MODIFIED HALF SAMPLE VARIANCE ESTIMATION FOR MEDIAN SALES PRICES OF SOLD HOUSES: EFFECTS OF DATA GROUPING METHODS

Katherine J. Thompson and Richard S. Sigman
Katherine J. Thompson, ESMPD, Room 3108-4, U.S. Census Bureau, Washington DC, 20233

Key Words: Variance of a Median; Modified BRR; Survey of Construction

I. Introduction

The U.S. Census Bureau publishes estimates of medians for several characteristics of new houses, with a key estimate being sales price of sold houses. These estimates are calculated from data acquired from interviews of home builders by the Survey of Construction (SOC). In the near future, the Survey of Construction (SOC) will move its current variance estimation system to the Census Bureau's re-engineered post-data-collection processing system, the Standardized Economic Processing System (StEPS). For sample designs that do not use Poisson sampling, the StEPS system uses replication methods to estimate standard errors. The SOC is a multi-stage probability survey whose sample design is well suited to the modified half sample (MHS) replication method[1] for reasons outlined in section III.B.

The literature supports the use of Balanced Half-Sample Replication (e.g. Rao, Wu, and Yue (1992); Rao and Shao (1996); Kovacevic and Yung (1997)) and MHS replication (Judkins (1990)) for estimating variances of medians from complex survey data. We considered two methods of median estimation for variance estimation purposes. The first method uses the replicate weights to estimate medians via replicated empirical cumulative-distribution functions (i.e., calculate the median of each half-sample). The second method uses linear interpolation of grouped continuous data to approximate the median of each half-sample. The latter method is implemented in VPLX (Variances from ComPLeX Survey, Fay (1995)), a variance estimation software package developed at the Census Bureau.

Direct calculation of sample medians can be computationally intensive because it requires separate sorts for each value of a given classification variable. An alternative estimation method is to group the continuous data into discrete intervals (called bins) and use linear interpolation over the interval containing the median. Provided that the data are approximately uniformly distributed over the interval containing the median, interpolation yields a good approximation. However, optimal bin widths and locations may change over time, as the sample distributions change. These considerations motivated our research.

In this paper, we compare six methods of median-estimation for MHS replication: the sample median and five variations using linear interpolation. Section II provides a brief overview of the SOC design. Section III presents general methodology. Section IV describes the empirical results from four months of SOC data that motivated the simulation study presented in Section V. Section VI provides our conclusions and recommendations.

II. SOC Sample Design

The SOC universe contains two sub-populations: local areas that require building permits and local areas that do not. The SOC sample units selected from the first sub-population comprise the Survey of the Use of Permits (SUP), and those selected from the second sub-population, the Nonpermit Survey (NP). The SUP sample comprises the majority of the SOC estimate. The two samples are multi-stage probability samples stratified by variables with high expected correlation with the survey's key statistics: housing starts, completions, and sales.

The first stage of the SUP and NP sample selection is a subsample of Current Population Survey (CPS) Primary Sampling Units (PSUs), which are contiguous areas of land with well-defined boundaries. Thus, both surveys are conducted in the same PSUs but are otherwise independent samples. One PSU per stratum was selected. Self-representing (SR) PSUs were included in the sample with certainty. Nonself-representing (NSR) PSUs were selected with probability proportional to size (PPS) from strata containing more than one PSU.

The second stage of SUP sample selection is a stratified systematic sample of permit-issuing places within sample PSUs (selected once a decade). In many cases, only one second stage unit was selected. The third stage of SUP sample selection is performed monthly: each month, Field Representatives (FRs) select a systematic sample of building permits from the permit offices in each sampled permit-issuing place. The third-stage samples are independent by month; the first and second stages are not.

The second stage of NP sample selection is a stratified systematic sample of small land areas (1980 Census Enumeration Districts, or EDs). For the third stage of NP sample selection, field representatives completely canvass all of the roads in the sampled EDs (called segments). All new housing units are included in the NP sample with certainty.

Median estimates are derived from the pooled SUP and NP samples and are calculated using a post-stratified weight for the SUP portion and an unbiased weight for the NP portion.

