Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Estimation and replicate variance estimation of deciles for complex survey data from positively skewed populations

by Stephen J. Kaputa and Katherine Jenny Thompson

January 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 351-368
Statistics Canada, Catalogue No. 12-001-X

Estimation and replicate variance estimation of deciles for complex survey data from positively skewed populations

Stephen J. Kaputa and Katherine Jenny Thompson[1]

Abstract

Thompson and Sigman (2000) introduced an estimation procedure for estimating medians from highly positively skewed population data. Their procedure uses interpolation over data-dependent intervals (bins). The earlier paper demonstrated that this procedure has good statistical properties for medians computed from a highly skewed sample. This research extends the previous work to decile estimation methods for a positively skewed population using complex survey data. We present three different interpolation methods along with the traditional decile estimation method (no bins) and evaluate each method empirically, using residential housing data from the Survey of Construction and via a simulation study. We found that a variant of the current procedure using the 95th percentile as a scaling factor produces decile estimates with the best statistical properties.

Key Words: Median; Modified half-sample replication; Interpolation; Deciles.

1 Introduction

Developing viable decile estimates for positively skewed populations from complex survey data poses interesting challenges. The literature supports two different approaches to percentile estimation with complex survey data. The first method (the "traditional" method) obtains decile estimates from empirical cumulative-distribution functions, selecting the item value that corresponds to the sample percentile computed by summing associated survey weights. This approach yields decile estimates that are "close to unbiased" but unstable. An alternative approach is to group the continuous data into disjoint intervals (bins), then use linear interpolation over the bin containing the decile. With appropriately defined bins, this approach also produces nearly unbiased decile estimates while improving their stability – at least for the percentiles that are far from the tail of the distribution. For the upper percentiles, often the binned data contain very few observations, with little or no uniformity. Hence, the reliability of the large decile estimates (*e.g.*, 90th percentile or greater) is rarely comparable to that of the other deciles.

Although the usage of interpolation is advantageous for developing stable estimates, developing an optimal set of bins for a given characteristic is not always an easy task. Often, the distributions change over time, and the bin widths/locations in the sample should reflect this change in scale. For example, the average sales price for single-family homes in a geographic area may increase over time due to inflation, but the population of single-family homes in that area is still characterized by a skewed distribution, with a few expensive homes located in the tail. Many economic data programs share this trait. Consequently, developing a fixed set of bins for interpolation with an ongoing survey is unwise. To address this, Thompson and Sigman (2000) introduced an estimation procedure for estimating *medians* from highly positively skewed population data. Their procedure uses interpolation over data-dependent intervals (bins), after scaling by the 75th percentile. The earlier paper examined the estimation and variance

1. Stephen J. Kaputa and Katherine Jenny Thompson, Office of Statistical Methods and Research for Economic Programs, US Census Bureau, 4600 Silver Hill RD, Washington, DC 20233. E-mail: Stephen.kaputa@census.gov.

estimation properties of the considered methods, using modified half sample (MHS) replication for variance estimation (Fay 1989; Judkins 1990).

This research extends the previous work to decile estimation methods using complex survey data sampled from a positively skewed population. We present three different interpolation methods along with the traditional decile estimation method (no bins) and evaluate each method empirically, using residential housing data from the Survey of Construction (SOC) conducted by the U.S. Census Bureau and via a simulation study. Our research was motivated by a recent request from the SOC data users to estimate and publish complete sets of decile estimates for several housing characteristics. Thus, our research was conducted under the constraints of maintaining comparably reliable median estimates as those currently published and using MHS replication for variance estimation.

Section 2 presents the candidate decile estimation methods and gives an overview of modified half-sample replication. Section 3 evaluates these procedures, using empirical and simulated data from the Survey of Construction (SOC). Finally, we conclude with recommendations in Section 4.

2 Methodology

2.1 Decile estimation

We consider two approaches to decile estimation for continuous data: the sample decile (SD) method and interpolation. The SD method uses ordered sample weights to locate the estimate (Rao and Shao 1996). For this, the characteristics values are sorted in ascending order, and the sample weights are accumulated until they exceed the desired decile's percent of the total weight.

Interpolation methods group the continuous data in bins and interpolate over the bin containing the decile. To obtain the decile estimate $\left(\xi^{d}\right)$, we use the Woodruff formula (Woodruff 1952) for interpolation provided below:

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

where

F = cumulative frequency of the characteristic using sample weights,

ll = lower limit of the bin containing the decile,

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

cf = cumulative frequency in all intervals preceding the bin containing the sample decile,

f_i = decile class frequency (estimated total number of elements in the population of the interval containing the sample decile),

i = width of the bin containing the sample decile,

d = desired decile (0.1, 0.2, 0.3, …, 0.9).

Notice that this formula does not require that each bin to be of equal length. However, it does require that the data within each bin be uniformly distributed. This later requirement poses the true challenge with a highly skewed population, especially in the upper tail.

Figure 2.1 below illustrates how to use the Woodruff method to estimate the 80th decile. The sample data have been grouped into twelve separate bins. The empirical CDF is produced from the complete set of weighted sample data (as one referee noted, the empirical CDF is extremely smooth for sample survey data; in practice, the curve would include discrete steps. The Woodruff method procedure would be the same, however). The decile estimate is located at the intersection of the empirical CDF curve and red asymptote at $Y = 0.80$. The 80th decile is $F^{-1}(0.80)$, contained in the 5th bin; the interpolated estimate of the 80th percentile would therefore be obtained by using (2.1) over the fifth bin.

