SDGB 7844 HW 3: Capture-Recapture Method

Instructor: Prof. Nagaraja

Due: 11/2 in class

Submit two files: (a) solutions (explanations, plots, tables, etc. in .docx or .pdf file) and (b) code (in a .R file). Both parts should be (a) printed and brought to class and (b) emailed to fordhamRcomputing@gmail.com by the start of class. (See the Lecture 1 exercises on Blackboard for an example of how to format your files.)

In your email, use the subject heading: "HW 3-[Your name]-[Time of Class]" and include HW 3 and your name in the file names (time of class is either 1:15 PM or 3:30 PM). Please email your solutions only once! Complete your work individually and comment your code for full credit.

In the beginning of the 17th century, John Graunt wanted to determine the effect of the plague on the population of England; two hundred years later, Pierre-Simon Laplace wanted to estimate the population of France. Both Graunt and Laplace implemented what is now called the *capture-recapture method.* This technique is used to not only count human populations (such as the homeless) but also animals in the wild.

In its simplest form, n_1 individuals are "captured," "tagged", and released. A while later, n_2 individuals are "captured" and the number of "tagged" individuals, m_2, is counted. If N is the true total population size, we can estimate it with $\hat{N}_{LP}$ as follows:

$$\hat{N}_{LP} = \frac{n_1 n_2}{m_2} \tag{1}$$

using the relation $\frac{n_1}{N} = \frac{m_2}{n_2}$. This is called the Lincoln-Peterson estimator[1].

We make several strong assumptions when we use this method: (a) each individual is independently captured, (b) each individual is equally likely to be captured, (c) there are no births, deaths, immigration, or emigration of individuals (i.e., a closed population), and (d) the tags do not wear off (if it is a physical mark) and no tag goes unnoticed by a researcher.

[1] Interestingly, this estimator is also the maximum likelihood estimate which you will learn about in DGGB 781A (MSSD students) and which we will talk about briefly later in the semester. As you probably guessed, more complex versions of this idea have been developed since the 1600s.

Goal: In this assignment, you will develop a Monte-Carlo simulation of the capture-recapture method and investigate the statistical properties of the Lincoln-Peterson and Chapman estimators of population size, N. (Since you are simulating your own data, you know the true value of the population size N allowing you to study how well these estimators work.)

Note: It is helpful to save your R workspace to an ".RData" file so that you don't have to keep running all of your code every time you work on this assignment. See Lecture 2, slide 17 for more details.

1. Simulate the capture-recapture method for a population of size $N = 5,000$ when $n_1 = 100$ and $n_2 = 100$ using the `sample()` function (we assume that each individual is equally likely to be "captured"). Determine m_2 and calculate $\hat{N}_{LP}$ using Eq.1. (Hint: think of everyone in your population as having an assigned number from 1 to 5,000, then when you sample from this population, you say you selected person 5, person 8, etc., for example.)

2. Write a function to simulate the capture-recapture procedure using the inputs: N, n_1, n_2, and the number of simulation runs. The function should output in list form (a) a data frame with two columns: the values of m_2 and $\hat{N}_{LP}$ for each iteration and (b) N. Run your simulation for 1,000 iterations for a population of size N =5,000 where $n_1 = n_2 = 100$ and make a histogram of the resulting $\hat{N}_{LP}$ vector[2]. Indicate N on your plot.

3. What percent of the estimated population values in question 2 were infinite? Why can this occur?

4. An alternative to the Lincoln-Peterson estimator is the Chapman estimator:

$$\hat{N}_C = \frac{(n_1 + 1)(n_2 + 1)}{m_2 + 1} - 1 \qquad (2)$$

Use the saved m_2 values from question 2 to compute the corresponding Chapman estimates for each iteration of your simulation. Construct a histogram of the resulting $\hat{N}_C$ estimates, indicating N on your plot.

5. An estimator is considered *unbiased* if, on average, the estimator equals the true population value. For example, the sample mean $\overline{x} = \sum_{i=1}^{n} x_i/n$ is unbiased because on

[2] Basically, you are empirically constructing the sampling distribution for $\hat{N}_{LP}$ here. Remember the Central Limit Theorem which tells us the sampling distribution of the sampling mean? Each statistic has a sampling distribution and we are simulating it here (but using frequency instead of probability on the y-axis).

average the sample mean $\overline{x}$ equals the population mean μ (i.e., the sampling distribution is centered around μ). This is a desirable property for an estimator to have because it means our estimator is not systematically wrong. To show that some estimator $\hat{\theta}$ is an unbiased estimate of the true value θ, we need to mathematically prove that $E[\hat{\theta}] - \theta = 0$ where $E[\cdot]$ is the expectation (i.e., theoretical average)[3]. We will check for this property empirically by replacing the theoretical average $E[\hat{\theta}]$ with the sample average of the $\hat{\theta}$ values from our simulation (i.e., $\sum_{i=1}^{n_{sim}} \hat{\theta}/n_{sim}$ where n_{sim} is the number of simulation runs; θ is N in this case, and $\hat{\theta}$ is either $\hat{N}_{LP}$ or $\hat{N}_C$ as both are ways to estimate N).

Estimate the bias of the Lincoln-Peterson and Chapman estimators, based on the results of your simulation. Is either estimator unbiased when $n_1, n_2 = 100$?

6. Based on your findings thus far, is the Lincoln-Peterson or Chapman estimator better? Justify your answer.

7. Can we get better estimates using larger sample sizes? Let's restrict $n = n_1 = n_2$. Write a function which computes the bias and variance of the Chapman estimator for varying sample sizes, n. The inputs for this function should be: N, number of simulation runs per sample size, and a vector of sample sizes. The function should return in list form (a) a data frame with three columns: n, bias of the Chapman estimator, and variance of the Chapman estimator for each sample size and (b) the true population size. Run this function using the following arguments: (a) N =100,000 for each sample size, (b) 1,000 simulation runs, and (c) n ranging from 100 to 5,000 (use `seq(from=100, to=5000, by=50)`).

   Based on your results, construct two plots: (a) bias versus n and (b) variance versus n (e.g., y-axis variable vs. x-axis variable). Indicate zero on both plots and connect the points with a line. Describe what you see.

8. An estimator $\hat{\theta}$ is called *consistent* if, as the sample size goes to infinity, the bias of the estimator goes to zero[4]. That is, $E[\hat{\theta}]$ gets closer to θ as the sample size increases (i.e., you collect more data). This is another desirable property for an estimator.

   For the simpler case where $n_1 = n_2$, could $\hat{N}_C$ be a consistent estimator? Justify your answer.

9. Explain why the assumptions (a), (b), and (c) listed on the first page are unrealistic.

[3] Everyone: note that the sample size n does not appear in this equation. For an estimator to be unbiased, this property cannot depend on sample size. MSSD students: $E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$..

[4] MSSD students: $\hat{\theta}_n$ is a consistent estimator of θ if $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$, $\varepsilon > 0$ (i.e., convergence in probability).