Under the RP-PPS method, the system of prorating is the same but the probabilities are different. Thus,

<table>
<thead>
<tr>
<th>Terminal branch</th>
<th>Prorated amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2-1-1</td>
<td>( \frac{4.13}{5.96} ) ( \frac{1.48}{3.80} ) (3) = 0.81</td>
</tr>
<tr>
<td>1-2-1-2</td>
<td>( \frac{4.13}{5.96} ) ( \frac{2.32}{3.80} ) (3) = 1.27</td>
</tr>
<tr>
<td>1-2-2</td>
<td>( \frac{1.83}{5.96} ) (3) = 0.92</td>
</tr>
</tbody>
</table>

Total = 3.00

The estimator \( \hat{Y}_i \), Eq. (3.2), can be written in a form that corresponds to the idea of prorating path fruit to terminal branches. Let \( p_i = (p_{o_1})... (p_{t_i}) \), which is the probability of selecting the \( i^{th} \) terminal branch. It follows that

\[
\hat{Y}_i = \frac{y_i}{p_i} 
\] (3.3)

where \( y_i = [(p_{l_1})... (p_{t_i})y_{o_i}] + \ldots + [(p_{(k+i)i})... (p_{t_i})y_{k_i}] + \ldots + [y_{t_i}] \)

Thus, \( y_i \) is the number of fruit "on" the \( i^{th} \) terminal branch including prorated amounts of path fruit. Assuming the RP-PPS method and terminal branch 1-2-1-1 as an example, the value of \( y_i \) is \( \frac{4.13}{5.96} \) \( \frac{1.48}{3.80} \) (3) + 73 = 73.81 and \( \hat{Y}_i \) is \( \frac{73.81}{.05103} = 2379 \) which gives the same result that was obtained when Eq. (3.2) was used.

Table 3.2, columns headed \( \hat{Y}_2 \) and \( \hat{Y}_4 \), present estimates of the total number of apples on the tree for the RP-EP and RP-PPS methods and each of the possible random paths. These estimates were obtained by using the technique of prorating path fruit,
Eq. (3.3). That is, estimates of the total number of apples were obtained by dividing the values of $Y_1$ (last two columns of Table 3.1) by the appropriate probabilities which are presented in Table 3.2, columns $P_2$ and $P_4$.

For comparison of the four methods we now need to decide how to include the path fruit for the DS-EP and DS-PPS methods. If the amount of path fruit is small, the best method might be to count all path fruit at the time the tree is mapped to determine terminal branches. In this case, assuming a sample of one terminal branch, the estimator, would be

$$\hat{Y}_1 = Y' + \frac{y'_i}{p_i}$$

(3.4)

where $Y'$ is the number of path fruit, $y'_i$ is the number of fruit on the $i^{th}$ terminal branch and $p_i$ is the probability of selecting the $i^{th}$ terminal branch. Alternatives are not considered in this illustration because, from a practical viewpoint, interest is in the random path methods. Thus, as a matter of expediency, the estimator (3.4) was used to obtain the estimates, $\hat{Y}_1$ and $\hat{Y}_3$, that are presented in Table 3.2 for the DS-EP and DS-PPS methods. Since only 51 apples out of 1901 were on path sections, the method of accounting for the apples on path sections probably has a very small impact on the sampling variance.
Exercise 3.3 For terminal branches 3-1-4-1 and 3-3, calculate estimates of the total number of apples on the tree for the DS-EP and DS-PPS methods using the estimator (3.4). Your answer should agree with the estimates that are presented in Table 3.1 for these two branches.

For each terminal branch and each of the four estimators (methods) there is a unique estimate of the total number of apples. All four estimators are unbiased. By definition, an estimator is unbiased if the expected (average) value of the estimates that might occur is equal to the population value. To find the expected value of an estimator, each estimate must be weighted by the probability of its occurrence.

Exercise 3.4 For the RP-EP and RP-PPS methods, compute the expected value of the estimates presented in Table 3.2. The answer, except for rounding error, should be exactly 1901, which is the total number of apples on the tree.

3.5 VARIANCES OF THE ESTIMATORS

With reference to the theory of expected values, the variance of a random variable, Y, is the average of the squared deviations of Y from its expected (average) value. To be more specific, suppose Y is a random variable that can equal one of a set of values \(Y_1, Y_2, \ldots, Y_N\) with probabilities \(P_1, P_2, \ldots, P_N\) where \(\sum P_i = 1\). By definition, the average value of Y is

\[
\bar{Y} = E(Y) = \sum P_i Y_i
\]
and the variance of \( Y \), which is the average value of \((Y - \bar{Y})^2\), is

\[
E(Y - \bar{Y})^2 = \sum_{i=1}^{N} p_i (Y_i - \bar{Y})^2
\]

**Exercise 3.5** Show that \( \sum_{i=1}^{N} p_i (Y_i - \bar{Y})^2 = \sum_{i=1}^{N} p_i Y_i^2 - \bar{Y}^2 \)

Consider the estimator for the RP-PPS method. It is a random variable that can equal any one of the set of values in column \( \hat{Y}_4 \) of Table 3.2. The set of probabilities is presented in column \( P_4 \). By definition, the variance of the estimator (or estimates) is

\[
(.05492)(3751-1901)^2 + \ldots + (.06163)(814-1901)^2 = 800,194
\]

or using the right hand side of the equation in exercise 3.5,

\[
(.05492)(3751)^2 + \ldots + (.06163)(814)^2 - (1901)^2 = 800,194
\]

The result, 800,194, is the sampling variance for the RP-PPS method when only one terminal branch is selected. If four terminal branches (or random paths) were selected with replacement, four estimates of the tree total would be computed, one for each branch, and the variance of the average of the four estimates would be \( \frac{800,194}{4} = 200,048 \).

The sampling variances (for a sample of one branch) are presented in Table 3.3 for each of the four methods and each of the six trees. The third tree is the one that was used above as an example. It is not expected that the four methods will always rank in the same order from one tree to another. However, the results illustrate some points that are of interest and importance.
3.6 DISCUSSION OF THE METHODS

The RP-EP method requires considerably less time than the RP-PPS method, but it has relatively high sampling variance because, at any given stage of branching, a large branch has the same probability of selection as a small one. That is, the RP-EP method is such that the probability of selecting a terminal branch has little or no relation to the number of fruit on the branch. The result, as shown by the sampling variances in Table 3.3, is a good illustration of a point that was made earlier. Compared to selecting sampling units with equal probability (as in the DS-EP method), the introduction of unequal probabilities of selection (as in the RP-EP method) will increase the sampling variance unless the selection probabilities are related to the values of the characteristic being measured in a way that will reduce sampling variance.

Figure 3.1 is a dot chart with the number of apples on a branch (column headed EP in Table 3.1) plotted against the values of $P_2$. The wide range in the selection probabilities and the lack of a relation explains the high sampling variance of the RP-EP method compared with the other methods. For comparison, Figure 3.2 is a dot chart for number of apples and the selection probabilities for the RP-PPS method. Compare Figure 3.2 with Figure 1.2 which showed a dot chart where sampling with pps would rank high.

After a branch has been identified and marked, the time required to obtain its csa, with a convenient instrument that
gives a reading directly in square inches (or square centimeters), is quite small. The use of csa as an auxiliary variable reduced sampling variance by a large amount. The reduction in variance in relation to cost is definitely advantageous. According to Table 3.3, the sampling variances for DS-PPS and RP-PPS are about the same and much less than the sampling variance for the DS-EP method. This indicates that RP-PPS is a good choice because it avoids the work of identifying all terminal branches before sampling. However, results in Table 3.3 should not be accepted as representative. The csa is not always an effective measure. Pruning and maintenance practices, age of trees, species or variety of trees, and other factors have some influence on the relation between csa and number of apples. The purposes of an intensive investigation limited to a few trees include testing different procedures for counting apples or measuring the size of branches, and acquiring ideas that seem to be worth exploring as possibilities for large scale application.

It is extremely important in the processes of sampling to understand the part played by randomization. Important biases sometimes occur even when strict attention is paid to details in making random selections. On the other hand, subjective evaluations or determinations in sampling are commonplace. With knowledge of how various factors effect sampling variance, the exercise of good judgement can be very effective in reducing sampling variance. But, there are points in the processes of
sampling where a determination should be strictly random. Some
design constraints may be determined subjectively but selections
of units for a sample should be in accord with rigorous, technical
interpretation of randomness. It is generally preferable to have
random selections made under competent supervision in an office,
but that is not always feasible. Thus, one advantage of taking
photographs of a sample of bare trees (assuming it is feasible)
is that sample branches can be selected in the office. The se-
lected branches are marked on photographs for enumerators. In this
situation an enumerator's work is subject to full verification.
Incidentally, the economics of sample surveys suggests that larger
investments in sample design and selection can often be justified
when the same sample is to be used for several surveys rather
than one.

