Evaluation of the 1990 Census Sample Tolerance Check

Carolyn Swan and Richard Griffin,* Census Bureau
Carolyn Swan, 3268 Gunston Rd. Alexandria, VA 22302

Key Words: *Contingency tables; Multiple comparisons; Sample selection bias.*

Introduction:

The 1990 Census Sample Tolerance Check (*STC*) tests for sample selection bias in list/enumerate Address Register Areas (*ARAs*), where traditional door-to-door Census enumeration persists, and the enumerators themselves select the Census systematic sample, distributing long form questionnaires to selected housing units as they list and interview. Under conventional list/enumeration, the enumerators who collect the data can become an intrusive presence in the Census itself, and this is particularly true for the sample survey component of the Census. Concern for the integrity of such samples is no recent phenomenon. The distortion effect interacting enumerators have on their respondents' data was first incorporated into Census error models by Morris Hansen in 1951.[1] At a more basic level, before the interviewing begins, enumerators can intervene to bias the list/enumerate sample at the initial stage of sample selection, injecting a radical kind of nonsampling error into the survey component of the Census. The STC is concerned with one manifestation of this sample selection bias, bias in the sample estimator of the population total. This kind of bias creates major discrepancies between the 100% count, the core statistic delivered by any census, and the sample estimate of the total population.

It is, of course, a primal function of the Census to provide a hundred percent count of the U.S. population; the Census sample was not designed to provide estimates of population totals. Yet any primary distortion of the sample total estimator has a secondary impact on sample characteristics correlated with household size. Total estimator bias has a potential to skew a wide range of sample estimates--among the vulnerable items from the sample (*long form*) questionnaire: income; health; fertility; employment; current school attendance; years of schooling completed; property taxes and mortgage[2]. We would expect these relationships to show intense local variation.

In 1990, there was powerful motivation for the most common form of total estimator bias--a financial incentive; enumerators received a bonus for number of cases in 1990, and previously had been paid on a piecework scale. It still paid to finish as many cases as possible, as soon as possible. Smaller households can be enumerated faster on long forms--and the smallest possible households, vacant units, go the fastest. There has been experimentation with modifying enumerator pay structure. Notably, the 1970 Census pretest in New Haven made enumerators' wages dependent on the number of Census person records brought in on long forms. The result in New Haven was merely a directional switch in total estimator bias--creating unanticipated overestimates of the 100 percent count.

Certainly, total estimator bias is not the only conceivable form of sample selection bias. But motivation for other kinds of sample selection bias appears more subjective, hence less common, less powerful. The biasing mechanisms involved probably require greater prior knowledge of local demography than the simplest forms of total estimator bias. Thus, we would expect to encounter alternate forms of sample selection bias less frequently; they should not have as wide-ranging an impact as total estimator bias. Also, many are far from easy to verify--e.g., skipping occasional individuals perceived to be "difficult," or "socially dangerous," or sometimes omitting certain kinds of housing units based on enumerator perception of household income.

Extant controls for selection bias are inadequate. The basic traditional controls include:

1. the address register, as a randomization guide for systematic sampling, either 1-in-6 or 1-in-2, for list-enumerate ARAs;

2. the vacant/delete check--primarily intended to catch fraudulent vacancy declarations;

3. the fixed starting point and directional indications for list/enumerate circuits, intended to guarantee random enumeration paths and a random systematic sample.

For systematic sampling, the traditional injunction has been to start in the northwest corner of the block and enumerate clockwise (enumerators, however, are not robotistic "random walkers," but thoughtful human agents; many of them worked in familiar territory and knew where the vacants were). Moreover, the northwest corner starting

*This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

point had been found so difficult to apply that it was dropped from Census field procedure for 1990.

