Survey Methodology, December 2000
Vol. 26, No. 2, pp. 153-162
Statistics Canada, Catalogue No. 12-001

Estimation and Replicate Variance Estimation of Median Sales Prices of Sold Houses

Katherine J. Thompson and Richard S. Sigman [1]

Abstract

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). The SOC is a multi-stage probability survey whose sample design is well suited to the modified half-sample replication (MHS) method of variance estimation. The literature supports applying the MHS method to replicate sample medians to estimate the sampling variance of a median. There are several computational advantages, however, to using grouped data to estimate medians, with linear interpolation being used within the grouped-data interval containing the median. Using survey data and simulated finite populations, we compared the effects of no grouping (*i.e.*, the sample median), grouping with fixed-size intervals, and grouping with data-dependent-sized intervals on medians and associated MHS variance estimates. We examined the mean squared errors and mean absolute errors of the median estimates and the relative bias and stability of the variance estimates and the coverage of the associated confidence intervals. We found that the data-dependent-sized intervals yielded variance estimates with the smallest bias, the best stability, and the best confidence intervals.

Key Words: Median; Modified half-sample replication; Survey of Construction.

1. 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). The SOC is a multi-stage probability survey whose sample design is well suited to the modified half-sample (MHS) replication method (balanced repeated replication with replicate weights of 1.5 and 0.5) for reasons outlined in section 3.B. In the near future, the SOC will move its current estimation and variance estimation systems to the Census Bureau's re-engineered post-data-collection system, the Standardized Economic Processing System (StEPS). When this occurs, SOC will change from its current non-replicate variance estimation procedure to the MHS replication variance estimation procedure (Thompson 1998). Because the SOC variance estimation methodology is changing, we decided to revisit the median-estimation methodology for continuous data. Our goal was to find a median-estimation method with good estimation and variance estimation properties, given the MHS replication.

We considered two methods of median-estimation. The first method uses the sample weights to estimate medians via empirical cumulative-distribution functions. The second method uses linear interpolation of grouped continuous data to approximate the median. The latter method is implemented in VPLX (Variances from ComPLeX Survey, Fay 1995), the replicate 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 while being considerably less computer resource-demanding. However, optimal bin widths and locations may differ by domain and may change over time as the sample distributions change.

In this paper, we compare six methods of median-estimation, given MHS replication: the sample median and five variations using linear interpolation. Section 2 provides a brief overview of the SOC design. Section 3 presents general methodology. Section 4 describes the empirical results from four months of SOC data that motivated the simulation study presented in section 5. Section 6 provides our conclusions and recommendations.

2. 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.

1. Katherine J. Thompson and Richard S. Sigman, Economic Statistical Methods and Programming Division, U.S. Census Bureau, Washington DC, 20233, U.S.A.

The first stage of the SUP and NP sample selection is a subsample of 1980 design 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. The PSUs were stratified within region by weighted 1980 population 16 years and older, weighted 1982 residential permit activity, and percent of housing in nonpermit areas. When possible, strata consisted of PSUs from the same state with the same metropolitan status. One PSU per stratum was selected. Self-representing (SR) PSUs were included in the sample with certainty (the stratum consists of one PSU). 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). These places were stratified by a weighted average of the ratio of permit-issuing activity in year i to the total US permit activity in year i (i = 78, 81, 82). 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. One-to-four-unit building permits are selected systematically in such a way that an overall one-in-forty sample is achieved; five-or-more-unit building permits are included with certainty. 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), stratified by 1980 Census population size. For the third stage of NP sample selection, field representatives completely canvass all of the roads in the sampled EDs (called segments). To reduce canvassing, a few of the larger EDs were subsegmented and one subsegment selected, or large EDs were 1-in-2 subsampled. Currently, there are a total of seventy-one active nonpermit 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.

3. Methodology

A. Median-Estimation Procedures

1. Sample Median

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

   – sort the sample observations in ascending order;

   – accumulate the sum of the associated survey weights;

   – select the first observation for which the associated sum of the weights exceeds fifty percent of the total weight.

