Empirical Comparison of Design- and Model-Based Estimators of Quantile Proportion in a Positively Skewed Distribution

Nancy E. Añez-Tandang and Zita VJ. Albacea*

Received January, 2004; Revised July, 2004

ABSTRACT

The study aims to compare the statistical properties of the design- and model-based estimators of quantile proportion as applied in a positively skewed distribution such as income data. Several estimators, namely: traditional design-based estimator, model-based estimator proposed by Chambers and Dunstan (1986) and difference design-based estimator reported by Rao, Kovar and Mantel (1990) were empirically compared and evaluated using the National Capital Region (NCR) data from the 1994 Family Income and Expenditure Survey (FIES). Treating the NCR data set as the population, random samples are generated following the sampling design of the FIES. In each of the generated samples, the design- and model-based estimates were computed and compared. In the model-based estimation procedure, a simple linear regression was assumed with household's total expenditure as predictor and household's total income as dependent variable. The relative mean errors and relative root mean square errors of the estimates were used as bases for the comparison of the estimates. The results showed that one could consider the use of the traditional design-based estimator, for the simplicity of the estimation procedure, in estimating the proportion of households in NCR with income value less than or equal to the 10% quantile income value of the population, while model-based estimator is better to use for the 30% and 50% quantiles.

Key Words: quantile proportion, design-based estimator, model-based estimator, auxiliary information, positively skewed distribution

1. INTRODUCTION

Estimation of distribution functions is often an important objective in survey practice. This is particularly so when the primary aim of the survey is to identify subgroups in the population whose values for a particular variable are below or above the population averages. In this regard, several studies were done to provide guidance for statisticians on efficient methods for estimating population distribution functions and associated quantiles from survey data. These literatures gave emphasis on estimation of means, totals and ratios defined by the survey variable. Chambers and Dunstan in 1986 proposed and discussed the statistical procedure of estimating distribution function in a model-based approach. They supported the proposition that many sampling problems can be analyzed usefully and realistically as prediction problems under appropriate superpopulation models. This is in contrast to the tendency, which has been dominant for years, to insist that the statistical inference theory should be derived not from models, but from the probability distribution created by the sampler's choice of a random sampling plan. In 1990, on the other hand, Rao, Kovar and Mantel reported an improved design-based estimator of distribution function to show the advantages of the design-based over the model-based estimator. Dorfman in 1994 compared

* Assistant and Associate Professor of Statistics, respectively, Institute of Statistics, University of the Philippines, Los Baños (UPLB), Los Baños, Laguna, Philippines. Email of second author: zvja@instat.uplb.edu.ph

the two reported estimators theoretically and empirically using the Australian agricultural data to further study the advantages and disadvantages of these estimators.

With the same objective, this study adopted the estimators developed by Chambers and Dunstan in 1986, and Rao, Kovar and Mantel in 1990 to estimate the distribution function or quantile defined as the proportion of units in the population with values less than or equal to a chosen population value. Further, the study considers the application of the proposed estimators in a positively skewed distribution. A distribution is said to be positively skewed if the scores tend to cluster toward the lower end of the scale with fewer scores at the upper end of the scale. Since income variable is one best example that exhibits a positively skewed distribution, it was considered as the characteristic of interest in this study. Hence, the estimators were applied to income data of the Philippines, in particular the National Capital Region income data, to evaluate the statistical properties of the design- and model-based estimators in their application to a positively skewed distribution.

2. METHODOLOGY

2.1 Different Estimators of Quantile Proportion

The function to be estimated is given by

$$F_p(t) = \mathrm{N}^{-1} \sum_{i=1}^{N} \mathrm{I}\{Y_j \le t\} \qquad (1)$$

Where

$\mathrm{I}\{\ \}$ (Standard indicator function)

$$= \begin{cases} 1 \ , \text{ if the condition in parenthesis is true} \\ 0 \ , \text{ if the condition in parenthesis is false} \end{cases}$$

N = population size;
Y = survey variable; and
t = population income value at a given percentile.

The study, aiming to estimate the quantile proportions of households in the National Capital Region (NCR) with total income less than or equal to a population quantile income, adopted the function given above as its parameter of interest to be estimated. The household's total income will serve as Y in the function and t will adopt the population value corresponding to the set quantile. The value of t is equal to Php60,450 for 10% quantile, Php89,390 for 30% quantile and Php125,480 for 50% quantile based on the 1994 FIES-NCR data set used in this study. In positively skewed distribution like income variable, the proportion distribution less than the 10th, 30th and 50th quantiles are usually of interest.

a. Traditional design-based estimator of quantile proportion

The design-based estimator of $F_p(t)$ is defined as

$$\hat{F}_T(t) = \mathrm{N}^{-1} \sum_{j=1}^{n} \pi_j^{-1}\ \mathrm{I}\{Y_j \le t\} \qquad (2)$$

Where Y = household total income;
n = sample size;
N = population size; and
π_j = probability that unit j is included in the sample.

This estimator is said to be biased in a particular sample but unbiased in overall samples. It is design-unbiased for $F_p(t)$ defined in equation (1) under any sampling scheme provided that $\sum \pi_j^{-1} = N$. It is also important to note that this estimator is advantageous for small sample sizes, in terms of having small mean square error (Dorfman AH, 1993).

b. Model-based estimator of quantile proportion

Chambers and Dunstan (1986) reported that the customary design-based estimator of a population distribution function does not make use of auxiliary population information at the estimation stage. To include population information on estimation procedure, Chambers and Dunstan proposed the use of model-based approach. A model-based procedure is described as one that allows relevant auxiliary information to be explicitly used in the estimation of a finite distribution function. This auxiliary variable, denoted as X, must be known for all elements of the population and assume to be linearly related to the survey variable, denoted as Y. In this study, the superpopulation model for Y that will be assumed to follow a linear regression model with heteroscedastic error is given by,

$$Y_i = X_i'\beta + \upsilon_i e_i \qquad \mathrm{i} = 1,2\ldots\mathrm{N} \qquad (3)$$

Where β is the unknown parameter with known υ(.).