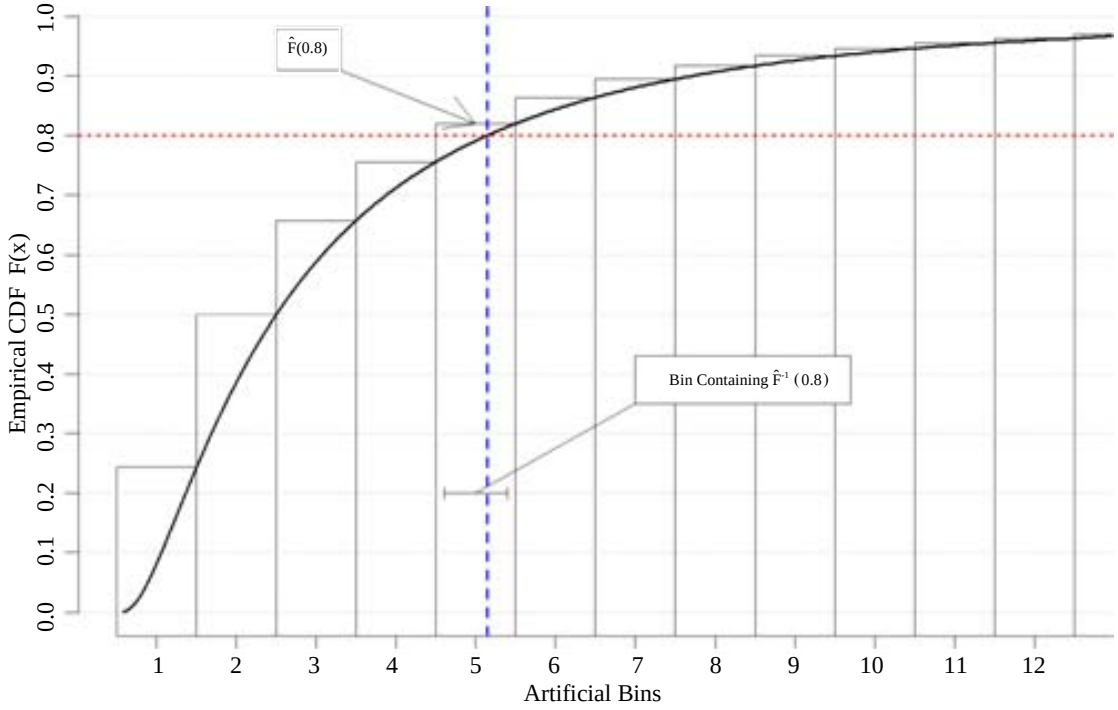

Figure 2.1 Illustration of the Woodruff method

Determining the optimal bin size for both estimation and variance estimation can be difficult. As the bins narrow (approaching width 1), then the variance estimates become more unstable. Smoothing the estimates via the interpolation reduces the instability of variance, but increases the bias in the estimate. The bias component increases as the bin widths increase.

Economic data generally have a positive skewed distribution. Moreover, the subdomains' characteristic distributions will vary, and their respective moments change over time as the economy changes. Consequently, developing a standard set of fixed bins for interpolation that work consistently over time is nearly impossible. Instead, Thompson and Sigman (2000) developed a "data-dependent" binning procedure, where the width of each bin is determined separately by the estimation cell. Their

recommended method linearly transforms each characteristic to a standard scale and then uses a standard set of bins for every characteristic. The authors use the following linear transformation

$$X_i' = X_i \times \frac{1{,}000}{Q_{75}}$$

where Q_{75} is the 75th percentile (3rd quartile) of the sample distribution, obtained using the SD method. The interpolated-median estimate of the X' is multiplied by $\left(Q_{75}/1{,}000\right)$ to obtain a value on the original scale. This procedure is equivalent to simply dividing the original sample in each estimation cell from 0 to Q_{75} into Z bins of equal width and placing the remainder of the sample into one bin, which, by design, is much larger than the others. With the highly positive skewed housing data, this transformation works well for estimating the median because it is far from the Q_{75} scaling parameter. However, it does not permit estimation of either the 80th or 90th deciles. Thus, if we wanted to continue using an interpolation method, we needed to consider alternative transformations.

The simplest approach is to use the original data-dependent bin method with a higher scaling parameter, *i.e.*, use any percentile value larger than 90%. We use the 95th percentile as the scaling factor and hereafter refer to this method as the "P95 method".

The P95 method does create uniform distributions within the majority of the bins but is still problematic at the upper end of the distribution for two reasons. First, the final bin contains only five percent of the sample distribution, and the values within this bin are generally very different. Second, the data-dependent binning procedure requires that each decile be "far from" the large final bin; if not, then the decile estimates exhibit the same instability as those obtained using the "SD method." Unfortunately, the bin containing the 90th percentile is often close to the final bin when using a scaling parameter of 95%.

To address the second issue, we considered another data-binning approach, denoted as the "P75 method." For this, we create two sets of bins per estimation cell, each with different widths above and below the cell's Q_{75} value. This requires two separate linear transformations per estimation cell, given by

$$X_i' = X_i \times \frac{1{,}000}{X_{75}} \text{ when } X_i < X_{75}$$

$$X_i'' = \left(X_i - X_{75}\right) \times \frac{1{,}000}{\left(X_{100} - X_{75}\right)}$$

$$\text{when } X_i \geq X_p \text{ and } X_{100} = \text{ maximun value in sample.}$$

The X_i' is then placed into Z equal length bins, and X_i'' into K equal length bins, where $Z \neq K$. The interpolation is performed independently for each decile, with the appropriate inverse transformation being applied to each interpolated decile. This procedure ensures that median estimates exactly match those obtained with the current procedure.

Our third considered interpolation approach makes parametric assumptions about the characteristics. Often, economic data are approximately log-normally distributed (*e.g.*, Steel and Fay 1995). The Normal Binning method (denoted "NB") uses the properties of the normal distribution applied to the log-transformed data to obtain data-dependent bins. The binning technique ensures that areas of high

probability have smaller bin widths to limit the amount of observations per bin and areas of low probability have larger bin width to increase the amount of observations per bin.

The NB method centers the log-transformed data around the weighted sample median, then scales the centered data by an estimate of the population standard deviation. We use the sample median because it is more outlier resistant than the sample mean. Of course, the mean and median are equivalent with normally distributed data. Given a standard normal distribution where $\mu = 0$ and $\sigma = 1$, then

$$IQR = Q_3 - Q_1$$
$$IQR = \left(0.67449*\sigma\right) - \left(-0.67449*\sigma\right) = \sigma\left(0.67449 + 0.67449\right) = \sigma * 1.34898.$$

We estimated the standard deviation (sigma) as the ratio $\sigma \approx IQR/1.34898$, where the IQR is obtained from the empirical CDF in the estimation cell. To normalize the data, we applied the following transformation

$$Y_i = \text{Log}\left(X_i\right) \quad Y_i' = \frac{Y_i - Y_{\text{med}}}{\sigma_y} = \frac{Y_i - Y_{\text{med}}}{IQR_y/1.34898}$$

where

Y_{med} = log-transformed sample median over domain i,

IQR_y = log-transformed sample interquartile range over domain i.

Again, the sample deciles and interquartile ranges are obtained via the SD method. If the data are log-normally distributed, Y_i' should have a standard normal distribution, so that roughly 68.3% of the data are within one standard deviation of the mean and 95.4% of the data are within two standard deviations of the mean. Using those properties, we split the transformed Y_i' into the five different zones and created the 45 bins shown in Table 2.1.

Table 2.1
Bins for the log-normal transformation (Normal method)

Zone	1	2	3	4	5
Range	[Low, -2)	[-2, -1)	[-1, 1)	[1, 2)	[2, High]
Percent in Zone	2.3	13.6	68.2	13.6	2.3
Bins	1	6	31	6	1
Average Percent of Sample Units per Bin	2.3	2.3	2.2	2.3	2.3

There are four different bin widths with roughly the same average percentage of sampled units per bin. Woodruff's method is applied to the transformed data to obtain the deciles and we exponentiate these decile estimates to obtain values on the original scale. Unlike the linear rescaling methods presented above, there is an additional induced estimation bias caused by the power transformation. It may have

been possible to make a bias adjustment for the transformation via a Taylor expansion, as suggested by a referee, but we did not consider this approach.