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:

$$\hat{M} = F^{-1}\left(\frac{1}{2}\hat{N}\right) \approx ll + \left(\frac{\frac{1}{2}\hat{N} - \text{cf}}{f_i}\right) * (i) \quad (3.1)$$

where

F = the cumulative frequency of the characteristic using sample weights

ll = lower limit of the bin containing the median

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

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

f_i = median class frequency (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:

$$X_i' = X_i * \frac{1{,}000}{Q_3} \quad (3.2)$$

where Q_3 is the third quartile of the sample distribution (estimated using the ordered observations and sample weight as outlined in section 2.A.1). The interpolated median of the X' is multiplied by $(Q_3/1{,}000)$ to obtain an estimated median on the original scale [If the distribution contains negative values (*e.g.*, a distribution of net income), then use $X_i'' = (X_i - X_{(1)}) * 1{,}000/Q_3(X_i - X_{(1)})$, where $X_{(1)}$ is

the first order statistic and $Q_3(X_i - X_{(1)})$ is calculated from the distribution of $(X_i - X_{(1)})$. To obtain an estimated median on the original scale, multiply the interpolated median by $(Q_3(X_i - X_{(1)})/1,000)$ and add $X_{(1)}$.] This procedure 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.

This procedure is designed for symmetric or positively skewed distributions (usually the case with economic data). 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. If we based the linear transformation on Q_1 (the first quartile), the bin containing the median might be very close to the lowest bin in the distribution. In this case, the difference in variability between an interpolated median and the sample median would be small.

Using the original data to develop medians has the advantage of producing production-ready estimates and SEs. Determining the appropriate fixed bin width is difficult, however. As the bin widths get small (approach width 1), 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 variance estimate stability and bias. 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. The "standard" bin widths used for the transformed data less than Q_3 are not standard on the untransformed scale: the bin width is data-dependent. 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.

Figures 1 through 4 illustrate the effect of the linear transformation on the bin widths and location for two distributions. Figures 1 and 2 present a distribution that has a large spread of data values, including a few very large observations. Figures 3 and 4 present a distribution consisting of primarily small data values.

Figure 1 presents a histogram of the original distribution for houses sold with conventional financing, with bin width of \$25,000 [Note: the bin size was selected purely for presentation convenience, since this is a long-tailed distribution]. The median of this distribution is \$167,130, and Q_3 is \$225,000. Figure 2 presents the histogram of the linearly transformed distribution with bin width of 50. In this example, the transformed bins of width 50 are equivalent to bins of width \$11,250 on the original scale ((\$225,000/1,000)*50). Recall that the original-data bin sizes considered are \$1,000 and \$2,000. Thus, the transformed-data bins of width 4 would have a width of \$900 on the original untransformed scale. Notice the large "spike" at the last bin, which contains all of the sample greater than Q_3.

These figures also illustrate the differences in distribution of sample sizes across bins between the two methods. Using fixed bin widths with the original data results in quite variable bin sample sizes (see Figure 1). In contrast, by design the sample sizes within the data-dependent bins are much more uniform for all but the last bin (see Figure 2).

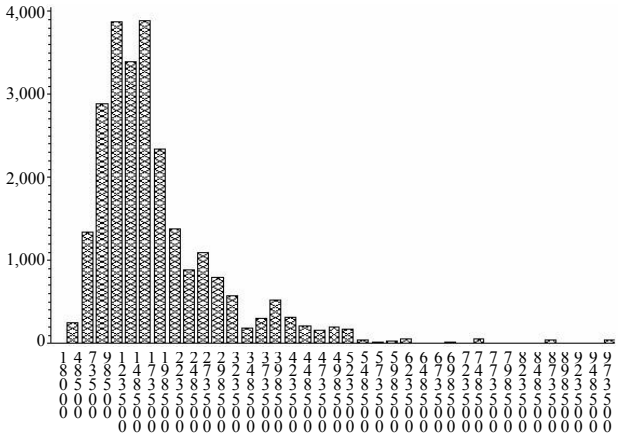

Figure 1 Original Distribution of Sales Price of Houses Sold with Conventional Financing Bin Width = \$25,000

