

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

1. A simple random sample of 210 households was chosen from a city containing 16,828 households to estimate the proportion of households in that area who owned their home. Of the 210 households sampled, 132 reported that they owned their home.
 - (a) Estimate the population proportion of households in this city who own their home and give a 99% confidence interval for this population proportion.
 - (b) What sample size is required to estimate the proportion of people who own their own home to within 0.03 of the true proportion with 99% probability? Compute this estimate two ways: 1) using the estimate from the first sample and 2) assuming the “worst-case.”
2. #3, p. 63 of Thompson
3. For a hypothetical survey to determine the number of pileated woodpecker nests, the study area is divided into $N = 5$ plots. For the i^{th} plot in the population, y_i is the number of nests, while x_i is the number of “snags” (old trees that provide nesting habitat). The values for each population unit follow: $y_1 = 3, x_1 = 20$; $y_2 = 2, x_2 = 22$; $y_3 = 0, x_3 = 0$; $y_4 = 1, x_4 = 12$; $y_5 = 1, x_5 = 6$. Consider a simple random sampling design with sample size $n = 2$.
 - (a) Make a table listing every possible sample of size 2, the probability of selecting each sample, the estimate $N\bar{y}$ of the population total for each sample, and the ratio estimator $\hat{\tau}_r$ for each sample.
 - (b) Compute the expected value and variance for the estimator $N\bar{y}$.
 - (c) Compute the expected value and mean square error of the ratio estimator $\hat{\tau}_r$.
 - (d) Compare the two estimators for this population.
4. In the handout illustrating PPS sampling and Hansen-Hurwitz estimators for the farm data, a sample of $n = 10$ farms was drawn and estimates of various parameters computed assuming N was unknown. The last two estimates were the estimated mean number of workers per farm (about 3.425) and the estimated average size of the farms (about 7.555 pixels). Compute SE’s for these two estimates using two methods: the delta method, and bootstrapping (using R function `boot`). You’re using the sample I drew; the indices of these ten farms are printed in the output.
5. The file `states.csv` lists the number of counties, land area, and 1992 population for the 50 states plus the District of Columbia. You can read it into R using the `read.csv` command. This will create a data frame in R (see handout).
 - (a) Draw an SRS of 10 states. Use the sample to estimate the number of counties in the U.S. and find the standard error of your estimate. How do your estimate and confidence interval compare to the true value (which you can compute)?
 - (b) Examine the relationship between land area and the number of counties with a scatterplot. Does it appear that using land area as an auxiliary variable through a ratio estimator might be helpful in estimating the total number of counties?

- (c) Use a ratio estimator to estimate the number of counties in the U.S. for your sample in part a) using land area as the auxiliary variable and assuming total land area is known. Also compute an approximate SE. Compare to part a).
6. Unequal probability sampling is often unavoidable, so it is important to recognize when it occurs and to collect adequate information from respondents to calculate inclusion probabilities. For example, one method of surveying anglers on a stretch of river is to place questionnaire postcards on all cars parked along the road. An important question is then the probability of inclusion for each respondent; an analysis which ignores the differing probabilities of inclusion can bias estimates of parameters of interest (such as proportion of fly fishers, average number of fish caught, average age of anglers, etc.).

So suppose we are interested in anglers who fish a certain stretch of a river over a 4-week period. The first decision is what is the sampling unit – is it an individual angler or a trip by an individual angler? For practical reasons in computing inclusion probabilities, it is easier if we consider each separate trip by an angler as a separate unit. Therefore, an angler who fishes on 4 separate days during this time period contributes 4 angler-trips to the population. Second, we have to decide how to deal with multiple people coming in one vehicle; we'll assume that every individual in a vehicle is asked to fill out a questionnaire. We should also recognize that we're ignoring people who fish this stretch of river, but don't park their vehicle there (floaters, for example). Perhaps we could restrict our population to bank anglers. These and related issues are all important to address before we design the survey.

Now, suppose I am designing a survey of this stretch of the river. I have decided that I will restrict my population to those anglers who arrive in a vehicle that is parked along this stretch sometime between 0600 and 2100 hours (6am and 9pm) each day during these 4 weeks. My plan is as follows: I will randomly select 3 of the 8 weekend days during this period and 5 of the 20 weekdays. On each selected day, I will randomly (and independently of other days) select one of two starting times: 6 am or 11 am.

After I select one of the two start times, I will select a random starting place for my route in the following way. It takes me 7 hours to drive the stretch from south to north at a steady pace, distributing postcards and 3 hours to drive from the north end to the south if I'm not distributing postcards. So imagine a loop starting at the south end at 0 hours, reaching the north end at 7 hours and returning to the south end at 10 hours. Since I go at a steady pace (assume the time to distribute postcards is negligible), after 3.5 hours, for example, I will be halfway between south and north; after 6 hours I will be 6/7 of the way from south to north, and after 8.5 hours I will be halfway along the route, heading south. Now I pick a random time from 0 to 10 hours and I start at the point on the route that that would put me. For example, if I picked 2.2 hours, then I would start 2.2/7 of the way from the south end. I would then continue north for 4.8 hours until I reached the north end, return to the south end (without distributing postcards) at 7.8 hours, then complete the rest of the route to my starting place. If I randomly picked 8.5 hours, then I would imagine starting 1.5 hours (halfway) into my return route from north to south. Therefore, I would reach the south end 1.5 hours after the randomly chosen start time, and complete the route from south to north.

I obviously wouldn't actually start halfway down the river at my start time; I'd just wait until 1.5 hours after my start time to actually start the route at the south end. In this way, the entire route is driven exactly once.

Assume the time to distribute postcards is negligible; Thus I will work either the period 0600-1600 or 1100-2100.

In order to calculate the inclusion probability for a selected angler, I will need to know the time his or her vehicle arrived and the time it left on the day the angler received a survey, so that will be one of the questions I ask on the questionnaire. Recall also that inclusion probabilities are computed from this perspective: before I select the days, starting points and starting times, what is the probability that an angler who is parked along the river between times t_1 and t_2 on a specific day will be included in the survey?

Compute the inclusion probabilities for the following anglers:

- (a) Rob: parked from 0500 to 0900 on Thursday of week 2.
- (b) Jennifer: parked from 0930 to 1600 on Sunday of week 3.
- (c) Lisa: parked from 1400 to 2200 on Monday of week 4.

It may help to draw a graph with time on the x-axis and location along the river on the y-axis. Then a parked car is a horizontal line segment and the worker (me) is a diagonal line segment with a random start time and random start point. The probability of inclusion is the probability that the worker's segment intersects the angler's segment times the probability that this is one of the chosen days. Actually, the worker will be represented by two separate line segments (why?).