STUDY OF A NEW PPZ ESTIMATOR OF THE POPULATION TOTAL

William D. Warde, Abbes Belgacem
William D. Warde, Oklahoma State University
Stillwater, Oklahoma 74075

Key Words: PPS, PPZ sampling, unequal probability of selection, updated standardized sizes.

INTRODUCTION

The use of unequal probabilities in sampling was first suggested by Hansen and Hurwitz (1943). They demonstrated that the use of Unequal Probabilities of Selection (UPS) frequently made for more efficient estimators of population totals than did equal probability sampling (Brewer and Hanif, 1983). In the UPS category, the Probability Proportional to Size (PPS or PPZ) sampling is the best known scheme. The size measure can be any arbitrary measure that suits the aims or the sample design. The size of all population units must be known a priori in order to use the PPS scheme.

THE STANDARD PPZ ESTIMATOR AND THE PROPOSED PPZ ESTIMATOR

Consider a population of N units with unit sizes X_i, $i = 1, 2, \ldots, N$, and define $Z_i = X_i/X_0$ where $X_0 = \sum_{i=1}^{N} X_i$. Hence we have $0 < Z_i < 1$ and $\sum_{i=1}^{N} Z_i = 1$.

Let y_i be the value of the characteristic of interest for the i^{th} unit. Hence, the population total is $Y = \sum_{i=1}^{N} y_i$.

When sampling is with replacement, the standard PPZ estimator of Y, using a sample of size n, is

$$\hat{Y}_{PPZ} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{Z_i},$$

$\hat{Y}_{PPZ}$ is unbiased for Y with variance

$$V(\hat{Y}_{PPZ}) = \frac{1}{n} \sum_{i=1}^{N} Z_i \left(\frac{y_i}{Z_i} - Y\right)^2.$$

This estimator of Y assumes that all unit sizes remained unchanged from those taken from the frame. In practice, for various reasons, the unit sizes may change; consequently, the set of probabilities of selection will change accordingly. Hence, by the time of the current survey, the old set of probabilities (Z_i, $i = 1, 2, \ldots, N$) will be outdated and therefore $\hat{Y}_{PPZ}$ will be biased for Y.

The proposed estimation method consists of selecting the sample based on the original probabilities of selection $Z_1, Z_2, \ldots, Z_N$. After getting the sample, we observe the actual sizes, X_i^*, for those units in the sample, and compute the corresponding updated standardized sizes, Z_i^*, $i = 1, 2, \ldots, n$. In calculating Z_i^*, we assume that $X_o = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} X_i^*$. However, if there are grounds to believe that the total size has increased or decreased by some proportion, then X_o^* can be estimated and used in calculating the Z_i^*. The proposed estimator for the population total Y is

$$\hat{Y}_{PPNZ} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{Z_i^*}.$$

In the event Z_i^* becomes 0, then the quantity y_i/Z_i^* is defined to be 0. Unless $Z_i/Z_i^* = 1$, $i = 1, 2, \ldots, N$, this estimator is biased with variance:

$$V(\hat{Y}_{PPNZ}) = \frac{1}{n} \sum_{i=1}^{N} \sum_{i<j}^{N} Z_i Z_j \left[\frac{y_i}{Z_i^*} - \frac{y_j}{Z_j^*} \right]^2 .$$

In order to compare the two estimators under study, their performances in terms of bias, variances and mean square errors have been examined using the Monte Carlo simulation technique.

METHODOLOGY AND SIMULATION

Data from the censuses of agriculture, industry, etc..., show that a realistic type of distribution for the unit sizes is one which is right hand skewed. We used the beta distribution with parameters 1.5 and 6, which simulates potential real life situations.

Using SAS, we generated ten finite populations as random samples of size N=100 from an assumed infinite superpopulation having a $\beta e(1.5, 6)$ distribution. For each of these populations, the Z_i's were computed from the sizes (the X_i's).