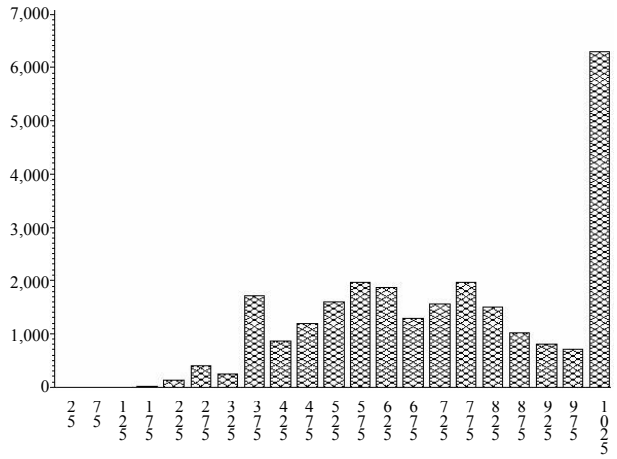

Figure 2 Transformed Distribution of Sales Price of Houses Sold With Conventional Financing Using Bin Width = 50 Bin Width on Untransformed Scale = \$11,250

Figure 3 presents a histogram of the original distribution of houses sold with FHA loans, with bin width of \$4,000 (again, the bin width is chosen for presentation convenience). The median of this distribution is \$108,280, and Q_3 is \$124,990. Figure 4 presents the histogram of the linearly transformed distribution with bin width of 50. In this example, the transformed bins of width 50 are

equivalent to bins of width \$6,250 on the original scale, and the transformed-data bins of width 4 would have approximate width \$500 on the original untransformed scale.

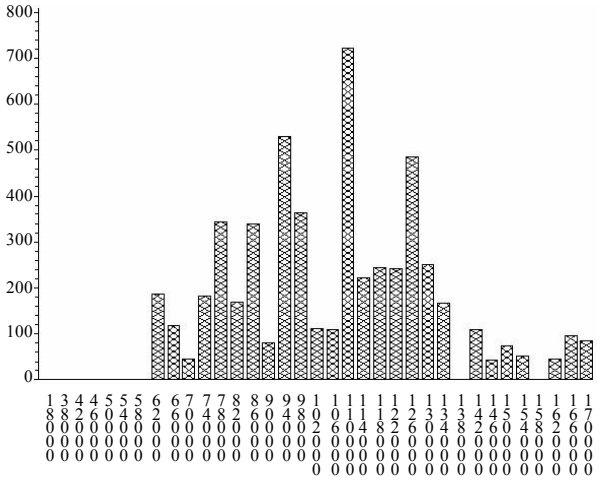

Figure 3 Original Distribution of Sales Price of Houses Sold with FHA Loans Bin Width = \$4,000

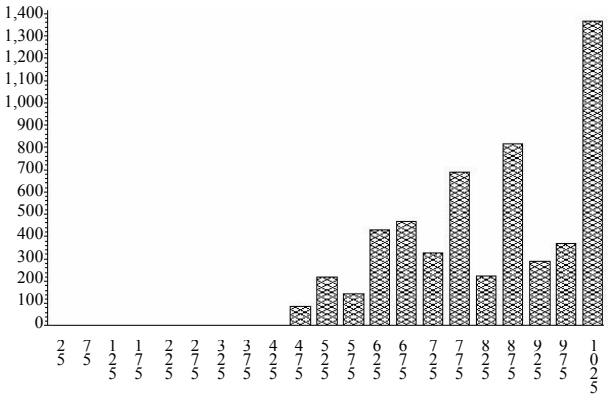

Figure 4 Transformed Distribution of Sales Price of Houses Sold With FHA Loans Using Bin Width = 50 Bin Width on Untransformed Scale = \$6,250

Figures 1 through 4 demonstrate the flexibility of the bins developed for linearly-transformed data. The bin size on the untransformed scale expands or contracts, depending on the spread of the data. Moreover, the data-dependent bin sample sizes are less variable compared to those associated with fixed bins.