With $G_i(t) = \Pr\{Y_i \le t\}$ and G(.) be the distribution function for e_i, the model-based estimator which Chambers and Dunstan defined in terms of model (3) is given by:

$$\hat{F}_{CD}(t) = \frac{1}{N}\left[\sum_{j=1}^{n} \mathrm{I}\{Y_j \le t\} + \frac{1}{n}\sum_{i=1}^{N-n}\ \sum_{j=1}^{n} \mathrm{I}\left\{\frac{(Y_j - X_j'\hat{\beta})}{\upsilon_j} \le \frac{(t - X_i'\hat{\beta})}{\upsilon_i}\right\}\right] \qquad (4)$$

where X is the auxiliary variable identified as household's total expenditure in this study, and $\hat{\beta}$ is the weighted least square estimator of β in model (2).

$\hat{F}_{CD}(t)$ will tend to be positively biased when υ(.) overstates the true variance,, and negatively biased when υ(.) understates the variance.

c. Design-based difference estimator of quantile proportion

In 1990, Rao, Kovar and Mantel (RKM) described several calibration-type design-based estimators. Their result showed that the most successful under model (3) is the design-based difference estimator. This estimator is defined as

$$\hat{F}_{RKM}(t) = N^{-1}\left[\sum_{j=1}^{n} \pi_j^{-1}\, \mathrm{I}\{Y_i \le t\} + \sum_{i=1}^{n} \tilde{G}_i - \sum_{j=1}^{n} \pi_j^{-1} \tilde{G}_{jc}\right] \qquad (5)$$

Where

$$\tilde{G}_i(t) = \left(\sum_{j=1}^{n} \frac{1}{\pi_j}\right)^{-1} \left[\sum_{j=1}^{n} \pi_j^{-1} I\left\{\frac{(Y_j - \hat{R}X_j)}{\upsilon_j} \le \frac{(t - \hat{R}X_i)}{\upsilon_i}\right\}\right] \text{ and } \qquad (6)$$

$$\tilde{G}_{jc}(t) = \left(\sum_{k\in s} \frac{\pi_j}{\pi_{jk}}\right)^{-1} \left[\sum_{k\in s} \left(\frac{\pi_j}{\pi_{jk}}\right) I\left\{\frac{(Y_k - \hat{R}X_j)}{\upsilon_j} \le \frac{(t - \hat{R}X_j)}{\upsilon_j}\right\}\right] \qquad (7)$$

with $\hat{R} = \left(\sum_{j=1}^{n} \frac{Yj}{\pi_j}\right)\left(\sum_{j=1}^{n} \frac{Xj}{\pi_j}\right)^{-1}$, the customary design-consistent ratio estimator of the population means of household's total income (Y) and total expenditure (X).

According to RKM, this estimator is bias-robust against model failure, where "bias" refers to the true model, not bias under repeated sampling. It could lead to sustainable gain in efficiency over $\hat{F}_T(t)$ defined in equation (2) when Y is approximately proportional to X. A key argument of RKM in adopting this approach to sample survey inference is that its design-unbiasedness guarantees 'robustness' of one's inference to misspecification of the model. It can also be used to more complex models involving multiple auxiliary variables.

2.2 Evaluation of the different estimators

Empirical analysis is done to evaluate the performance of the three estimators described in the previous section. Due to limitation of computing resources, only a large number of possible samples instead of all possible samples were generated. Simulation was done using Statistical Analysis System (SAS) language. In generating the samples, a random start was provided by the SAS system. A stratified random sample was drawn from the population with sample size ranging from 178 to 305. Five hundred of such stratified samples were randomly generated. For each sample generated, estimates of $F_p(t)$ for 10 percent, 30 percent, and 50 percent quantiles were calculated using the traditional design-based, model-based, and difference design-based estimators. The different estimators are then computed and evaluated by computing their relative mean error (RME), absolute bias and relative root mean square error (RRMSE).

With 500 generated samples, RME, which measures the relative biasedness of the estimator, is computed as

$$RME = \frac{1}{F_p(t)} \left[\sum_{q=1}^{500} \frac{F_{p*}^{q}(t) - F_p(t)}{500} \right] \qquad (8)$$

while RRMSE, which is used to measure efficiency of the estimators, is given by

$$RRMSE = \frac{1}{F_p(t)} \left[\sum_{q=1}^{500} \frac{(F_{p*}^{q}(t) - F_p(t))^2}{500} \right]^{1/2} \qquad (9)$$

with $F_p(t)$ = population distribution function;
$F_{p*}^{q}(t)$ = value of the estimate for the qth run; and
$t=F^{-1}_{p}(\alpha)$.

Estimator with lowest RRMSE is considered to be the 'best' estimator in estimating the quantiles of the income distribution. In addition, the estimator with absolute bias of less than 0.1 is said to have a negligible relative bias. Gain in efficiency of the estimators relative to the traditional design-based was also computed by dividing the RRMSE of an estimator to the RRMSE of the traditional design-based estimator.

2.3 Limitation of the Study

This study considers income data, with annual expenditure as the auxiliary variable, to evaluate the performance of the proposed estimators as applied to positively skewed distribution. Income of those who are residing in the National Capital Region (NCR) was the one considered in this study. The design- and model-based quantile proportion estimators were empirically analyzed and compared. Using the survey variable income in its application to poverty measurement, one may try to compare these estimators to usual estimates of poverty incidence.

Only simple regression model was considered in the estimation. While a best estimator may exist for an assumed superpopulation model, other models consistent with the observed sample may lead to a different estimator so that a unique best estimator is still not available. Thus, it is suggested that studies on rival models should also be conducted.

Further, the study focused only on the low quantiles of positively skewed distribution. It is also recommended that the behavior and potential of these estimators be evaluated in the other quantiles with varying sample size. Such concerns are also of great interest to the authors, but due to limited resources, the study was not able to work on these concerns.