Exercise 3.6 Suppose the RP-PPS method is being applied
and in the process you come to the following situation:

Assume that branch A, which has a csa equal to 3.2 square inches,
has already been selected. It divides into two branches A-1 and
A-2 with csa's equal to 1.4 and 1.6. With regard to size, the two
branches, A-1 and A-2, qualify as terminal branches and ordinarily
A-1 and A-2 would be accepted as terminal branches. But, before
selecting one of the two, you happen to notice that A-2 has no
apples on it and that A-1 appears to have approximately an average
amount. Consider the following alternatives:
(1) Accept A, which includes A-1 and A-2, as the terminal branch, and expand the count of apples by \( \frac{1}{P_A} \), where \( P_A \) was the probability of selecting A.

(2) Accept A-1 and A-2 as terminal branches and select one with pps. Expand the count on A-1 or A-2 by \( \frac{1}{P_A} \)(\(3.0\)) or \( \frac{1}{P_A} \)(\(3.0\) \(1.0\)), depending on whether A-1 or A-2 is selected.

(3) Discard A-2 since it has no apples on it and take A-1 as the terminal branch using \( \frac{1}{P_A} \)(\(3.0\)) as the expansion factor.

Discuss the alternatives with regard to bias and sampling variance.

Exercise 3.7 Refer to exercise 3.6 and as a variation of the situation assume that branch A-2 has been selected at random in accord with the instructions for the random path method. The enumerator prepares to count the apples on A-2 but finds there are no apples. He recognizes, since a sample of only one branch is to be selected for the sample from this tree, that the estimate of the number of apples on the tree will be zero (assuming no path fruit on the path to A-2). There is obviously a large number of apples on the tree, so he might have a strong opinion that something should be done that would give a better sample. How would you respond to each of the following possibilities:

(1) Accept A-2 as a terminal branch, which means using zero as an estimate of the number of apples on the tree. Remember A-2 has already been selected.

101
(2) Reject A-2 as a sample. Start at the beginning and
select another terminal branch to replace A-2.

(3) Accept A which includes A-1 and A-2, as the terminal
branch for the sample.

Discuss the three possibilities with regard to bias and sampling
variance.

Exercise 3.8 In application of the RP-PPS method would it
be advisable to be looking forward, as one approaches the terminal
branch stage, for branches that are large enough to be terminal
branches but clearly have a very small number of apples on them.

With reference to the diagram in exercise 3.6 as an example, an
counter looking forward, and considering what was ahead, could
have stopped when A was selected and accepted A as a terminal
branch. Otherwise, he would normally have followed the selection
procedure one stage further. In application of the random path
method, what is your opinion of the feasibility of looking ahead
and taking eye estimates of numbers of apples into account in
determining the terminal branch. Can it be used to reduce sampling
error without risk of introducing bias? Think about the matter
with regard to instructions that would be given to enumerators.

Exercise 3.9 It is not likely that there would be an interest
in estimating the average number of terminal branches per tree.
However, as an exercise, suppose the RP-PPS method is applied to
the tree for which data are presented in Table 3.1. Assume that
the following four terminal branches are selected as a sample:
1-1-2, 1-2-1-2, 2-4, and 3-2-1. From this sample, estimate the number of terminal branches on the tree. (The selection probabilities have already been computed, see Table 3.2). The parameter being estimated is 26. Ans. 33.4.

Exercise 3.10 Suppose a sample of 25 apple trees has been selected and that four enumerators have been trained in the application of the RP-PPS method. Assume that each enumerator, working independently and using the RP-PPS method, selects a sample of one terminal branch from each of the 25 trees. It is unlikely that enumerators will interpret terminal branches in exactly the same way. For example, one enumerator might have a tendency to follow the random path to terminal branches of the smallest permissible size, whereas another might stop as soon as he obtains a branch that is small enough to qualify as a terminal branch. Or, a branch along a path might be treated as a terminal branch by one enumerator and as path fruit by another. However, for each enumerator an estimate of the total number of apples on each tree is made using either (3.2) or (3.3) as the estimator. The 25 estimates are added together to obtain an estimate of the total number of apples on the 25 trees. This gives four estimates, one for each enumerator, of the total number of apples on the 25 trees.

(a) Assume that random selection is performed correctly at each stage of branching (after all branches at the stage have been completely identified and measured), and assume that apples have
been correctly counted. Do the four estimates of the total number of apples all have the same expected value and the same variance?

(b) Suppose four estimates, one for each enumerator, of the total number of terminal branches on the 25 trees are made. Do these estimates have the same expected value? Why?

(c) Two enumerators measuring the csa's of any given set of branches are not likely to obtain exactly the same numerical values. Is this important? Discuss.

(d) The assumptions made in (a) are subject to question. Try listing some differences among enumerators that will, and will not, have an effect on the expected value of an estimate of the total number of apples on the 25 trees.

Exercise 3.11 Suppose, owing to pruning practices, that many cases like the following are found:

Assume the instructions were to always measure the csa at the base (point A) of a branch. Would you expect the csa measurements under the RP-PPS method to be ineffective, or even increase the sampling variance, compared with the DS-EP method? In cases like the above drawing, perhaps measuring the csa at position B would be more effective. What is your opinion? Incidentally, this is
a good example of why it is essential that a research and development staff should have actual experience with practical operations and decisions that must be made by enumerators. Do not expect high quality results when instructions are not well adapted. Agreement between concepts (the theoretical model) and operations as actually performed is of fundamental importance.
Table 3.1--Data by Branches for Apple Tree No. 3

<table>
<thead>
<tr>
<th>Branch identification</th>
<th>1st stage csa</th>
<th>1st stage No.</th>
<th>2nd stage csa</th>
<th>2nd stage No.</th>
<th>3rd stage csa</th>
<th>3rd stage No.</th>
<th>4th stage csa</th>
<th>4th stage No.</th>
<th>No. of apples on or assigned to a terminal branch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1-1</td>
<td>11.60</td>
<td>3.65</td>
<td>2.68</td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206</td>
</tr>
<tr>
<td>1-1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.97</td>
<td>32</td>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>1-1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206</td>
</tr>
<tr>
<td>1-2-1-1</td>
<td>5.61</td>
<td>(3)</td>
<td>4.13</td>
<td>1.48</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td>73.8</td>
</tr>
<tr>
<td>1-2-1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.32</td>
<td>138</td>
<td></td>
<td></td>
<td>138.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.80</td>
</tr>
<tr>
<td>1-2-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.83</td>
<td>133</td>
<td></td>
<td></td>
<td>134.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.96</td>
</tr>
<tr>
<td>1-3-1</td>
<td>2.01</td>
<td></td>
<td>.97</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>1-3-2</td>
<td></td>
<td></td>
<td>1.03</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.00</td>
</tr>
<tr>
<td>1-4</td>
<td>1.43</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>1-5</td>
<td></td>
<td></td>
<td>2.24</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.94</td>
</tr>
<tr>
<td>2-1-1</td>
<td>13.45</td>
<td>(6)</td>
<td>3.36</td>
<td></td>
<td>.92</td>
<td>42</td>
<td></td>
<td></td>
<td>42.8</td>
</tr>
<tr>
<td>2-1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.99</td>
<td>109</td>
<td></td>
<td></td>
<td>109.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.91</td>
</tr>
<tr>
<td>2-2-1</td>
<td>5.09</td>
<td></td>
<td>1.47</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>74.7</td>
</tr>
<tr>
<td>2-2-2-1</td>
<td></td>
<td></td>
<td>3.47</td>
<td>(16)</td>
<td>1.64</td>
<td>56</td>
<td></td>
<td></td>
<td>64.4</td>
</tr>
<tr>
<td>2-2-2-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.54</td>
<td>116</td>
<td></td>
<td></td>
<td>124.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.18</td>
</tr>
<tr>
<td>2-3</td>
<td>1.99</td>
<td>124</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>125.5</td>
</tr>
<tr>
<td>2-4</td>
<td>1.83</td>
<td>79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.27</td>
</tr>
<tr>
<td>3-1-1</td>
<td>12.84</td>
<td>(1)</td>
<td>6.30</td>
<td>(2)</td>
<td>1.47</td>
<td>30</td>
<td></td>
<td></td>
<td>30.6</td>
</tr>
<tr>
<td>3-1-2</td>
<td></td>
<td></td>
<td>1.21</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.6</td>
</tr>
<tr>
<td>3-1-3</td>
<td></td>
<td></td>
<td>1.91</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.6</td>
</tr>
<tr>
<td>3-1-4-1</td>
<td></td>
<td></td>
<td>4.13</td>
<td>(23)</td>
<td>1.47</td>
<td>16</td>
<td></td>
<td></td>
<td>27.7</td>
</tr>
<tr>
<td>3-1-4-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.15</td>
<td>23</td>
<td></td>
<td></td>
<td>34.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.62</td>
</tr>
<tr>
<td>3-2-1</td>
<td>5.35</td>
<td></td>
<td>1.40</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35.1</td>
</tr>
<tr>
<td>3-2-2</td>
<td></td>
<td></td>
<td>1.42</td>
<td>61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>61.1</td>
</tr>
<tr>
<td>3-2-3</td>
<td></td>
<td></td>
<td>1.76</td>
<td>116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>116.1</td>
</tr>
<tr>
<td>3-2-4</td>
<td></td>
<td></td>
<td>3.26</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.84</td>
</tr>
<tr>
<td>Total number of apples on terminal branches</td>
<td>1850</td>
<td>Total number of apples on path sections</td>
<td>51</td>
<td>Grand total</td>
<td>1901</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 Probabilities of Selection and Estimates of the Total Number of Apples on Tree No. 3