2.2 Variance estimation

The MHS replication method (aka "Fay's method") is a "compromise" between the stratified jackknife and the BRR method (Fay 1989). Rao and Shao (1999) demonstrate that the MHS variance estimator is asymptotically consistent for both smooth statistics such as ratio estimators and for non-smooth statistics such as sample quantiles estimated using the SD method outlined in 2.1. Their paper does not extend this property to interpolated decile estimates, although it does follow that these variance estimates should be consistent as the bin width approaches width 1. Like BRR, MHS replication uses a Hadamard matrix to form replicates, but uses replicate weights of 1.5 and 0.5 instead of the values of 2 and 0 used in BRR. The MHS formula for standard error estimation of any estimate $\hat{\theta}$ is

$$\hat{S}\left(\hat{\theta}\right) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}_0\right)^2} \tag{2.2}$$

where $\hat{\theta}_r$ is the r^{th} replicate estimate $\left(r = 1, 2, \ldots, R\right)$ and $\hat{\theta}_0$ is the full sample estimate. The sum of squared error term is adjusted by a factor of $4 = 1\Big/\left(1 - 0.5\right)^2$ to prevent negative bias in the variance estimate (Judkins 1990).

3 Empirical analysis

3.1 SOC sample design

As mentioned in the introduction, our research was motivated by a request from data users of the Survey of Construction (SOC). The SOC is a national survey that collects information on characteristics of new residential housing in the United States. The SOC data are used to produce three principle economic indicators published each month by the U.S. Census Bureau: housing starts, housing completions, and housing sales (single-family homes only). In addition, SOC publishes monthly, quarterly, and annual estimates on a variety of housing characteristics, such as sales price and average sales price per square foot of sold houses, length of time from permit authorization to housing start, and length of time from start to completion of housing construction. This paper examines two key housing characteristics that are published annually: sales price and price per square foot of sold homes. Both characteristics are collected monthly as they become available from the homebuilders. Currently, average and median estimates for both characteristics are included in the annual reports; average and median sales price of sold homes are also published monthly.

The SOC universe comprises two sub-populations: areas that require building permits and areas that do not. Areas that require building permits are covered in the Survey of the Use of Permits (SUP) and non permit issuing areas are covered by the Nonpermit Survey (NP). The vast majority of the sample comes from SUP. Both populations are sampled from the same PSUs, but are independent samples at subsequent

stages. Since the majority of the SOC sample consists of sampled permits, we focus entirely on the SUP portion of the SOC in our research.

The SUP is selected in three stages. The first stage selects a probability proportional to size (PPS) subsample of Primary Sampling Units (PSUs) from the 2000 Current Population Survey design (CPS) and is performed once every ten years. The CPS PSUs are land areas such as counties or townships. The second stage of the SUP sample is a stratified systematic sample of permit-issuing places within sampled PSUs, also performed once every ten years. The third stage of sampling is performed monthly in each of the sampled permit-issuing places. Each month, the Field Representatives develop complete lists of new building permits from the permit offices in the sampled places and select a systematic sample of building permits. Sampling rates are assigned to permit offices to obtain an overall sampling rate of one-in-fifty for one to four unit structures. Larger buildings of five or more units are included with certainty (*i.e.*, are self-representing).

The SOC uses the MHS replication method to estimate variances with a 200×200 Hadamard matrix, assigning a total of 198 rows to replicate groups. Since SOC does not have a two-PSU per stratum design, SOC uses a collapsed stratum approach for creating replicates: see Thompson (1998) for details.

3.2 Empirical data analysis of SOC data

Our empirical analyses uses SOC data collected from 2006 through 2009. SOC uses the data-dependent binning method described in Section 2.1 with 41 bins to produce median estimates. We use 51 bins for the P95 method, with 95% of the sample spread over 50 equally sized bins; the 51st bin contains any data greater than the 95th percentile. The P75 method uses 40 equal-sized bins for all values below the 75th percentile and 10 equal-sized bins for all values between the 75th percentile and the maximum value of the sample distribution. Finally, the NB method uses a total number of 45 bins.

For sales price and price per square foot, all of the decile estimates obtained via the P75, P95, and SD methods are quite comparable to each other for the 10th through 70th deciles; the NB deciles were generally slightly larger than their other counterparts. However, the P75 method decile estimates for the 80th and 90th deciles were consistently larger than the other three methods' estimates. The explanation is straightforward: with the P75 method, both the 80th and 90th deciles are almost always located in the same bin. Both characteristic's distributions are quite skewed. Consequently, the majority of the upper 25 percent of the sample is contained in the bin closest to the 75 percentile.

The patterns displayed by the decile estimates for each method were extremely consistent at both the national and regional levels. Unlike the estimate comparisons, there are fewer clear patterns with the variance estimates. The variance estimates for 80th decile obtained using the P75 method were considerably larger than those obtained by the three other methods and the 90th was likewise considerably smaller. With respect to the other three methods, the NB variances tend to be smaller than the corresponding P95 method and SD variances; these differences are more pronounced with the sales price per square foot decile estimates.

Three of the four considered methods yielded comparable sets of decile estimates. The P75 method proved intractable given the highly skewed distributions considered; we simply could not find an adequate "bin width" for the upper quartile of data. Thus, the empirical data evaluation reduced our candidate set of estimation methods to three. However, although the corresponding estimates were quite similar, the

variance estimates were clearly different. Consequently, we decided to conduct a simulation study to evaluate the statistical properties of the alternative estimators over repeated samples.

3.3 Simulation study

3.3.1 Modeling and sample selection procedure

For our simulation, we developed a population that mimics the qualities of the majority of the SUP population. That is, we developed stratified populations of permit offices (PSUs), from which we selected permit samples (SSUs). There were several advantages to developing such a complex simulation set up. From a practical immediate perspective, it was beneficial for interpreting the empirical results presented in Section 3.2. More important, the earlier research conducted by Thompson and Sigman (2000) obtained nearly perfect results for the data-dependent interpolated medians on a simulated population that did not include clustering; the distinctions between the statistical properties of each method only became apparent when clustering was incorporated into the design.

We used a "bottom-up" approach to develop viable simulated population data. First, we modeled multivariate populations of permit data within each region. Next, we combined the modeled permit data to form "clusters" representing the permit offices (the primary sampling units). To guard against model misspecification, we independently developed two artificial populations of permit data with each permit record containing sales price and price per square foot, modeling one population as log-normally distributed within region using the algorithm outlined in Lienhard (2004) and modeling the other using a nonparametric SIMDATA algorithm (Thompson 2000) in each region.