3. RESULTS AND DISCUSSION

3.1 The Population data

Three thousand eight hundred eighty six (3,886) households from the National Capital Region were considered in the 1994 Family Income Expenditure Survey (FIES). Household's annual income is considered as the survey variable in this study and it is measured as total annual household income in pesos at current prices. Table 1 below shows the income of those residing in NCR corresponding to a given quantile. There are 389 households in NCR whose annual income is less than or equal to 60,450 pesos, 50% of those in NCR has annual income less than or equal to 125, 480 pesos, and 90% of those in NCR or 3,498 households in NCR has annual income less than or equal to 343, 190 pesos.

Table 1. Income Quantile Distribution Of Households in NCR, 1994

QUANTILE	ANNUAL INCOME (PESOS)
10 percent	60, 450
20 percent	75, 747
30 percent	89, 390
40 percent	106, 672
50 percent	125, 480
60 percent	149, 030
70 percent	180, 473
80 percent	229, 500
90 percent	343, 190

Moreover, household's annual income in NCR tends to cluster at Php219,975.85 (refer to Table 2). The median value is Php125,480. The income distribution is highly positively skewed with skewness coefficient of 20.54 and a standard deviation of Php534,387.97. The minimum income observed is Php12,702 while the maximum income is Php21,635,333. The income data in NCR exhibits large variability with a coefficient of variation of 242.93%.

Table 2. Summary Statistics Of Household's Annual Income (in pesos) in NCR, 1994. N=3,886

Mean	219, 975.85
Median	125, 480.00
Standard Deviation	534, 387.97
Minimum	12, 702.00
Maximum	21, 635, 333.00
Coefficient of Variation(%)	242.93
Skewness	20.54

Household's annual expenditure and household size are possible auxiliary variables that are related to household's annual income. Thus, they were correlated with household's annual income to identify an auxiliary variable known to all households in NCR that will be used in the model-based and difference design-based estimation procedures. Based on the correlation matrix given in Table 3, there is a strong positive linear relationship between annual income and expenditure with a correlation coefficient of 0.9156, while there is a positive weak linear relationship between annual income and number of household members with a correlation coefficient of 0.0291. Since annual expenditure is the one which is highly correlated with annual income, it was chosen as the auxiliary variable for the estimators in need of this information.

Table 3. Correlation Matrix Of Annual Income, Annual Expenditure And Household Size in NCR, 1994. N=3886

	EXPENDITURE	HOUSEHOLD SIZE
INCOME	0.9156	0.0291
EXPENDITURE		0.0379

Furthermore, characteristics of households in NCR whose income belong to the lower 10%, 30% and 50% income quantiles were determined. As stated earlier, there are 389 households in NCR whose annual income is less than or equal to Php60,450. These households can be considered as low-income households. Based on Table 4, this group of households in NCR has an average income of Php48,934 and a standard deviation of Php9,055. In this group, household's annual income is strongly related with annual expenditure with a correlation coefficient of 0.6244, and weakly related with the household size with a correlation coefficient of 0.2496. (Refer to Table 5).

Table 4. Summary Statistics Of Household's Annual Income (in pesos) in the 10%, 30% and 50% Quantiles of NCR, 1994

	QUANTILES		
	10%	30%	50%
Mean	48, 934	66, 509	82, 729
Standard Deviation	9, 055	15, 121	24, 010
Minimum	12, 702	12, 702	12, 702
Maximum	60, 450	89, 390	125, 460
Coefficient of Variation(%)	18.50	22.76	29.02
Number of Household	389	1, 166	1, 943

Table 5. Correlation Matrix Of Annual Income, Annual Expenditure And Household Size in the 10% quantile of NCR, 1994. N=389

	EXPENDITURE	HOUSEHOLD SIZE
INCOME	0.6244	0.2496
EXPENDITURE		0.3242

There are 1,166 and 1,943 households in NCR whose annual income are less than or equal to Php89,390 and Php125,480, respectively. In 30% quantile group, the average annual income is Php66,509 while, in the 50% quantile, the households' income tend to cluster at Php82,729. There is an increase in the variability of income as the number of households in a group increases. Also, as shown in Tables 6 and 7, the degree of linear relationship between household's annual income and annual expenditure increases as the quantile being considered gets larger, while the degree of relationship between annual income and household size gets weaker.

Table 6. Correlation Matrix Of Annual Income, Annual Expenditure And Household Size in the 30% Quantile of NCR, 1994. N=1166

	EXPENDITURE	HOUSEHOLD SIZE
INCOME	0.6957	0.2293
EXPENDITURE		0.2902

Table 7. Correlation Matrix Of Annual Income, Annual Expenditure And Household Size in the 50% quantile of NCR, 1994. N=1943

	EXPENDITURE	HOUSEHOLD SIZE
INCOME	0.7569	0.2052
EXPENDITURE		0.2481

3.2 Estimates of the 10% Quantile Proportion

The traditional design-based estimates tend to cluster at 0.0999 and deviate from this, on the average, by 0.0273 (see Table 8). The minimum estimate computed is 0.0319, while 0.1974 is the maximum estimate. The simulated distribution of the traditional design-based estimator is reasonably bell shaped with coefficient of skewness equal to 0.270.

Mean of the proportional estimates using model-based estimator in this quantile is 0.1394 with a standard deviation of 0.0274. The mean overestimated the population proportion of 0.10. The derived minimum and maximum estimates are 0.0935 and 0.2770, respectively. The distribution of the estimates is positively skewed, with coefficient of skewness equal to 1.499, implying that there exist few extremely high model-based estimates in the distribution.

The difference design-based estimator, on the other hand, has an average estimate of 0.0910 and a standard deviation of 0.0196. The estimates have a minimum value of 0.0144, and a maximum value of 0.1846. It also has a reasonably bell-shaped distribution with a coefficient of skewness equal to 0.168.

Among the three estimators, the traditional design-based estimates exhibit large variability with coefficient of variation equal to 27.28%.