We considered three ways to model the change of the Z_i's to the Z_i^*'s:

<u>MODEL I</u>: This model allows for change in all units of the population:

$$Z_i^* = Z_i + e_i, \quad i = 1, 2, \ldots N$$

where $e_i \sim N(0, .000001)$. This variance of the errors was chosen considering the size of the Z values so that the Z*'s be within a realistic range.

<u>MODEL II</u>: This model allows for change in only a given portion of the population:

$$Z_i^* = Z_i + \gamma e_i, \quad i = 1, 2, \ldots, N$$

where the e_i's are as in Model I, and

$$\gamma = \begin{cases} 1 & \text{if } u \leq c \\ 0 & \text{if } u > c \end{cases},$$

where $u \sim Un(0, 1)$ and c specifies the proportion of units in the population allowed to change. We considered three cases; c = 0.1, 0.2, and 0.3.

<u>MODEL III</u>: This model allows for situations of "gone out of business" or "moved out", etc,...:

$$Z_i^* = \delta (Z_i + \gamma e_i), \quad i = 1, 2, \ldots, N$$

where e_i: is as in Models I and II,

γ: is as in Models II, and

$$\delta = \begin{cases} 1 & \text{if } u \leq c_0 \\ 0 & \text{if } u > c_0 \end{cases},$$

where $u \sim Un(0, 1)$ and c_0 specifies the proportion of the "gone out of business" situation. We considered a c_0 of 0.1 combined with each of the three rates of change used in Model II.

In all three models, we set Z_i^* equal to zero whenever it was negative. Overall, we examined seven variants of change from the Z's to the Z*'s.

With respect to the variable of interest, y, we used the example of wheat production in the State of Oklahoma as a framework for our simulation. In our infinite superpopulation, we assumed the model:

$$y_i = \beta Z_i^* + e_i,$$

where: y_i: total for unit i,

Z_i^*: standardized size of unit i,

β: regression coefficient,

e_i: $e_i \sim NID(0, \sigma^2)$ and independent of the Z_i^*.

In our population, the average size measurement is 0.01 which yields, based on data from the 1980 – 1989 decade in the State of Oklahoma, a β equal to 1,395,000. The errors represent the fluctuations in the yield per acre. Their standard deviation was set equal to 400 in order to get a yield per acre within two bushels 95% of the time based on the average size.

A hundred samples each of size n=10 were created from each of the ten populations.

ANALYSIS

For each population, we computed and compared:

i) $\hat{Y}_{PPZ}$ and $\hat{Y}_{PPNZ}$ from each of the hundred samples,

ii) the mean of the hundred estimates from both estimators,

iii) the biases, $(\hat{Y}_{PPZ} - Y)$ and $(\hat{Y}_{PPNZ} - Y)$, from each sample and their respective means over the hundred samples,

iv) the absolute values of the biases ,

v) the estimates of the variances of both estimators from each sample,

vi) the estimates of $V(\hat{Y}_{PPZ})$ and $V(\hat{Y}_{PPNZ})$ based on the hundred samples,

vii) the estimates of $MSE(\hat{Y}_{PPZ})$ and $MSE(\hat{Y}_{PPNZ})$ based on the hundred samples.

RESULTS AND DISCUSSION

The results are presented in tables which contain the following information:

i) the population total, Y, (the parameter of interest),

ii) the mean of the hundred $\hat{Y}_{PPZ}$ estimates, denoted YHZBAR,

iii) the mean of the hundred $\hat{Y}_{PPNZ}$ estimates, denoted YHNZBAR,

iv) the estimates of the variances of $\hat{Y}_{PPZ}$ and $\hat{Y}_{PPNZ}$ based on the hundred samples, denoted by VYHZ and VYHNZ respectively,

v) a comparison of the sample variances of $\hat{Y}_{PPZ}$ and $\hat{Y}_{PPNZ}$ from each sample, denoted v and v* respectively. This comparison shows the number of times where v* was smaller than v.

