Stat 549

Fall 2014 Homework 2: due Fri, Oct. 10

Note: Report results from R, but please also include R scripts (including output).

1. A simple random sample of 250 households was chosen from a city containing 13,828 households to estimate the proportion of households in that area who owned their home. Of the 250 households sampled, 158 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. Chapter 6, problem 3 (p. 89, 2nd ed: p. 63) (answers in back; show how you get these).

3. Chapter 6, problem 3 with a modification: suppose you don't know the total number of lakes in the study region nor the total area covered by the lakes. Estimate the mean pollution concentration per lake and the mean size of the lakes in the region. Compute standard errors and confidence intervals for both estimates by two methods: linearization and bootstrapping (note: the sample size here may be too small for either of these methods to work well).

4. MATH ONLY: Chapter 7, problem 5 (p. 113; 2nd ed: p. 86)

5. NON-MATH ONLY: 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: $y_1 = 3, x_1 = 20$; $y_2 = 2, x_2 = 23$; $y_3 = 0, x_3 = 0$; $y_4 = 1, x_4 = 12$; $y_5 = 1, x_5 = 7$. 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 $N\overline{y}$ of the population total for each sample, and the ratio estimate $\widehat{\tau}_r$ for each sample.

   (b) Compute the expected value and variance for the estimator $N\overline{y}$.

   (c) Compute the expected value and mean square error of the ratio estimator $\widehat{\tau}_r$.

   (d) Compare the two estimators for this population.

6. The file census.csv has data on a number of variables for all 3143 counties in the U.S. from the 2010 U.S. Census. You can read it into R using the `read.csv` command. This will create a data frame in R (see Sec. 9 of the R intro document on the website). A separate file on the webpage lists all the variables and their definitions if you are interested. For this exercise, we

are only interested in two variables: TotalPop, the total number of persons counted in the county, and HousingUnits, the total number of housing units (both owned and rented) in the county. Suppose we are interested in seeing if we can accurately estimate the total number of people in the U.S. by counting the number of housing units in every county, which we think will be much easier than trying to count every person (including the homeless). Our plan is then to take a sample of counties where we will do a careful count of all the people living there and then extrapolate to the U.S. using ratio or regression estimation. Note that we have the entire population so we can see well these methods work for possible use in the next decennial Census in 2020.

(a) Examine the relationship between number of housing units and total population for all the counties in the U.S. with a scatterplot. Does it appear that using number of housing units as an auxiliary variable through a ratio or regression estimator might be helpful in estimating the total population of the U.S.?

(b) Set up a script in R to draw an SRS of 100 counties and compute three estimates of the total population of the U.S.: the simple SRS estimate without using an auxiliary variable, and the ratio and regression estimates by using the total number of housing units in the country as an auxiliary variable. Draw a few different samples and see how the estimates compare to the true value (which you can calculate from the data frame using the `sum` command). Which appears to be doing best? You do not need to report numerical results for this part; just report what you observe.

(c) To mimic what would happen in practice, take an SRS of size $n = 100$ and compute the three estimators and their standard errors (using the finite population correction). Before you select your sample, use the R command `set.seed()` where you put any integer inside the parentheses to set the random number generator so that I can reproduce your sample and check your calculations (e.g., `set.seed(45519)`).

(d) Since we have the entire population, we can calculate the theoretical standard deviation of the sample mean and the approximate theoretical standard deviations of the ratio estimator and regression estimators. Go ahead and do so.

(e) Imagine that you will be using this method in the next census in 2020. Estimate the sample size needed to estimate the total population of the U.S. to within 5,000,000 with probability 95% with this method (using the finite population correction). Use the value of σ_r^2 from the whole 2010 census to estimate σ_r^2 for the next census.

(f) MATH ONLY: Do a simulation to estimate the bias and standard deviations of the ratio and regression estimators for sample size $n = 100$. Compare the bias of the ratio estimator to the estimate derived in problem 5 from Chapter 7 (above). Compare the standard deviations to the approximate theoretical values in part (d).

(g) MATH ONLY: This looks like a situation where PPS sampling with replacement would do well (where the size variable is the number of housing units). Compute the theoretical

standard deviation of the PPS estimator for $n = 100$. What would be the disadvantage of this sampling plan versus SRS?

7. 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) Jeff: parked from 0500 to 1030 on Thursday of week 2.

(b) Ellen: parked from 0930 to 1700 on Sunday of week 3.

(c) Iris: parked from 1700 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?).

8. Extra Credit: In Section 6.4 of Thompson, he gives a small example ($N = 3, n = 2$) of sampling with replacement with unequal probabilities where the Hansen-Hurwitz estimator ($\widehat{\tau}_p$) of the population total had much smaller variance than the Horvitz-Thompson estimator ($\widehat{\tau}_\pi$). Create an example of this same size, where the reverse is true.