Stratified Random Sampling Homework

Estimation and Allocation

Educational psychologists wanted to know the mean performance on a cognitive test for the 1000 students at a school. The test is expensive and time-consuming and so it was decided to administer the test to only a sample of students and to use these data to estimate the mean test score for all students at the school. A sample of 200 students was selected using stratified random sampling. The strata were defined by classifying the students as high-, medium-, or low-achieving students based on other test scores that were on record for all students at the school. The 200 sampled students were given administered the cognitive test. The table below shows the number of students in each stratum, the number of students sampled from each stratum, and the mean and standard deviation of the test scores of the students that were sampled from each stratum.

Level	N_j	n_j	$\bar{y}_j$	s_j
High	300	50	80	10
Medium	500	100	70	15
Low	200	50	50	20

Use this information to answer the following questions.

1. Confirm that the estimate of the mean test score for all of the students at the school is 69, that the estimate of the variance of the estimator used to obtain this estimate is approximately 0.84, that the standard error is approximately 0.92, that the bound on the error of estimation is approximately 1.83, and so the confidence interval for the mean test score for all students at the school is approximately 69 $\pm$ 1.83.

2. The survey described above was conducted several years ago. The educational psychologists are planning a new survey this year. The composition of the school has changed slightly. The school has grown to a total of 1300 students. Of these, 400 are classified as high-achieving, 600 are classified as medium-achieving, and 300 are classified as low-achieving students. Assume that it costs \$20 to test one student, regardless of their achievement level. Confirm that the optimum allocation for a fixed bound on the error of estimation of one point (i.e., $B = 1$) would be to sample approximately 106 high-achieving students, approximately 238 medium-achieving students, and approximately 159 low-achieving students.

3. It was determined that the allocation found in the previous problem was too expensive. In addition to a cost of \$20 per tested student, there is an overhead cost of \$2000 for the specialists administering the test. Confirm that the optimum allocation for a fixed total cost of \$10000 would be to sample approximately 84 high-achieving students, approximately 189 medium-achieving students, and approximately 126 low-achieving students.

Forestry researchers conducted a survey to estimate the number of trees of a particular species in a region of forest. They used a stratified random sampling design. The region had been divided into 500 units, each with an area of one hectare. Each unit had also been classified as either "low-elevation" or "high-elevation" based on the average elevation within each unit. A total of 400 units were classified as low-elevation, and 100 units were classified as high-elevation. The researchers selected 50 low-elevation units using simple random sampling, and sent teams out to count the number of trees in each sampled unit. The average number of trees per sampled low-elevation unit was 100 trees, and the standard deviation was 10 trees. The researchers also selected 50 high-elevation units using simple random sampling, and sent teams out to count the number

of trees in each of these sampled units. The average number of trees per sampled high-elevation unit was 25 trees, and the standard deviation was 5 trees.

1. Confirm that the estimate of the total number of trees in the region is 42500 trees, that the estimate of the variance of the estimator used to obtain this estimate is 282500, that the standard error is approximately 532 trees, that the bound on the error of estimation is approximately 1063 trees, and that the confidence interval for the total number of trees in the region is approximately 42500 $\pm$ 1063 trees.

2. The forestry researchers are planning another survey in a similar forest region using stratified random sampling. This region has a total of 1000 one-hectare units, of which 300 are low-elevation and 700 are high-elevation. They estimate that the cost to survey a low-elevation unit is \$20 and the cost to survey a high-elevation unit is \$10 (high-elevation units are cheaper despite being less accessible because they tend to be more sparse with respect to flora and so faster to survey). Confirm that the optimum allocation for a fixed bound on the error of estimation of 1000 trees is to sample approximately 55 low-elevation units and approximately 91 high-elevation units.

3. Consider the previous problem where the forestry researchers are planning a new survey. Assume that the researchers can spend \$5000 on the survey, but they only need to pay for surveying the units so there is no overhead cost (any overhead cost will be paid for out of a separate budget). Confirm that the optimum allocation for a fixed total cost of \$5000 is to sample approximately 137 low-elevation units and approximately 226 high-elevation units.

Design Effect and Effective Sample Size

Consider the problem of estimating the parameter μ using one of three designs: simple random sampling, stratified random sampling, and cluster sampling (which we will talk about later in the course). Each are based on a total sample size of $n = 1000$ elements. Assume we can compute (or at least estimate) the variance of the estimator of μ for each of the three designs. The variances of the estimator under simple random sampling, stratified random sampling, and cluster sampling are $V_{\text{srs}}(\hat{\mu}) = 10$, $V_{\text{strat}}(\hat{\mu}) = 8$, and $V_{\text{clus}}(\hat{\mu}) = 16$, respectively. The stratified and cluster sampling designs are the *complex* sampling designs. Confirm that the *design effects* of the stratified and cluster sampling designs are 0.8 and 1.6, respectively. Also confirm that the *effective sample sizes* of the stratified and cluster sampling designs are 1250 and 625, respectively.

Double Sampling

Consider the previous example where educational psychologists wanted to know the mean performance on a cognitive test for the 1000 students at a school. In that example they had used the students' scores on other tests to stratify them into high-, medium-, and low-achieving students. To avoid confusion such a test will be referred to as the *stratifying test* as opposed to the *cognitive test* which is the focus of the survey. But now suppose that the educational psychologists did not have the scores on a stratifying test. This might be because they had not taken such a test, or because the students had taken such a test but the scores were not easily obtainable because they were in separate files and so would require excessive effort to obtain for all students at the school. So the researchers use a double sampling design. In the first phase of sampling a sample of 500 students was selected using simple random sampling. Stratifying test scores were then obtained for these students (either by testing these students or by going through their files) to classify them as high-, medium-, and low-achieving students. In that sample 145 students were classified as high-achieving, 246 were classified as medium-achieving, and 109 were classified as low-achieving. Then a stratified random sampling design was applied to the sample of 500 students, using approximate proportional allocation. The results of this survey are summarized in the table below.

Level	n_j	$\bar{y}_j$	s_j
High	58	78	11
Medium	98	72	14
Low	44	55	22

Confirm that the estimate of the mean score on the cognitive test of all students at the school is approximately 70.

Consider the previous example were forestry researchers were conducting a survey to estimate the number of trees of a particular species in a region of forest divided into a total of 500 units. Recall that they used elevation to stratify the units of forest. Now suppose that the elevation data were not available, so the researchers used a double sampling design to observe the elevation data in the field. First a simple random sample of 100 units was selected and field crews were sent to each unit to quickly inspect it and classify it as low-elevation or high-elevation. They classified 77 units as low-elevation and 23 as high-elevation. They then obtained a stratified random sample of these sampled units by obtaining a simple random sample of 40 of the low-elevation units, and a simple random sample of 20 of the high-elevation units. For these 60 sampled units field crews counted the number of trees. The mean number of trees per unit was 96 trees in the low-elevation units, and 28 trees in the high-elevation units. Confirm that the estimate of the total number of trees in the region is then 40180 trees.