Table 8. Summary Statistics Of The Estimates Derived Using Traditional Design-Based, Model-Based, And Difference Design-Based Estimators for t=P60, 450. M=500

	ESTIMATES		
	Traditional Design-based	Model-based	Difference Design-Based
Mean	0.0999	0.1394	0.0910
Standard Deviation	0.0273	0.0274	0.0196
Minimum	0.0319	0.0935	0.0114
Maximum	0.1974	0.2770	0.1846
Coefficient of Variation(%)	27.280	19.670	21.570
Skewness	0.2700	1.4990	0.1680

The difference design-based estimator ($\hat{F}_{RKM}(t)$) in the 10% quantile has a smaller relative bias than the model-based estimator ($\hat{F}_{CD}(t)$), while the relative bias of the traditional design estimator ($\hat{F}_{T}(t)$) is almost negligible. (Refer to Table 9). This is supported by the absolute biases of the three estimators. The absolute bias of the traditional design-based estimators, which is 0.0037, is less than 0.1. Hence, in this quantile, it can be said that only the relative bias of the traditional design-based estimator of the quantile proportion is negligible. Further, Figures 1 and 2 revealed that the relative bias of the traditional design-based estimator is almost equal to zero. Model-based estimator is positively biased, while, difference design-based estimator is negatively biased.

In terms of the relative root mean square error, the difference estimator has the smallest value among the three estimators (refer to Figure 3). Hence, it can be considered that the difference design-based estimator is more efficient than the traditional design-based estimator, which in turn is more efficient than the model-based estimator. Therefore, it can be said that the difference design-based estimator is the most efficient estimator in the 10% quantile. However, it is also important to note that the gain in efficiency of this estimator relative to the traditional estimator is 0.7919, indicating a small difference in efficiency between the two.

Table 9. Relative Mean Error, Aboslute Bias, Relative Root Mean Square Error and Gain In Efficiency Of Estimators Of Fp(T) For 10% Quantile Proportion. M=500.

	ESTIMATORS		
	$\hat{F}_T(t)$	$\hat{F}_{CD}(t)$	$\hat{F}_{RKM}(t)$
RME	-0.0001	0.3939	-0.0897
Absolute Bias	0.0037	1.4380	0.4592
RRMSE	0.2725	0.4798	0.2158
Gain in Efficiency Relative to $\hat{F}_T(t)$	1.0000	1.7607	0.7919

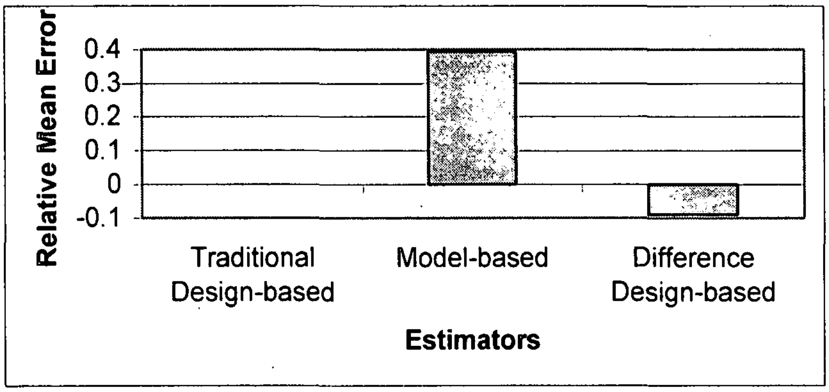

Figure 1. Relative Mean Error Of The Different Estimators Of The 10% Quantile Proportion

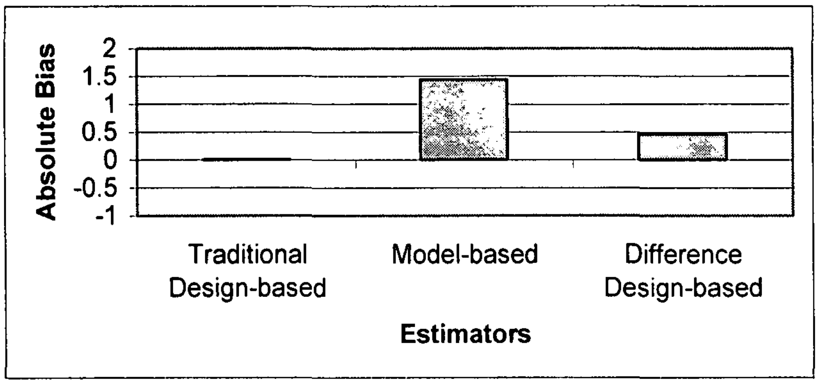

Figure 2. Absolute Bias Of The Different Estimators Of The 10% Quantile Proportion

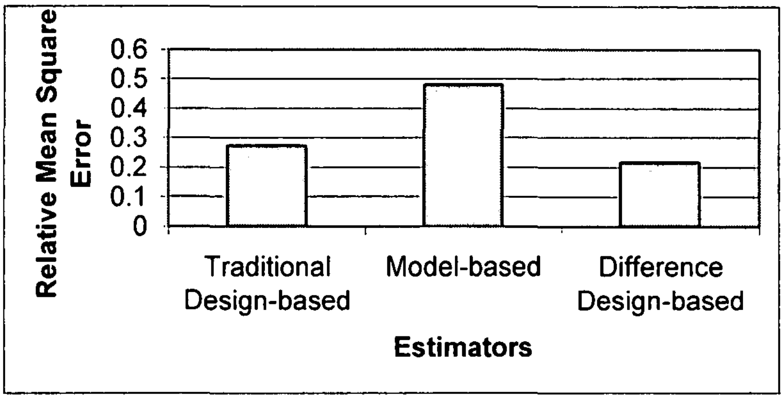

Figure 3. Relative Mean Square Error Of The Different Estimators Of The 10% Quantile Proportion.