III. Methodology

A. Median-Estimation Procedures

1. Sample Median

One procedure for estimating the median of a population is calculate the sample median from ungrouped data, using the sample weight to locate the median as recommended in Kovar, Rao, and Wu (1988) and Rao and Shao (1996).

2. Linear Interpolation

Another approach for estimating the median of a population is to group the sample data and interpolate for the sample median. Woodruff (1952) provides the following formula for linear interpolation of a sample median:

$$\hat{M} = F^{-1}\left(\frac{1}{2}\hat{N}\right) \approx ll + \left(\frac{\frac{1}{2}\hat{N} - cf}{f_i}\right)*(i) \qquad (2.1)$$

where

F = the cumulative frequency of the characteristic using sample weights

ll = lower limit of the bin containing the median

[1]Balanced repeated replication with replicate weights of 1.5 and 0.5.

$\hat{N}$ = estimated total number of elements in the population

cf = cumulative frequency in all intervals preceding the bin containing the median

f_i = estimated total number of elements in the population of the interval containing the median

i = width of the bin containing the median

This is the method used by the current SOC production **variance estimation** system for monthly estimates and is also the linear interpolation method employed by VPLX.

We considered two options for setting the class size (bin widths) for the interpolation. The first option develops bins based on the specific characteristic under consideration using the original data. The second option linearly transforms the data to a standard scale and then uses a standard set of bins for every characteristic. We used the following linear transformation:

$$X' = X_{original} * (1{,}000/Q_3) \qquad (2.2)$$

where Q_3 is the third quartile of the sample distribution (estimated using the sample weight). The interpolated median of the X' is multiplied by $(Q_3/1000)$ to obtain an estimated median of $X_{original}$.

Using the original data to develop medians has the advantage of producing production ready estimates and SEs. Determining the appropriate bin width is difficult, however. As the bin widths get small, the variance estimates become more unstable. As the bin widths increase, the bias of the estimate due to interpolation increases. The "optimal" bin size balances estimate bias and variance-estimate stability . Unfortunately, the optimal bin width may not remain constant between samples. Often, the distributions change over time, and the bins widths/locations in the sample should reflect this change in scale. Moreover, the optimal bin width may be different for different values of a classification variable: for example, the optimal bin width for the Midwest's sales price is probably different from the optimal bin width for the South's sales price.

The desire to have the width of the bin depend on the sample motivated the linear transformation. Our procedure of linearly transforming the data and then using standard bin widths is equivalent to simply dividing the original sample from 0 to Q_3 into x bins of equal width and placing the remainder of the data into one bin, which, by design, is much larger than the others (containing up to 25% of the sample). Indeed, the "standard" bin widths used on the transformed data are not standard on the untransformed scale: they are data dependent. As the distribution changes, the bin widths on the untransformed (original) scale also change. Using the linearly transformed data requires more bookkeeping in terms of scaling constants but easily allows for changes in the scale and shape of the distribution.

The procedure described above was designed for non-negative data. If the distribution contains negative values (e.g., a distribution of net income), then a modification of the linear transformation described in (2.2) is required. To make all of the observations in the sample non-negative, replace $X_{original}$ with $X'' = (X_{original} - X_{(1)})$, where $X_{(1)}$ is the smallest observation in the sample. Calculate Q_3 from the distribution of X'' (using the sample weight associated with $X_{original}$), and apply (2.2) to the X''.

To evaluate the first option, we used two different sets of bin widths (classification sizes): bins of size \$2000 (the same bin width used in the current production variance estimation system) and bins of size \$1000. [Note: The VPLX variance estimation software would not allow any bin size smaller than 1000 because the number of classes exceeded the allowable array range.] Based on our data analysis, we assumed that median sales price would always be larger than \$36,000 and smaller than \$550,000, so the first original-data classification is always (low - 35,999) and the last original-data classification is always (550,000 - high): this yields 257 bins of size \$2000 or 514 bins of size \$1000, plus one bin of size \$36,000 and one bin whose width depends on the largest observation in the sample.