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p_1$</td>
<td>$y_1$</td>
<td>$p_2$</td>
<td>$y_2$</td>
</tr>
<tr>
<td>1-1-1</td>
<td>0.03846</td>
<td>5407</td>
<td>0.03333</td>
<td>6180</td>
</tr>
<tr>
<td>1-1-2</td>
<td>0.03846</td>
<td>883</td>
<td>0.03333</td>
<td>960</td>
</tr>
<tr>
<td>1-2-1</td>
<td>0.03846</td>
<td>1949</td>
<td>0.01667</td>
<td>4425</td>
</tr>
<tr>
<td>1-2-2</td>
<td>0.03846</td>
<td>3639</td>
<td>0.01667</td>
<td>8325</td>
</tr>
<tr>
<td>1-3-1</td>
<td>0.03846</td>
<td>883</td>
<td>0.03333</td>
<td>960</td>
</tr>
<tr>
<td>1-3-2</td>
<td>0.03846</td>
<td>831</td>
<td>0.03333</td>
<td>900</td>
</tr>
<tr>
<td>1-4</td>
<td>0.03846</td>
<td>753</td>
<td>0.06667</td>
<td>405</td>
</tr>
<tr>
<td>1-5</td>
<td>0.03846</td>
<td>2339</td>
<td>0.06667</td>
<td>1320</td>
</tr>
<tr>
<td>2-1-1</td>
<td>0.03846</td>
<td>1143</td>
<td>0.04167</td>
<td>1026</td>
</tr>
<tr>
<td>2-1-2</td>
<td>0.03846</td>
<td>2885</td>
<td>0.04167</td>
<td>2634</td>
</tr>
<tr>
<td>2-2-1</td>
<td>0.03846</td>
<td>1975</td>
<td>0.04167</td>
<td>1794</td>
</tr>
<tr>
<td>2-2-2-1</td>
<td>0.03846</td>
<td>1507</td>
<td>0.02083</td>
<td>3090</td>
</tr>
<tr>
<td>2-2-2-2</td>
<td>0.03846</td>
<td>3067</td>
<td>0.02083</td>
<td>5972</td>
</tr>
<tr>
<td>2-3</td>
<td>0.03846</td>
<td>3275</td>
<td>0.08333</td>
<td>1506</td>
</tr>
<tr>
<td>2-4</td>
<td>0.03846</td>
<td>2105</td>
<td>0.08333</td>
<td>966</td>
</tr>
<tr>
<td>3-1-1</td>
<td>0.03846</td>
<td>831</td>
<td>0.02778</td>
<td>1101</td>
</tr>
<tr>
<td>3-1-2</td>
<td>0.03846</td>
<td>857</td>
<td>0.02778</td>
<td>1137</td>
</tr>
<tr>
<td>3-1-3</td>
<td>0.03846</td>
<td>1117</td>
<td>0.02778</td>
<td>1497</td>
</tr>
<tr>
<td>3-1-4-1</td>
<td>0.03846</td>
<td>467</td>
<td>0.01389</td>
<td>2001</td>
</tr>
<tr>
<td>3-1-4-2</td>
<td>0.03846</td>
<td>649</td>
<td>0.01389</td>
<td>2505</td>
</tr>
<tr>
<td>3-2-1</td>
<td>0.03846</td>
<td>961</td>
<td>0.02778</td>
<td>1263</td>
</tr>
<tr>
<td>3-2-2</td>
<td>0.03846</td>
<td>1637</td>
<td>0.02778</td>
<td>2199</td>
</tr>
<tr>
<td>3-2-3</td>
<td>0.03846</td>
<td>3067</td>
<td>0.02778</td>
<td>4179</td>
</tr>
<tr>
<td>3-2-4</td>
<td>0.03846</td>
<td>2339</td>
<td>0.02778</td>
<td>3171</td>
</tr>
<tr>
<td>3-3</td>
<td>0.03846</td>
<td>1351</td>
<td>0.11111</td>
<td>453</td>
</tr>
<tr>
<td>0.99996</td>
<td>1.00001</td>
<td>0.99997</td>
<td>1.00001</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3 Variances of Estimates of the Total Number of Apples on Each of Six Trees from a Sample of One Terminal Branch

<table>
<thead>
<tr>
<th>Tree</th>
<th>No. of terminal branches</th>
<th>csa of trunk</th>
<th>No. of apples on tree</th>
<th>Variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>7.0</td>
<td>214</td>
<td>40 (000)</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>20.0</td>
<td>1448</td>
<td>882 (000)</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>23.0</td>
<td>1901</td>
<td>1419 (000)</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>16.5</td>
<td>1658</td>
<td>1148 (000)</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>13.5</td>
<td>403</td>
<td>82 (000)</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>19.5</td>
<td>1575</td>
<td>894 (000)</td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>99.5</td>
<td>7199</td>
<td>4465 (000)</td>
</tr>
</tbody>
</table>
Figure 3.1 Dot Chart---Number of Apples vs Selection Probabilities for RP-EP
Figure 3.2 Dot Chart---Number of Apples vs Selection Probabilities for RP-PPS
4.1 INTRODUCTION

Most sampling plans for estimating or forecasting tree-crop production will involve three or four stages of sampling. Typically, there will be a sample of orchards (fields), a sample of trees in selected orchards, and a sample of branches from a sample of trees. Fruit on the sample branches would be counted and a small sample of fruit on the sample branches might be selected for measurements of size of fruit.

This chapter illustrates some alternative two-stage sampling plans using data for the six apple trees. Trees are the psu's (primary sampling units) and terminal branches or "paths" are the ssu's (secondary sampling units). The six trees will be treated as a population to be sampled and population variance formulas will be used to find the first and second-stage components of variance. Incidentally, the problem of making accurate counts of numbers of fruit on sample branches needs serious consideration. However, in the illustrations that follow, attention is limited to matters of sampling.

In the application of two-stage sampling, psu's are often selected with probabilities proportional to \( N_i \), where \( N_i \) is the number of ssu's in the \( i^{th} \) psu. For some surveys, sampling with
probability proportional to $N_i$ has important advantages. When the $N_i$ are not known, approximations of $N_i$ are often used.

With regard to sampling trees, the $N_i$ (number of branches on trees) are not known and it is not feasible to determine the $N_i$ for trees in an orchard. Some other effective measure of size must be found or the sample trees will need to be selected with equal probability. One possibility is to use a double sampling procedure. For example, a "large" sample of trees might be selected with equal probabilities. For each tree in the large sample a measurement of size, that takes relatively little time, might be made and used in the selection of a small sample of trees from the large sample. Possible measures of size are the csa of the trunk, the sum of the csa's of primary branches, and eye estimates of the amount of fruit. The feasibility of double sampling would depend upon the cost of obtaining the measurements of size and the relation between the measure of size and the amount of fruit on the trees. Stratification of trees within an orchard also needs to be considered. Sometimes strata within an orchard are readily recognized; for example, differences in age or variety. Perhaps a relation between size of trunk and number of apples will be found to be effective only within strata comprised of trees of the same variety and of a uniform condition.

Stratification, systematic sampling, or other techniques might be applied at any stage of sampling. However, for simplicity, the discussion will be limited to: (1) simple random sampling of psu's (selection with equal probability and without replacement)
and (2) sampling the psu's with pps (sampling with unequal probabilities of selection and replacement). Within each selected psu we will assume that a simple random sample of \( n_i \) ssu's is selected. The number of psu's in the sample is \( m \) and the number of ssu's in the sample is \( n = \sum n_i \).

Refer to Table 4.1 for an exposition of the notation that will be used for representing data for a population. Examine the notation carefully. Sample data are represented in the same way except that lower case letters are used.

Since a general mathematical formulation of estimators and their variances is rather complex for two-stage sampling, we will proceed from specific cases to more general description. The primary purpose of the next section is to present an elementary view of two-stage sampling.

4.2 PRIMARY SAMPLING UNITS EQUAL IN SIZE

The simplest case of two-stage sampling is one where all psu's have the same number of ssu's, where simple random is applied at both stages, and where the same number of ssu's is selected from each psu in the sample. In this case, and with reference to the notation in Table 4.1, the \( N_i \) all equal \( N \) and the \( n_i \) all equal \( \tilde{n} \). To summarize, the sampling plan under consideration is to select a simple random sample of \( m \) psu's from a population of \( M \) psu's and a simple random sample of \( \tilde{n} \) ssu's from each of the \( m \) psu's, which gives a total sample of \( n = m \tilde{n} \) ssu's.
For illustration a hypothetical population of 4 psu's with 5 ssu's in each is assumed. The 20 values of $Y_{ij}$ are presented in the top part of Table 4.2. Deviations of $Y_{ij}$ from $\bar{Y}$ are also presented. In single-stage sampling, there is one component of variance, namely the variance of $(Y_{ij} - \bar{Y})$ which in the illustration is 487.053.