3.3 Estimates of the 30% Quantile Proportion

Summary statistics of the estimates of 30% quantile proportion using the different estimators considered in this study are given in Table 10. The traditional design-based estimates' average is 0.3018 and deviate from this value, on the average, by 0.0408. The minimum and maximum estimates are 0.1925 and 0.4116, respectively. The mean of the estimates is almost equal to the median, with skewness coefficient equivalent to 0.009, indicating a symmetric distribution

Table 10. Summary Statistics Of The Estimates Derived Using Traditional Design-Based, Model-Based, And Difference Design-Based Estimators for t=P89, 390. M=500

	ESTIMATES		
	Traditional Design-based	Model-based	Difference Design-Based
Mean	0.3018	0.3003	0.2811
Standard Deviation	0.0408	0.0206	0.0246
Minimum	0.1925	0.2558	0.2173
Maximum	0.4116	0.4085	0.4257
Coefficient of Variation(%)	13.530	8.6800	8.7500
Skewness	0.0090	1.3120	0.8670

The model-based estimator is also unbiased with the estimates averaging at 0.3003. The estimates deviate from the mean by 0.0206, on the average. The computed minimum estimate is 0.2558, while 0.4085 is the maximum value. It has a positive skewed distribution with coefficient of skewness of 1.312.

The difference design-based estimates have a mean of 0.2811 and a standard deviation of 0.0246. The average of the estimates underestimated the population proportion of 0.30. It has a minimum value of 0.2173, and a maximum value of 0.4257. It also has few extremely high estimates in the distribution with coefficient of skewness equal to 0.867.

The estimates were found to be least dispersed in the model-based procedure with a coefficient of variation of 8.68% and most dispersed in the traditional design-based procedure with a coefficient of variation of 13.53%.

It can be seen from Table 11 that the values of the absolute bias of the traditional design-based ($\hat{F}_T(t)$) and the model-based ($\hat{F}_{CD}(t)$) estimators are less than 0.1. This is further shown in Figures 4 and 5. This implied that the relative biases of the traditional design-based and model-based estimators, which are 0.0058 and 0.0010, respectively, are negligible. The difference design-based, on the other hand, has a relative mean error of -0.0631. This indicates that this estimator is slightly negatively biased as further shown in its absolute bias of 0.7683.

Among the three estimators (see Figure 6), traditional design-based estimator has the largest value of RRMSE. Hence, it can be said that in the 30% quantile, it is the traditional estimator of the quantile proportion which is the least efficient. While the model-based estimator, having the smallest value of RRMSE, can be considered as the most efficient among the three estimators. Model-based estimator's gain in efficiency relative to the traditional design-based estimator supported the model-based estimator being the most efficient in this quantile.

Table 11. Relative Mean Error, Aboslute Bias, Relative Root Mean Square Error And Gain In Efficiency Of Estimators Of Fp(T) For 30% Quantile Proportion M=500

	ESTIMATORS		
	$\hat{F}_T(t)$	$\hat{F}_{CD}(t)$	$\hat{F}_{RKM}(t)$
RME	0.0058	0.0010	-0.0631
Absolute Bias	0.0441	0.0146	0.7683
RRMSE	0.1361	0.0688	0.1034
Gain in Efficiency relative to $\hat{F}_T(t)$	1.0000	0.5055	0.7597

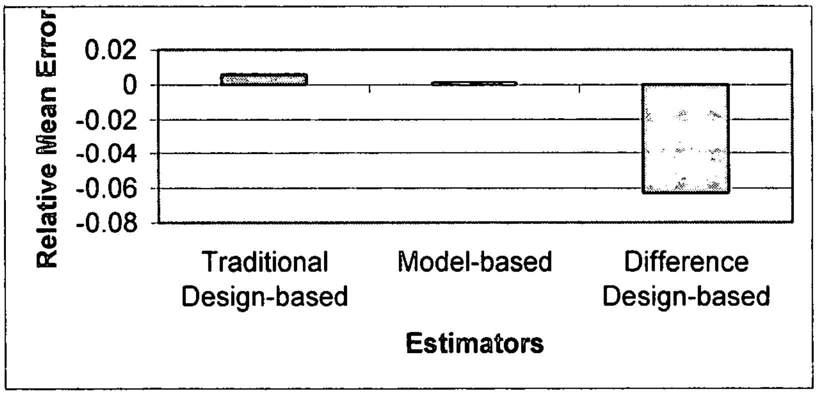

Figure 4. Relative Mean Error Of The Different Estimators Of The 30% Quantile Proportion

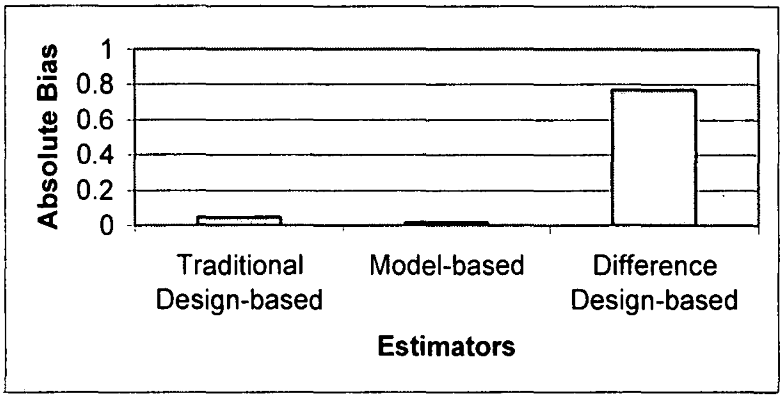

Figure 5. Absolute Bias Of The Different Estimators Of The 30% Quantile Proportion

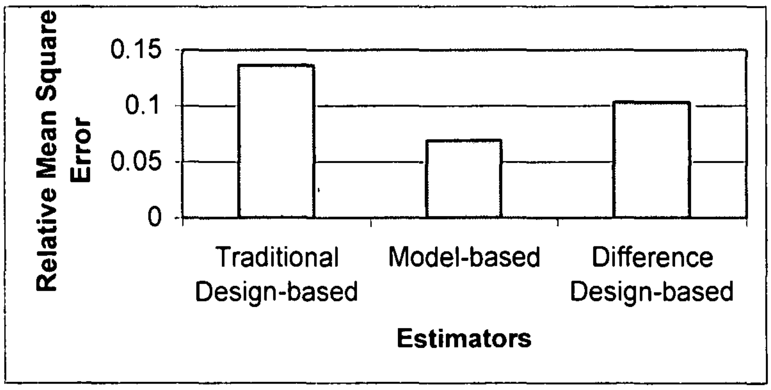

Figure 6. Relative Mean Square Error Of The Different Estimators Of The 30% Quantile Proportion.