In general, the modeled permit data in the nonparametric population is a better representation of the corresponding levels at each decile of the training data. However, the distributions of permits in the log-normal population are quite smooth, whereas the nonparametric population distributions are "choppy," with large breaks (steps) between adjacent point estimates. Table 3.1 presents key percentiles and means from the simulated populations for both modeled characteristics, comparing them to the empirical values from the SOC data (denoted by the "Training Data" column).

Our simulation procedure developed populations of permits (the SSUs), then created the artificial first stage clusters (permit offices) and stratified them. The two-step cluster creation and stratification process described below assumes that the permits within population strata are heterogeneous with respect to sales price and price per square foot and that permits issued from the same permit office have similar housing characteristics. These criteria were obtained from the subject matter experts, who believe that modeling variation between permit offices was more realistic than modeling variation within permit offices. The multivariate log-normal population lends itself better to this than the nonparametric population; the assigned elements within each cluster tend to be homogeneous because of the smoother distribution.

We used discriminant analysis to group the simulated permits into disjoint strata. After applying the same discriminant function to each simulated permit data population (log-normal and nonparametric), we clustered the permits within strata to form approximately 14,000 active permit offices per population. The cluster analysis application created permit offices of variable size with homogenous characteristics within office.

Table 3.1
Simulated populations' statistics and empirical data statistics

			Simulated Population		Training Data (Weighted)
			Log-normal	Nonparametric	
Sales Price	Decile	1	74,747.74	94,323.27	95,000.00
		5	105,009.64	120,073.33	120,000.00
		10	126,492.43	136,009.85	140,000.00
		20	158,517.61	158,931.57	160,000.00
		30	186,927.58	181,941.12	180,000.00
		40	215,346.44	205,003.46	210,000.00
		50	245,502.91	230,188.69	230,000.00
		60	280,064.08	260,524.68	260,000.00
		70	322,790.68	301,374.90	300,000.00
		80	381,501.24	359,138.64	360,000.00
		90	482,209.62	488,517.45	490,000.00
		95	586,359.10	622,185.68	630,000.00
		99	855,983.47	1,167,704.85	1,300,000.00
		Mean	283,085.94	287,134.16	290,000.00
			Simulated Population		Training Data (Weighted)
			Log-normal	Nonparametric	
Sales Price per Square Foot	Decile	1	32.76	32.68	35.00
		5	43.11	47.27	47.00
		10	49.86	54.70	55.00
		20	59.26	63.74	64.00
		30	67.22	70.71	72.00
		40	74.91	77.14	78.00
		50	83.11	83.60	84.00
		60	92.42	90.60	91.00
		70	103.75	98.70	99.00
		80	119.61	109.57	110.00
		90	147.62	130.02	130.00
		95	178.94	155.08	160.00
		99	265.10	262.68	270.00
		Mean	93.58	94.90	96.00

We selected 5,000 repeated samples from our simulated population using a much simplified version of the SOC design described above. The first stage of sampling selects permit offices. The largest 250 offices at the US level were selected with a probability of one (certainty), so that each repeated sample contains the same self-representing offices. Then, we selected a probability proportionate to size sample of two non self-representing permit offices in each stratum, with each office receiving its own permit office weight.

At the second sampling stage, we selected permit records from each sampled permit office. We selected a simple random sample (SRS) of permits from each office, with an office sampling rate obtained by dividing the permit office's weight by 50, thus obtaining an overall one-in-fifty sample of permits (if the permit office weight is greater than fifty, all permits within the office are sampled). Final weights for each record were calculated by multiplying the permit office weight and the permit weight. The permits selected from the certainty offices vary in each repeated sample due to the independent sampling unless the office contained more than 50 permits.

Finally, in each sample, we assigned permits or permit offices to replicates. We did not mimic the SOC partially balanced half sample application described in Section 3.1. Collapsing strata induces bias in the variance estimates. To eliminate this bias component from our simulation, we used a two-PSU per stratum design and 572 replicates (*i.e.*, a 572×572 Hadamard matrix), so that each of the 250 self-representing offices and each of the 321 non self-representing strata (pairs of sampled permit offices) received its own Hadamard matrix row. Mimicking the SOC production method, each self-representing office is treated as a "pseudo stratum," and replicate panels are obtained by randomly splitting the permits within each office.

Within each sample, a set of estimates at the US and regional level were calculated for the three considered decile estimation methods in each replicate and the MHS replicate variance estimates for each decile were computed using (2.2) with $R = 572$.

3.3.2 Evaluation methodology

The simulation study examines the statistical properties of each decile estimation method and associated variance estimates over repeated samples. Let ζ_m^d represent the decile d estimate calculated by using method $m\,(m = \text{SD, P95, NB})$.

To assess *estimation* properties of method m for decile d over repeated samples, we computed the relative bias and the empirical mean squared error. The relative bias of each decile estimate for each estimation procedure is given by

$$\hat{B}\left(\xi_m^d\right) = 100 \times \left[\overline{\xi}_m^d \Big/ \xi_p^d\right] - 1$$

where ξ_{ms}^d is the estimated decile d estimate from method m in sample s, $\overline{\xi}_m^d$ is the average over the 5,000 samples, and ξ_p^d is the population decile (evaluation measures for estimates and variance estimates for domain i (Northeast, Midwest, South, and West in the simulation study presented in Section 4) are available upon request to the authors, but are omitted for brevity).

The empirical mean squared error (MSE) of each decile estimate for each estimation method is given by

$$\text{MSE}\left(\xi_m^d\right) = \frac{1}{5{,}000}\sum_{s=1}^{5{,}000}\left(\xi_{ms}^d - \xi_p^d\right)^2 = \frac{1}{5{,}000}\sum_{s=1}^{5{,}000}\left(\xi_{ms}^d - \overline{\xi}_m^d\right)^2 + \left(\overline{\xi}_m^d - \xi_p^d\right)^2$$

$$\text{MSE}\left(\xi_m^d\right) = \hat{\sigma}\left(\xi_m^d\right) + \hat{B}^2\left(\xi_m^d\right).$$

To assess the *variance estimation* properties of the estimation method for decile d over repeated samples, we computed the following statistics:

Relative bias of the variance $100 \times \left[\hat{\bar{v}}\left(\xi_m^d\right) \Big/ \text{MSE}\left(\xi_m^d\right)\right] - 1$ where $\hat{\bar{v}}\left(\xi_m^d\right)$ is the average *variance* estimate of decile d from method m over 5,000 samples *i.e.*, $\hat{\bar{v}}\left(\xi_m^d\right) = \hat{v}\left(\xi_{ms}^d\right) \Big/ 5{,}000$. In our case

study, the variance estimates in each sample $s, \hat{v}\left(\xi_{ms}^{d}\right)$, are modified half-sample replicate variance estimates described in Section 3.3.1.