To evaluate the second option, we used three different sets of bin widths: bins of size 4, 25, and 50. The bins of size 4 were chosen to be analogous to the bins of size 2000 in terms of the **number** of bins: 251 bins total. The selection of widths 25 and 50 was somewhat arbitrary: we chose bin size 50 to get a total of twenty bins for the data less than Q_3; and we chose bin size 25 to examine the effect of doubling the number of bins/halving the width of the bins for data less than Q_3. The transformed-data median will always be less than 1,000, so the last transformed-data classification is always (1,000 - high).

This procedure is designed for symmetric or positively skewed distributions. The data in the last bin is not used to estimate the median because it is greater than Q_3, which is expected to be far from the median. We guarantee that the first and last bins are not immediately below or above the bin containing the median by the standard bins sizes: 6.7 bins per quartile for bin size 50; 13.3 bins per quartile for bin size 25; and 83.3 bins per quartile for bin size 4. Consequently, there is no loss in precision in making the last bin so much larger than the others.

B. Variance Estimation

We used the Modified Half Sample replication method (Fay, 1989 and Judkins, 1990) to estimate the variance of a median. Modified half-sample replication is a variation of the "traditional" balanced half-sample (BRR) variance estimation described in Wolter (1985, Chapter 3), using same replicate assignment methodology as BRR (a Hadamard matrix) with replicate weights of 1.5 and 0.5 in place of the 2 and 0. The SE for a median estimate using MHS replication is given by

$$\hat{SE}(\hat{Med}) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R} (\hat{Med}_r - \hat{Med}_0)^2} \quad . \qquad (2.3)$$

where the r subscript refers to the replicate median estimate (r = 1, 2,...,R) and the 0 subscript refers to the full sample median estimate. This expression contains a four (4) in the numerator because the MSE of the replicate estimates is too small by a factor of $1/(1-0.5)^2$. See Judkins (1990).

As stated in Section II, neither the SUP nor the NP designs are two-sample-unit-per-stratum designs. To address the one sample unit per stratum problem , we "split" the SR sample-units into two panels per sample unit using the original sampling methodology and form collapsed strata by pairing two (or three) "similar" NSR sample-units. We then apply the half-sample approach in such a way that the elements contributing to the half samples are panels within sample units for SR sample units and are sample units within collapsed strata for NSR sample units.

The current SOC production variance system uses a Keyfitz estimator (a paired difference estimator) for NSR sample and a design-based estimator for SR sample to produce **level** estimate variances (Luery, 1990). Because SOC methodologists had already collapsed NSR strata for their paired difference estimator, a BRR-like application was a logical extension of the pre-existing variance estimation

structure. For the SR cases, we sort permits within predetermined sample-unit groups by geography and permit authorization date and systematically split the ordered sample into two panels as suggested in Wolter (1985, p. 131). This method of assigning units to panels is referred to as the grouped balanced half sample (GBHS) method in Rao and Shao (1996) and is discussed further in Section V. For more details on the replicate assignments, see Thompson (1998).

The SOC production system uses the Woodruff method (Woodruff, 1952) to estimate the SE of a median. This is not a replicate variance estimation method. This methodology has had mixed success in the past according to survey analysts.

IV. Empirical Data Results

Initially, we used four months of SOC sample data to examine the variances of the median-estimation methods for sales price of sold houses: March 1997, May 1997, June 1997, and July 1997. We produced medians by region and by type of financing. We used the same weight used by the SOC production estimation and variance systems (post-stratified for SUP sample and unbiased for NP sample), pooling both surveys' data to obtain medians. Each set of variance estimates was produced using 200 replicates.

We found that the six median-estimation methods produced three distinct sets of SEs: one set for the sample median, one set for the original-data-interpolated medians, and one set for the transformed-data-interpolated medians. There was no clear relationship between bin width and SE estimates for the two sets of interpolated medians. Indeed, within type of data (original or transformed), the SEs were all very close. Clearly, there was a linear transformation and an interpolation effect. None of the median-estimation methods yielded SEs resembling the published SEs, so there was no available argument for publication consistency.

The empirical results left us in a quandary. We had three distinct sets of variance estimates, and no "gold standard" against which to measure them. Because our empirical results were inconclusive, we conducted a Monte Carlo simulation study to evaluate the properties of the MHS variance estimates produced from the different median estimators.