We began documenting operating mechanisms for sample selection bias in Washington State, after the 1988 Census Dress Rehearsal. In the most common scenario, the enumerator derandomizes--plans the enumeration circuit before listing, setting the starting point to target the smallest households available (vacant units, if any), as sample housing units. Repeated applications of this minimizing tactic bias the sample total estimator. Alternatively, large households may be targeted for short forms (exclusion from sample), with the same biasing effect. Enumerators may falsely claim long form housing units vacant. Or a simple mix-up may occur, (wrong address register, set for 1-in-6 sampling, not 1-in-2) with the same downwardly biasing effect.
Although the phenomenon is much less frequent, we have come across cases of systematic *upward* biasing of the population estimator, possibly intended as displays of sample collecting virtuosity (1990 evidence from Oregon).

Historically, the Census response to enumerator-induced nonsampling error has been to reduce the list/enumerate modality. Although it cannot be totally eliminated, list/enumeration (L/E) has been largely supplanted by mail, from the initial mailing experiments of 1960 to the present. From constituting 100% of Census enumeration areas, the L/E areas shrank to 30% of U.S. housing units by 1970. By 1990, list/enumerate covered only 5.4% of U.S. households and the same percentage of the U.S. population. Still, list/enumerate territory constitutes more than 50% of the land area of the continental U.S.; excluding the southeast, it includes most of the U.S. west of 100° W. longitude, outside metropolitan areas, along with the upper Midwest, and a strip of northern New England), Alaska (except Anchorage), as well as the island of Hawaii.

Most important: L/E areas contain significant subpopulations, a distinct demography. Eliminating these areas could exclude from the Census:

- most American Indians and Alaska native peoples (58% of this population in households, according to 1990 Census hundred percent count data);
- 15% of the elderly in households, again based on 1990 Census hundred percent count data ;
- much of rural America: 21% of persons reporting at least $1000 worth of agricultural produce outside the southeast, and 24% of such persons reporting at least $10,000 of agricultural produce, are our estimates from the 1990 Census sample.

An Operational Overview:

The test was conducted for L/E ARAs in 78 of the 79 type 3 District Offices (*DOs*) in the US and Puerto Rico; Alaska was excluded after cancellation of resampling in remote ARAs. The 1990 Census Sample Tolerance Check was the first automated STC. The test was run on headquarters software in each DO's processing section. The basic unit for testing is the ARA, an enumerator's geographic workload.

Automation allowed much greater control over STC procedures than in the past, when any testing had to be administered by clerks in the DOs. We could now achieve uniformity in testing and aspire to a less simplistic methodology. The basic numeric data appear in the table below, with continental U.S. figures in the column to the right.

STC OPERATIONAL DATA

32,761 ARAs tested:

6,745,305 total HUs	5,564,708 US
1,920,870 sample Hus	1,735,782 US
13,401,626 total pop	10,075,220 US
3,481,123 sample pop	2,970,228 US

6.9% failure rate

The global failure rate was 6.9%, with many more ARAs failing than would be expected to fail randomly, in the absence of selection bias, under our distributional assumptions. ARAs failing the test were resampled.

The test statistics for the STC, the total estimator, the test score, and the STC test itself, appear below. The test statistic for the STC is the absolute difference between the sample estimator of the total and its expected value, divided by an estimator of the standard deviation of the sample total estimator:

$$\hat{Y} = (N/n)\, y_L;$$
$$z = (\hat{Y} - Y) \,/\, (NS/\sqrt{n})\sqrt{(1 - n/N)}\ .$$

The test is derived in large sample theory, the asymptotic normality of a scaled sample total, divided by its (appropriately scaled) standard deviation. The standard deviation estimator posits simple random sampling; its accuracy in a systematic sampling environment depends on low intraclass correlation of the sample. The null hypothesis posits the equality of the sample total estimator and the hundred percent count; the alternative asserts their inequality. The test is two-sided, allowing for underestimation and overestimation. The critical value (cut-off score) was 2.5, (standard normal α level, .0124).