Stability of the variance estimate $\sqrt{\sum_{s=1}^{5,000}\left(\hat{v}\left(\xi_{ms}^{d}\right) - \mathrm{MSE}\left(\xi_{m}^{d}\right)\right)^{2} \Big/ 5,000} \Big/ \mathrm{MSE}\left(\xi_{m}^{d}\right)$.

Coverage rates (CR) = the proportion of 90% confidence intervals for a given method that contain the true population decile ξ_{p}^{d}.

The stability of the variance is a measure of the variance of the variance estimates. Ideally, both the relative bias and stability measures should be near zero. Coverage rates demonstrate the combined effect of the estimate and variance estimate on inference.

3.3.3 Simulation study results

The following sections summarize our simulation study results, presenting in illustrative graphs (tables available upon request to the authors).

3.3.3.1 Estimation properties of each method

Figure 3.1 plots the relative biases of the national level decile estimates by estimation method in the log-normal population. Recall that unbiased estimates will have a relative bias of zero indicated by the grey horizontal asymptote on each figure. Caution is advised in visual comparisons of bias levels, as the two characterics' graphs may not be on the same scale.

The SD method produces the least biased decile estimates for both sales price and price per square foot. That said, the biases of the decile estimates for both characteristics obtained using the P95 and NB methods are trivial. The largest biases can be found at the 10th percentile and the 90th percentile, that is, near the tails of the distribution where the sample is expected to be less stable. Although the SD decile estimates are less biased than their P95 and NB counterparts, they are less precise. In general, the P95 deciles have the minimum MSE among the three competing methods, although in many cases, the differences between the P95 and NB MSEs are negligible.

Figure 3.2 plots the relative biases of the national level decile estimates by estimation method in the nonparametric population. The bias patterns for price per square foot follow the same patterns as above, as do the MSEs. However, the pattern of the bias and MSE of sales price is different. Here, the SD estimates are the least biased, but the largest bias occurs at the median (0.005). This is also true for the two interpolation methods, with the P95 and NB medians each having a relative positive bias of seven tenths of a percent. For the 50th and 60th deciles, the MSE of the P95 estimates is somewhat larger than the other corresponding estimates, reflecting the impact of this estimator.

Some of the biases from the nonparametric population are large enough to warrant concern, especially for the median estimate. That said, the log-normal population does appear to more closely mimic the true SOC data, so the nonparmetric results are not necessarily reflective of SOC's "reality." These results do reflect the impact of the constant bias term in the decile estimates caused by interpolation.

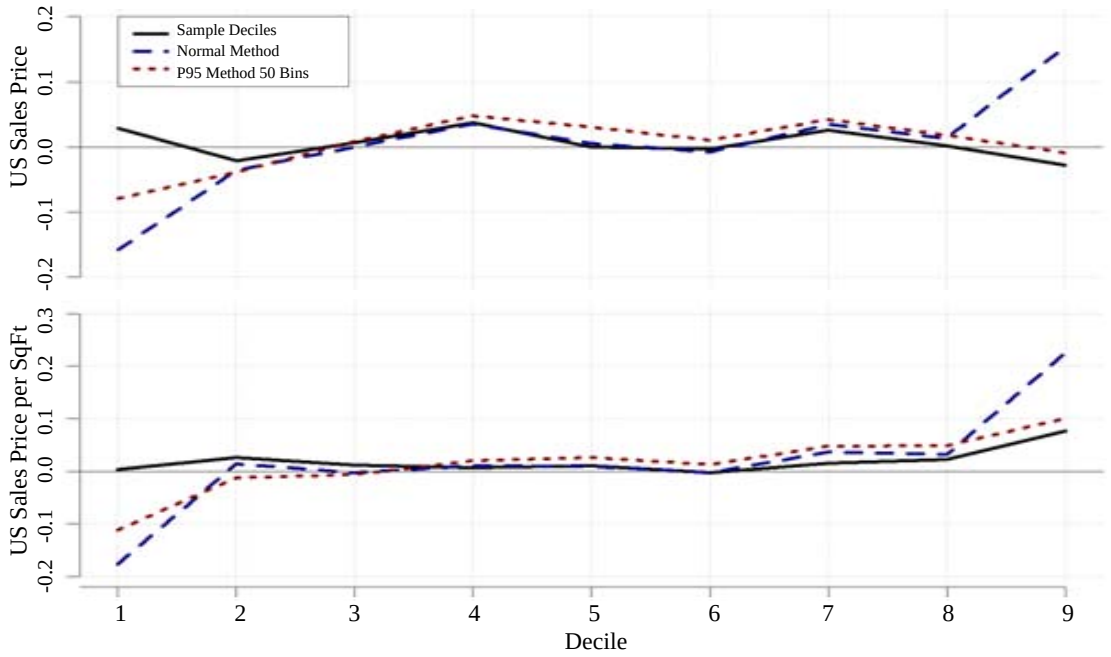

Figure 3.1 Relative bias of sales price and sales price per square foot estimates from the log-normal population (Expressed in percentages)