V. Simulation Study Comparison

A. Procedure for Simulation Study

We created four finite artificial populations based on a data analysis of four SOC sample populations: one type-of-financing population (Conventional Financing) and three regional populations (Midwest (Region 2), South (Region 3), and West (Region 4)). These populations represented a variety of the types of SOC populations from which estimates are produced. Note that the SOC type-of-financing population is <u>not</u> independent of the SOC-region populations.

To approximate the finite population of sales price for houses sold, we generated w_i records for each sample unit i, where w_i is the sample weight associated with unit i. The distributions of sales price for single-unit sold houses could be approximated by lognormal distributions. The lognormal distribution has the probability density function

$$f(y) = \frac{1}{y-\theta}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{(\log(y-\theta)-\zeta)}{\sigma}\right)^2\right) \quad for\ \theta < y < \infty$$

where θ is the threshold parameter, ζ is the scale parameter, and σ is the shape parameter.

After performing this data analysis, we generated four artificial finite populations of bivariate random normal variables with expected correlation ρ=0.6 using the method outlined in Naylor et al (1968). One of the two variables represented sales price for houses sold and is generated using the parameters determined above. This variable was exponentiated and shifted by the appropriate location parameters to obtain the sales price variable. The second variable was distributed as a standard normal and is used to form strata. Each population's size was the estimated population total in the given category rounded to the nearest 50. The sample size is the original sample size rounded to the nearest 50. Model parameters and sample correlations (between simulated sales price and stratifying variable) are reported in Table 1.

We compared the percentiles, sample skewness, and sample kurtosis of each simulated population to its corresponding original population, and they were quite close. To examine the effect of outliers in the original population on the model, we removed outliers using the resistant outer fences rule described in Hoaglin and Iglewicz (1987) and found that this improved agreement between the two populations for the 90%, 95%, and 99% percentiles.

Table 1: Population Parameters and Sample Sizes

Population	θ	σ	ζ	ρ	N	n
Con. Financing	27578	0.4895	11.84	0.5703	25150	500
Midwest	31801	0.5957	11.69	0.5584	6500	150
South	29414	0.5549	11.55	0.5593	14550	300
West	53781	0.5822	11.59	0.5553	11550	250

After generating the finite populations, we sorted them by the stratifying variable and formed 50 equal sized strata in each population. From these strata, we selected 5000 stratified without-replacement random samples from each artificial population using the same sampling rate in each stratum (self-weighting design). To perform the MHS replication, we sorted the sample within each stratum by stratifying variable and then systematically split the sample into two panels. Thus, the simulation study captures some of the stratification properties of the SOC design and mimics the panel assignment for SR permit sample but does not take the multistage sample and PPS sampling into account.

We determined the median of each finite population (ζ_p). Using the 5000 samples, we estimated empirical Mean Square Errors (MSE) and Mean Absolute Errors (MAE) for the following six median-estimation procedures:

SM: the sample median of each half-sample

IO2000: interpolated medians using original data, bins of size 2000 (fixed bin width)

IO1000: interpolated medians using original data, bins of size 1000 (fixed bin width)

IT4: interpolated medians using linearly transformed data, bins of size 4 (data dependent bin width)

IT25: interpolated medians using linearly transformed data, bins of size 25 (data dependent bin width)

IT50: interpolated medians using linearly transformed data, bins of size 50 (data dependent bin width)

The linear transformation was performed **once** for procedures IT4, IT25, and IT50. The original data were transformed using the full sample Q_3, and these transformed data were assigned to the half-samples. Table 2 provides the median and third quartile of each finite population, along with the bin widths on the original scale for the transformed data.