In two-stage sampling, each deviation $(Y_{ij} - \bar{Y})$ divides into two deviations as follows:

$$(Y_{ij} - \bar{Y}) = (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)$$

The values of $(\bar{Y}_i - \bar{Y})$ are a set of deviations which reflect the variation among psu's and the values of $(Y_{ij} - \bar{Y}_i)$ form the other set which reflects variation among ssu's within psu's.

Turn to Table 4.2 and verify the deviations (components) $(\bar{Y}_i - \bar{Y})$ and $(Y_{ij} - \bar{Y}_i)$. Notice that the between psu component, $(\bar{Y}_i - \bar{Y})$, varies from one psu to another but is constant within a psu. There are only $M$ different values of $(\bar{Y}_i - \bar{Y})$ and selecting a sample of $m$ psu's is equivalent to selecting a sample of $m$ values of $(\bar{Y}_i - \bar{Y})$.

Also, study the values of the within psu component, $(Y_{ij} - \bar{Y}_i)$. It varies from one ssu to another within a psu, but its average value is zero for each psu. Therefore, these deviations reflect only variation within psu's. The second stage of sampling is equivalent to selecting $mn$ of the deviations, $(Y_{ij} - \bar{Y}_i)$.

Now consider the variance of $\bar{y}$, the mean of a two-stage sample. The difference between $\bar{y}$ and $\bar{Y}$ may be expressed as follows:

$$\bar{y} - \bar{Y} = d_1 + d_2$$
where $\bar{d}_1$ is the average value of $(\bar{Y}_i - \bar{Y})$ for the $m$ psu's in the sample and $d_2$ is the average value of $(Y_{ij} - \bar{Y}_i)$ for the $mn$ ssu's in the sample.

**Exercise 4.1** With reference to Table 4.2, suppose that psu's 1 and 3 are selected at the first stage and that ssu’s 1 and 4 are selected within psu No. 1 and ssu’s 3 and 5 are selected within psu No. 3. Find the values of $\bar{y}$, $\bar{d}_1$, and $\bar{d}_2$. Verify that $\bar{y} - \bar{Y} = \bar{d}_1 + \bar{d}_2$. Ans. 34-43 = -9.4 + 0.4.

Since $\bar{d}_1$ is the average of $m$ random values of $(\bar{Y}_i - \bar{Y})$ and $\bar{d}_2$ is the average of $mn$ random values of $(Y_{ij} - \bar{Y}_i)$, it follows that $\bar{d}_1$ and $\bar{d}_2$ are random variables. It happens that $\bar{d}_1$ and $\bar{d}_2$ are independent. Therefore, the variance of $\bar{y}$ is equal to the variance of $\bar{d}_1$ plus the variance of $\bar{d}_2$. From knowledge of the variance of the mean of a simple random sample, one might anticipate what the variances of $\bar{d}_1$ and $\bar{d}_2$ are and hence the formula for the variance of $\bar{y}$ which is:

$$V(\bar{y}) = \frac{M - m}{M} \frac{S_1^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{S_2^2}{mn}$$  \hspace{1cm} (4.1)$$

where $S_1^2$ is the variance of $(\bar{Y}_i - \bar{Y})$ and $S_2^2$ is the variance of the deviations $(Y_{ij} - \bar{Y}_i)$. In this case, $S_2^2$ is a simple average of the within psu variances, $S_{2i}^2$, which is logical since the psu's are equal in size and are selected with equal probabilities. Moreover, the within psu sample size is constant.

For the illustration, values of $S_1^2$ and $S_2^2$ as functions of the deviations $(\bar{Y}_i - \bar{Y})$ and $(Y_{ij} - \bar{Y}_i)$ are shown at the bottom of Table 4.2.
In practice, the two sets of deviations \( (Y_1 - \bar{Y}) \) and \( (Y_{ij} - \bar{Y}_1) \), would not be computed. The variances, \( S_1^2 \) and \( S_2^2 \), could be calculated as follows:

\[
S_1^2 = \frac{1}{N^2} \left[ \frac{\sum Y_i^2 - \frac{Y^2}{M}}{M - 1} \right] \tag{4.2}
\]

\[
S_2^2 = \frac{1}{M(N-1)} \left[ \frac{\sum \sum Y_{ij}^2 - \frac{\sum Y_{ij}^2}{N}}{M - 1} \right] \tag{4.3}
\]

**Exercise 4.2** Use Eq.'s 4.2 and 4.3 to find the values of \( S_1^2 \) and \( S_2^2 \) in the numerical example. Explain why \( N^2 \) appears as a divisor in Eq. 4.2.

**Exercise 4.3** For \( m=2 \) and \( n=2 \) find the variance of \( \bar{Y} \) using Eq. 4.1. Ans.118.9

**Exercise 4.4** Show algebraically, that the right hand side of Eq. 4.3 is equal to \( \frac{\sum S_{2i}^2}{M} \), where \( S_{2i}^2 \) is the variance among ssu's within the \( i^{th} \) psu.

One partial check on a variance formula is to determine whether it reduces to known formulas for special cases. Two special cases are of interest: (1) When \( m = M \), two-stage sampling becomes stratified random sampling. That is, the psu's become strata. Observe, when \( m = M \), that the first term on the right side of Eq. 4.1 vanishes and the second term becomes the variance for a stratified random sample of \( n \) units from each stratum (psu). (2) When \( n = N \), two-stage sampling reduces to single-stage cluster sampling. In this case the last term in Eq. 4.1 vanishes, leaving the first...
Exercise 4.5 Suppose \( m=1 \) and \( n=1 \). In this case the selection of one psu at random and the selection of one ssu within it is equivalent to a single-stage sample of one ssu. Therefore, the variance of \( \bar{y} \) given by Eq. 4.1 when \( m=1 \) and \( n=1 \) should be equal to the variance of \( \bar{y} \) for a single-stage random sample when \( n=1 \). Verify this using the data in Table 4.2. Remember the appropriate variance formula for the single-stage sample is Eq. 1.4.

It is important to study the structure of the variance formula, Eq. 4.1, for the variance of \( \bar{y} \). When the number of psu's in the sample is fixed, increasing the size of the sample in each psu reduces only the second component of variance. As \( \bar{n} \) increases, a point is reached where the among-psu variance is the major component and further increases in \( \bar{n} \) contributes very little to reducing the variance of \( \bar{y} \). Notice that increasing \( m \) reduces both components when \( n_i \) is constant for all psu's.

4.3 PRIMARY SAMPLING UNITS UNEQUAL IN SIZE

Populations having psu's with equal numbers of ssu's are relatively infrequent. In this section, it is assumed that the numbers, \( N_i \), of ssu's vary and that simple random sampling (without replacement) is applied at both stages.

As discussed in Chapter I, Sec. 1.1.2, "P" or "p" with appropriate subscripts refer to selection probabilities on the occasion of a particular random draw and "f" with an appropriate subscript refers to the probability that a particular unit has of being in the sample.
A general expression for the probability, $f_{ij}$, which any given ssu has of being included in a two-stage sample is:

$$f_{ij} = f_i f(j|i)$$

(4.4)

where $f_i$ is the probability which the $i^{th}$ psu has of being in the sample, and

$f(j|i)$ is the conditional probability which the $j^{th}$ ssu in the $i^{th}$ psu has of being in the sample, given that the $i^{th}$ psu is in the sample of psu's.

With simple random sampling at both stages, $f_i = \frac{m}{M}$, and $f(j|i) = \frac{n_i}{N_i}$.

Since $f_i$ is constant for the case under consideration, let $f_i = f_1$ which is the sampling fraction at the first stage. Also, let $f(j|i) = f_2i$ which is the sampling fraction at the second stage within the $i^{th}$ psu. Then Eq. 4.4 reduces to:

$$f_{ij} = f_1 f_2i$$

(4.5)

If the $f_2i$ (the sampling fractions at the second stage) are constant, $f_{ij}$ is constant and every ssu in the population has the same chance of being in the sample. Then, Eq. 4.5 becomes:

$$f = f_1 f_2$$

where $f_2$ is the constant second-stage sampling fraction. However, in the interest of generality, a requirement that $f_2i$ be constant will not be specified at this point in the discussion.

An estimator of the population mean, $\bar{Y}$, is

$$\tilde{Y} = \left(\frac{1}{N}\right)(M) \sum_{i}^{m} \frac{N_i \bar{Y}_i}{m}$$

(4.6)
where \( \bar{y}_i \) is the average of \( n_i \) ssu's in the sample from the \( i^{th} \) psu in the sample. Study the estimator 4.7 and observe that:

\[ \frac{N_i \bar{y}_i}{n} \]

is an estimate of \( Y_i \), the total for the \( i^{th} \) psu;

\[ \frac{\sum_i N_i \bar{y}_i}{m} \]

is an average of the estimated totals for the \( m \) psu's in the sample; therefore,

\[ (M \sum_i \frac{N_i \bar{y}_i}{m}) \]

is an estimate of the population total and \( \left( \frac{1}{N} \right) \) in Eq. 5.6, changes the estimated total to an estimate of \( \bar{Y} \).

