useful, the loss function should lead to a tractable result. Without taking this consideration into account, the result could be a computational problem of some magnitude. Bayesian principles require the minimization of expected loss (equivalent to the maximization of expected utility). Here that would mean the minimization over the set A of the expectation of the 50-dimensional posterior distribution of the expected loss, considered as a function of the state apportionments. The set A is finite, but quite large (a simple occupancy argument from Feller 1957 pp. 36, 37, shows that $|A| = \binom{434}{385}$). For each member of A, the expected utility would have to be determined by numerical integration of some kind. This mammoth calculation seems unlikely to yield theoretical insight that might be used to help design the sampling for the census of 2000. Hence tractability is a rather pressing concern. Some discussion of loss functions in this problem can be found in Spencer (1985), Citro and Cohen (1985), Cressie (1989), Fienberg (1986), and Zaslavsky (1993).

An important consideration for these loss functions is that the decision space be modeled correctly. The focus of this article is the set A of possible apportionments, but in other countries an estimation focus might be more useful.

To establish some further notation, let $\psi_i = h\phi_i$ be state i's quota, the number of seats in the House of Representatives state i "deserves" absent the constraints that the allocations a_i must be positive integers.

One commonly used loss function in statistics is squared error, which here would be $\sum(a_i - \psi_i)^2$. In analogy to chi-square like distances, it might make sense to divide by a_i, yielding the loss function

$$L_1(\boldsymbol{a}, \boldsymbol{\psi}) = \sum_{i=1}^{50} \frac{(a_i - \psi_i)^2}{a_i} = \sum \psi_i^2/a_i - h. \tag{3.1}$$

Thus the choice of a to minimize L_1 (a, ψ) over A minimizes $\sum \psi_i^2/a_i$. In the case of known populations (or population shares), the minimizing A is precisely the apportionment found by the Hill algorithm (see Balinski and Young (1980, p. 105) and Huntington 1928).

This property suggests that L_1 may be treated as a loss function, and its expectation with respect to the distribution of the random variables ψ_i may be minimized.

Thus I seek

$$L' = \min_{\{a_i\} \in A} E\left(\sum_{i=1}^{50} \psi_i^2/a_i\right) = \min_{\{a_i\} \in A} \sum_{i=1}^{50} E(\psi_i^2)/a_i. \tag{3.2}$$

Now let $\psi_i^* = \sqrt{E(\psi_i^2)}$. Under the assumptions made above, although the ψ_i's are random variables, their distribution is known, so the ψ_i^*'s are (in principle) a known set of numbers. Then (3.2) becomes

$$L' = \min_{\{a_i\} \in A} \sum_{i=1}^{50} (\psi_i^*)^2/a_i \tag{3.3}$$

which is exactly in the form that the Hill algorithm minimizes. Consequently the Hill algorithm may be used again, substituting ψ_i^* for P_i, which proves the following theorem.

Theorem 3.1 *When $L_1(\boldsymbol{a}, \boldsymbol{\psi})$ is the loss function, the apportionment found by applying the Hill algorithm to the "quotas" $\psi_i^* = \sqrt{E(\psi_i^2)}$ minimizes expected loss.*

Suppose that ψ_i has mean μ_i and variance σ_i^2. Then $\psi_i^* = \sqrt{\mu_i^2 + \sigma_i^2}$. As $\sigma_i^2 \to 0$, that is, as one becomes more and more sure that state i's share of the nation's population is μ_i, $\psi_i^* \to \mu_i$, so the allocations under (3.1) approach those that would be obtained by applying the Hill algorithm to μ_i, which is the same as applying the Hill algorithm to known state populations P_i. Hence this method satisfies the first desiderata mentioned above, that it should reduce to current practice when state populations are known with certainty. Additionally since it applies a known algorithm (Hill's) to a simple functional of the first two mements of ψ_i, it is tractable, satisfying the second requirement. Interestingly, the apportionments obtained in this way do not depend on the covariances between ψ_i and ψ_j for different states i and j.