Table 2: Median, Third Quartile, and Bin Widths on Original Scale for Transformed Simulated Data

Population	Median	Q_3	Bin Width		
			4	25	50
Con. Financing	167173	222263	889	5557	11113
Midwest (Region 2)	151312	210647	843	5266	10532
South (Region 3)	133745	180868	723	4522	9043
West (Region 4)	162130	214320	857	5358	10716

To measure the precision of the six median-estimation procedures over repeated samples, we calculated empirical MSEs and Mean Absolute Errors (MAEs) for each procedure in each population. $M(\zeta_i)$, the empirical MSE of median-estimation procedure i, was calculated as $M(\zeta_i) = \frac{\sum_r (\zeta_{ri} - \bar{\zeta}_i)^2}{5000} + (\bar{\zeta}_i - \zeta_p)^2$, where ζ_{ri} is the estimated median for sample r and estimator i, $\bar{\zeta}_i$ is the average of the ζ_{ri}, and ζ_p is the population median. This is the empirical MSE described in Judkins (1990).The Mean Absolute Error (MAE) of each median-estimation procedure i was calculated as $MAE(\zeta_i) = [\sum|\zeta_{ri} - \zeta_p|]/5000$ as defined in DeGroot (1986).

To compare the variance estimation properties of the different median-estimation procedures, we calculated an MHS variance estimate (v_{ij}) corresponding to each median-estimation procedure i from 1000 of the 5000 samples. These variance estimates were compared in terms of relative bias $[(\sum v_{ij}/1000)/M(\zeta_i) - 1]$; relative stability $[[(\sum v_{ij} - M(\zeta_i))^2/1000]^{1/2}/M(\zeta_i)]$; and error rate [(the number of samples where $\zeta_p < \theta_{Li}$ or $\zeta_p > \theta_{Ui}$)/1000 where θ_{Li} is the lower end of a 90% confidence interval, and θ_{Ui} is the upper end of a 90% confidence interval]. These criterion are used in Kovar, Rao, and Wu (1988) and in Rao and Shao (1996). With an "optimal" variance estimator, both the relative bias and relative stability will be near zero, and the error rate will be ten percent.

B. Results

Table 3 presents the empirical root MSE, SE, the bias, and the MAE for each median-estimation procedure. Each of these statistics was calculated from 5000 independent samples. The results from Table 3 can be summarized as follows:

- The transformed-data-interpolated medians with bins of width 50 have the smallest root-MSE in three of the four populations (all but Region 3), with the transformed-data-interpolated medians with bins of width 25 a close second. However, the root-MSEs of all six procedures are very close in each population, so there is no dramatic loss in overall precision with the choice of any particular estimator.
- Similarly, the transformed-data-interpolated medians with bins of width 50 have the smallest SE in each population, with the transformed-data-interpolated medians with bins of width 25 a close second. Again, the differences in SE are very close between all six procedures (within approximately 3% of each other in all populations).
- The bias of the estimation procedures does not have much influence on overall error. In all populations, the bias as a percentage of the MSE is very small.
- The six sets of MAEs in each population are very close, reinforcing the conclusion above regarding the equally-good performance of the different median-estimation methods.

Table 3: Precision of Median-Estimation Procedures

Population	Median-Estimation Procedure	Root MSE	SE	Bias	MAE
Conventional Financing	SM	3345	3345	-12	2671
	IO2000	3320	3316	161	2698
	IO1000	3387	3368	-354	2642
	IT4	3351	3340	273	2673
	IT25	3304	3293	276	2617
	IT50	3282	3265	329	2606
Region 2 Midwest	SM	6316	6287	-598	4966
	IO2000	6276	6275	-127	4992
	IO1000	6343	6297	-767	4939
	IT4	6372	6363	328	5004
	IT25	6273	6272	127	4937
	IT50	6220	6218	160	4936
Region 3 South	SM	3670	3658	301	2931
	IO2000	3708	3669	539	2998
	IO1000	3742	3740	101	2941
	IT4	3718	3662	639	2951
	IT25	3699	3638	669	2924
	IT50	3692	3616	745	2912
Region 4 West	SM	4385	4382	-140	3509
	IO2000	4425	4421	185	3578
	IO1000	4477	4469	-258	3530
	IT4	4414	4403	318	3514
	IT25	4376	4364	315	3460
	IT50	4367	4350	391	3455

Table 4 summarizes the three different comparison measures for the variance estimates in the four populations. The numerators for the relative bias and stability and the coverage rates are based on 1000 samples. The denominator for the relative bias and stability ("truth") are based on 5000 samples. An asterisk (*) in the last column of Table 4 indicates that the error rate is significantly different from the nominal error rate of 0.10 using the normal approximation to the binomial distribution at the 90% confidence level.

