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 of 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 $z_i = \frac{X_i}{X_0}$, where $X_0 = \sum x_i$. Hence we have

$$0 < z_i < 1 \text{ and } \sum_{i=1}^{N} z_i = 1.$$

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 Hence, by the time of the current survey, the old set of probabilities $(z_i, i = 1, 2, \ldots, N)$ will be outdated and therefore $Y_{PPZ}$ will be biased for $Y$.

The proposed estimation method consists of selecting the sample based on the original probabilities of selection $Z_i, Z_2, \ldots, Z_N$. 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 $x_o = \sum x_i = \sum x_i$.

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
$Z_i^*/Z_i = 1$, $i = 1, 2, ..., N$, this estimator is
biased with variance:

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

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 $\beta(1.5, 6)$
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:

MODEL I: This model allows for
change in all units of the population:

$$z_i^* = z_i + e_i, \quad i = 1, 2, ..., N$$

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

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

$$z_i^* = z_i + \gamma e_i, \quad i = 1, 2, ..., N$$

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

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

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

where $e_i$ is as in Models I and II,
$\gamma$ 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 $u \sim \text{Un}(0, 1)$ 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 $z_i$'s to the $z_i^*$'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 $i$,
$\beta$: regression coefficient,
$e_i$: $e_i \sim \text{NID}(0, \sigma^2)$ 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 $\beta$ 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 \( \text{MSE}(\hat{Y}_{PPZ}) \) and \( \text{MSE}(\hat{Y}_{PPNZ}) \) 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 \( Y_{HZBAR} \),

iii) the mean of the hundred \( \hat{Y}_{PPNZ} \) estimates, denoted \( Y_{HNZBAR} \),

iv) the estimates of the variances of \( \hat{Y}_{PPZ} \) and \( \hat{Y}_{PPNZ} \) based on the hundred samples, denoted by \( V_{YHZ} \) and \( V_{YHNZ} \) 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 \( \text{MSE}(\hat{Y}_{PPZ}) \) and \( \text{MSE}(\hat{Y}_{PPNZ}) \) based on the hundred samples, denoted, respectively, by \( \text{MSE}_{YHZ} \) and \( \text{MSE}_{YHNZ} \).

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 \( B_{ZBAR} \) and \( B_{NZBAR} \) 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.

**MODEL I:** 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


Table I: Summary Results for Model I.