4. Designing Sampling for the Census of 2000

There are (at least) two objections to using (3.1) as the basis for apportionment during the census of 2000. First, the Census Bureau has proposed, and the National Research Council has endorsed, a "single number" census, not a "single-distribution" census. While the Census Bureau anticipates publishing standard errors for the state population estimates, this will probably reflect only sampling errors, and hence

underestimate the real uncertainty. Not to include other sources of error would be an unfortunate omission, in my view. The notations μ_i and σ_i^2 are used here to represent the Census Bureau's posterior mean and variance on the population of state i including errors from all sources, sampling and non-sampling.

Second, it might be alleged that use of (3.1) could create inequity between states. Suppose that the posterior quotas of the two states i and j are such that $\mu_i = \mu_j$ and $\sigma_i^2 > \sigma_j^2$. Then $\psi_i^* > \psi_j^*$, and state i may get more congressional representation than state j, because uncertainty about its population share is greater. I take this to be a very serious matter, since the history of this subject is full of equity arguments (see Balinski and Young 1980).

The perspective I suggest on both these matters lies in the thought that the variances σ_i^2 are to at least some extent design variables in the control of the Census Bureau as it prepares for the census of 2000. Suppose the sampling for the census of 2000 were designed for equal coefficients of variation in state shares, or equivalently, state quotas, i.e.,

$$\sigma_i = k\mu_i \tag{4.1}$$

for some $k > 0$. In this case

$$\psi_i^* = \mu_i\sqrt{1 + k^2} \tag{4.2}$$

so the Hill algorithm applied to the means μ_i would result in the same apportionment as would the Hill algorithm applied to the ψ_i^*'s. Thus if the Census Bureau reported state population estimates proportional to μ_i, it would ensure that the Hill algorithm would have the same effect as if loss function L_1 were used in conjunction with Theorem 3.1. Designed this way, the "single number" census would be sufficient. Secondly, the equity argument would not apply, because of the operation of (4.1).

In planning the sampling structure of the PES following the 1990 census, the Census Bureau chose a plan "to achieve a minimum constant coefficient of variation for the estimate of the population of each area" (Hogan 1992, p. 263), where area is a cross-classification of census division (nine in number, groups of states) by place/size (six categories). After the PES was conducted, poststrata were constructed reflecting demographic variables as they related to the probability of not being counted. Stratification, both pre- and post, seem to me to be an entirely appropriate way to use prior knowledge about important variables concerning miscounting. Woltman, Alberti, and Moriarity (1988) discuss the Census Bureau's reasoning in choosing to sample so as to emphasize areas, as defined above.

In planning the census of 2000, it is easier for politicians and the public to understand an adjustment in which each state's final estimated population depends only on data from that state. A plan along these lines would require large sample sizes from each state for reasonably accurate results, and hence is likely to be very costly or very inaccurate. Even if it were feasible to create "raw" estimates of this kind, it is obvious that the states with the lowest and highest rates of estimated undercount are not likely to be that low and high, respectively. Thus some kind of smoothing is likely to be advantageous in the sense of being more accurate. While it might be desirable from a public relations viewpoint to segregate data by state, I think the losses of doing so may well exceed the gains.

The purpose proposed in this article, to attend to accuracy at the state level by sampling to achieve minimum constant posterior coefficient of variation for state share estimates, could be achieved whether smoothing is used or not. Certainly the current methods, involving dual system estimation, hierarchical Bayes modeling, and synthetic or other demographic analysis, complicate variance estimation at the state level, especially before the sample results are known. The use of strata that cross state boundaries is another source of complication. Any census design would have to deal with these complications, of course, whether the goal of the design is the one suggested here, or some other goal.

There are two ways in which sampling is likely to be used in connection with the census of 2000. As an innovation, sampling is likely to be used in the later stages of non-response follow-up (as argued for in Ericksen and Kadane 1985). Not all localities have to be sampled at the same rate for this purpose. The discussion above suggests a criterion for deciding the allocation of sample (i.e., resources) among localities. Second, sampling is likely to be used for coverage completion, in some kind of new PES. Again, my suggestion is to sample in order to achieve equal state posterior coefficients of variation. I do not believe that this goal can be achieved exactly, of course, since a sample usually does not come out as anticipated. However, planning aimed at achieving that goal will come closer to achieving it than will planning aimed at achieving some other goal, for example, equal area sampling coefficients of variation, as was used in the 1990 PES.