The variance of \( \hat{y} \) is given by:

\[
V(\hat{y}) = \frac{1}{m} \left[ (1-f_1)S_1^2 + \frac{1}{N^2} \sum_i M(1-f_2i) \frac{N_i^2S_{2i}^2}{n_i} \right]
\]

(4.7)

where \( S_1^2 = \frac{1}{N^2} \sum_i (Y_i - \bar{Y})^2 \) is the variance among psu totals divided by \( N^2 \) so \( S_1^2 \) will be expressed on the basis of one ssu, and

\[
S_{2i}^2 = \frac{1}{n_i} \sum_j (Y_{ij} - \bar{y}_i)^2
\]

is the variance among ssu's within the \( i^{th} \) psu.

The first part of 4.7, \( \frac{1}{m} (1-f_1)S_1^2 \), is the variance of \( \hat{y} \) assuming all of the \( m \) psu's are enumerated completely. That is, the theory for single-stage sampling applies to the first stage.
The quantity:

\[(1 - f_{2i}) \frac{N_i^2 S^2_{2i}}{n_i}\]

in Eq. 4.7 is recognizable as the variance of \(N_i \bar{y}_i\) where \(\bar{y}_i\) is the mean of a simple random sample of \(n_i\) ssu's in the \(i^{th}\) psu.

Eq. 4.7 was written in the above form for comparison with other variance formulas given later for two-stage sampling. The second term within \(\left[\right]\) could be written as follows:

\[
\left(\frac{1}{N^2}\right)\left(\frac{1}{M}\right)^M \sum_{i} (1 - f_{2i}) \frac{N_i^2 S^2_{2i}}{n_i}
\]

(4.8)

because \(\frac{M}{N^2} = \frac{1}{N^2} \frac{1}{M}\). Expression 4.8 shows that the variances of \(N_i \bar{y}_i\) are summed over all psu's in the population and the sum is divided by \(M\) giving an average of such variances. The variances of \(N_i \bar{y}_i\) receive equal weight in the average because the psu's are selected with equal probabilities. Since the average variance of \(N_i \bar{y}_i\) pertains to psu totals, the divisor \(N^2\) appears in 4.8 to convert the variance to a basis of one ssu. Such analysis of a formula is helpful in determining whether one has the right formula for a particular purpose.

Exercise 4.6 If the variance formulas (4.1) and (4.7) are correct, formula (4.7) should reduce to (4.1) when \(N_i = \bar{N}\) and \(n_i = \bar{n}\). Show that this is true.
When the second-stage sampling fractions $\frac{n_i}{N_i}$, are constant and equal to $f_2$, the estimator, (4.6), reduces to:

$$\hat{y} = \frac{\sum y_{ij}}{f_2 m \bar{N}} \quad (4.9)$$

and its variance, (4.7), reduces to:

$$V(\hat{y}) = (1 - f_1) \frac{S_1^2}{m} + (1 - f_2) \frac{S_2^2}{m \bar{n}} \quad (4.10)$$

where $S_1^2$ is the same as in 4.7,

$$S_2^2 = \frac{M N_i}{\bar{N}} S_{2i}$$

and

$$\bar{n} = \frac{\sum n_i}{M} = \frac{\sum f_i N_i}{M} = f_2 \bar{N}$$

Exercise 4.7. Show that Eq.'s (4.9) and (4.10) follow from (4.6) and (4.7) when $f_2 = \frac{n_i}{N_i}$

Exercise 4.8. Show that $f_2 m \bar{N}$, in Eq. 4.9, is equal to the expected sample size. That is, show that $E(n) = f_2 m \bar{N}$ where $n = \sum n_i$. In practice one would probably use $n$, the actual sample size, in the estimator instead of the expected size, $f_2 m \bar{N}$. Moreover, $\bar{N}$ is not known in most practical applications.
4.3.1 NUMERICAL EXAMPLE

As a numerical example, the apple tree data presented in Table 2.1 will be treated as a population to be sampled. The psu's are trees and ssu's are terminal branches. The number of trees in an orchard is usually large and in practice the number of sample trees selected from an orchard would be relatively small, that is \((1 - f_1)\) would be nearly equal to 1. Accordingly, for this illustration, \((1 - f_1)\) is assumed to be 1 even though \(M = 6\) and \((1 - f_1) = \frac{M - m}{M}\) is considerably less than 1.

Suppose we are interested in knowing what the sampling variance is for the following three allocations of a sample of four terminal branches assuming simple random sampling at both stages:

<table>
<thead>
<tr>
<th>Allocation</th>
<th>No. of Trees</th>
<th>(n_i = \bar{n})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

To find the variances for the three allocations we need part of the results in Table 2.6. The relevant results, \(N_i\), \(Y_i\), and \(S^2_{2i}\), from Table 2.6 are included in Table 4.3 along with some other information that will be used later.
In each allocation, \( n_i \) is constant (the same for all trees) which means that \( \frac{n_i}{N_i} \) is not constant and the branches do not have equal probability of being in the sample. Thus, the estimator, Eq. 4.6, and its variance, Eq. 4.7, are applicable. The variances for the three allocations are presented in Table 4.4.

**Exercise 4.9** Refer to the data presented in Table 4.3, columns \( N_i, Y_i, \) and \( S^2_{2i} \) and perform the calculations that are needed to obtain the results presented in Table 4.4 for \( m = 2 \) and \( n_i = \bar{n} = 2 \). Assume that \( f_1 \) is negligible.

**Exercise 4.10** Complete the following table:

<table>
<thead>
<tr>
<th>( n )</th>
<th>( m )</th>
<th>( \bar{n} )</th>
<th>( V(\hat{y}) )</th>
<th>Among psu's</th>
<th>Within psu's</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1167.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>4</td>
<td>306.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If you understand the variance formula 4.7 and the results in Table 4.3, this table can be completed very easily. First, fill in the "Among psu's" column by copying the appropriate numbers from Table 4.4. Consider how to fill in the "Within psu's" column by making simple changes in the within psu components in Table 4.4. Study the results. For a constant value of \( \bar{n} \) and an increase in \( m \) from 1 to 4 there is a 75 percent reduction in the variance of \( \hat{y} \); but, for a constant \( m \), increasing \( \bar{n} \) from 1 to 4 reduces the variance of \( \hat{y} \) by less than 50 percent. 123
Exercise 4.11  One of the numbers in Table 4.4 is the sampling variance for \( m = 2 \) and \( n_i = N_i \). What is the number?

Exercise 4.12  Find the probability that any given terminal branch on tree No. 1 has of being in the sample when \( m = 2 \) and \( n_i = 2 \) for all trees. What is the probability for tree No. 3? Is the unequal probability something to be concerned about? In what ways?

It is of interest to compare the variance for a simple random (single-stage) sample of 4 branches with the variances of \( \hat{y} \) in Table 4.4. The variance among the 135 branches is 1,762 (see Table 2.6). Hence, the variance of the mean of a sample of 4 branches is \( \frac{1762}{4} = 440 \), disregarding the fpc. The answer, 440, is less than the variances of \( \hat{y} \) in Table 4.4. This is expected with the possible exception of the allocation \( m = 4 \) and \( n_i = 1 \), which has a variance equal to 583.5. However, when one recognizes in the specified two-stage plans that all branches do not have the same probabilities of selection, it is reasonable to expect that the answer for simple random sampling would be less than 583.5.

Suppose we wish to give every branch an equal chance of being in the sample. Considering samples of 4 branches the overall sampling fraction would be \( \frac{4}{135} \). If we specify that \( m = 2 \), then \( f_1 = \frac{1}{3} \) and all \( \frac{n_i}{N_i} \) (or \( f_2 \)) should equal \( \frac{4}{45} \). Since the \( N_i \) are small and the \( n_i \) must be integers, it is not possible to have all \( \frac{n_i}{N_i} \) exactly equal to \( \frac{4}{45} \). This presents a type of practical problem that often occurs when working with small integers. Ways
of dealing with this problem will not be discussed at this point. Instead, we will proceed as though the fraction \( \frac{n_i}{N_i} \) is sufficiently close to \( \frac{4}{45} \) to warrant use of the unweighted average of the sample data as the estimator and the variance formula 4.10. Assuming \( (1 - f_1) = 1 \), for reasons explained above, and substituting the numerical values of \( S_1^2 \) and \( S_2^2 \) in 4.10, we have:

\[
V(\hat{y}) = \frac{1}{m} \left[ 917.1 + (1 - f_2) \frac{1367}{\bar{n}} \right] \quad (4.11)
\]

When \( m = 2 \) and \( f_2 = \frac{4}{45} \), the value of \( \bar{n} \) is 2 and the variance of \( \hat{y} \) is 769.9. This answer compares with 797.6 in Table 4.4.