To evaluate the first interpolation option (original-data-interpolated medians), we used two different sets of bin widths (classification sizes): bins of size \$2,000 (the same bin width used in the current production variance estimation system) and bins of size \$1,000. [Note: The VPLX variance estimation software would not allow any bin size smaller than 1,000 because the number of classes exceeded the allowable array range.] After examining several months of sales price estimates for the total U.S., 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 \$2,000 or 514 bins of size \$1,000, plus one bin of size \$36,000 and one bin whose width depends on the largest observation in the sample. One obvious problem with the locations of these bins is the potential effect of inflation. It is conceivable that within special financing categories or certain regions, the median sales price for houses sold could approach \$550,000, and the interpolation would fail as a consequence.

To evaluate the second interpolation option (transformed-data-interpolated-medians), 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 2,000 in terms of the number of bins. There are 250 bins of size 4 for the transformed data less than Q_3, and one larger bin containing all data greater than Q_3. 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 is always less than 1,000, so the last transformed-data classification is always (1,000 – high). Thus, by definition the last bin contains up to twenty-five percent of the data and is considerably wider than the other bins.

B. Variance Estimation

We used the Modified Half-Sample (MHS) replication method (Fay 1989 and Judkins 1990) to estimate the variance of a median as supported in the literature (*e.g.*, Rao, Wu, and Yue (1992); Rao and Shao (1996); Kovacevic and Yung (1997) for balanced repeated replication; and Judkins (1990) for MHS replication). MHS replication is a variation of the "traditional" balanced half-sample variance estimation described in Wolter (1985, 110-152). Balanced half-sample replication (BRR) is a variance estimation method designed for a two-PSU per stratum design. With BRR, a half-sample replicate is formed by selecting one unit from each pair and weighting the selected unit by 2 (so that it represents both units). Thus, estimates for every PSU are included in each replicate although half are weighted by zero. Replicates (half-samples) are specified using a Hadamard matrix. See Wolter (1985, 114-115) for a detailed description of the replicate formation procedure using Hadamard matrices. MHS replication uses replicate weights of 1.5 and 0.5 in place of the 2 and 0. The standard error for a median estimate using MHS replication is given by

$$\hat{\mathrm{SE}}\left(\hat{\mathrm{Med}}\right) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R} \left(\hat{\mathrm{Med}}_r - \hat{\mathrm{Med}}_0\right)^2}$$

where the $\underline{r}$ subscript refers to the replicate $\underline{r}$ median estimate $(r = 1, 2, \ldots, R)$ and the $\underline{0}$ subscript refers to the full sample the 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).

Neither the SUP nor the NP designs are two-sample-unit-per-stratum designs. At the first stage, one PSU per stratum is selected. The second and third stages are systematic samples, and often only one unit per stratum was selected at the second stage. A common approach used to address the one sample-unit per stratum problem is to

- "split" the SR sample-units into two panels per sample-unit using the original sampling methodology;
- form collapsed strata by pairing two (or three) "similar" NSR sample-units; and
- 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 the first stage sample-units (PSUs) 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 an approximate sampling-formula 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 MHS replication, we sort permits within predetermined sample-unit groups in SR units by geography and authorization date and systematically split the ordered sample into two panels as suggested in Wolter (1985, 131). Although this is essentially the only approach available for the SOC design, this method may not provide the correct variance estimates since units in both panels are correlated (in the original half-sample method, the two PSUs in the stratum are assumed independent). For more details on the replicate assignments, see Thompson (1998).

The SOC production system uses the Woodruff method (Woodruff 1952) to estimate the standard error of a median. The Woodruff method uses the estimated SE of a proportion $\underline{p}$ ($\underline{p} = 0.50$ for median-estimation) and projects the interval $(\underline{p} \pm \mathrm{SE}(\underline{p}))$ through the cumulative frequency distribution to obtain the lower limit of a 62.86 percent confidence interval for the median (the $\mathrm{SE}(\underline{p})$ can be estimated using replicate methods). The SE of the median is then estimated by subtraction. This methodology has had mixed success in the past according to SOC survey analysts.

4. 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 very similar estimates, but yielded three distinct sets of SEs: one set for the sample median, one set for the original-data-interpolated medians (fixed bin width), and one set for the transformed-data-interpolated medians (data-dependent bin width). There was no clear relationship between bin width and SE estimates within 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 standard errors resembling the published standard errors, so there was no available argument for publication consistency.