The variance estimates of the transformed-data-interpolated medians perform best in terms of relative bias, stability, and coverage (error rates). Specifically,

- The variance estimates of the transformed-data-interpolated medians (IT4, IT25, IT50) have the smallest relative bias. The difference in estimation method is quite pronounced in three of the four populations, where the **largest** relative bias of the transformed-data-interpolated medians is less than one-half the size of the **smallest** relative bias of the original-data-interpolated and sample medians. In all four populations, using bins of width 50 on the transformed data yielded the smallest relative bias;
- The variance estimates of the interpolated medians had the best stability. The sample median had the poorest stability in all four populations. This result was expected due to the smoothing effect of interpolation. The transformed-data-interpolated medians generally performed slightly better than the original-data-interpolated medians;
- The confidence intervals constructed from transformed-data-interpolated medians and SEs have the best coverage: in each population, the data dependent bins (all widths) yield statistically nominal coverage [Note: there is no clear relationship between size of bin width on the transformed scale and improved/reduced error rates]. The coverage for the confidence intervals constructed from original-data-

interpolated medians and SEs is very poor, yielding very conservative intervals, and the coverage with the sample median is erratic.

Table 4: Relative Bias and Stability for Variance Estimates and Error Rates and Coverage Error Rates

Population	Median-Estimation Procedure	Relative Bias	Relative Stability	Error Rate
Conventional Financing	SM	0.19	0.69	11.0%
	I02000	0.25	0.35	6.9%*
	IO1000	0.21	0.32	7.0%*
	IT4	0.06	0.25	10.0%
	IT25	0.07	0.25	10.9%
	IT50	0.05	0.26	9.5%
Region 2 Midwest	SM	0.57	1.24	7.3%*
	IO2000	0.33	0.44	6.9%*
	I01000	0.30	0.42	7.0%*
	IT4	0.15	0.41	10.1%
	IT25	0.16	0.40	9.8%
	IT50	0.15	0.42	9.0%
Region 3 South	SM	0.30	0.88	12.4%*
	IO2000	0.31	0.42	6.7%*
	IO1000	0.29	0.40	6.7%*
	IT4	0.04	0.29	11.0%
	IT25	0.02	0.28	11.0%
	IT50	0.01	0.29	11.1%
Region 4 West	SM	0.39	0.98	8.9%
	IO2000	0.32	0.42	6.2%*
	IO1000	0.29	0.39	6.2%*
	IT4	0.11	0.32	8.6%
	IT25	0.10	0.31	9.4%
	IT50	0.08	0.31	9.5%

These tests of error rates have good power, as verified through a simple power analysis. Let P_A = binomial error rate probability under the alternative hypotheses ($P_A \neq 0.10$). Using the normal approximation to the binomial, for $P_A > 0.10$, we have 90% confidence and x-percent power when the upper limit of a 90% confidence interval equals the x-percent lower limit (one sided) under the alternative hypothesis. For $P_A < 0.10$, we have 90% confidence and x-percent power when the lower limit of a 90% confidence interval equals the x-percent upper limit under the alternative hypothesis. Solving for P_A, we find that we have 90% confidence and at least 70% power when $P_A \leq 0.079$ or $P_A \geq 0.121$ (when $|P_A - P_0| \geq 0.021$). The power increases to 80% when $P_A \leq 0.075$ or $P_A \geq 0.125$.

To determine whether the differences in error rates between estimators was significant, we performed a one-way ANOVA in each population modelling each median estimator as a treatment effect using the variance stabilizing arcsin-square root transformation on the error rates. Because the error sums of squares for the transformed binomial random variables is $821/n$ (Snedecor and Cochran, 1980), we tested for overall fit using a chi-square(5) critical value. All tests are highly significant: p-values of 0.0007 for Conventional Financing; 0.0168 for Region 2; 0.0000 for Region 3; and 0.0053 for Region 4. Thus, we can conclude that the six treatments yield different results.