vi) the estimates of $MSE(\hat{Y}_{PPZ})$ and $MSE(\hat{Y}_{PPNZ})$ based on the hundred samples, denoted, respectively, by MSEYHZ and MSEYHNZ.

vii) a comparison of the absolute values of the estimated biases. This comparison shows the number of times the bias of $\hat{Y}_{PPNZ}$ was smaller than that of $\hat{Y}_{PPZ}$, denoted by BNZ < BZ.

viii) the mean over the hundred samples of the absolute value of the estimated biases, denoted BZBAR and BNZBAR for $\hat{Y}_{PPZ}$ and $\hat{Y}_{PPNZ}$ respectively.

ix) the average over all ten populations for those of the above quantities where the average has meaning.

<u>MODEL I</u>: Table I summarizes the results of this model.

The mean absolute value of the estimated biases of $\hat{Y}_{PPNZ}$ was smaller than that of $\hat{Y}_{PPZ}$ in all ten populations, and the average ratio between the two was one to two. The absolute value of the estimated bias of $\hat{Y}_{PPNZ}$ was smaller 72% of the time over all populations. However, in terms of means of the hundred estimates, $\hat{Y}_{PPZ}$ yielded better estimates of Y in all ten populations, though the difference between the two estimators was practically insignificant in most cases. This can be explained by the signs of the deviations from Y canceling out to give such a result.

Based on the hundred samples, estimates of $V(\hat{Y}_{PPNZ})$ were smaller than those of $V(\hat{Y}_{PPZ})$ in nine out of the ten populations, and on average, the ratio was two to five. The same comparison is noted with respect to the estimates of the mean square errors of the two estimators. Also, $\hat{Y}_{PPNZ}$ had smaller sample variances 98% of the time over all populations. These results, clearly favor $\hat{Y}_{PPNZ}$ over $\hat{Y}_{PPZ}$ as estimator of Y under the conditions of Model I.

MODEL II: Three variants were considered for this model: $c = 0.1$, $c = 0.2$, and $c = 0.3$. In Table II, we present the results for $c = 0.3$ only.

$\hat{Y}_{PPNZ}$ had a smaller mean absolute value of the estimated bias in eight of the ten populations for both the $c = 0.1$ and $c = 0.3$, and in nine populations at $c = 0.2$. However, the magnitude of the average difference between the two estimated biases increases with the rate of change in the populations. This pattern is also observed for the absolute value of the estimated bias of $\hat{Y}_{PPNZ}$ being smaller than that of $\hat{Y}_{PPZ}$: 37% of the time at $c = 0.1$, 57% of the time at $c = 0.2$ and 64% at $c = 0.3$.

In terms of means of the hundred estimates, $\hat{Y}_{PPZ}$ generally performed better than $\hat{Y}_{PPNZ}$, yielding a mean closer to the parameter in 19 of the 30 populations. We note that both means are frequently very close to each other.

Estimates of $V(\hat{Y}_{PPNZ})$ are smaller than those of $V(\hat{Y}_{PPZ})$ in eight populations at $c = 0.1$, nine populations at $c = 0.2$ and eight populations at $c = 0.3$, and the decrease in variance is large in almost all cases. This pattern is also true with respect to estimates of $MSE(\hat{Y}_{PPNZ})$ and $MSE(\hat{Y}_{PPZ})$.

The sample variances of $\hat{Y}_{PPNZ}$ compared better to those of $\hat{Y}_{PPZ}$ at higher rates of change. On the average, those of $\hat{Y}_{PPNZ}$ were smaller than those of $\hat{Y}_{PPZ}$ 53% of the time at $c = 0.1$, 76% of the time at $c = 0.2$ and 86% of the time at $c = 0.3$.

These results indicate that, under conditions of Model II, higher rates of change in the population units tend to favor $\hat{Y}_{PPNZ}$ over $\hat{Y}_{PPZ}$, in terms of estimated variance and estimated absolute bias.