3.4 Estimates of the 50% Quantile Proportion

Based on Table 12, the traditional design-based estimates of the 50% quantile proportion tend to cluster at 0.5015 and deviate from this value, on the average, by 0.0506. The minimum estimate amounted to 0.3619, while the maximum amounted to 0.6492. Again, it has a bell-shaped distribution with skewness coefficient of 0.024.

Table 12. Summary Statistics Of The Estimates Derived Using Traditional Design-Based, Model-Based, And Difference Design-Based Estimator S For T=P125, 480. M=500

	ESTIMATES		
	Traditional Design-based	Model-based	Difference Design-Based
Mean	0.5015	0.4855	0.4699
Standard Deviation	0.0506	0.0144	0.0230
Minimum	0.3619	0.4508	0.3923
Maximum	0.6492	0.5503	0.5793
Coefficient of Variation (%)	10.100	2.9700	4.8900
Skewness	0.0240	0.4440	0.5740

The proportion estimates using the model-based estimator has a mean equal to 0.4855 with a standard deviation of 0.0144. The minimum and maximum estimates are computed as 0.4508 and 0.5503, respectively. It has a skewness coefficient of 0.444 indicating a small difference between the mean and median estimates in favor of the mean.

The difference design-based estimates, on the other hand, have a mean of 0.4699 and deviate from this amount by 0.0230, on the average. The minimum estimate computed using the design-based estimator is 0.3923, while the maximum is 0.5793. The distribution is positively skewed with coefficient of skewness of 0.574.

In this quantile, the estimates were found to be least disperse in the model-based procedure with a coefficient of variation of 2.97%, and most dispersed in the traditional design-based procedure with a coefficient of variation of 10.1%.

The relative biases (see Table 13) of the three estimators in the 50% quantile are small, but only the traditional design-based estimator has a negligible relative bias with its absolute bias of 0.0296. The difference design-based estimates, also, underestimated the population proportion of 0.50. This is further shown in Figures 7 and 8.

In terms of relative root mean square error, the difference design-based estimator is slightly more efficient than traditional design-based estimator. Moreover, model-based estimator, having the smallest RRMSE (see Figure 9), is considered as the most efficient among the three estimators of this quantile proportion.

Table 13. Relative Mean Error, Aboslute Bias, Relative Root Mean Square Error And Gain In Efficiency Of Estimators Of Fp(t) For 50% Quantile Proportion. M=500

	ESTIMATORS		
	$\hat{F}_T(t)$	$\hat{F}_{CD}(t)$	$\hat{F}_{RKM}(t)$
RME	0.0030	0.0290	-0.0601
Absolute Bias	0.0296	1.0069	1.3087
RRMSE	0.1012	0.0408	0.0756
Gain in Efficiency relative to $\hat{F}_T(t)$	1.0000	0.4032	0.7470

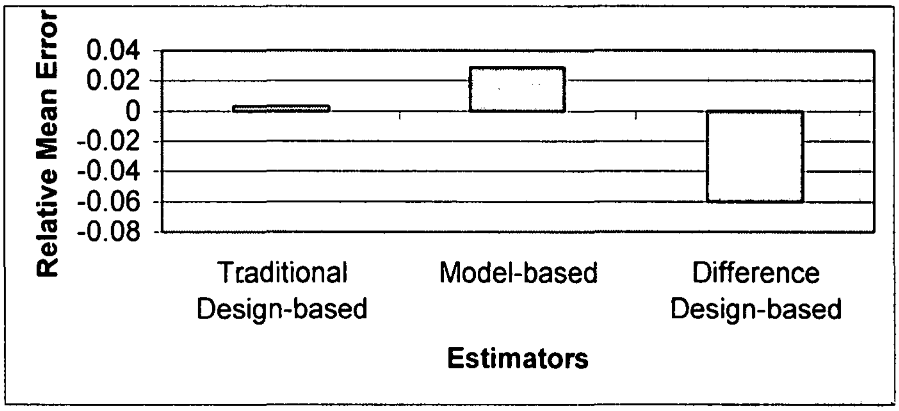

Figure 7. Relative Mean Error Of The Different Estimators Of The 50% Quantile Proportion

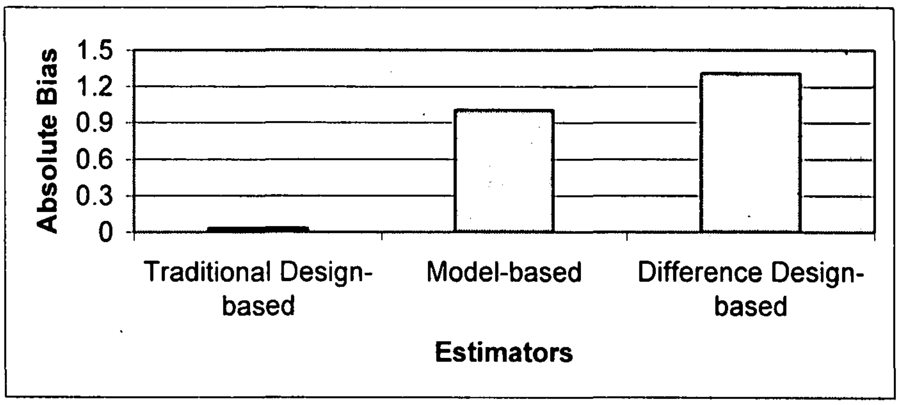

Figure 8. Absolute Bias Of The Different Estimators Of The 50% Quantile Proportion