<table>
<thead>
<tr>
<th>Y</th>
<th>YNH</th>
<th>YNHBAR</th>
<th>VNH</th>
<th>VNHBAR</th>
<th>v&lt;-&gt;v</th>
<th>MSEY</th>
<th>MSEYBAR</th>
<th>BM&lt;-&gt;BI</th>
<th>BISR</th>
<th>BMISR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1388564</td>
<td>1390100</td>
<td>1395143</td>
<td>462322906</td>
<td>890787834</td>
<td>97</td>
<td>4628245662</td>
<td>1034244230</td>
<td>81</td>
<td>51371</td>
<td>19784</td>
</tr>
<tr>
<td>1382269</td>
<td>1380012</td>
<td>1402429</td>
<td>2050555996</td>
<td>15557811906</td>
<td>96</td>
<td>2007579732</td>
<td>15956502942</td>
<td>79</td>
<td>379124</td>
<td>26839</td>
</tr>
<tr>
<td>1395080</td>
<td>1391533</td>
<td>1390207</td>
<td>3753137435</td>
<td>1423951036</td>
<td>99</td>
<td>3765849449</td>
<td>1474937276</td>
<td>78</td>
<td>46389</td>
<td>20191</td>
</tr>
<tr>
<td>1402499</td>
<td>1402074</td>
<td>1391230</td>
<td>3144797825</td>
<td>256020127</td>
<td>99</td>
<td>3149490750</td>
<td>381810057</td>
<td>80</td>
<td>44050</td>
<td>15630</td>
</tr>
<tr>
<td>1375113</td>
<td>1378873</td>
<td>1393515</td>
<td>261954215</td>
<td>202395333</td>
<td>100</td>
<td>2633834753</td>
<td>545327084</td>
<td>61</td>
<td>37328</td>
<td>19998</td>
</tr>
<tr>
<td>1412308</td>
<td>1404657</td>
<td>1398482</td>
<td>2416681483</td>
<td>443164139</td>
<td>94</td>
<td>2473193180</td>
<td>636258066</td>
<td>66</td>
<td>37741</td>
<td>20728</td>
</tr>
<tr>
<td>1417632</td>
<td>1409877</td>
<td>1404980</td>
<td>4616494762</td>
<td>216363837</td>
<td>98</td>
<td>4678193280</td>
<td>380387065</td>
<td>70</td>
<td>41401</td>
<td>19423</td>
</tr>
<tr>
<td>1375962</td>
<td>1376116</td>
<td>1387558</td>
<td>3578112517</td>
<td>364241709</td>
<td>100</td>
<td>3578320251</td>
<td>500061893</td>
<td>73</td>
<td>45629</td>
<td>18244</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Y</th>
<th>YNH</th>
<th>YNHBAR</th>
<th>VNH</th>
<th>VNHBAR</th>
<th>v&lt;-&gt;v</th>
<th>MSEY</th>
<th>MSEYBAR</th>
<th>BM&lt;-&gt;BI</th>
<th>BISR</th>
<th>BMISR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1398758</td>
<td>1403356</td>
<td>1405114</td>
<td>647387475</td>
<td>07</td>
<td>698743570</td>
<td>315659717</td>
<td>60</td>
<td>17635</td>
<td>11901</td>
<td></td>
</tr>
<tr>
<td>1395207</td>
<td>1393510</td>
<td>1392056</td>
<td>1180306124</td>
<td>07</td>
<td>1186655288</td>
<td>298418484</td>
<td>76</td>
<td>25035</td>
<td>11040</td>
<td></td>
</tr>
<tr>
<td>1406286</td>
<td>1410471</td>
<td>1398974</td>
<td>2970157013</td>
<td>08</td>
<td>2978705268</td>
<td>295783255</td>
<td>79</td>
<td>28699</td>
<td>13729</td>
<td></td>
</tr>
<tr>
<td>1390076</td>
<td>1398196</td>
<td>1397708</td>
<td>1658138064</td>
<td>08</td>
<td>1658945609</td>
<td>383665754</td>
<td>68</td>
<td>28548</td>
<td>15150</td>
<td></td>
</tr>
<tr>
<td>1408059</td>
<td>1409955</td>
<td>1368330</td>
<td>21164291040</td>
<td>07</td>
<td>2119924318</td>
<td>752917150</td>
<td>64</td>
<td>34781</td>
<td>22078</td>
<td></td>
</tr>
<tr>
<td>1396710</td>
<td>1399962</td>
<td>1396991</td>
<td>814959380</td>
<td>07</td>
<td>821301066</td>
<td>474296244</td>
<td>63</td>
<td>20655</td>
<td>13461</td>
<td></td>
</tr>
<tr>
<td>1387632</td>
<td>1386849</td>
<td>1413314</td>
<td>0228178641</td>
<td>07</td>
<td>0228893877</td>
<td>693458280</td>
<td>57</td>
<td>23348</td>
<td>29991</td>
<td></td>
</tr>
<tr>
<td>1395327</td>
<td>1390234</td>
<td>1417774</td>
<td>1253113344</td>
<td>08</td>
<td>1279315245</td>
<td>8144918075</td>
<td>59</td>
<td>24867</td>
<td>41378</td>
<td></td>
</tr>
<tr>
<td>1390753</td>
<td>1394517</td>
<td>1393429</td>
<td>795690099</td>
<td>08</td>
<td>795975410</td>
<td>271234580</td>
<td>52</td>
<td>20290</td>
<td>12738</td>
<td></td>
</tr>
<tr>
<td>1396892</td>
<td>1391596</td>
<td>1393401</td>
<td>2080250468</td>
<td>07</td>
<td>208592555</td>
<td>450207433</td>
<td>70</td>
<td>31331</td>
<td>14538</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Y</th>
<th>YNH</th>
<th>YNHBAR</th>
<th>VNH</th>
<th>VNHBAR</th>
<th>v&lt;-&gt;v</th>
<th>MSEY</th>
<th>MSEYBAR</th>
<th>BM&lt;-&gt;BI</th>
<th>BISR</th>
<th>BMISR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1301109</td>
<td>1282622</td>
<td>1283917</td>
<td>17768468948</td>
<td>08</td>
<td>17933198682</td>
<td>16822527262</td>
<td>50</td>
<td>113790</td>
<td>110466</td>
<td></td>
</tr>
<tr>
<td>1160713</td>
<td>1141065</td>
<td>1137836</td>
<td>62707045198</td>
<td>07</td>
<td>62497341782</td>
<td>27026574323</td>
<td>50</td>
<td>132871</td>
<td>129043</td>
<td></td>
</tr>
<tr>
<td>1262870</td>
<td>1284503</td>
<td>1286410</td>
<td>16759764281</td>
<td>56</td>
<td>17235214941</td>
<td>1599105430</td>
<td>57</td>
<td>102211</td>
<td>96272</td>
<td></td>
</tr>
<tr>
<td>1175666</td>
<td>1166296</td>
<td>1176979</td>
<td>24089543909</td>
<td>57</td>
<td>24178142248</td>
<td>22786255361</td>
<td>51</td>
<td>131240</td>
<td>128382</td>
<td></td>
</tr>
<tr>
<td>1175246</td>
<td>1159394</td>
<td>1160652</td>
<td>62732050683</td>
<td>58</td>
<td>62646458372</td>
<td>23469865873</td>
<td>50</td>
<td>125508</td>
<td>123455</td>
<td></td>
</tr>
<tr>
<td>1210733</td>
<td>1204293</td>
<td>1211347</td>
<td>3621288723</td>
<td>56</td>
<td>36249612115</td>
<td>25706378144</td>
<td>45</td>
<td>128560</td>
<td>129121</td>
<td></td>
</tr>
<tr>
<td>1244950</td>
<td>1267415</td>
<td>1273811</td>
<td>17593054455</td>
<td>100</td>
<td>17919706570</td>
<td>17439386980</td>
<td>61</td>
<td>197021</td>
<td>104155</td>
<td></td>
</tr>
<tr>
<td>1322136</td>
<td>1328590</td>
<td>1322236</td>
<td>10818063336</td>
<td>79</td>
<td>10860141196</td>
<td>8226571310</td>
<td>47</td>
<td>87358</td>
<td>63782</td>
<td></td>
</tr>
<tr>
<td>1327137</td>
<td>1322005</td>
<td>1341103</td>
<td>9351041865</td>
<td>84</td>
<td>9378004665</td>
<td>9095239218</td>
<td>36</td>
<td>13184</td>
<td>79047</td>
<td></td>
</tr>
<tr>
<td>1139184</td>
<td>1140999</td>
<td>1140577</td>
<td>32706619288</td>
<td>64</td>
<td>32710056789</td>
<td>30688292593</td>
<td>64</td>
<td>147756</td>
<td>143439</td>
<td></td>
</tr>
</tbody>
</table>

AVERAGE 64 48 114280 12710