

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

1. In the Exxon-Valdez oil spill trial, one important issue was the amount of lost harvest of fish, shellfish, seals, etc. suffered by native subsistence harvesters in the region. In particular, it was desired to estimate the per capita harvest loss and the total loss. Some data had been collected over the years on such harvests, both prespill and postspill. You will analyze data from just one community for one year (1989, the spill year) and estimate total harvest and per capita harvest for this community in 1989. A random sample of 42 households out of 67 was selected from the community. The number of individuals living in each household and the total household harvest (in pounds, estimated through detailed questioning about harvests of many individual species) were recorded.

The data are available on the website in the file `harvest.csv` which can be read into R using the `read.csv` command. The resulting data frame has two variables: `size` (number of people in household) and `harvest` (in pounds).

- (a) Estimate the per capita harvest for this community. Attach a standard error to your estimate and calculate a 95% confidence interval.
 - (b) Estimate the total harvest for the community. Attach a standard error to your estimate and calculate a 95% confidence interval.
 - (c) Now suppose the total number of individuals in the community is known to be 190 (do not use this information in part b). Estimate total harvest using both ratio and regression estimation and calculate SE's. Compare these results with your result in b); is there much of an improvement? Which seems more appropriate based on a plot of the data: ratio or regression estimation?
2. Kruuk et al. (1989) used a stratified sample to estimate the number of otter dens (called holts) along the coastline of Shetland, UK. The coastline (except for parts that were predominantly buildings) was divided into 237 5-km sections and each section was assigned to one of four terrain types. A random sample of sections within each stratum were chosen for counting. In each section chosen, researchers counted the number of otter dens in a 110-m-wide strip along the coast. The data are in the file `otters.csv` which can be read into R using `read.csv`. The population and sample sizes for the strata are given in the table below. Estimate the total number of otters along the coast in Shetland, along with a standard error and a 95% confidence interval. Note: use the `tapply` command in R to compute the variance (or any other function) of the observations by stratum. If `df` is a data frame with a quantitative variable `y` and a categorical variable `x`, then `tapply(dfy,dfx,var)` will give the variance of `y` for each of the categories of `x`.

Stratum	Total Sections	Sections Counted
1 Cliffs over 10 m	89	19
2 Agriculture	61	20
3 Not 1 or 2, peat	40	22
4 Not 1 or 2, nonpeat	47	21

3. A manufacturer of band saws wants to estimate the average repair cost per month for the saws he has sold to certain industries. He cannot obtain a repair cost for each saw, but he can obtain the total amount spent for saw repairs and the total number of saws owned by each industry. Thus he decides to use cluster sampling, with each industry as a cluster. The manufacturer selects a simple random sample of size $n = 20$ from the $N = 82$ industries he services. The data on total cost of repairs per industry and the number of saws per industry are as given in the accompanying table. [Problem taken from pages 274-275, Scheaffer, Mendenhall, & Ott.]
- Estimate the average repair cost per saw for the past month, and give the standard error of this estimate.
 - Estimate the total amount spent by the 82 industries on band saw repairs and give the standard error of this estimate.
 - After checking his sales records, the manufacturer finds that he sold a total of 690 band saws to these industries. Using this additional information, estimate the total amount spent on saw repairs by these industries, and give the standard error.
 - The manufacturer wants to estimate the average repair cost per saw for next month. How many clusters should he select for his sample if he wants to estimate this average cost to within \$2.00 with 95% confidence?

Industry	Number of Saws	Total Repair Cost for Past Month (\$)	Industry	Number of Saws	Total Repair Cost for Past Month (\$)
1	3	50	11	8	140
2	7	110	12	6	120
3	11	230	13	3	70
4	9	140	14	2	50
5	2	50	15	1	10
6	12	260	16	4	60
7	14	240	17	12	280
8	3	45	18	6	150
9	5	60	19	5	110
10	9	230	20	8	120

4. Show that in two-stage random sampling with equal-sized primary units, that if m secondary

units will be sampled within each primary unit, then the number of primary units n required to estimate the population mean per secondary unit to within d with probability $100(1 - \alpha)\%$ is

$$n = \frac{\sigma_b^2 + \left(\frac{\bar{M} - m}{\bar{M}}\right) \frac{\sigma_w^2}{m}}{\frac{d^2}{z^2} + \frac{\sigma_b^2}{N}}.$$

5. The data file `coots.txt` is a csv file containing data from Arnold's (1991) work on egg size and volume in American coot eggs in Minnedosa, Manitoba, as cited in Lohr, (1999), *Sampling Design and Analysis*, Brooks-Cole. The length and breadth (in mm.) of two randomly selected eggs from each of 184 nests (clutches) were recorded. The number of eggs in each clutch was also recorded. The volume (in cm^3) of each egg can be estimated by the formula $V = 0.000507 \times \text{length} \times \text{breadth}^2$ (where length and breadth are in mm). Estimate the mean volume per egg for coots in the study area along with an SE and a 99% confidence interval. Assume that these 184 nests are a random sample from the nests in the study area, although the total number of nests in the population is unknown (assume it's large).

Notes:

- The `tapply` function described in problem 2 may be useful for this problem also.
 - You will also have to think creatively in estimating the variance of the estimator since N is unknown (but assumed large) and since we don't know \bar{M} (the average clutch size for the population).
6. Use the coot data to estimate the total number of nests and number of eggs within each nest that should be sampled in order to estimate the mean volume per egg to within $.25 \text{ cm}^3$ with 99% probability. Assume for the purposes of this calculation that the clutches are all the same size, 10 eggs each, and that you will sample equal numbers of eggs within each selected clutch. Find the two-stage allocation that will achieve this goal with minimum cost, with cost measured as time. Assume that the time to find a nest is 20 times the time to randomly select and measure an egg carefully. (See also problem 5.)