Exercise 4.13 Verify the numbers, 917.1 and 1367, in Eq. 4.11.

It is often desirable to specify that all ssu's in the population have an equal chance of being in the sample. As discussed above, one way of fulfilling this requirement is to select psu's with equal probability and apply a constant sampling fraction at the second stage of sampling. But, when the sizes of the psu's vary widely, this method often has two important disadvantages: (1) Variance associated with variation in the sizes of the psu's is included in the variance of an estimate unless such variation is reduced by design. Notice that \( S_1^2 \) in 4.7 is the variance among psu totals rather than the variance among psu means. Incidentally, an auxiliary variable(s) might be useful in reducing the sampling variance associated with the first stage of sampling. (2) When the second-stage sampling
fraction, \( \frac{n_i}{N_i} \), is constant, \( n_i \) is proportional to \( N_i \) and the workload varies from one PSU to another. For many surveys, it is important for reasons of economy that \( n_i \), rather than \( \frac{n_i}{N_i} \) be constant. Selecting PSUs with PPS is often very helpful in overcoming these disadvantages.

**Exercise 4.14** Under the plan of applying a sampling fraction of \( \frac{4}{45} \) to each tree that is selected, suppose that trees numbered 1 and 3 are selected. Find the values of \( n_i \) for these two trees where \( n_i = \frac{4}{45} N_i \) rounded to the nearest integer. Also, find \( n = \sum n_i \). Do the same assuming trees numbered 2 and 4 are selected. This illustrates that the size of the sample, \( n = \sum n_i \), is a random variable. Also, in this case, \( \frac{n_i}{N_i} \) cannot be exactly constant. One should consider whether there is an appreciable bias in the estimator (4.9). Use (4.6) instead of (4.9) unless there is assurance that any bias in (4.9), owing to unequal probabilities of the PSU's being in the sample, is negligible.

### 4.4 SELECTION OF PSU'S WITH PPS

Consider a sample of \( m \) PSUs selected with replacement and with selection probabilities \( P_1, P_2, \ldots, P_m \) (See section 1.1.2 in Chapter I). Let \( n_i \) be the size of a simple random sample of SSUs that is to be selected from the \( i^{th} \) PSU in the event that it is selected. If, by chance, the \( i^{th} \) PSU is selected a second time another sample of \( n_i \) SSUs is selected. For a sample of \( n \) PSUs the estimator is:

\[
y = \left( \frac{1}{N} \right) \left( \frac{1}{m} \right) \sum_{i}^{m} \frac{N_i y_i}{P_i}
\]  

(4.12)
Remember to interpret "i" as an index of the psu's selected by the m random draws. Notice that $N_i \bar{y}_i$ is an estimate of a psu total and that $\frac{N_i \bar{y}_i}{p_i}$ is an estimate of the population total, $Y$, based on a sample of one psu and a simple random sample of $n_i$ ssu's within it. Thus, there are $m$ estimates of $Y$ and $\left(\frac{1}{m}\right) \sum_i \frac{N_i \bar{y}_i}{p_i}$ is an average of these estimates. The factor $\frac{1}{N}$ makes $\hat{y}$ an estimator of $\bar{Y}$. The variance of $\hat{y}$, in Eq. 4.12, is:

\[
V(\hat{y}) = \frac{1}{m} \left[ \sigma_1^2 + \frac{1}{N^2} \sum_i \left( \frac{1}{p_i} \right)(1-f_{2i}) \frac{N_i S_{2i}^2}{n_i} \right]
\]  

where \[
\sigma_1^2 = \frac{1}{N^2} \sum_i p_i (\frac{Y_i}{p_i} - Y)^2
\]
and \[
S_{2i}^2 = \frac{1}{N_i - 1} \sum_j (Y_{ij} - \bar{y}_i)^2
\]

**Exercise 4.15** Compare $\frac{\sigma_1^2}{m}$ in 4.13 with the variance of $\hat{y}_4$ in Table 1.1, using the alternative expression for $\sigma_4^2$ in the variance of $\hat{y}_4$. Change the notation used in Chapter 1 to conform to the notation used for psu's. This gives:

\[
V(\hat{y}_4) = \left(\frac{1}{m}\right) \left(\frac{1}{M^2}\right) \sum_i \frac{Y_i}{p_i} (\frac{1}{p_i} - Y)^2
\]
Why is this expression for \( V(\hat{Y}_4) \) different from the between psu part of the variance in 4.13? In terms of the notation for two-stage sampling \( \hat{Y}_4 \) is an estimate of \( \bar{Y} \) rather than \( \bar{Y} \). Change \( \hat{Y}_4 \) so it will be an estimator of \( \bar{Y} \) and make the corresponding change in \( V(\hat{Y}_4) \). Your answer should agree exactly with \( \frac{\sigma_1^2}{m} \) in (4.13).

Notice the correspondence between \( S_1^2 \) in Eq. 4.7 and the variance of \( \hat{Y}_1 \), plan 1, in Chapter I; also, notice the correspondence between \( \sigma_1^2 \) in Eq. 4.13 and the variance of \( \hat{Y}_4 \), plan 4, in Chapter I. The discussion in Chapter I of the efficiency of plan 4 compared to plan 1 is relevant to the first stage of sampling. If \( N_i \) is a good measure of size, \( \sigma_1^2 \) will be considerably less than \( S_1^2 \).

Compare the components of variance in Eq. 4.7 and Eq. 4.13 which pertain to the second stage of sampling. The only difference is a reflection of the difference in the probabilities of selection at the first stage. When the probabilities are equal, \( P_i = \frac{1}{M} \) and substituting \( \frac{1}{M} \) for \( P_i \) in 4.13 gives 4.7.

In Eq. 4.4, \( f_{ij} \) was expressed as the probability that any given ssu has of being in a sample assuming the sample at both stages was simple random sampling without replacement. This equation now needs modification to be in accord with sampling at the first stage with unequal probability and with replacement. An appropriate probability equation is:

\[
f_{ij}' = P_i f_{2i} = P_i n_i \frac{n_i}{iN_i}
\]  
(4.14)
where $P_i$ is the selection probability, at any given random draw, for the $i^{th}$ psu in the population,

$f_{2i}$ as defined before, is the sampling fraction within the $i^{th}$ psu of the population, and

$f_{ij}$ is the probability which the $j^{th}$ ssu in the $i^{th}$ psu of the population has of being in a sample obtained by selecting one psu with pps and selecting a simple random sample of $n_i$ ssu's within the selected psu.

It is in the context of the probability Eq. 4.14 that the estimator, 4.12 and its variance, 4.13, are applicable, assuming $m$ independent random selections of psu's.

The estimator, Eq. 4.12, and its variance, Eq. 4.13, are for any given set of selection probabilities at the first stage and any given set of sample sizes, $n_i$, at the second stage. An important special case exists when $f_{ij}$, in Eq. 4.14 is held constant and when the psu's are selected with probabilities proportional to $N_i$, that is, when $P_i = \frac{N_i}{N}$. By letting $f'$ be the constant value of $f_{ij}$, we obtain the following results from Eq. 4.14:

$$n_i = f'N = \hat{n}$$

and

$$f_{2i} = \frac{\hat{n}}{N_i}$$

That is, the sample size within a psu is constant, and, since $f'_{ij}$ is also constant, the sample is self-weighted.
The estimator and its variance become:

\[ \hat{y} = \frac{\sum y_{ij}}{n} \]  

(4.15)

and

\[ V(\hat{y}) = \frac{1}{m} \left[ \sigma_1^2 + \frac{1}{n} \sum_{i=1}^{M} N_i (1 - f_{2i}) S_{2i}^2 \right] \]  

(4.16)

where \( \sigma_1^2 = \frac{1}{N} \sum_{i=1}^{M} N_i (\bar{y}_i - \bar{y})^2 \)

For computational purposes one might use:

\[ \sigma_1^2 = \frac{1}{N} \sum_{i=1}^{M} \frac{Y_i^2}{N_i} - \left( \frac{\bar{Y}}{N} \right)^2 \]

and

\[ \frac{1}{N} \sum_{i=1}^{M} N_i (1 - f_{2i}) S_{2i}^2 = \frac{1}{N} \sum_{i=1}^{M} S_{2i}^2 - \frac{n}{N} \sum_{i=1}^{M} S_{2i}^2 \]

Exercise 4.16  Show that Eqs. 4.12 and 4.13 reduce to 4.15 and 4.16 when \( P_i = \frac{N_i}{N} \) and \( n_i = \bar{n} \).