Evaluation of the 1990 Census STC

Because of space constraints, we took a 1/6 representative sample from the housing unit-level files available for every type 3 DO (13 of the 78 list/enumerate DOs[3] in the continental US-yielding 4492 sample ARAs). These files were used to analyze resampling and to re-examine the

original sampling. We submitted all ARAs testable to three kinds of tests: sample tolerance check reconstructions; contingency table tests for homogeneity of distribution of household size by form type; and tests of the fit of the reconstructed sample tolerance check statistics to a standard normal distribution. A total of 4319 of our sample ARAs yielded a 100% count of at least 25 Hus, not all vacant; these were subjected to STC testing. Initial failure rate among them was 5.4%.

We applied our own STC procedures to the housing unit level sample file at two stages: after initial data collection, and at conclusion of field follow-up. Since the original sample tolerance check assumed that prescribed sampling rates of 1/2 or 1/6 were attained, so did our initial reconstruction. In the subsequent check (that is, after field follow-up), we used the empirical sampling rate to compute an STC statistic, since the prescribed sampling rate was often unrealized. After resampling, the global failure rate of the sample ARAs sank to 1.5%. On the DO level, with α set at .05, 8 of the 13 sample DOs passed Wilk-Shapiro tests for normality--we still had local undersampling with persistent extreme values among the z scores.

A major analytic tool for the evaluation was provided by contingency table testing. We applied two contingency table models to examine various distributions of household sizes in the sample, comparing them with the analogous distributions over the nonsample (*short form*) households.[4] Homogeneity tests for the distribution of housing unit size groups by questionnaire form type (sample vs. nonsample) were performed on the initial housing unit-level data and on the housing unit-level data after resampling, for ARAs with a sample size of at least 25--smaller samples generate excessively small cell sizes, invalidating the test. Size distributions analyzed for each ARA in the sample DOs were a vacant/nonvacant dichotomy, and a more comprehensive trichotomy consisting of households of no more than one person, households containing two or three persons, and households of more than three persons. The test statistics applied were the usual Pearson chi-squares, 1 degree of freedom for the 2x2 table (vacancy status by form type), 2 degrees of freedom for the 3 x 2 (household size by form type). We checked specifically for the most common ways of biasing an ARA's sample estimate, e.g., inclusion or exclusion of a disproportionate number of vacants. If the probability of a random occurrence of this event is sufficiently small, the ARA in question will fail the test for homogeneous distribution of vacants over form type--analogously, disproportionate counts of household size groups in sample, compared to the nonsample distribution, could result in failure on the household size test. Testing was performed at two α-levels, .05 and .01. ARAs were studied by DO, as well as pooled, by STC score group. These tests provide valuable corroboration for the STC--interpretive evidence. Still, to understand precisely how the disparate distributions arose, it would be necessary to consult the address registers themselves, and, in some cases, to physically retrace enumeration paths. The pooled test results are shown below.

Address Register Areas' Homogeneity Test Failure Rates by Sample Tolerance Check Score Group**

ARAs by Sample Tolerance Check Test Score	ARA Count in STC Score Group		Homogeneity Test Size α =.05		Homogeneity Test Size at α =.01	
			# of Failed ARAS	Fail Rate	# of Failed ARAs	Fail Rate
z>2.5	192	5.8%	115	.599	79	.411
Resample	194		37	.191	13	.067
1.96 <z <=2.5	230	6.9%	97	.422	37	.161
1.65<z <=1.96	225	6.8%	48	.213	17	.076
z<=1.65	2668	80.5%	202	.076	38	.014
Total	3315	100%				

The table displays failure rates on one or both of the two homogeneity tests, at levels .05 and at .01 for the pooled sample ARAs, by STC score group. Thus, z-scores over 2.5, the STC failures, 5.8% of homogeneity testable sample ARAs (192): 59.9% of these failed the homogeneity test at α=.05; 41.1% of them failed at α=.01. The STC failures were resampled. After resampling, we had 194 testable initial STC failures, whose failure rate on homogeneity testing at α=.05 dropped to 19.1%. At both α levels, the sample data show a consistent positive association between STC score-group magnitude and homogeneity failure rate; the higher the score group, the higher the homogeneity failure rate.

We performed further analysis by DO, on homogeneity test failure rates of the four STC score groups. One-sided sign and paired t tests confirmed a post-resampling drop in the average homogeneity test failure rate of the DOs' STC failures. Kruskal-Wallis[5] and median[6] tests established that, for the sample DOs, the expected homogeneity test failure rates for ARAs in the four STC score groups were not all equal. In addition, four different Anova multiple comparison tests were run on both sets of transformed[7] scores, those resulting from homogeneity testing at each of the two α levels. These tests established that for the four initial STC score groups, the average DO homogeneity test failure rates represent distinct subpopulations. Though they defined the subpopulations differently, the four tests concurred in setting the first, highest STC score group apart for both data sets (homogeneity testing at both α levels), concluding that the expected homogeneity test failure rate is highest for ARAs failing the initial sample tolerance check.[8] For the α=.05 data, none of the tests could equate the second scoring group (initial STC scores between 1.96 and 2.5) with the fourth, safely passing, low-scoring group, and three testing procedures classified the second highest score group as unique--in which case, its failure rate seems dangerously high. Two of four multiple comparison tests of failure rates, from homogeneity testing at α=.01, concur.

Conclusions

**The base for any rate consists of all testable ARAs in the respective STC score category, regardless of DO affiliation. Thus, there were 192 ARAs in the failing STC category (STC scores above 2.5), 230 ARAs in the passing category, with 1.96<z≤2.5, 225 in the passing category with 1.65<z≤ 1.96, and 2668 ARAs in the passing category with z not exceeding 1.65. Here, "z" refers to the absolute-valued STC score. Adjustments in housing unit population at resampling, deletions and additions of HUs in ARAs whose sample size, near 25 HUs, placed them at the borderline of eligibility, accounted for the disparate 194 (not 192) testable ARAs failing the original STC, and eligible for homogeneity testing after the resampling operation.

Homogeneity testing supports the sample tolerance check results. We may conclude that the highest STC scores, the STC failures, are associated with the highest homogeneity test failure rates, indicating biased sampling. Resampling reduces total estimator bias. However, the multiple comparison data suggest that the sample tolerance check cut-off score of 2.5 may have been overly generous; lowering the 1990 STC's passing score could improve the test, making it more stringent. Still, homogeneity-failing ARAs in the doubtful second-highest STC score range constitute fewer than 3% of all ARAs in our homogeneity testing sample; the workload increase, from about 6% of ARAs to about 13%, may not be justified. Ultimately, we prefer preventive to corrective measures against bias. Certain modifications in I/e operations could result in major reductions in the resampling workload. We feel that improved training for enumerators, and the use of Census computerized mapping to set random starts for block enumeration, would substantially improve sampling in list/enumerate ARAs.

We have found contingency table analysis extremely useful for obtaining answers to specific questions on sampling execution. We plan to continue using this methodology in future censuses. Homogeneity testing could be automated for District Office use and applied alone or in conjunction with the Sample Tolerance Check to determine whether an ARA should be resampled and assigned for follow-up interviewing.

For evaluation purposes, contingency table analysis could be applicable beyond list/enumerate ARAs. Bias related to household size may affect sampling and resampling in areas of computerized mailing lists (TAR areas).

1. See Hansen, Hurwitz, Marks, Mauldin, 'Response Errors in Surveys', *JASA*, 6, 147-190.

2. Item numbers for these specifically long form questions are P32, P18-19, P20, P21-27, P11, P12, and H23-H25, respectively.

3. District Offices selected in sample were Hyannis, Massachussetts; Portsmouth, New Hampshire; State College, Pennsylvania; Green Bay, Wisconsin; Hays, Kansas; Boise, Idaho; Bend, Oregon; Yakima, Washington; Mesa, Arizona; Pueblo, Colorado; Santa Fe, New Mexico; Ogden, Utah; and Bakersfield, California.