MODEL III: We considered three cases. We set the rate of "gone out of business" at 10%, that is $c_0 = 0.1$, combined with rates of change in the population of 10%, 20% and 30%, that is $c = 0.1$, 0.2, and 0.3 respectively. Results for $c = 0.3$ are presented in Table III.

Looking at the mean absolute values of the bias estimates, the comparison is less evident in this model than in the first two models, especially at the lower rates of change where both estimators had very close mean absolute bias estimates in most cases, roughly yielding equal over–all–averages for both estimators. However, at the two higher rates of change, this criterion seems to favor $\hat{Y}_{PPNZ}$. This latter gave a smaller mean absolute estimates of bias in six populations at $c = 0.2$ and in nine populations at $c = 0.3$ with slightly smaller over–all averages than those of $\hat{Y}_{PPZ}$ in both cases. This pattern is also noted for the absolute value of bias per sample: $\hat{Y}_{PPNZ}$ yielded a smaller absolute value of bias only 26% of the time at $c = 0.1$ but improved to 48% of the time at $c = 0.2$ and $c = 0.3$. To explain these percentages in the light of the comparison of the mean absolute biases, we suspect that the difference between the absolute biases when the standard estimator has a smaller bias is much smaller than when the new estimator has a smaller one.

These comparisons suggest that in situations of Model III, higher rates of change tend to favor $\hat{Y}_{PPNZ}$ in terms of bias. However, we notice here that these results are even lower than those at the same rates of Model II.

In terms of means of the hundred estimates, both estimators performed similarly in all three variants with $\hat{Y}_{PPZ}$ having means closer to Y in six of the ten populations for each of the three cases.

At the low rate of change, estimates of $V(\hat{Y}_{PPZ})$ and $V(\hat{Y}_{PPNZ})$ from the hundred samples were very much larger than those in the two previous models. This comparison is also valid for estimates of $MSE(\hat{Y}_{PPZ})$ and $MSE(\hat{Y}_{PPNZ})$. At the two other rates of change, we still notice very large variance estimates for both estimators, ranging from 3 to 30 billion, with slightly smaller estimates for $V(\hat{Y}_{PPNZ})$ in seven populations with c = 0.2 and in all populations with c = 0.3.

The sample variances of $\hat{Y}_{PPNZ}$ compared better to those of $\hat{Y}_{PPZ}$ at the two higher values of change. On the average, those of $\hat{Y}_{PPNZ}$ were smaller than those of $\hat{Y}_{PPZ}$ only 38% of the time with c = 0.1, and improved to 62% and 64% of the time with c = 0.2 and c = 0.3 respectively.

CONCLUSION

Under conditions of Model I, the new estimator compared very favorably to the standard PPZ estimator; it resulted in important reductions in the estimates of variance and absolute bias.

Under conditions of Model II, the new estimator is shown to be clearly favored by higher rates of change in the population. At low rates of change, around c = 0.1, there is little reason to choose between the two estimators. However, at higher rates of change, the new estimator appears to result in a smaller estimated bias and a markedly smaller estimated variance.

Under conditions of Model III, we generally found the same conclusions as in Model II, but in this case even higher rates of change are required in order for the new estimator to achieve better results than the standard estimator. Even at c = 0.2, the new estimator seems not to perform better than the standard one. However, at c = 0.3, it showed an interesting reduction in the estimates of bias and especially in the estimates of variance.

In this study we have considered only three rates of change in the population units, and only one rate of the going out of business type of situation. More rates, in both senses, need to be investigated to validate our conclusions.

REFERENCES

Brewer, K.R.W. and Hanif, M. Sampling With Unequal Probabilities. Springer–Verlag, New York, 1983.

Hansen, M.H., and Hurwitz, W.N. (1943), "On the theory of sampling from finite populations," *Ann. Math. Stat.*, 20;333–362.