Moreover, in all four populations, all pairwise differences between error rates greater than 0.10% are significant at the 95% joint confidence level (based on Scheffé 95% joint confidence intervals for all pairwise contrasts, using the 95% confidence level due to the conservative nature of the procedure). Absolute differences between two error rates is greater than or equal to 0.0010 are significant. Consequently, error rate comparisons between median-estimation-method variances are statistically meaningful.

C. Validation of Simulation Results Using Randomly Grouped Balanced Half Sample Replication

Inferences from this simulation study are as valid as the variance estimates used. Rao and Shao (1996) establish the asymptotic inconsistency of the grouped balanced half sample (GBHS) estimator for estimating the SE of quantiles from samples with a fixed number of strata as the strata sample sizes $n_h \to \infty$. Instead, they recommend a repeatedly grouped balanced half sample (RGBHS) estimator, i.e. repeating the random panel assignment T times and using the average of the T GBHS estimators. Because we used the MHS variance estimator in all our applications, so GMHS refers to GBHS with replicate weights of 1.5 and 0.5, and RGMHS refers to RGBHS with replicate weights of 1.5 and 0.5.

We performed a small simulation study (300 samples per population) comparing GMHS and RGMHS (T = 15) variance estimation for the six median-estimation procedures. Because the RGMHS estimator requires a great deal of computer overhead (4,500 runs per procedure for T=15), we restricted our comparisons to two of the four sample populations (the largest and smallest). Table 5 presents the relative bias, stability, and error rates for 90% confidence intervals calculated from the first 300 samples for each median-estimation procedure i for the GMHS and RGMHS in the Conventional Financing and in the Region 2 (Midwest) populations. An asterisk indicates that an error rate is significantly different from the nominal error rate of 10%. The results in Table 5 can be summarized as follows:

- The relative biases are generally the same using GMHS and RGMHS for each treatment, although the RGMHS variance estimate does reduce the relative bias for the SM procedure by twenty-five percent in the Region 2 population;
- As expected, the RGMHS procedure yields more stable variance estimates. In the Conventional Financing population, the reduction is as great as thirty-five percent for three of the six median-estimation procedures. However, the improvements in stability for all median-estimation procedures are less pronounced in the Region 2 population, and neither the RGMHS and GHMS variances have good stability;
- In both populations, the error rates for the SM confidence intervals constructed from the GMHS and RGMHS SEs are the same and are indeed nominal. Error rates constructed from the GMHS and RGMHS SEs for other treatments are close, and for most treatments these error rates are not significantly different from 10%. The error rates for RGMHS original-data-interpolated medians (IO2000 and IO1000) in the Conventional Financing population are significantly less than 10%, providing more evidence that the original-data-interpolation procedures are too conservative (although this pattern is not seen in the Region 2 population). In Region 2, the error rates for both the GMHS and RGMHS transformed-data-interpolated medians with bins of width 4 and the error rates for the GMHS transformed-data-interpolated medians with bins of width 25 are significantly higher than nominal. We believe that the conflicting results between Tables 4 and 5 for confidence interval coverage for the different median-estimation

procedures in Region 2 is caused by the inherent instability in the variance estimates due to small sample size in that population, since the Table 5 error rates within median-estimation procedure are very similar for the GMHS and RGMHS SEs.

In terms of relative bias and error rates, the results in Table 5 are fairly consistent for the two variance estimates for each median-estimation procedure. The RGMHS estimator does improve the stability, but improved stability does not appear to be reflected in confidence interval coverage (at least for these samples). The consistency between the GMHS and RGMHS results reinforces our earlier conclusions vis-à-vis the different estimation procedures. Moreover, it supports the variance estimation methodology used in the larger simulation study and in SOC production: comparable results are achieved with 1/15 the replicate estimates.

