Introduction: Several estimators for \( V_{HT} \), the variance of Horvitz-Thompson estimator have been proposed in the recent years. For comparing variance estimators, the criterion of mean square error is appealing, but not necessarily best. An estimator with a relatively large mean square error may come fairly close to the true value in a large proportion of samples. Therefore, within the framework of conventional sampling theory, one is most interested in the accuracy of confidence intervals and the average width of confidence intervals based on a given variance estimator. Therefore, in this paper, we have empirically compared the distribution of standardized (studentized) estimates and average widths of confidence intervals obtained using different estimators of the variance of the Horvitz-Thompson estimator of the population total.

Formulae: The Horvitz-Thompson estimator of the population total is given by

\[
e_{HT} = \sum_{i \in s} z_i,
\]

where \( z_i = y_i/\pi_i \).

The variance of \( e_{HT} \) can be expressed in two equivalent forms as

\[
V_{HT} = \sum_{i < j} (\pi_i \pi_j - \pi_{ij}) z_i z_j
\]

and

\[
V_{HT} = \sum_{i < j} (\pi_i - \pi_{ij})(z_i - z_j)^2
\]

Some of the proposed estimators for the variance of the Horvitz and Thompson estimator, \( V_{HT} \), are defined below:

\[
V_{HT}(1) = \sum_{i \in s} (1-\pi_i)z_i^2 + 2 \sum_{i < j \in s} \pi_{ij} z_i z_j
\]

(due to Horvitz-Thompson (1952))

\[
V_{HT}(2) = \sum_{i < j \in s} (\pi_i \pi_j - \pi_{ij}) (z_i - z_j)^2
\]

(due to Yates and Grundy (1953))

\[
V_{HT}(3) = \sum_{i \in s} \left( \sum_{i < j \in s} c_{ij} f_{ij} \right) \left( \sum_{i < j \in s} c_{ij} \right)^{-1} \sum_{i < j} c_{ij}
\]

(due to Fuller (1970))

\[
V_{HT}(4) = \sum_{i < j \in s} c_{ij} f_{ij} \sum_{i \in s} c_{s1}, f_{is}(n-1) c_{s} f_{s} - f_{is}
\]

(due to Biyani (1978))

\[
V_{HT}(5) = \sum_{i < j \in s} c_{ij} f_{ij} + \frac{1}{n} \sum_{i \in s} c_{s1} [(n-1) f_{is} + f_{s}] c_{s} f_{s}
\]

(due to Biyani (1970))

\[
V_{HT}(6) = \sum_{i < j \in s} \left( \sum_{i \in s} c_{ij} f_{ij} \right) \sum_{i \in s} c_{ij} \left( \sum_{i \in s} c_{ij} \right)^{-1} \sum_{i < j} c_{ij}
\]

(Ratio type estimator)

\[
V_{HT}(7) = \frac{2}{n(n-1)} \sum_{i < j < s} c_{ij} \sum_{i < j < s} f_{ij}
\]

(Simplified version of \( V_{HT}(5) \))

where, the sum \( \sum_{i \in s} \) is over all distinct units of samples \( s \), and \( f_{ij} = (z_i - z_j)^2 \),

\[
f_{is} = \frac{1}{n} \sum_{i \in s} f_{is},
\]

\[
c_{ij} = \pi_i \pi_j - \pi_{ij},
\]

\[
f_{is} = \frac{1}{n} \sum_{i \in s} f_{is},
\]

\[
c_{si} = \sum_{k \in s} c_{ik} - \pi_i \sum_{k \in s} c_{ik},
\]

\[
c_{s} = \sum_{k \in s} c_{ik} = \sum_{i = 1}^{N} \pi_i (1-\pi_i) - \sum_{i \in s} c_{ij},
\]

\[
c_{ss} = \sum_{i \in s} c_{si},
\]

where, \( \sum_{i < j < s} \) is over all distinct units of samples \( s \), and \( f_{ij} = (z_i - z_j)^2 \),

\[
f_{is} = \frac{1}{n} \sum_{i \in s} f_{is},
\]

\[
c_{ij} = \pi_i \pi_j - \pi_{ij},
\]

\[
f_{is} = \frac{1}{n} \sum_{i \in s} f_{is},
\]

\[
c_{si} = \sum_{k \in s} c_{ik} - \pi_i \sum_{k \in s} c_{ik},
\]

\[
c_{s} = \sum_{k \in s} c_{ik} = \sum_{i = 1}^{N} \pi_i (1-\pi_i) - \sum_{i \in s} c_{ij},
\]

\[
c_{ss} = \sum_{i \in s} c_{si},
\]