When the \( N_i \) are not known, estimates of \( N_i \) or a suitable measure of size might be used in place of \( N_i \). In this case, assuming \( f_{ij}' = f' \), the sampler would choose a value of \( f' \) such that \( f'N \) is the desired average size of sample from a psu. Since the selection probabilities for psu's are known, the second-stage sampling fraction \( f_{2i} = \frac{f'}{P_i} \) would be calculated for each selected psu. Application of these second-stage sampling fractions gives a self-weighted sample. The \( n_i \) will be nearly equal if the measure of size is close to being proportional to \( N_i \). The estimator, Eq. 4.12, and its variance, Eq. 4.13, are applicable. They could be modified by making use of the fact that

\[ P_i f_{2i} = P_i \frac{n_i}{N_i} = f' \]
4.4.1 NUMERICAL EXAMPLE

Exercise 4.17 With reference to the apple tree example, we found for simple random sampling at both stages that the sampling variance was 797.6 when \( m=2 \) and \( n_1 = \tilde{n}=2 \) (See Table 4.4). For comparative purposes, find the sampling variance for \( m=2 \) and \( \tilde{n}=2 \) when the trees are selected with probabilities proportional to \( N_1 \). The data needed are found in Table 4.3, columns headed \( N_1, Y_1 \), and \( S^2_{2i} \). Find the values of \( \sigma^2_1, \frac{1}{N_1}N_1S^2_{2i} \), and \( \frac{1}{N_1}S^2_{2i} \), then compute the variance of \( \hat{y} \) for \( m=2 \) and \( \tilde{n}=2 \). Ans. 532.6.

Substituting results from exercise 4.16 in Eq. 4.16 gives:

\[
V(\hat{y}) = \frac{439.7}{m} + \frac{1367 - \tilde{n}(57.99)}{m\tilde{n}}
\]

For \( m=2 \) the between psu variance, \( \frac{439.74}{2} = 219.9 \), compares with 458.6 (see Table 4.4) when two psu's are selected with equal probability. As indicated by this result, selecting psu's with pps is often very important in reducing the between psu component of variance. For \( m=2 \) and \( \tilde{n}=2 \) the within psu component in Eq. 4.17 is equal to 312.8 which compares with two other results that were obtained when the trees are selected with equal probabilities: 339.0 when \( n_1 = 2 \), and 311.4 when \( \frac{n_1}{N_1} \) is constant and \( \tilde{n}=2 \). The first result, 339.0, was recorded in Table 4.4 and the second, 311.4, is readily obtained by Eq. 4.11.
Suppose that one tree is selected with probability \( \frac{N_i}{N} \) and that one branch is selected from it with equal probability. In this case, \( m=1 \), and \( \hat{n}=1 \), and the variance of \( \hat{y} \) according to variance formula 4.17 is 1748.8. The probability of selecting any given branch in the population is \( \left( \frac{1}{N} \right) \left( \frac{1}{N} \right) \). This is a special case of two-stage sampling that is the same as a single-stage, simple random sample of one branch. We found earlier that the variance among the 135 branches was 1762. The exact variance for a simple random sample of one branch is:

\[
\frac{135 - 1}{135} \cdot \frac{1762}{1} = 1748.8
\]

4.5 UNEQUAL PROBABILITY OF SELECTION AT BOTH STAGES

As a further exposition of the theory for two-stage sampling, suppose a sample of trees is selected with replacement and with selection probabilities proportional to trunk size. Also, suppose that the method of sampling at the second stage is the random-path method, RP-PPS, that was discussed in Chapter III. You may recall that the random-path method was presented in the context of sampling with replacement.

When the sampling at both stages is with unequal probability, the estimator of the population total \( Y \) is:

\[
\hat{y}_t = \sum_{i} \sum_{j} \frac{m n_i \left( \frac{y_{ij}}{p_{ij}} \right)}{n}
\]

(4.18)

where

\[
m = \sum n_i
\]

\[
p_{ij} = p_i p(j|i)
\]
\( p_i \) is the selection probability for the \( i^{th} \) psu in the sample, and
\( p(j|i) \) is the selection probability for the \( j^{th} \) ssu given that its psu has been selected.

Consider the quantity \( \frac{y_{ij}}{p_{ij}} \) in the estimator. When the value for a unit in the sample (in this case, \( y_{ij} \)) is divided by its selection probability (in this case, \( p_{ij} \)) the quotient is an estimate of the population total. Therefore, \( \hat{y}_t \) in Eq. 4.18 is an average of \( n \) estimates of \( Y \), one estimate from each branch in the sample.

The subscript "t" was added to \( \hat{y} \) because it is an estimator of \( Y \) rather than \( \bar{y} \). Notice that the estimator does not contain \( N \). In practice, one finds many populations to be sampled where \( N \) is unknown. An estimate, \( \hat{N} \), of \( N \) might be made from a sample and, if needed, \( \hat{Y} \) could be used as an estimate of \( N \). An estimator of \( N \) is obtained by substituting "1" for \( y_{ij} \) in 4.18.

**Exercise 4.18** Suppose, for \( m=3 \), and \( n_i=2 \), that application of the above method to the apple tree population gives the following sample:

<table>
<thead>
<tr>
<th>Population index values of ( i ) and ( j )</th>
<th>Sample index values of ( i ) and ( j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Path</td>
<td>Tree Path</td>
</tr>
<tr>
<td>1  2-2-3</td>
<td>3  1</td>
</tr>
<tr>
<td>1  4-1</td>
<td>3  2</td>
</tr>
<tr>
<td>3  1-2-1-2</td>
<td>1  1</td>
</tr>
<tr>
<td>3  2-3</td>
<td>1  2</td>
</tr>
<tr>
<td>3  3-1-2</td>
<td>2  1</td>
</tr>
<tr>
<td>3  1-2-1-2</td>
<td>2  2</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccc}
\text{Tree} & \text{Path} & \text{Tree} & \text{Path} \\
1 & 2-2-3 & 3 & 1 \\
1 & 4-1 & 3 & 2 \\
3 & 1-2-1-2 & 1 & 1 \\
3 & 2-3 & 1 & 2 \\
3 & 3-1-2 & 2 & 1 \\
3 & 1-2-1-2 & 2 & 2 \\
\end{array}
\]

| \( p_i \) | \( p(j|i) \) | \( y_{ij} \) |
|------------|--------------|--------------|
| 0.07035    | 0.15996     | 59.5         |
| 0.07035    | 0.07779     | 7.0          |
| 0.2312     | 0.04864     | 139.3        |
| 0.2312     | 0.05757     | 125.0        |
| 0.2312     | 0.02081     | 31.3         |
| 0.2312     | 0.04864     | 139.3        |
Tree No. 3 and path 1-2-1-2 were selected twice. The selection probabilities \( P_i \) were proportional to \( X_i \), the trunk sizes which are presented in Table 4.3. Verify the values of \( P_i \). For tree No. 3 in the population, the conditional probabilities, \( P(j|i) \), are the probabilities in Column \( P_4 \) of Table 3.2. Thus, the above values of \( p(j|i) \) and \( y_{ij} \) for the branches in the sample from this tree were taken from Table 3.2. The values of \( p(j|i) \) and \( y_{ij} \) for tree No. 1 are from records not reproduced herein. Using Eq. 4.18 as the estimator, calculate the estimate of the total number of apples. The answer is 7873, which is an estimate of 7199, the total number of apples including "path" apples (See Table 4.3).

To find the variance of \( \hat{Y}_t \), refer to Eq. 4.13 and make two modifications:

1. For the first stage we want \( N^2 \sigma_1^2 \) instead of \( \sigma_1^2 \) because \( \hat{Y}_t \) is an estimator of \( Y \) rather than \( \bar{Y} \), and

2. For the second stage, the part of the formula representing the variance of an estimate of \( Y_i \) for a simple random sample of \( n_i \) needs to be changed. That is, \( \frac{N_i^2 S_i^2}{n_i} \) needs to be replaced by the corresponding variance for sampling within the \( i^{th} \) psu with pps. Also, \( \frac{1}{N} \) needs to be dropped. This gives:

\[
V(\hat{Y}_t) = \frac{1}{m} \sum_{i=1}^{M} \frac{Y_i}{P_i} \left( \frac{Y_i}{P_i} - Y \right)^2 + \sum_{i=1}^{M} \frac{1}{P_i} \frac{S_i^2}{n_i} \tag{4.19}
\]
where

\[ S_{ri}^2 = \frac{1}{N} \sum_j P(j|i) \frac{Y_{ij}}{P(j|i)} - Y_i^2 \]

The subscript \( r \) signifies random path.

For Tree No. 3 the values of \( P(j|i) \) and the values of \( \frac{Y_{ij}}{P(j|i)} \) are in columns \( P_4 \) and \( \hat{Y}_4 \) of Table 3.2. From these two columns the value of \( S_{ri}^2 \) for tree number 3 can be computed. The answer, 800,000 is recorded along with other values of \( S_{ri}^2 \) in the last column of Table 4.3.

Exercise 4.19 When \( n_i = \bar{n} \), the second term in \([\]\) of Eq. 4.19 becomes \( \frac{1}{n} \sum_i \frac{S_{ri}^2}{P_i} \). In the problem under consideration, \( P_i = \frac{X_i}{X} \) where \( X_i \) is trunk size. From the data in Table 4.3, find the value of \( \frac{S_{ri}^2}{P_i} \) and of \( \sum_i P_i \frac{Y_i}{P_i} - Y \). When your results are substituted in 4.19, you should have:

\[ V(\hat{y}_t) = \frac{1}{m} \left[ 5,322,000 + \frac{11,462,000}{n} \right] \]

Exercise 4.20 From the sample data given in Exercise 4.16 estimate the total number of terminal branches on the six trees. Ans. 122.0.

When the equation in Exercise 4.19 is divided by \( N^2 \) we obtain:

\[ V(\hat{y}) = \frac{1}{m} \left[ 292 + \frac{629}{n} \right] \]

135
To summarize, the following variance equations have been obtained for three alternative two-stage plans for sampling the small population of apple trees:

\[
V(\hat{y}) = \frac{1}{m} \left[ 917.1 + \frac{(1-f_2) \times 1367}{\bar{n}} \right]
\]

for simple random sampling at both stages, where \( \frac{n_i}{N_i} \) is constant and equal to \( f_2 \) and \( 1-f_1 \) was assumed to be equal to 1,

\[
V(\hat{y}) = \frac{1}{m} \left[ 439.7 + \frac{1367 - \bar{n}(58.0)}{\bar{n}} \right]
\]

for sampling trees with probability proportional to \( N_i \) and a simple random sample of \( \bar{n} \) branches from each selected tree, and

\[
V(\hat{y}) = \frac{1}{m} \left[ 292 + \frac{629}{\bar{n}} \right]
\]

for sampling trees with probability proportional to \( X_i \) (trunk size) and application of the RP-PPS method to the sample trees.

The results are too limited to provide a basis for generalization.
Table 4.1 Representation of Population Data for Two Stage Sampling

<table>
<thead>
<tr>
<th>psu</th>
<th>ssu</th>
<th>psu total</th>
<th>psu mean</th>
<th>Within psu variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Y_{11}, \ldots, Y_{1j}, \ldots, Y_{1N_1})</td>
<td>(Y_1)</td>
<td>(\bar{Y}_1)</td>
<td>(S^2_{21} = \frac{\sum_{j} (Y_{1j} - \bar{Y}_1)^2}{N_1 - 1})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>i</td>
<td>(Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{iN_i})</td>
<td>(Y_i)</td>
<td>(\bar{Y}_i)</td>
<td>(S^2_{21} = \frac{\sum_{j} (Y_{ij} - \bar{Y}_i)^2}{N_i - 1})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>M</td>
<td>(Y_{M1}, \ldots, Y_{Mj}, \ldots, Y_{MN_M})</td>
<td>(Y_M)</td>
<td>(\bar{Y}_M)</td>
<td>(S^2_{2M} = \frac{\sum_{j} (Y_{Mj} - \bar{Y}_M)^2}{N_M - 1})</td>
</tr>
</tbody>
</table>

\(1/\)

A single bar "\(\cdot\)" is used for an average of psu totals and a double bar "\(\bar{\cdot}\)" indicates an average of secondary units. A subscript 1 or 2 affixed to \(S^2\) indicates first or second stage variance. See definitions below.

- \(Y_{ij}\) is the value of the characteristic \(Y\) for the \(j^{th}\) ssu in the \(i^{th}\) psu,
- \(N_i\) is the population number of ssu's in the \(i^{th}\) psu,
- \(N = \sum_i N_i\) is the number of ssu's in the population,
- \(\bar{Y}_1 = \frac{\sum_j Y_{ij}}{N_1}\) is the average number of ssu's per psu,
- \(\bar{Y}_M = \frac{\sum_j Y_{Mj}}{N_M}\) is the average number of ssu's per psu,
- \(S^2_{21} = \frac{\sum_j (Y_{1j} - \bar{Y}_1)^2}{N_1 - 1}\) is the variance among ssu's in the \(i^{th}\) psu, and
- \(S^2_{2M} = \frac{\sum_j (Y_{Mj} - \bar{Y}_M)^2}{N_M - 1}\) is the variance among psu's on the basis of one ssu.
Table 4.2 Components of Variation for a Hypothetical Population

<table>
<thead>
<tr>
<th>psu</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>(Y_i)</th>
<th>(\bar{y}_i)</th>
<th>(S^2_{2i})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>45</td>
<td>51</td>
<td>20</td>
<td>35</td>
<td>218</td>
<td>43.6</td>
<td>308.8</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>27</td>
<td>82</td>
<td>39</td>
<td>18</td>
<td>198</td>
<td>39.6</td>
<td>620.3</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>25</td>
<td>21</td>
<td>30</td>
<td>28</td>
<td>118</td>
<td>23.6</td>
<td>40.3</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>48</td>
<td>72</td>
<td>63</td>
<td>88</td>
<td>326</td>
<td>65.2</td>
<td>242.7</td>
</tr>
</tbody>
</table>

Values of \((Y_{ij} - \bar{Y})\)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>2</td>
<td>8</td>
<td>-23</td>
<td>-8</td>
</tr>
<tr>
<td>2</td>
<td>-11</td>
<td>-16</td>
<td>39</td>
<td>-4</td>
<td>-25</td>
</tr>
<tr>
<td>3</td>
<td>-29</td>
<td>-18</td>
<td>-22</td>
<td>-13</td>
<td>-15</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>5</td>
<td>29</td>
<td>20</td>
<td>45</td>
</tr>
</tbody>
</table>

Values of \((\bar{y}_i - \bar{Y})\)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
</tr>
<tr>
<td>3</td>
<td>-19.4</td>
<td>-19.4</td>
<td>-19.4</td>
<td>-19.4</td>
<td>-19.4</td>
</tr>
<tr>
<td>4</td>
<td>22.2</td>
<td>22.2</td>
<td>22.2</td>
<td>22.2</td>
<td>22.2</td>
</tr>
</tbody>
</table>

Values of \((Y_{ij} - \bar{y}_i)\)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.4</td>
<td>1.4</td>
<td>7.4</td>
<td>-23.6</td>
<td>-8.6</td>
</tr>
<tr>
<td>2</td>
<td>-7.6</td>
<td>-12.6</td>
<td>42.4</td>
<td>-0.6</td>
<td>-21.6</td>
</tr>
<tr>
<td>3</td>
<td>-9.6</td>
<td>1.4</td>
<td>-2.6</td>
<td>6.4</td>
<td>4.4</td>
</tr>
<tr>
<td>4</td>
<td>-10.2</td>
<td>-17.2</td>
<td>6.8</td>
<td>-2.2</td>
<td>22.8</td>
</tr>
</tbody>
</table>

\[S^2 = 487.053\] is the variance among the 20 values of \((Y_{ij} - \bar{Y})\)

\[S^2_1 = 293.707\] is the variance among the 4 values of \((\bar{y}_i - \bar{Y})\)

\[S^2_2 = 303.025\] is the average of the variances of \((Y_{ij} - \bar{y}_i)\) within psu's. Within the first psu the variance is:

\[
\frac{23.4^2 + 1.4^2 + 7.4^2 + (-23.6)^2 + (-8.6)^2}{4} = 308.8.
\]
Table 4.3 Summary Data for Six Apple Trees

<table>
<thead>
<tr>
<th>Tree</th>
<th>N_i</th>
<th>Y_i</th>
<th>S^2_{zi}</th>
<th>X_i</th>
<th>Y'_i</th>
<th>S^2_{ri}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>213</td>
<td>259</td>
<td>7.0</td>
<td>214</td>
<td>22,000</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>1,388</td>
<td>1,147</td>
<td>20.0</td>
<td>1,448</td>
<td>478,000</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>1,850</td>
<td>2,184</td>
<td>23.0</td>
<td>1,901</td>
<td>800,000</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>1,592</td>
<td>3,106</td>
<td>16.5</td>
<td>1,658</td>
<td>350,000</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>402</td>
<td>241</td>
<td>13.5</td>
<td>403</td>
<td>79,000</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>1,528</td>
<td>892</td>
<td>19.5</td>
<td>1,575</td>
<td>513,000</td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>6,973</td>
<td>99.5</td>
<td>7,199</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The values of N_i, Y_i, and S^2_{zi} are from Table 2.6. The values of Y_i and S^2_{zi} are labeled Y_h and S^2_{Yh} in Table 2.6. "Path apples" are not included in Y_i and S^2_{zi}. The values of Y'_i and S^2_{ri} include the path apples and are taken from Table 3.3. The subscript "r" refers to random path.

DS-EP and RP-PPS refer to the method of sampling a tree as discussed in Chapter III.

Table 4.4 Variances for Alternative Sample Allocations

<table>
<thead>
<tr>
<th>Components</th>
<th>Among psu's</th>
<th>Within psu's</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>n</td>
<td>V(\hat{y})</td>
</tr>
<tr>
<td>(1) 1</td>
<td>4</td>
<td>1225.8</td>
</tr>
<tr>
<td>(2) 2</td>
<td>2</td>
<td>797.6</td>
</tr>
<tr>
<td>(3) 4</td>
<td>1</td>
<td>583.5</td>
</tr>
</tbody>
</table>

1/ Assumes f_1 is negligible.