Moreover, there is some evidence that the Woodruff method publication SEs are underestimates or are at least inappropriate for the sample design used. Kovar, Rao, and Wu (1988) compared Woodruff SEs and BRR standard errors and found that the two methods had similar properties except for the case of stratified samples, where the strata are based on highly correlated separate variables (such as the SOC design). In this case, the Woodruff SE is often too small, and they concluded that "the BRR... methods (sic) are more robust to different population structures, since the error is extracted directly from the replicates." When the production system Woodruff SEs used the directly-calculated $\mathrm{SE}(\underline{p})$, the Woodruff SEs were generally smaller than the replicate SEs.

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.

5. 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 not 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 $\underline{i}$, where w_i is the sample weight associated with unit $\underline{i}$. The distributions of sales price for single-unit sold houses could

be approximated by lognormal distributions. The lognormal distribution has the probability density function

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

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

From our models, we generated four simulated finite bivariate populations with expected correlation $\rho = 0.6$ using the method outlined in Naylor, Balintfy, Burdick and Chu (1968, 99). The first of the two variables in each population represented sales price of sold houses and was obtained by generating a random normal variable with mean ζ and variance σ^2 using the parameters determined above, then exponentiating and shifting by the appropriate location parameters (θ). The second variable was used to form strata and first stage clusters. This variable had a marginal standard normal distribution and was obtained by independently generating a second standard random normal value, multiplying it by 0.8, and adding this term to $0.6\times$ the standard normal random variable used to generate the sales price variable. Percentiles, sample skewness, and sample kurtosis of each simulated population's sales price variable were very close to the corresponding statistics in the original population, especially when outliers were deleted using the resistant outer fences rule described in Hoaglin and Iglewicz (1987). Each population's size was the $\hat{N}$ estimated from the sample populations. Model parameters, sample correlations (between simulated sales price and stratifying variable), population size (N), and sample sizes (n) are reported in Table 1.

After generating the finite populations, we formed 50 equal sized strata in each population, then selected two sets of samples for two different survey designs:

- The first design is patterned after the SUP sample of permits for four-or-less-housing units in SR permit offices in SR PSUs (approximately 28% of the SOC sample). In this study, we selected 5,000 stratified without-replacement random samples from each simulated population using the same sampling rate in each stratum. To perform MHS replication, we sorted the sample within each stratum by stratifying variable and then systematically split the sample into two panels.
- The second design is patterned after the SUP sample of permits for four-or-less-housing units in NSR permit offices in SR PSUs and in SR permit offices in NSR PSUs (approximately 40% of the SOC sample). In this study, we selected 5,000 two-stage samples from each simulated population. The first stage is stratified without-replacement random sample of two PSUs per stratum ($N_h = 5$). The second stage is a systematic sample of units within PSUs. Because all PSUs are the same size, this study does not take the SOC PPS sampling into account and does not include the collapsing of first-stage units. The MHS replication uses the first-stage sample units (PSUs) within the same strata. The replicate weights do not account for large sampling fractions at the first stage of selection as recommended in Wolter (1985, 122), so all of the variance estimates are probably upwardly biased.

We did not attempt to simulate the SUP sample of permits for four-or-less-housing units in NSR PSUs and NSR permit offices (a three-stage sample, approximately 25% of the SOC sample); the SUP sample of permits for five-or-more housing units (approximately 2% of the SOC sample); or the NP sample of EDs (approximately 5% of the SOC sample). The three-stage sample, although non-negligible in SOC, is rarely used by other surveys at the Census Bureau, and the other two sectors of the SOC design do not contribute enough to the estimates to warrant a separate investigation.