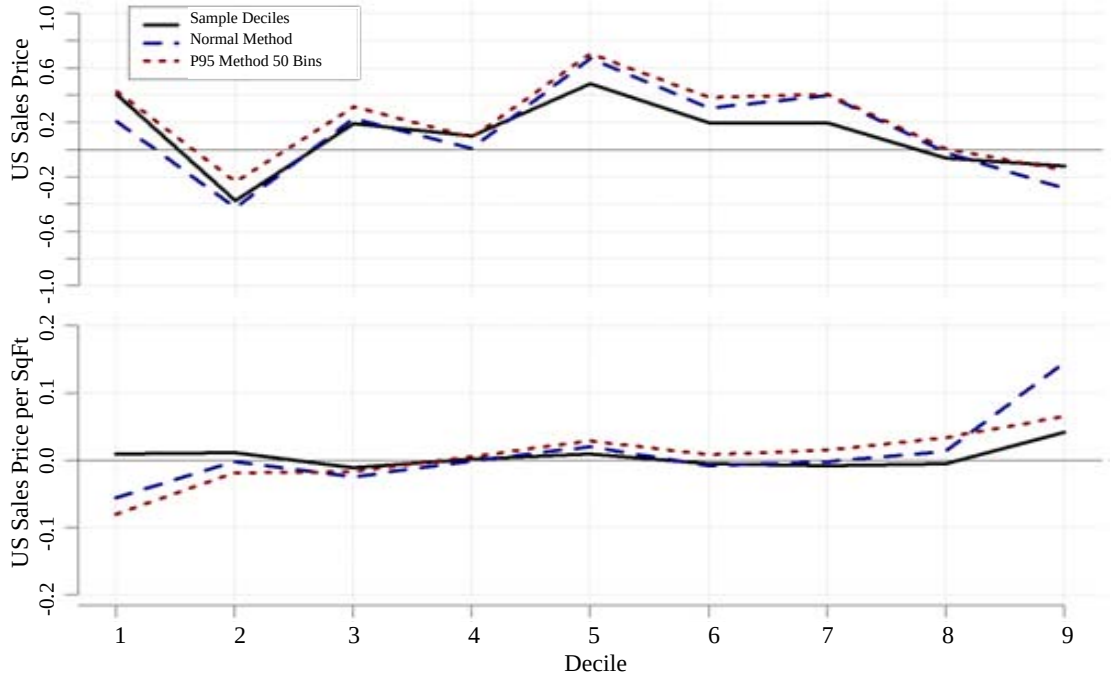

Figure 3.2 Relative bias of the sales price and sales price per square foot estimates from the nonparametric population (Expressed in percentages)

Overall, the MSEs follow similar patterns for both populations and both characteristics. Minimal MSEs can be found around the center deciles. The lower tail deciles have slightly higher MSEs and the upper tail decile MSEs increase at a rapid rate.

3.3.3.2 Variance estimation properties of each method (Given MHS replication)

As demonstrated in Figure 3.3, *all* relative variance biases of the MHS variance estimates in the log-normal population are positive regardless of estimator, and the SD variance estimates are the most biased. The P95 and NB methods have similar relative biases for all characteristics, with all being less biased than those obtained via the SD method. Overall, the P95 variance estimates are the least biased. Notice that *all* of the variance estimates for sales price are positively biased with all estimators, to save space, the y-axis begins at 5%. The same cautions about visual comparisons stated in Section 3.3.3.1 apply to the figures in this section.

The SD variance estimates for both characteristics are by far the least stable. This result is expected, since the interpolation variance estimates benefit from smoothing. Of the two interpolation methods, the P95 method yields more stable variance estimates for all deciles except for a handful of the upper tail deciles. The more stable variances in the NB upper deciles are likely a result of using properties of a normal distribution to obtain equal percentages of the sample in each bin.

None of the three considered methods yielded 90% coverage rates for sales price or price per square foot (Figure 3.4). Most of the coverage rates are slightly anti-conservative (below the 90% horizontal asymptote), and no decile estimation method appears to exhibit superior coverage properties over the others.

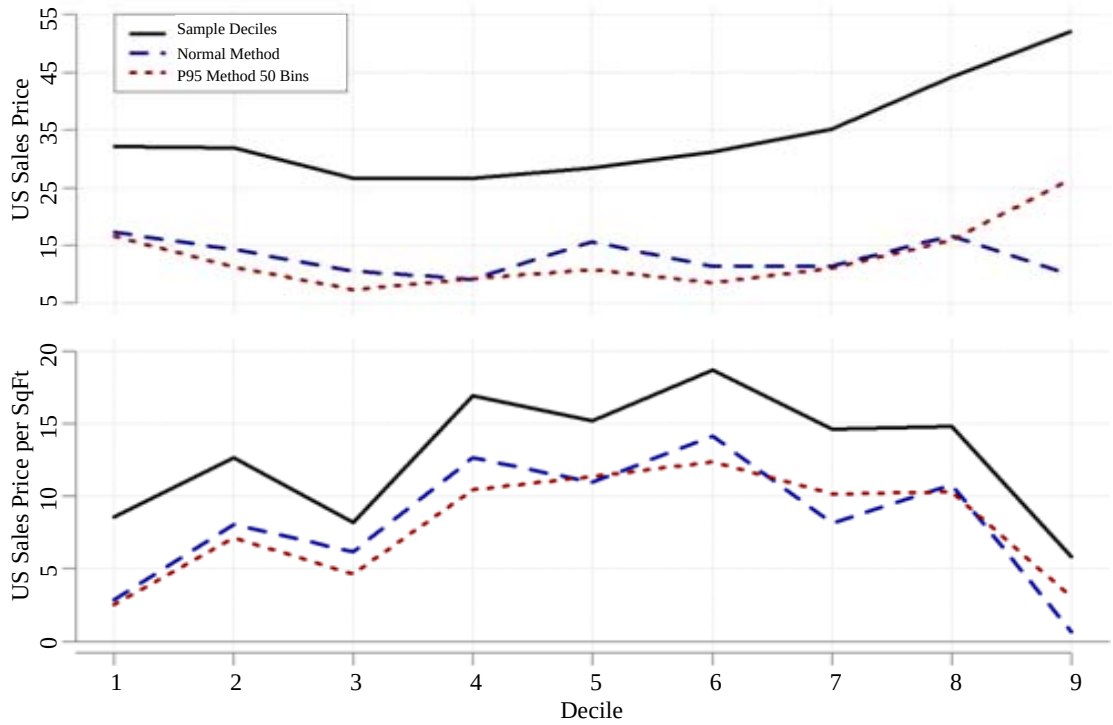

Figure 3.3 Relative bias of the variance for sales price and sales price per square foot from the log-normal population (Expressed in percentages)