Method: Define \( t_i = \frac{e_{HT} - \pi}{\sqrt{V_{HT}(i)}}, \ i = 1,2 \ldots 7 \)

where, \( \pi \) = true population total
The populations used in the study are listed in Table 1.

<table>
<thead>
<tr>
<th>Popn.</th>
<th>Source</th>
<th>Year</th>
<th>Source</th>
<th>Year</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cochran</td>
<td>1920</td>
<td>area of farm</td>
<td>1930</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>Jessen</td>
<td>1978</td>
<td>area of farm</td>
<td>1978</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Scheaffer,</td>
<td>1979</td>
<td>estate value</td>
<td>2 yrs.</td>
<td>20</td>
</tr>
</tbody>
</table>

The sampling scheme of Sampford (1967) was used to draw samples with inclusion probabilities proportional to \( x \). For this design, all estimators, except \( V_{HT}(1) \), are non-negative, as the relation \( \pi_i \pi_j > 0 \) holds. Since \( V_{HT}(1) \) can take negative values, \( V_{HT}(1) \) was not included in the study. One thousand independent samples were drawn from each population, for each of the sample sizes, 3, 5, and 7, the last one being restricted to the last two populations. For first population, sample size 7 was not used because the condition \( \pi_i^2 x_i \) would have forced the largest \( \pi_i \) to exceed unity for \( n = 7 \). For each sample and each variance estimator \( V_{HT}(i) \), the quantities above were calculated. From the resulting empirical distribution of each \( t_i \), and for selected percentile, \( p \), the lower \( (1000p) \)th percentile of the distribution of \( t_i \) is

\[
t_i(p) = (1000p)\text{th observation in increasing order, and the upper (1000p)th percentile is}
\]

\[
t_i(1-p) = (1000p)\text{th observation in decreasing order}
\]

A 100(1-2p)% confidence interval for \( y \) based on the \( i \)-th variance estimator is

\[
( e_{HT} - t_i(p) \sqrt{V_{HT}(i)}, e_{HT} + t_i(1-p) \sqrt{V_{HT}(i)} )
\]

The width of the above confidence interval is

\[
[t_i(1-p) - t_i(p)] \sqrt{V_{HT}(i)}
\]

and hence the average width over 1000 samples can be computed by multiplying the average of \( \sqrt{V_{HT}(i)} \) with \( [t_i(1-p) - t_i(p)] \).

The percentiles of the distribution of the \( t_i \)'s are shown in Table 2 and the average widths of the confidence intervals are shown in Table 3.

### Table 2 Percentiles of the distribution of Studentized estimates with \( p = .05 \)

<table>
<thead>
<tr>
<th>Popn.</th>
<th>Sample Size</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
<th>t7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>-6.23</td>
<td>-5.81</td>
<td>-7.82</td>
<td>-6.08</td>
<td>-6.21</td>
<td>-6.64</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-6.72</td>
<td>-5.37</td>
<td>-5.92</td>
<td>-5.27</td>
<td>-5.09</td>
<td>-5.31</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-2.56</td>
<td>-2.39</td>
<td>-3.04</td>
<td>-2.32</td>
<td>-2.24</td>
<td>-2.35</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-2.19</td>
<td>-2.18</td>
<td>-2.17</td>
<td>-1.89</td>
<td>-1.86</td>
<td>-1.9</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-2.07</td>
<td>-1.96</td>
<td>-1.79</td>
<td>-1.63</td>
<td>-1.72</td>
<td>-1.58</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-7.29</td>
<td>-7.2</td>
<td>-9.66</td>
<td>-7.2</td>
<td>-7.2</td>
<td>-7.2</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-1.94</td>
<td>-1.94</td>
<td>-2.16</td>
<td>-1.96</td>
<td>-1.97</td>
<td>-1.96</td>
</tr>
</tbody>
</table>
Table 3: Average widths of confidence intervals based on different variance estimators (.95 Confidence Coefficients).