5. Conclusion

This article proposes that sampling for the census of 2000 be designed to achieve a minimum constant coefficient of variation among state shares, or equivalently, state quotas. There is no reason given here why the same units used in 1990 should not be used in the census of 2000 for stratification. However, the suggestion here is that sample sizes be assigned to those strata to equalize posterior state share coefficients of variation, not strata coefficients of variation. Doing so would permit use of means in a "single-number" census to reproduce the Hill apportionments that would be obtained by minimizing the loss function (4.1) in conjunction with a "single distribution" census report.

In sampling theory approaches to the effects of uncertainty on apportionment, a sample of state populations is drawn from the distribution, and the apportionment is calculated. Then another sample is imagined drawn, etc. The resulting dispersion of apportionments might be taken as evidence of lack of robustness of the method. By contrast, the Bayesian approach, minimizing the expected loss (here 4.1), chooses an apportionment that takes the entire distribution into account at once. It therefore has a kind of built-in robustness. The result of this article offers a way for the Census Bureau to obtain this robustness by designing the sampling in the way specified.

The design of sampling for the census 2000 must take into account each source of variability: sampling variability in both the neo-PES and in sampling for follow-up, detection of gross errors by the PES, etc. Also it must take into account the myriad uses of the census. In this consideration, accuracy and fairness for reapportionment is likely to be regarded as quite important in the U.S. context. For other countries,

where estimation is possibly more important than apportionment, other loss functions might be more appropriate.

Appendix: Webster's Method

The Hill algorithm is not the only method that might be used to apportion Congressional representation among the states. The main alternative is known as Webster's method, and uses the same algorithm as Hill's except that the divisor $\{a_j(a_j+1)\}^{1/2}$ is replaced by $(a_j+1/2)$. The Webster algorithm was used until 1941, when it was replaced by Hill's. Balinski and Young (1980) argue that Webster's method is less biased against large states, a conclusion challenged by Spencer (1985). Without taking sides in this matter, I wish to indicate how the analysis above can be extended to Webster's method.

Suppose that, instead of dividing $(a_i-\psi_i)^2$ by a_i, as in (3.1), one divides by ψ_i, to obtain

$$L_2(a,\psi)=\sum_{i=1}^{50}\frac{(a_i-\psi_i)^2}{\psi_i} \tag{A.1}$$

$$=\sum a_i^2/\psi_i - h. \tag{A.2}$$

The member of A minimizing (A.1) is found by applying the Webster algorithm to the numbers ψ_i (Balinski and Young 1980, p.103 and Huntington 1928). Treating L_2 as a loss function, its minimization yields

$$L''=\min_{\{a_i\in A\}} E\Big(\sum a_i^2/\psi_i\Big)=\min_{\{a_i\in A\}}\sum_{i=1}^{50} a_i^2 E(1/\psi_i). \tag{A.3}$$

Let $\tilde{\psi}_i=1/E(1/\psi_i)$. Then

$$L''=\min_{\{a_i\in A\}}\sum_{i=1}^{50} a_i^2/\tilde{\psi}_i \tag{A.4}$$

and the minimization is achieved by the application of the Webster algorithm to the numbers $\tilde{\psi}_i$, which proves the following theorem.

Theorem 2 *When $L_2(a,\psi)$ is the loss function, the apportionment found by applying the Webster algorithm to the "quotas" $\tilde{\psi}_i=1/E(1/\psi_i)$ minimizes expected loss.*

Note that, again, as $\sigma_i^2\to 0, \tilde{\psi}_i\to\mu_i$, so the result of using Theorem 2 in the case of known populations reduces, as it should, to the Webster method applied to those known population means. The variances σ_i^2 are, presumably small compared to the means μ_i, so it makes sense to study the behavior of $\tilde{\psi}_i$ asymptotically as $\sigma_i\to 0$ (Kadane 1970, 1971). Expanding $1/\psi_i$ in a Taylor series

$$\frac{1}{\psi_i}=\frac{1}{\mu_i+\epsilon\sigma_i}=\frac{1}{\mu_i}-\frac{\epsilon\sigma_i}{\mu_i^2}+\frac{\epsilon^2\sigma_i^2}{\mu_i^3}+O_p(\sigma_i^3)$$

where ϵ has mean 0 and standard deviation 1. State populations are bounded from below, so that even the least populous state has positive population and thus positive

ψ_i. Hence

$$E\left(\frac{1}{\psi_i}\right) = \frac{1}{\mu_i} + \frac{\sigma_i^2}{\mu_i^3} + O(\sigma_i^3)$$

and

$$\tilde{\psi}_i = 1/E(1/\psi_i) = \frac{1}{\frac{1}{\mu_i} + \frac{\sigma_i^2}{\mu_i^3} + O(\sigma_i^3)} = \mu_i - \sigma_i^2/\mu_i + O(\sigma_i^3). \tag{A.5}$$

If the census of 2000 is designed so that $\sigma_i = k\mu_i$, then A.5 shows

$$\tilde{\psi}_i = \mu_i[1 - k^2] + O(\sigma_i^3)$$

so, to this order of approximation once again the numbers μ_i may be reported and used in the algorithm. Thus the design suggestion of this article, equation (4.1), does not depend on the use of the Hill algorithm. Asymptotically as $\sigma_i \to 0$, the design minimizing equal state posterior coefficients of variation permits certainty - equivalent numbers to be used in the Webster algorithm as well.

6. References

Balinski, M.L. and Young, H.P. (1982). Fair Representation. New Haven: Yale University Press.

Citro, C. and Cohen, M. (1985). The Bicentennial Census: New Directions for Methodology in 1990. Washington, DC: Panel on Decenniel Census Methodology, National Academy Press.

Cressie, N. (1989). Empirical Bayes Estimation of Undercount in the Decennial Census. Journal of the American Statistical Association, 84, 1033–1044.

Department of Commerce (1990). 1990 Census Population for the United States is 249,632,692; Reapportionment will shift 19 seats in U.S. House of Representatives. Press Release of Dec. 26, 1990.

Diamond, I. and Skinner, C. (1994). Comment. Statistical Science, 9, 508–510.

Erickson, E. and Kadane, J. (1985). Estimating the Population in a Census Year: 1980 and Beyond. Journal of the American Statistical Association, 80, 98–131 (with discussion).

Ericksen, E., Kadane, J., and Tukey, J. (1989). Adjusting the 1980 Census of Housing and Population. Journal of the American Statistical Association, 84, 927–944.

Feller, W. (1957). An Introduction to Probability Theory and Its Applications, Vol I, 2nd ed. New York: John Wiley.

Fienberg, S. (1986). Adjusting the Census: Statistical Methodology for Going Beyond the Count. U.S. Census Bureau Annual Research Conference, 570–577.

Gilford, L. (1983). Affidavit Cuomo, et al. v. Baldridge, et al.

Gilford, L., Causey, B.D., and Rothwell, N.D. (1982). How Adjusting Census Counts Could Affect Congress. American Demographics, 4, 30–44.

Hill, J. (1911). Letter to William C. Huston, Chairman, House Committee on the Census, dated April 25, 1911. In the U.S. Congress, House Apportionment of Representatives. House Report 12, 62nd Congress, First Session, April 25, 1911.

Hogan, H. (1992). The 1990 Post-Enumeration Program: An Overview. The American Statistician, 46, 261–269.

Huntington, E. (1928). The Apportionment of Representatives in Congress. Transactions of the American Mathematical Society, 30, 85–110.

Kadane, J. (1970). Testing Overidentifying Restrictions when the Disturbances are Small. Journal of the American Statistical Association, 65, 182–185.

Kadane, J. (1971). Comparison of k-Class Estimators when the Disturbances are Small. Econometrica, 39, 723–737.

National Research Council (1993). A Census that Mirrors America. Interim Report, Panel to Evaluate Alternative Census Methods.

Schirm, A.L. and Preston, S.H. (1987). Census Undercount Adjustment and the Quality of Geographic Population Distributions. Journal of the American Statistical Association, 82, 965–990 (with discussion).

Spencer, B. (1985). Statistical Aspects of Equitable Apportionment. Journal of the American Statistical Association, 80, 815–822.

Steel, D. (1994). Comment. Statistical Science, 9, 517–519.

Strauss, S. (1993). Statscan Devises a New Math for Counting Warm Bodies. Chance, 6, 23, 49.

Woltman, H., Alberti, N., and Moriarity, C. (1988). Sample Design for the 1990 Census Enumeration Survey. Presented at the Joint Statistical Meeting, American Statistical Association, New Orleans.

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

Received May 1994
Revised July 1995