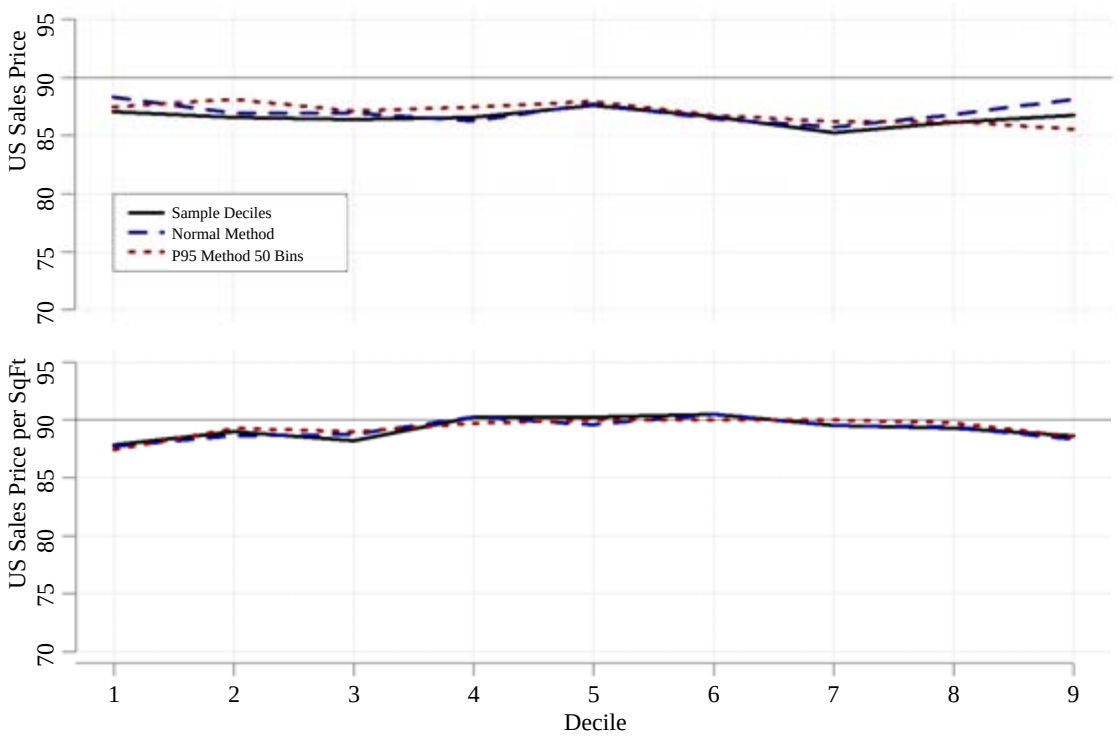

Figure 3.4 Coverage rates for sales price and sales price per square foot (Log-normal population)

Figure 3.5 illustrates the relative biases of the MHS variances obtained from the nonparametric population where both the P95 and NB interpolation methods are generally much smaller than their SD counterparts, and the P95 methods tend to produce less biased variance estimates for both characteristics. The relative bias of sales price for the nonparametric population does not follow the same pattern as the log-normal characteristics. The SD's relative biases are always positive and higher than the two interpolation methods. The two interpolation methods produced different results across populations of sales price. The nonparametric population contains many negative relative biases as opposed to all positive biases. The relative biases for sales price per square foot does follow the same pattern as the log-normal population, with large positive biases for the SD method, and lower similar positive biases for the two interpolation methods.

The stability of the nonparametric variance estimates matched up well with those obtained from the log-normal population, except for a few differences with sales price decile estimates. Sales price stability estimates for the SD method are still always larger than the other two interpolation methods, but follow a more erratic pattern. The NB method has a large stability estimate for the 40^{th} decile, which does not follow the expected trend.

The coverage rates for sales price per square foot follow the same pattern as in the log-normal population (Figure 3.6). However, the pattern of the sales price rates is more variable.

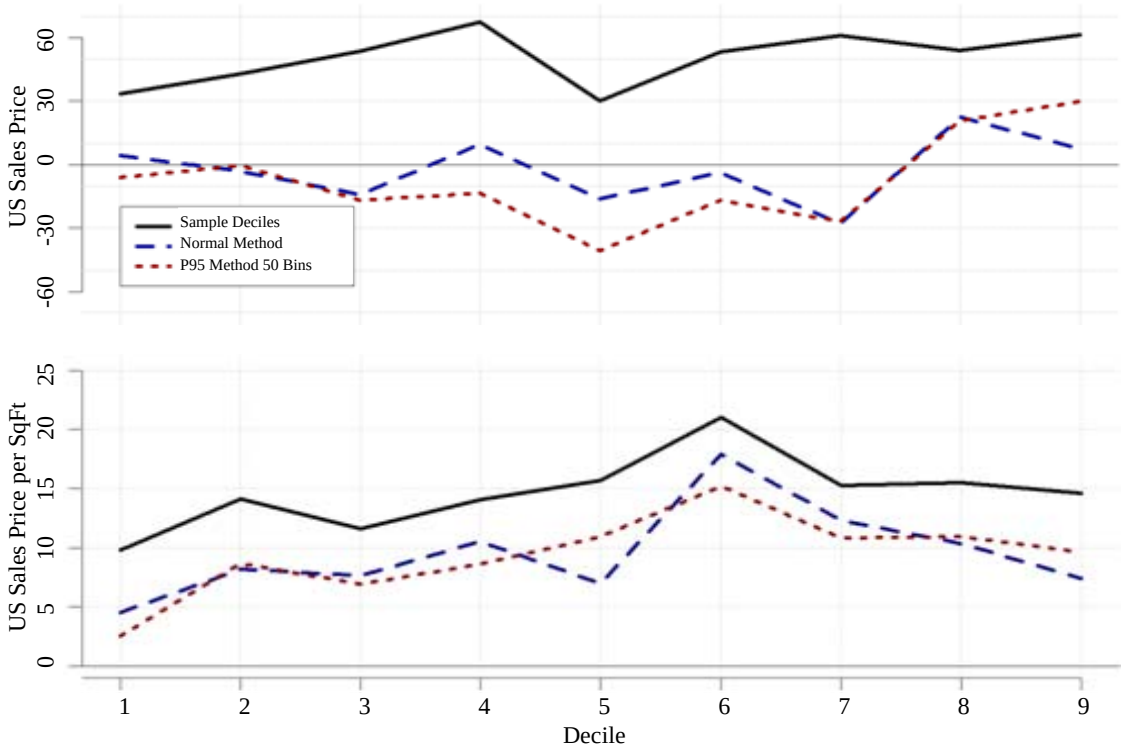

Figure 3.5 Relative bias of the variance of sales price and sales price per square foot (Nonparametric population)