To examine the precision of each median-estimation procedure over repeated samples, we estimated empirical Mean Squared Errors (MSE) and Mean Absolute Errors (MAE) from the 5,000 samples for:

SM: the sample median of each half-sample

IO2000: interpolated medians using original data, bins of size 2,000 (fixed bin width)

IO1000: interpolated medians using original data, bins of size 1,000 (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)

Table 1
Characteristics of Simulated Populations and Sample Sizes of Stratified Samples

		Sales Price Parameters			Correlation (Stratifier, Sales Price)	Population Size	Sample Size
Population	Distribution	θ	σ	ζ	ρ	N	n
Conventional Financing	lognormal	27,578	0.4895	11.84	0.57030	25,150	500
Midwest	lognormal	31,801	0.5957	11.69	0.55835	6,500	150
South	lognormal	29,414	0.5549	11.55	0.55929	14,550	300
West	lognormal	53,781	0.5822	11.59	0.55525	11,550	250

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

Population	Median	Q_3	Bin Width 4	25	50
Conventional Financing	167,173	222,263	889	5,557	11,113
Midwest (Region 2)	151,312	210,647	843	5,266	10,532
South (Region 3)	133,745	180,868	723	4,522	9,043
West (Region 4)	162,130	214,320	857	5,358	10,716

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 (including replicate 0, the full sample). 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.

We calculated $M(\zeta_i)$, the empirical MSE of median-estimation procedure i as

$$M(\zeta_i) = \frac{\sum_r (\zeta_{ri} - \bar{\zeta}_i)^2}{5,000} + (\bar{\zeta}_i - \zeta_p)^2$$
$$= \hat{\sigma}^2(\zeta_i) + \hat{bias}^2(\zeta_i) \qquad (5.1)$$

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).

We calculated the Mean Absolute Error (MAE) of each median-estimation procedure i as

$$MAE(\zeta_i) = \frac{\sum_r |\zeta_{ri} - \bar{\zeta}_p|}{5,000} \qquad (5.2)$$

as defined in DeGroot (1986, 209-211).

To compare the variance estimation properties of the different median-estimation methods, we calculated an MHS variance estimate (v_{ij}) corresponding to each median-estimation procedure i from 1,000 of the 5,000 samples. These variance estimates were compared in terms of

Relative bias $\left(\sum_{j=1} v_{ij}/1,000\right)/M(\zeta_i) - 1$

Relative stability $\left[\left(\sum_{j=1} (v_{ij} - M(\zeta_i))^2/1,000\right)\right]^{1/2} \Big/ M(\zeta_i)$

Error Rate Number of samples where ($\zeta_p < \theta_{Li}$ or $\zeta_p > \theta_{Ui}$)/1,000 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 criteria are used in Kovar, Rao, and Wu (1988) and in Rao and Shao (1996). The relative bias is a measure of the bias of the variance estimate as a proportion of the true MSE. The stability is a measure of the variance of the variance estimates; it approximates a c.v. of the variance estimate v_i. Note that the relative stability is not the relative MSE defined in Wolter (1985, 297) which uses the squared-MSE in the denominator. 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

1. Comparison of Median-estimation Procedures

   Table 3 presents the empirical root MSE, standard error, the bias, and the MAE for each median-estimation procedure from both simulation studies. Each of these statistics was calculated from 5,000 samples.

   These results reinforced our suspicions from the empirical data analysis described earlier. At least for sales price, all six median-estimation procedures perform approximately equally well, with approximately equal root-MSEs and MAEs between procedures in each population.

2. Comparison of MHS Replication Variance Estimation Properties of Median-Estimation Procedures

When we examined the variance estimation properties for each procedure, the results were quite different. As with our empirical data analysis, we had three very distinctive sets of results. 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 1,000 samples. The denominator for the relative bias and stability ("truth") are based on 5,000 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.

Table 3
Empirical Root MSE, Standard Error, Bias, and MAE for Median-Estimation Procedures

