Stat 549

Fall 2014 Homework 4: due Mon, Nov. 17

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

1. Consider the results from our in-class penny sampling project. The sampling plans were:

   - SRS: Zor, Vernon
   - Double sampling, every 2nd: Eli and Guedem, Kevin and Lia
   - Double sampling, every 4th: Caitlin and Mike, Ann and Jacob
   - Double sampling, every 5th: Grant and Jon

   The resulting seven data sets are given in the **R** script file `PennyData.R` on the course web page. The data have been entered in the same format as for the biomass example on page 128 of the class notes: the vector `y` contains the $n$ actual counts, the first $n$ elements of the vector `x` are the visual estimates corresponding to the actual counts, and the remaining elements of `x` are the visual only counts done before time was called. The vector `x.all` contains the visual counts for all 80 jars which is used only in part (c) below.

   (a) For each of the seven sets of data above, estimate the total number of pennies in the population of 80 jars, and give the corresponding SE. For the first two sets, the unbiased SRS estimator should be used. For the latter five sets of data, a ratio estimator with double sampling should be used (using only the visual estimates in `x`). Show your work for just your own data set - you do not need to show me the calculations for all seven sets of data. A table with the estimates and SE's would make a nice visual summary.

   (b) Keeping in mind that the sampling effort (in terms of time) was roughly the same for all four sampling schemes, what, if anything, can you say about which method was most effective and why? Which group wins the dubious title of "Least efficient estimators?" Which group was most efficient? Write a well-constructed paragraph or two to address this question.

   (c) Just to see how much difference it makes, for the five double sampling schemes, calculate the ratio estimate and SE of the total number of pennies using the visual estimates for all 80 jars (`x.all`), as in Chapter 7. Does using all 80 visual estimates reduce the corresponding SE's from part (a) much?

   (d) Some of the "actual" counts were off – mostly by 1 or 2, but in two instances by 10. I replaced these mistaken counts by the true counts so that we could make a fairer comparison of the sampling methods. But suppose, I had retained the incorrect actual counts. More generally, suppose there was significant random measurement error in the actual counts. How would that affect estimates and SE's? (Think about what would happen to the regression and to $s_r^2$.)

(e) Estimate the optimal sampling ratio $n/n'$ by using estimates of $\sigma_r^2$ and $\sigma^2$ from any one of the double sampling data sets (tell me which one you used). Use three different estimates of the cost ratio $c'/c$; a 'best guess' based on your experience in the classroom and a smaller value and a bigger value.

2. For this problem, you will need to retrieve the data file `fraser.txt` from the course web page. This files contains the mean monthly flows ($m^3$/sec) of the Fraser River at Hope, B.C. from March 1912 through December 2012 (100+ years, 1210 observations). These data can be read into **R** by `scan("fraser.txt")`. A function called `systematic` can also be found on the web page and is required for this problem.

   (a) Plot the flows over time: use the command `ts.plot(x)` (time series plot). It may be hard to see patterns since a lot of data (1210 flows) is compressed into the plot. To get around this, you might plot just the first 250 observations by `ts.plot(x[1:250])` or any other subset of the data by using `ts.plot(x[a:b])`. Describe any patterns you see in a few sentences. (You do not need to include the plots.)

   (b) Consider the 1210 monthly flows as the population and suppose that you want to estimate the mean of the mean monthly flows for this period from a sample of months. Compare systematic sampling (with a random starting point) with intervals of 3, 6, 9, 11, 12, 13, 23, 24 to each other and to simple random samples (SRS's) of the same sample sizes. Do this b computing the theoretical standard deviations of the estimators (you can use the function `systematic` as discussed below). Discuss your results. Are the results what you would expect? How big an effect can matching the data pattern have on the systematic estimator? How do the systematic SE's compare to those for SRS of the same sample size?

   Note: For a systematic sample with interval $k$, there are only $k$ possible samples corresponding to the $k$ possible starting points. The sampling distribution of the mean therefore contains only $k$ equally likely values and the theoretical variance of the estimator can be easily computed. The function **systematic(x,k)** automates this; it computes the expected value and standard deviation of the sampling distribution of the sample mean based on a systematic sample with interval $k$ (with random starting point) from the population vector `x`.

3. In class, we discussed the following nonresponse example. 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, estimate the proportion of all 400 individuals in the population who would respond "Yes" to the question of interest and attach

a standard error to this estimate. Compare the 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.

4. 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 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.

5. Andrew McDonald, a student in this class some years ago, did a project to compare various sampling methods for estimating the total number of sweetclover shrubs in a square plot 108 feet on a side. One of his methods was a line-intercept survey. The resulting data 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 (use `read.csv` to import the data). 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, calculate the SE and compute a 95% confidence interval.

   (b) How many transects would be required to estimate the total number of plants to within $\pm 100$ with 95% probability?

6. 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 $\mathrm{SE}(\widehat{p})$. Using the data given in problem 3, estimate the standard error of the mean number of moose when $\mathrm{SE}(\widehat{p})$ 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 values for $\mathrm{SE}(\widehat{p})$. Does the SE of the detection probability have much effect on the value of $\mathrm{SE}(\widehat{p})$?

7. **NON-MATH ONLY**: The data file `coots.csv` contains data from Arnold's (1991) work on egg size and volume in American coot eggs in Minnedosa, Manitoba, as cited in Lohr's text *Sampling Design and Analysis*. 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 $V = .000507\times$ length $\times$ 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 95% 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).

   Note: 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 $\overline{M}$ (the average clutch size for the population).

8. **MATH STUDENTS ONLY**: Problem 4 of Homework 3 was a two-stage sampling plan to estimate the average height of seedlings in a field. Bootstrap this problem. On each bootstrap replication, take a random sample with replacement of $n = 5$ plots and then within each of the selected plots taking a random sample with replacement of the appropriate number of seedlings (which varies from plot to plot). However, there are only 25 plots in the population so, at the first stage, do what is often suggested for a finite population: create a bootstrap population of 5 copies each of the 5 selected plots, then each bootstrap sample will consist of a sample of 5 plots without replacement from the bootstrap population. At the second stage, just ignore the finite population (since 10% is a relatively small proportion) and sample with replacement from the seedlings within a plot. (If we wanted to take population size of each plot into account, it would be a little involved since the population size is not an even multiple of the sample size. So, for example, for plot 1, where there are 52 seedlings and a sample of size 5, the bootstrap population would consist of 10 copies of each seedling plus an 11th copy of two of the seedlings, selected randomly. Except that the 11th copies would not stay the same from bootstrap iteration to iteration; we would select them anew randomly on each iteration.)

   You will have to program the bootstrap yourself. At the end, you will have at least 10,000 bootstrapped values of your statistic. Examine the sampling distribution of the estimate of the population mean, compute the bootstrap SE and compute the percentile confidence interval. Comment.