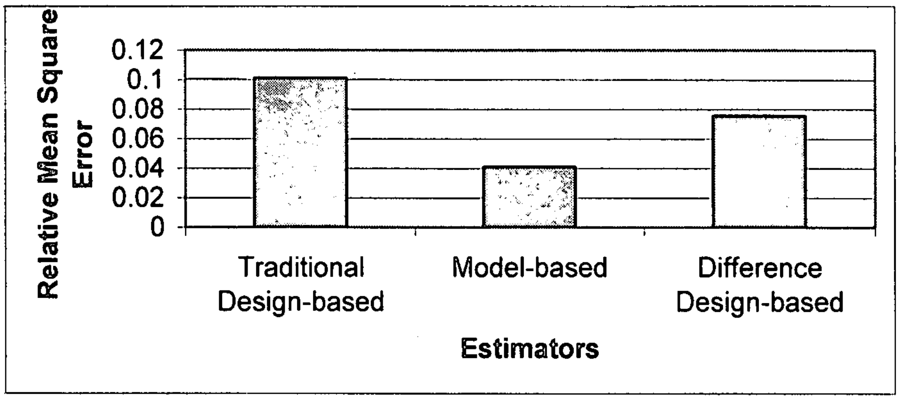

Figure 9. Relative Mean Square Error Of The Different Estimators Of The 50% Quantile Proportion

3.4 Comparison Across Estimators of the 10%, 30% and 50% Quantile Proportions

The values of the relative mean error (RME) of the 10%, 30% and 50% quantile proportions are small for all the estimators, except for the model-based estimator, which has a reasonable large relative bias in the 10% quantile. The traditional design-based estimator is generally unbiased, as can be seen in Figure 10 where the RME for all the quantiles are almost equal to zero. Furthermore, in the 30% quantile, the RMEs of the traditional and model-based estimators are close to zero indicating that these two estimators in the 30% quantile are almost unbiased. The difference estimator in all the quantiles is slightly negatively biased as shown in Figure 10.

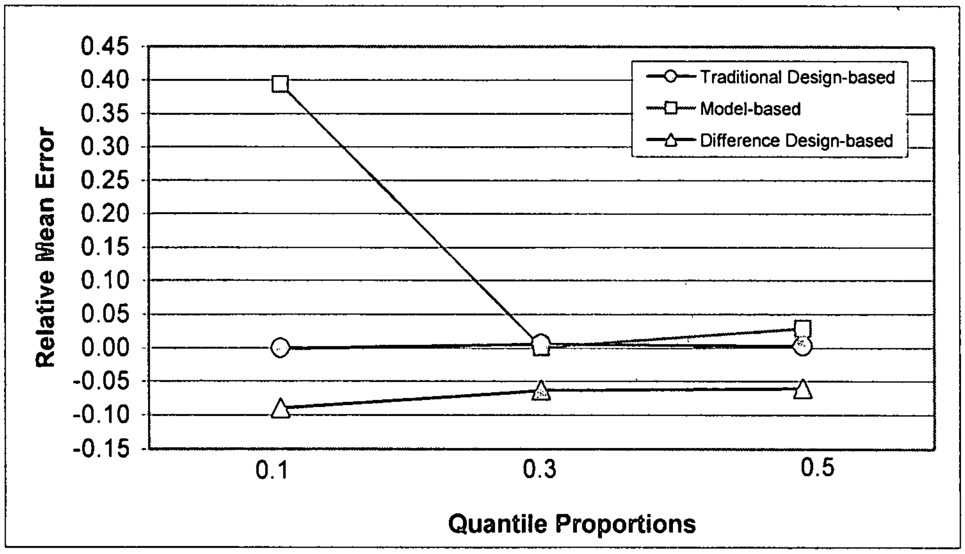

Figure 10. Relative Mean Errors Of Estimates Of The 10%, 30% And 50% Quantile Proportions

Observations in the performance of the different estimators in the different quantiles in terms of the relative mean error are supported by the observations in the absolute values. The relative biases of the traditional design-based estimator in all the quantiles are negligible, as expected. The model-based estimator's relative bias is very large in the 10% quantile, but became negligible in the 30% quantile. Though the relative bias of the model-based estimator in the 50% quantile is not negligible, there is a great improvement in the model-based estimator in this quantile compared in the 10% quantile, as shown in Figure 11. Moreover, the difference design-based estimator is biased in all the quantiles, and its bias continue to increase as the quantile gets larger.

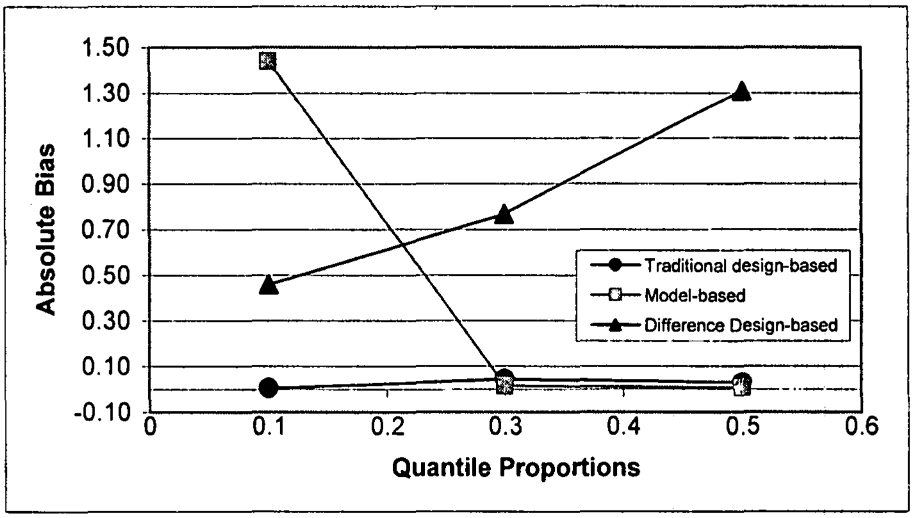

Figure 11. Absolute Bias Of Estimates Of The 10%, 30% And 50% Quantile Proportions