<table>
<thead>
<tr>
<th>Popn</th>
<th>Sample Size</th>
<th>( V_{HT}(2) )</th>
<th>( V_{HT}(3) )</th>
<th>( V_{HT}(4) )</th>
<th>( V_{HT}(5) )</th>
<th>( V_{HT}(6) )</th>
<th>( V_{HT}(7) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2124</td>
<td>1820</td>
<td>1600</td>
<td>1654</td>
<td>1469</td>
<td>1688</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1396</td>
<td>1321</td>
<td>1014</td>
<td>1024</td>
<td>950.9</td>
<td>1048</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1638</td>
<td>1636</td>
<td>1667</td>
<td>1659</td>
<td>1813</td>
<td>1643</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>863.2</td>
<td>885.5</td>
<td>893.1</td>
<td>898</td>
<td>892.7</td>
<td>901.8</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>533.6</td>
<td>543.3</td>
<td>549.1</td>
<td>552.7</td>
<td>540.8</td>
<td>567.6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>97.63</td>
<td>97.61</td>
<td>97.62</td>
<td>97.66</td>
<td>98.17</td>
<td>97.61</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>14.15</td>
<td>14.10</td>
<td>14.02</td>
<td>14.02</td>
<td>14.16</td>
<td>13.94</td>
</tr>
</tbody>
</table>

Table 4: Coverage probabilities for 95% confidence interval based on Student's t.

<table>
<thead>
<tr>
<th>Popn No.</th>
<th>Sample Size</th>
<th>( V_{HT}(2) )</th>
<th>( V_{HT}(3) )</th>
<th>( V_{HT}(4) )</th>
<th>( V_{HT}(5) )</th>
<th>( V_{HT}(6) )</th>
<th>( V_{HT}(7) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>.86</td>
<td>.90</td>
<td>.81</td>
<td>.95</td>
<td>.95</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.76</td>
<td>.79</td>
<td>.70</td>
<td>.76</td>
<td>.75</td>
<td>.76</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>.96</td>
<td>.96</td>
<td>.94</td>
<td>.96</td>
<td>.97</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.95</td>
<td>.95</td>
<td>.94</td>
<td>.95</td>
<td>.97</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
<td>.97</td>
<td>.97</td>
<td>.97</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>.88</td>
<td>.88</td>
<td>.85</td>
<td>.87</td>
<td>.87</td>
<td>.87</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.95</td>
<td>.95</td>
<td>.93</td>
<td>.95</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.95</td>
<td>.95</td>
<td>.94</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
</tr>
</tbody>
</table>

Discussion: Based on this study, the following conclusions may be drawn. The distributions of \( t_2, t_3, t_5, t_6, \) and \( t_7 \) are very similar, and are somewhat closer to the \( t \)-distribution than those of \( t_4 \). This may be expected as the estimators \( V_{HT}(2), V_{HT}(3) \) and \( V_{HT}(5) \) through \( V_{HT}(7) \) are either design-unbiased or model-unbiased, whereas \( V_{HT}(4) \) is a shrinkage estimator and inflate the value of \( t_4 \) by appearing in the denominator.

Although the relative efficiencies of different variance estimators differ considerably, rather surprisingly, we find few major differences among the average widths of confidence intervals based on them. However, there are a few cases where the ratio estimator, \( V_{HT}(6) \), gives considerably narrower confidence intervals than other estimators. However, these comparisons depend on the knowledge of the distribution of the \( t_i \)'s which may not be available in practice. The construction of confidence intervals based on unequal probability sampling, when we do not know the distribution of the \( t_i \)'s is an open problem.

In case Student's t is used for constructing confidence intervals, the coverage probabilities would not generally differ much for different variance estimators provided shrinkage is compensated for in case of shrinkage estimator. However, the coverage probabilities may differ considerably from the nominal confidence coefficient as shown in Table 4.
References:


