

Note: please use R for all calculations. Include your R scripts with output.

1. Two dentists A and B make a survey of the teeth of 200 children in a village. Dr. A selects an SRS of 20 children and counts the number of decayed teeth for each child, with the following results: 0 0 0 0 0 0 0 1 1 1 1 2 2 3 3 4 5 9 10.

Dr. B examines all 200 children, recording merely whether each child has any decayed teeth or not. She finds 60 children have no decayed teeth.

Give an estimate along with an SE for the total number of decayed teeth among all 200 children in the village,

- (a) using only A's results.
 - (b) using post-stratification of A's results by B's results.
2. Pages 126-7 of the notes on double sampling and pages 166-7 of the text discuss a way to handle non-response in surveys by viewing the population as composed of two strata: those in the population who would respond to the survey if asked, and those who would not respond if asked. If a follow-up survey of a random sample of non-respondents is conducted (often by a different method), then we can view the entire process as a double-sample within each stratum. In the example in the notes, a survey is mailed to a random sample of 120 individuals in a population of 400 individuals. 30 out of the 120 people respond to the survey and of these 30, 20 respond "Yes" to some particular question of interest. A followup telephone survey is done on a random sample of 25 of the 90 nonrespondents. 20 of the 25 respond to the phone survey; of these 20, 4 answer "Yes" to the question of interest. Ignoring the 5 who didn't respond even to the phone survey, the notes show how to estimate the proportion of all 400 individuals in the population who would respond "Yes" to the question of interest. Calculate the SE of this estimate (use eq. (4) on p. 167 of the text). Compare this stratified double-sampling estimate and SE to the estimate and SE we would get if we simply treated the total of 50 respondents as an SRS from the population.
 3. Problem 3, page 197, with modifications:
 - (a) Do Problem 3, but give the standard error of the mean number of moose instead of the variance.
 - (b) Now suppose that the detection probability of $p = 0.89$ was estimated from another study independent of this one with a standard error of $SE(\hat{p})$. Using the data given in problem 3, estimate the standard error of the mean number of moose when $SE(\hat{p})$ is 0.01, 0.02, 0.03, 0.04, 0.05, 0.08, and 0.10 (note: you can do all these values simultaneously in R by creating a vector of these values and using this vector in the formula for the SE).

Write a couple of sentences comparing the standard errors resulting from the different choices for $SE(\hat{p})$.

4. Andrew McDonald's sweetclover plot data is in a file called `SweetCloverPlots.csv`. It contains data for the 36 plots in his population. There are 4 variables: the plot number (`Plot`), the 2-second visual estimate (`V2`), the 30-second visual estimate (`V30`), and the actual count (`Actual`). Andrew also provided the following information:

- It takes exactly 2 minutes to obtain the actual number of sweetclovers on any of the 36 plots.
- It takes exactly 4 minutes to obtain the 2 second visual estimates of all 36 plots.
- It takes exactly 18 minutes to obtain the 30 second visual estimates of all 36 plots.
- It takes exactly 8 minutes to obtain the data from one randomly chosen transect.

Note that “2-second” visual estimates take more than 2 seconds on average (the time to walk between plots is apparently included in his calculations)

- (a) Investigate the relationship between the two sets of visual counts and the actual counts. Does it appear that ratio estimation will be helpful? Would regression estimation be better?
- (b) Consider this a population of $N = 36$ plots so that all population parameters are known. Compare the sample sizes required to estimate the total number of plants to within ± 50 with 95% probability for (i) simple random sample (SRS) of n plots, (ii) ratio estimation using 2-second visual estimates for all plots and actual counts for SRS of n plots, and (iii) ratio estimation using 30-second visual estimates for all plots and actual counts for SRS of n plots.
- (c) Which of the plans in (a) has the smallest cost in terms of time?
- (d) Consider a double-sampling scheme where visual estimates are obtained on a random sample of n' plots and actual counts on a random subsample of n plots (on these plots, you would do the visual estimate first, then the actual count). What are the optimal ratios n/n' for the 2-second and 30-second visual estimates? Is it possible to estimate the total number of plants to within ± 50 with 95% probability with a double-sampling plan with these optimal ratios? (Remember that n' cannot be bigger than N .)
- (e) Suppose this were a pilot study for estimating the density of plants (mean plants per plot) in a much larger area where the number of plots N is essentially infinite (say, an irregular area roughly a mile square). Consider again using double sampling. Now, however, assume that there is a non-trivial fixed cost to locate a site (GPS?) and delineate it (the plots are 18 feet by 18 feet). Perhaps plots are taken systematically to minimize the travel time between plots. In any case, estimate a fixed cost for locating each site to add to the cost of surveying the site. Then recalculate the optimal double-sampling

ratios for the 2-second and 30-second visual estimates, and use these to calculate the sample sizes (n' and n) needed to estimate the mean count per plot to within ± 2 with 95% confidence (this is roughly $\pm 10\%$ based on the pilot study mean of 22 plants per plot). Which plan (2-sec or 30-sec) costs less?

5. Andrew's data from his line-intercept survey is in the file `SweetCloverTransects.csv`. It contains the "widths" (in inches, perpendicular to the transect) of the intercepted plants on each of 12 random parallel transects. The area is a square 108 feet on a side and the transects are parallel to one pair of sides. Note that there is no overlap data either within or between transects so that the "separate transects" estimates of the population total and SE must be used.
- (a) Estimate the total number of plants using the data from all 12 transects and calculate an SE.
 - (b) How many transects would be required to estimate the total number of plants to within ± 50 within 95% probability? Compare the cost of this plan to the plans from problem 3(b). Which plan is most efficient and which is least? Why do you think transects are so inefficient compared to plots?