Table I: Summary Results for Model I.

Y	YHZBAR	YHNZBAR	VYHZ	VYHNZ	v*<v	MSEYHZ	MSEYHNZ	BNZ<BZ	BZBAR	BNZBAR
1388584	1390100	1395143	4623922906	990787834	97	4626245662	1034244230	81	51371	19784
1382269	1381002	1402629	2805955988	15537811906	96	2807579732	15956502982	79	39124	28638
1395080	1391533	1390207	3753137435	1423951036	99	3765849449	1447943726	78	46388	20919
1402499	1402074	1391340	3144797825	256020127	99	3144980789	381811057	80	44050	15830
1375113	1378873	1393514	2619554215	203295333	100	2633834753	545327084	64	37328	19996
1412308	1404697	1398482	2414684843	443164134	94	2473199180	636250866	66	37741	20728
1417632	1409817	1404980	4616494762	218693837	96	4678193980	380387065	70	41401	15943
1375962	1376416	1387558	3578112517	364241709	100	3578320511	500061183	73	45629	18244
1319536	1432643	1398182	18135542944	422858475	98	18309048853	883478521	50	43153	25932
1385969	1385630	1394369	3118208166	430678303	99	3118324382	501954079	79	42120	17383
AVERAGE					98			72	42830	20320

Table II: Summary Results for Model II, c=0.3.

Y	YHZBAR	YHNZBAR	VYHZ	VYHNZ	v*<v	MSEYHZ	MSEYHNZ	BNZ<BZ	BZBAR	BNZBAR
1398758	1403356	1405114	667387475	274881336	68	688743870	315690717	60	17635	11901
1395387	1393510	1392056	1183096424	287209135	90	1186655288	298418465	76	25034	11040
1406286	1410471	1398974	2970157013	239778492	93	2987852688	293783255	72	28699	13729
1390076	1389196	1397785	1658163984	323635509	87	1658945609	383665754	68	28548	15158
1408059	1409955	1388330	2186291040	359783579	98	2189924318	752931758	64	34781	22078
1396710	1399982	1396991	831495938	474216847	85	842310666	474296244	63	20655	13461
1387362	1386849	1413314	1228718481	6263270900	81	1228983874	6943582805	57	23448	29991
1395327	1390234	1417774	1253113344	80940250998	84	1279315425	81449180750	59	24687	41378
1390753	1394517	1394279	795690098	259575410	81	809996963	272134580	52	20290	12738
1396892	1391595	1393401	2080250468	437898689	94	2108592559	450207433	70	31331	14538
A V E R A G E					86			64	25511	18601

Table III: Summary Results for Model III, c=0.3.

Y	YHZBAR	YHNZBAR	VYHZ	VYHNZ	v*<v	MSEYHZ	MSEYHNZ	BNZ<BZ	BZBAR	BNZBAR
1301109	1288262	1283917	17766468948	16523975083	83	17933198682	16822527262	50	113790	110486
1160713	1141065	1137836	28707045198	26497934782	45	29096995710	27026574323	50	132871	129044
1262870	1284503	1286410	16759764281	15431320485	68	17232514941	15991054340	57	102211	96272
1175666	1166296	1176979	24089459389	22784514753	57	24178142248	22786255361	51	131346	128382
1175246	1169194	1160652	23973206833	23234758241	48	24010205870	23449896873	50	125508	123455
1218733	1206293	1211347	26213288723	25648572958	65	26369611215	25703673814	45	128560	129121
1249450	1267415	1273811	17593054455	16839898235	47	17919076570	17439388690	51	106702	104155
1322136	1328590	1322236	10818069336	9826561198	79	10860141961	9826571310	47	87358	83782
1327137	1322005	1341103	9351401486	8898315313	84	9378004665	9095329318	36	73184	79047
1139164	1140999	1140577	32706619288	30686249286	64	32710056789	30688295293	46	147758	143439
A V E R A G E					64			48	114928	112718