4. One caution: the chi-sqare statistics for these tests presuppose independent observations, multinomial sampling, but the Census sample design is systematic. Still, this sample design is not complex; we are dealing with single-stage sampling and unweighted data. Under these conditions, the fit of the contingency table model has generally been considered acceptable, since the chi-square testing yields conservative results--i.e., the true p-value of our test statistics should be smaller than the nominal p-values of the contingency table approximations. The authors wish to express appreciation to Robert Fay, for a very helpful discussion of the topic of chi-square adjustments. For situations in which adjustment is crucial, and for adjustment methodologies, see Rao and Scott, 'The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, *JASA*, 1981: 221-230; Fay, 'A Jackknifed Chi-Squared Test for Complex Samples,' *JASA*, 1985: 149-157, and Singh and Kumar, 'Categorical Data Analysis for Complex Surveys,' *Proceedings of the American Statistical Association, Survey Research Methods 1986*, 252-257.

5. The Kruskal-Wallis test, based on rank sums, tests the null hypothesis that all target subpopulations share the same distribution, versus the alternative that at least two distributions are different.

6. The median test and the Kruskal-Wallis test share the same null and alternative hypotheses. The median test compares the observations in each subgroup with the median of the pooled observations.

7. The arcsine transformation was applied to stabilize variances for Anova testing, after which multiple comparison tests, the Bonferroni test and the Ryan-Einot-Gabriel-Welsch multiple range test were run on the transformed scores.

8. The multiple comparison tests performed were: the Least Significant Difference Test, which does not control experimentwise type I error--for this test, we ran the pairwise comparisons at $\alpha = .01$; the Bonferroni Test, and the Ryan-Einot-Gabriel-Welsch Test, which do control experimentwise type I error, although the Bonferroni Test generally has a higher type II error rate than Ryan-Einot-Gabriel-Welsch (the experimentwise α for these analyses was set at .01); and, finally, the Bayesian Waller-Duncan test, for which the k-ratio, or ratio of seriousness of type I to type II error, was set, successively, at 100 and 500. For the data from the $\alpha = .05$ homogeneity testing, The Ryan-Einot-Gabriel-Welsch Multiple Range Test, with experimentwise α fixed at .01 could not distinguish between the second sample tolerance check scoring class (above 1.96 and not exceeding 2.5) and the third (above 1.65 and not exceeding 1.96), or between the third and the fourth (not exceeding 1.65). Mathematical transitivity does not hold in this situation; thus, the second highest scoring group remains distinct from the lowest scoring group. At experimentwise $\alpha = .05$, this procedure classified the two highest groups as unique, while pooling the two lower ones. For the same data, the Bonferroni test yielded the same results as Ryan-Einot-Gabriel-Welsch at .01, at the experimentwise α levels of .01 and .05. For the same data, conventional pairwise t tests, Least Significant Difference testing with α set at .01 for any individual pairwise comparison, set the highest and the second highest score groups apart, as two distinct categories, while combining the two lowest scoring groups--when α was set at .05 each group was classified as unique. The Waller-Duncan test, which minimizes Bayesian risk under additive loss, distinguished each of the four scoring groups as a different population at the k-ratio of 500 and 100, successively.

For the failure rate data from homogeneity testing with α at .01, Bonferroni and Ryan-Einot-Gabriel-Welsch procedures, also run with experimentwise type I error set at .01, concurred in isolating the highest scoring sample tolerance check group while pooling the three other groups as indistinguishable. The Waller-Duncan Test (k-ratio set at 500) isolated the highest scoring group, but could not detect a significant difference between the second and the third group, or between the third and the fourth; here, transitivity does not hold, and we cannot conclude that the second highest scoring group represents the same population as the safe, lowest-scoring group. The Least Significant Difference procedures, with pairwise comparison α set at .01 concurred with the results of the Waller-Duncan Test.