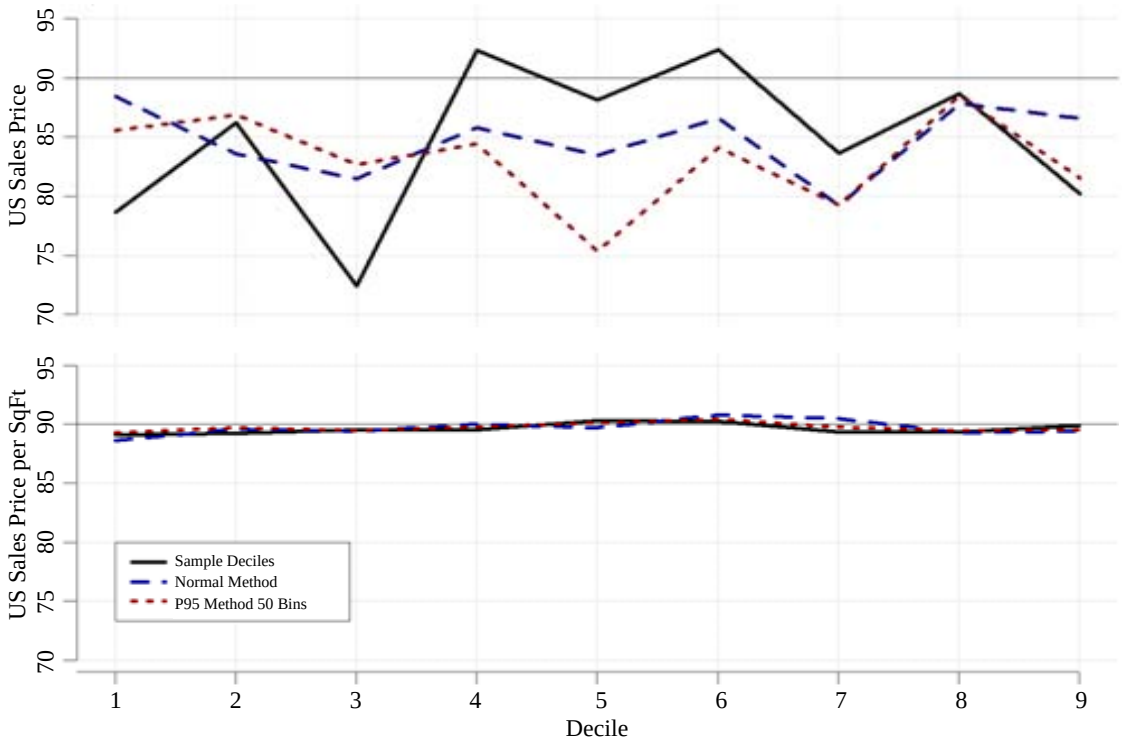

Figure 3.6 Coverage rates of sales price and sales price per square foot (Nonparametric population)

3.3.3.3 Additional simulations to assess bin size effects

Overall, the statistical properties of the P95 estimates and variance estimates obtained from the log-normal population (both characteristics) and from the nonparametric population for sales price per square foot are quite promising. However, none of the considered methods have nearly as solid properties for sales price in the nonparametric population. This is troubling, despite the previous caveats about the nonparametric population modeling.

In the nonparametric population, the sales price bins near the median contained more observations than bins in the distribution's tail [Note: this is true for both the P95 and NB methods.] These large bins can over-smooth the distribution, resulting in very stable estimates. The over-smoothing manifests itself in the replication variance method as underestimation due to lack of variability between replicate estimates for the "middle" deciles. Conversely, as expected, the SD method produces unstable estimates throughout and consequently overestimates the variance (positive bias).

The purpose of transforming the data before binning is to obtain uniform distributions within bins. For sales price, neither transformation achieves a uniform distribution within the bins, resulting in non-negligible interpolation bias, in turn affecting the MSE estimates.

To better understand how the estimation procedure affects the variance estimation procedure, recall that our variance estimates are evaluated with respect to the mean squared error (MSE) obtained under the estimation procedure. The SD procedure yields essentially unbiased estimates, but the trade-off is a large and unstable variance. Using interpolation reduces the sampling variance and improves its stability, but can substantively increase the bias squared term of the MSE.

Ultimately, the P95 method had the most promising results for most characteristics. However, there were still several concerns about the bias of the median estimate for nonparametric sales price. To address these concerns, we conducted additional simulations on both populations and characteristics, using the P95 method with 50 bins, 75 bins, and 100 bins.

In most cases, using 75 bins with the P95 method generally reduces the estimate bias and MSE without detrimentally affecting the variance estimate properties. This is definitely a balancing act. As the number of bins increases, the corresponding evaluation statistics begin to mimic those obtained with the SD method. This improves the interpolated estimates in the cases where the SD deciles had better statistical properties. However, increasing the number of bins has a detrimental effect on the statistical properties of the decile estimates and variance estimates when the P95 method with 50 bins yielded less biased estimates or more stable variance estimates.

4 Conclusion

The fundamental finding from Thompson and Sigman (2000) was that interpolation methods can be used to produce stable *median* estimates for samples from positively skewed populations, but the effectiveness of the interpolation was highly dependent on both the width of the bins and their location in the sample. Their primary contribution was to develop a data-dependent binning approach that used each individual estimation cell distribution.

Our approach to determining a decile estimation method for complex samples from a positively skewed population builds on these earlier findings, recognizing both that data-dependent binning is a necessity and that the binning method selected must account for a positively skewed distribution to facilitate the complete set of decile estimates. We considered three interpolation methods, each of which took a different approach to resolving the sparse data problem at the 90^{th} decile posed by the skewed distributions. Our empirical analysis showed that all of the studied approaches yielded complete sets of decile estimates with reasonable statistical properties, at least for the Survey of Construction. However, the properties of the corresponding MHS variance estimates were not as good and exhibited different patterns. At the U.S. level, our simulation results demonstrate consistently good statistical properties for decile estimation and variance estimation using the P95 transformation and 75 bins in one simulated population in terms of estimate bias, MSE, bias of the variance estimates, and stability, while rarely achieving a 90% coverage rate.

Of course, it is much more challenging to estimate a complete set of deciles than a single median, especially from positively skewed distributions. However, our recommended method appears to work quite well for most decile estimates and could certainly be modified to produce viable quartile estimates if the production decile estimates prove too unstable. In the meantime, the SOC program has decided to implement the P95 interpolation method and produce complete sets of decile estimates for selected annual characteristics in future reports.

Although we believe that our findings can be extended to other survey designs, we recognize that our research is conducted under extremely restrictive conditions, namely multi-stage cluster sampling from a highly skewed population, with a two-PSU per stratum design at the first stage. In other applications, interpolation with data-dependent bins could be combined with the variance estimator proposed in the 1952 Woodruff paper, as suggested by J.N.K. Rao. For surveys that are not well suited to BRR or MHS replication that publish decile estimates, our data-dependent binning and interpolation approach could be used in conjunction with a bootstrap replication method such as the Rao-Wu bootstrap (Rao and Wu 1988).

Acknowledgements

This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, or technical issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors acknowledge Erica Filipek, Bonnie Kegan, Amy Newman-Smith for their valuable contributions to this research project. In addition, we thank Wan-Ying Chang, Laura Bechtel, Xijian Liu, J.N.K. Rao, the Associate Editor, and two anonymous referees for their helpful comments of earlier drafts of this manuscript.

References

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

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

Lienhard, S. (2004). Multivariate Lognormal Simulation with Correlation. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=6426&objectType=File.

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.

Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

Rao, J.N.K., and Wu, C.F.J. (1988). Re-sampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Steel, P., and Fay, R.W. (1995). Variance estimation for finite populations with imputed data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Thompson, J.R. (2000). Simulation: A Modeler's Approach. New York: John Wiley & Sons, Inc., 87-110.

Thompson, K.J. (1998). Evaluation of Modified Half Sample Replication for Estimating Variances for the Survey of Construction (SOC). Technical Report #ESM9801, available upon request to the Office of Statistical Methods and Research for Economic Programs from the U.S. Census Bureau.

Thompson, K.J., and Sigman, R.S. (2000). Estimation and replicate variance estimation of median sales prices of sold houses. *Survey Methodology*, 26, 2, 153-162.

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