In terms of efficiency, as measured by the relative root mean square error (RRMSE), model-based estimator is considered to be the least efficient estimator in the 10% quantile. It has the largest value RRMSE, while there is a small difference between the traditional and difference design-based estimators. Model-based estimator did not perform well in the 10% quantile. As noted previously, the definition of model-based estimator ($\hat{F}_{CD}(t)$) depends on model (3) being the assumed model for the population. Thus, the behavior of being less efficient of the model-based estimator of 10% quantile proportion may be accounted to the misspecification of the model. The weighted least square of income as a function of expenditure was the one considered in this study without considering other rival models. However, the performance of the model-based estimator improves in the 30% quantile (see Figure 12). The efficiency of the difference design-based estimator diminishes relative to the model-based estimator. Model-based estimator does well in this quantile compared to others. Hence, it may be regarded that in the 30% quantile, the samples seem to obey the assumed model. The efficiency of the model-based estimator increases further in the 50% quantile. The smallest value of RRMSE is in the 50% quantile, and that is using the model-based estimator. The improvement in the model-based estimator may also be accounted to the relationship of household's annual income with household's annual expenditure. It can be seen from Table 14 that relationship of the two variables gets stronger as the quantile gets larger. The model-based estimator can be considerably more efficient than the design-based estimator when a strong linear relationship between the annual income and annual expenditure exists.

Table 14 . Correlation Coefficient Of Annual Income With Annual Expenditure And Household Size By Quantile Group

	QUANTILES		
	10%	30%	50%
Expenditure	0.6244	0.6957	0.7569
Household Size	0.2496	0.2293	0.2052

The traditional design-based estimator ($\hat{F}_T(t)$) is invariably less efficient than the other two estimators. The efficiency of the difference design-based estimator over the traditional design-based estimator does not change in all the quantiles considered. In general, the efficiency of the different estimators improves as the quantile gets larger as further shown in Figure 12.

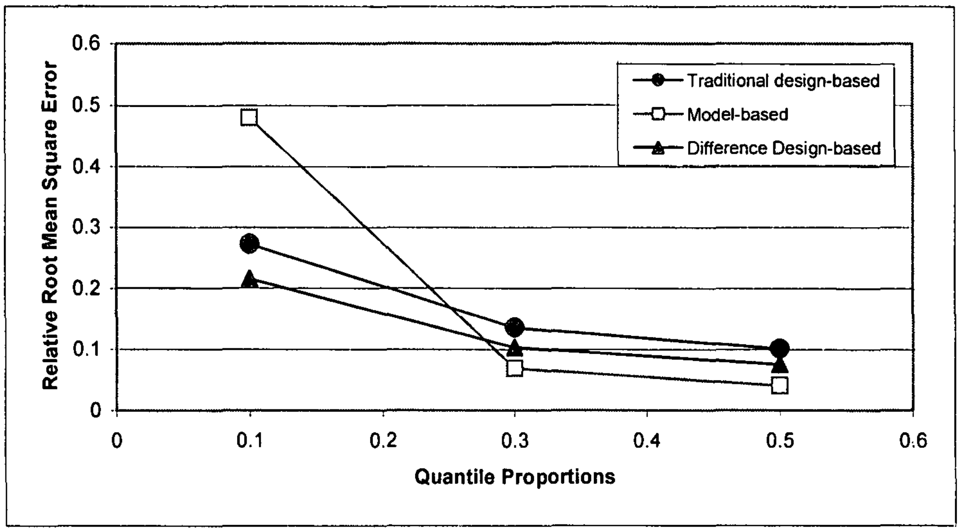

Figure 12. Relative Root Mean Square Errors Of Estimates Of The 10%, 30% And 50% Quantile Proportions.

4. SUMMARY AND CONCLUSION

Traditional design-based estimator is unbiased, as expected, of the 10%, 30% and 50% quantile proportions. It also has a bell-shaped distribution with a skewness coefficient of almost equal to zero. Model-based and difference design-based estimators, on the other hand, have positive skewed distributions. In all quantiles considered, the model-based estimator is least variable as measured by the relative root mean square error (RRMSE) compared to the other two estimators.

The values of the relative mean error (RME) of all the estimates are small for all the quantiles, except for the model-based estimator which is considerably more biased than the design-based estimators in the 10% quantile, and considerably the least efficient in terms of the relative root mean square error. This poor performance of the model-based estimator may be due to the misspecification of the model at this level. The assumed model does not accurately represent the relationship that exists in the population data, hence, the estimate of the population parameter is substantially biased. However, in the 30% quantile, model-based estimator became far better than the other estimators. It was clear from the results, as measured by the relative mean error and supported by its absolute bias, that the relative bias of the model-based estimator is negligible. The efficiency of the model-based estimator improves further in the 50% quantile. This is because the relationship of household's annual income with household's annual expenditure gets stronger as the quantile gets larger. Also, the behavior of the model-based estimator is dependent on the correctness of the assumed model. Hence, the improvement on the properties of this estimator may also be accounted to the model assumed in this study. Model-based estimator may have substantial advantage if the model assumed accurately represents the relationship that exists in the population data.

Therefore, based on the simulated results, to estimate the proportion of households in NCR with income value less than or equal to the 10% quantile income value of the population, one could use the traditional design-based estimator for the simplicity of the estimation procedure. Model-based estimator, on the other hand, is the best one for the 30% and 50% quantile proportions.

ACKNOWLEDGMENT

We wish to thank the editors of the Philippine Statistician and the anonymous referee for their helpful comments and suggestions.

References

CHAMBERS, RL and DUNSTAN (1986). 'Estimating distribution functions from survey data'. Biometrika, Vol.73, No.3, pp.597-604.

DORFMAN, AH.(1993). 'A comparison of design-based and model-based estimators of the finite distribution function'. Australian Journal Statistics, Vol.35, No.1, pp.29-41.

RAO, JNK, KOVAR JG., and MANTEL, HJ.(1990). 'On Estimating distribution functions and quantiles from survey data using auxiliary information'. Biometrika, Vol.77, No.2, pp.365-375.