# Two-Stage Cluster Sampling

## Forest Survey

Forestry researchers used a two-stage cluster sampling design to estimate the total volume of 9000 black cherry trees in a forest region. The region was first divided into 100 transects. Then 3 of these transects were selected by a probability sampling design. Finally within each of these transects some of the trees were selected using simple random sampling. The total number of trees within each of these three transects as well as the number of sampled trees and the mean volume (in cubic feet) of those sampled trees are shown in the table below.

| Transect | Number of Trees | | Mean Volume |
| --- | --- | --- | --- |
| | Total Trees | Sampled Trees | |
| 1 | 100 | 20 | 10 |
| 2 | 110 | 30 | 12 |
| 3 | 90 | 25 | 11 |

Use the information given above to answer in the following concerning the estimation of the total volume of all the black cherry trees in the region.

1. Assume that the transects were selected using *simple random sampling*. Confirm that the estimate of the total volume of all the trees in the region using the *unbiased* estimator is 110333.3 cubic feet, and that the estimate using the *ratio* estimator is 99300 cubic feet.

2. Assume that the transects were selected using *sampling with replacement* with *probabilities proportional to size*. Confirm that the estimate of the total volume of all the trees in the region using the *Hansen-Hurwitz* estimator is 99000 cubic feet.

## Educational Testing

Consider a cluster sampling design using a population of 2000 second grade students in a school district. Researchers would like to estimate the average score of these students on a particular test based on a sample of students since the testing is expensive. They decided to use a two-stage cluster sampling design where they selected a sample of 3 classrooms of students out of the 100 classrooms in the district, and then selected some of the students from each of these selected classrooms using simple random sampling. The table below shows the mean test score for the sampled students from each of the sampled classrooms. It also shows the number of students in the classroom as well as the number of students that were sampled for testing.

| Classroom | Number of Students | | Mean Score |
| --- | --- | --- | --- |
| | Total Students | Sampled Students | |
| 1 | 16 | 5 | 70 |
| 2 | 20 | 7 | 80 |
| 3 | 19 | 6 | 75 |

Use the information given above to answer in the following concerning the estimation of the mean test score for all the second grade students in the school district.

1. Assume that the classrooms were selected using *simple random sampling.* Confirm that the estimate of the mean test score for all second grade students in the district using the *unbiased* estimator is approximately 69.1, and that the estimate using the *ratio* estimator is approximately 75.4.

2. Assume that the classrooms were selected using *sampling with replacement* with *probabilities proportional to size.* Confirm that the estimate of the mean test score for all second grade students in the district using the *Hansen-Hurwitz* estimator is 75.

## Optimum Sample Sizes for Educational Testing

Suppose that based on either a previous survey or a pilot survey the researchers in the previous problem had estimates of the between-classroom and within-classroom mean squares. Assume that the researchers also have a fixed budget of 500 units, and that the estimated cost per classroom selected is 20 units, and the estimated cost per student selected and tested is 5 units (here the cost units represents both financial cost as well as time). The table below shows the approximate optimal sample sizes for the number of students per classroom to sample as well as the number of classrooms to sample (assuming simple random sampling at both stages) for four hypothetical populations.

| | Mean Square | | | |
| Population | between | within | $m_{\mathrm{opt}}$ | $n_{\mathrm{opt}}$ |
| --- | --- | --- | --- | --- |
| A | 62.5 | 50.00 | 17.89 | 4.57 |
| B | 250.0 | 25.00 | 2.98 | 14.33 |
| C | 500.0 | 15.00 | 1.57 | 17.95 |
| D | 750.0 | 6.25 | 0.82 | 20.75 |

Not all classrooms are the same size, but they are usually around 20 students. So the calculations use $\bar{M} = 20$ in their calculations. Confirm the optimal sample sizes shown above (within rounding error).