Population	Median-Estimation Procedure	Unclustered Single-Stage Sample				Clustered Two-Stage Sample			
		Root MSE	SE	Bias	MAE	Root MSE	SE	Bias	MAE
Conventional Financing	SM	3,345	3,345	-12	2,671	3,389	3,374	324	2,733
	IO2000	3,320	3,316	161	2,698	3,346	3,341	189	2,685
	IO1000	3,387	3,368	-354	2,642	3,431	3,420	-278	2,774
	IT4	3,351	3,340	273	2,673	3,378	3,364	311	2,719
	IT25	3,304	3,293	276	2,617	3,337	3,321	322	2,664
	IT50	3,282	3,265	329	2,606	3,305	3,283	375	2,636
Region 2 Midwest	SM	6,316	6,287	-598	4,966	6,273	6,228	-753	4,959
	IO2000	6,276	6,275	-127	4,992	6,335	6,207	-1,271	5,029
	IO1000	6,343	6,297	-767	4,939	6,526	6,280	-1,774	5,204
	IT4	6,372	6,363	328	5,004	6,294	6,228	-908	4,979
	IT25	6,273	6,272	127	4,937	6,270	6,154	-1,199	4,971
	IT50	6,220	6,218	160	4,936	6,224	6,114	-1,164	4,966
Region 3 South	SM	3,670	3,658	301	2,931	3,835	3,752	796	3,054
	IO2000	3,708	3,669	539	2,998	3,796	3,739	656	3,011
	IO1000	3,742	3,740	101	2,941	3,809	3,804	212	3,066
	IT4	3,718	3,662	639	2,951	3,814	3,736	766	3,028
	IT25	3,699	3,638	669	2,924	3,793	3,711	787	2,992
	IT50	3,692	3,616	745	2,912	3,778	3,680	856	2,970
Region 4 West	SM	4,385	4,382	-140	3,509	4,394	4,351	616	3,506
	IO2000	4,425	4,421	185	3,578	4,362	4,339	449	3,487
	IO1000	4,477	4,469	-258	3,530	4,411	4,410	-57	3,535
	IT4	4,414	4,403	318	3,514	4,383	4,342	599	3,494
	IT25	4,376	4,364	315	3,460	4,334	4,296	573	3,439
	IT50	4,367	4,350	391	3,455	4,320	4,271	644	3,436

In both studies, the variance estimates of the transformed-data-interpolated medians perform best in terms of relative bias and stability. 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. These results are surprisingly strong for the two-stage clustered design, since the variance estimates are expected to be biased upwards (see section 5.A);
- The variance estimates of the interpolated medians had the best stability. The variance estimates of the sample median had the poorest stability in all four populations. This result was expected due to the smoothing effect of interpolation. Again, the transformed-data-interpolated medians generally performed better than the original-data-interpolated medians, although the difference is not as pronounced as in the case of relative bias. Generally, the stability is close with all three bin widths for the transformed-data-interpolated medians.

The results for each median-estimation procedure's confidence interval coverage are not as consistent, varying by design. With the single-stage unclustered design, 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 close to nominal or better coverage; in fact, none of these error rates is statistically different from the nominal 10%. The confidence intervals constructed from original-data-interpolated medians and SEs are extremely conservative. Here, the positive bias in the variance estimates makes these intervals unnecessarily wide, thereby reducing the power to make interesting findings. The coverage with the sample median is erratic.

Some of these coverage patterns are repeated in the two-stage clustered design. Again, the coverage with the sample median is erratic, and the coverage rates for the confidence intervals constructed from original-data-interpolated medians are better than nominal (although only significantly better than nominal in two populations). The error rate pattern is quite different for the transformed-data-interpolated medians. In all but the Region 4 population, the coverages rates for the three procedures are worse than nominal. However, with bins of widths 4 and 25, only one error rate is significantly larger than 10%; for bins of width 50, two of these three error rates are significantly larger than 10%. All of the interpolated-data-medians have significantly smaller than nominal error rates in the Region 4 population; consistent with the other population's results, the error rates for the original-data-interpolated medians are the farthest from 10%.

Table 4
Relative Bias and Relative Stability for Variance Estimates, and Error Rates for 90% Confidence Intervals