Table 5: Relative Bias, Stability , and Error Rates Using GMHS and RGMHS Variance Estimation

Population	Median-Estimation Procedure	Relative Bias		Stability		Error Rate	
		GMHS	RGMHS	GMHS	RGMHS	GMHS	RGMHS
Con. Financing	SM	0.18	0.18	0.67	0.62	11.3	11.3
	IO2000	0.27	0.26	0.37	0.28	8.0	6.7*
	IO1000	0.22	0.21	0.33	0.23	8.0	6.3*
	IT4	0.08	0.07	0.26	0.17	9.3	9.7
	IT25	0.09	0.07	0.26	0.17	10.7	8.7
	IT50	0.07	0.06	0.26	0.17	9.3	8.3
Region 2 (Midwest)	SM	0.64	0.48	1.33	1.06	10.0	10.0
	IO2000	0.32	0.31	0.44	0.37	10.7	10.3
	IO1000	0.30	0.28	0.41	0.34	11.3	10.0
	IT4	0.14	0.14	0.39	0.36	13.3*	14.0*
	IT25	0.16	0.16	0.39	0.35	13.0*	12.7
	IT50	0.17	0.17	0.43	0.38	12.7	12.0

VI. Conclusion

We explored the effect of using variations of two different methods of estimating the median of continuous data on MHS variance estimation: direct estimation versus linear interpolation. Linear interpolation requires classifying continuous data into bins of standard width. This width can be arbitrary, and "optimal" widths may change as the sample distribution changes over time. The linear transformation based on the third quartile appeared to correct this problem. With the transformed data, the bins' locations change depending on the data.

Our empirical results indicated that the choice of method has a pronounced impact on the variance estimates given modified half sample replication. Our simulation study results examined the properties of the different median-estimation procedures on the variance estimates, using the grouped MHS variance estimator. In all four simulated populations, the transformed-data-interpolated medians performed the best, usually by a wide margin. Since all three bins widths considered with transformed data appeared to have the same variance estimation properties, we recommend using the fewest number of bins examined, i.e. use twenty-one bins (bins of size 50 on the transformed scale).

The recommended method has several advantages. First, it takes the scale of the different distributions into account through the linear rescaling. Second, the larger bin size should ameliorate some of the sampling effects. Finally, using linear interpolation saves computing resources by avoiding sorting each half-sample.

Acknowledgments

The authors would like to thank Elizabeth Huang and James Fagan of the U.S. Census Bureau for their helpful comments on earlier versions of this manuscript, and J.N.K Rao for his useful comments on the original simulation study. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to inform discussion.

References

DeGroot, Morris (1986). Probability and Statistics. Reading, MA: Addison-Wesley Publishing, Inc.

Fay, Robert E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Fay, Robert E. (1995), "VPLX: Variance Estimation for Complex Surveys, Program Documentation," unpublished Bureau of the Census Report.

Hoaglin, D.C. and Iglewicz, B. (1987). Fine-tuning Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, **83**, pp. 1147-1149.

Judkins, David R. (1990). Fay's Method for Variance Estimation. *Journal of Official Statistics,* **6**, pp. 223-239.

Kovar, J.G, Rao, J.N.K, and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. *The Canadian Journal of Statistics,* **16**, pp. 25-45.

Kovacevic, Milorad and Yung, Wesley (1997). Variance Estimation for Measures of Income Inequality and Polarization -- An Empirical Study. *Survey Methodology*, **23**, pp. 41-52.

Luery, Donald M (1990). Survey of Construction Technical Paper. Unpublished draft Bureau of the Census internal documentation.

Naylor, Thomas H., Balintfy, Joseph L., Burdick, Donald S., and Chu, Kong (1968). Computer Simulation Techniques. New York: John Wiley and Sons, Inc.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, **18**, pp. 209-217.

Rao, J.N.K. and Shao, J. (1996). On Balanced Half-Sample Variance Estimation in Stratified Random Sampling. *Journal of the American Statistical Association,* **91**, pp. 343-348.

Snedecor, George W. and Cochran, William G. (1980). Statistical Methods. Iowa: The Iowa State University Press.

Thompson, Katherine J. (forthcoming in 1998). *Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC)*. Washington, DC: U.S. Bureau of the Census. (Technical Report #ESM-9801, available from the Economic Statistical Methods and Programming Division).

Wolter, Kirk M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag, Inc.

Woodruff, Ralph S. (1952). Confidence Intervals for Medians and Other Position Measures. *Journal of the American Statistical Association*, **47**, pp. 635-646.