Population	Median-Estimation Procedure	Unclustered Single Stage Design			Clustered Two-Stage Design		
		Relative Bias	Relative Stability	Error Rate	Relative Bias	Relative Stability	Error Rate
Conventional Financing	SM	0.19	0.69	11.0%	0.11	0.58	15.1%*
	IO2000	0.25	0.35	6.9%*	0.25	0.37	9.0%
	IO1000	0.21	0.32	7.0%*	0.19	0.33	9.3%
	IT4	0.06	0.25	10.0%	0.06	0.27	11.3%
	IT25	0.07	0.25	10.9%	0.06	0.27	11.8%*
	IT50	0.05	0.26	9.5%	0.05	0.28	12.1%*
Region 2 Midwest	SM	0.57	1.24	7.3%*	0.41	1.07	7.9%*
	IO2000	0.33	0.44	6.9%*	0.23	0.35	8.6%
	IO1000	0.30	0.42	7.0%*	0.17	0.30	8.7%
	IT4	0.15	0.41	10.1%	0.14	0.41	11.5%*
	IT25	0.16	0.40	9.8%	0.11	0.37	10.4%
	IT50	0.15	0.42	9.0%	0.11	0.40	10.4%
Region 3 South	SM	0.30	0.88	12.4%*	0.15	0.71	11.1%
	IO2000	0.31	0.42	6.7%*	0.28	0.39	7.5%*
	IO1000	0.29	0.40	6.7%*	0.27	0.38	7.3%*
	IT4	0.04	0.29	11.0%	0.01	0.28	10.8%
	IT25	0.02	0.28	11.0%	-0.01	0.27	11.3%
	IT50	0.01	0.29	11.1%	-0.02	0.28	11.9%*
Region 4 West	SM	0.39	0.98	8.9%	0.25	0.79	8.6%
	IO2000	0.32	0.42	6.2%*	0.31	0.41	5.2%*
	IO1000	0.29	0.39	6.2%*	0.28	0.38	5.2%*
	IT4	0.11	0.32	8.6%	0.10	0.31	7.6%*
	IT25	0.10	0.31	9.4%	0.09	0.30	7.5%*
	IT50	0.08	0.31	9.5%	0.08	0.31	8.3%*

In both studies, the transformed-data-interpolated medians have the best variance estimation properties in terms of relative bias and relative stability by a large margin, regardless of bin width. And, in both studies, the transformed-data-interpolated medians using bins of width 4 or width 25 have excellent confidence interval coverage. Since the transformed-data-interpolated-medians using bins of width 50 or width 25 yielded the "best" estimators in terms of root-MSE and MAE in both studies, using linear interpolation on transformed data with bins of width 25 appears to be the best median-estimation procedure in terms of estimation and variance estimation properties.

6. Conclusion

We explored the effect of using variations of two different methods of estimating the median sales price of sold houses: direct estimation versus linear interpolation. Linear interpolation requires classifying continuous data into bins of standard width. This width can be arbitrary, can differ greatly by domain, and 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' widths and locations in the distribution change depending on the data.

Our empirical results indicated that the choice of method has a pronounced impact on the variance estimates given MHS replication. Our simulation study examined the properties of the different median-estimation procedures on the MHS replicate variance estimates. In all four simulated populations, the transformed-data-interpolated medians (data dependent bin widths) performed the best, usually by a wide margin. Most critically, this method greatly reduces the overestimation of the variance. Using bins of width 25 on the transformed scale (41 bins total) yielded the best median sales price estimates and variance estimates, given MHS replication and is our recommended method for the Survey of Construction.

The recommended method has several advantages. First, it is adaptive. It works well for a variety of distributions, because the bin widths themselves depend on the distribution at hand. Second, it saves computing resources by avoiding sorting half-samples. Third, the data-dependent-intervals can be easily incorporated into generalized survey-processing software. Finally, it gives better estimates and

MHS replicate variance estimates (at least for sales price of sold houses). We expect that these results are generalizable for other continuous distributions as well, although obviously this conjecture should be tested on other data sets. Other areas for future research include examining the relationship between sample size and precision of the median estimates, examining alternative bin sizes, and exploring the robustness of the recommended procedure with different replicate variance estimation procedures.

Acknowledgements

The authors would like to thank Elizabeth Huang and James Fagan of the U.S. Census Bureau, two anonymous referees, and the associate editor 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 encourage discussion.

References

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

Fay, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Fay, R.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, 1147-1149.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 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, 25-45.

Kovačević, M., and Yung, W. (1997). Variance estimation for measures of income inequality and polarization – An empirical Study. *Survey Methodology*, 23, 41-52.

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

Naylor, T.H., Balintfy, J.L., Burdick, D. S. and Chu, K. (1968). *Computer Simulation Techniques*. New York: John Wiley & 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, 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, 343-348.

Thompson, K.J. (1998). Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC). Technical Report #ESM-9801, available from the Economic Statistical Methods and Programming Division.

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

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.