SDGB 7844 HW 3: Capture-Recapture Method

Instructor: Prof. Nagaraja

Due: 11/1

Submit two files through Blackboard: (a) .Rmd R Markdown file with answers and code and (b) Word document of knitted R Markdown file. Your file should be named as follows: "HW3-[Full Name]-[Class Time]" and include those details in the body of your file.

Please submit your solutions only once! Complete your work individually and comment your code for full credit. For an example of how to format your homework see the files related to the Lecture 1 Exercises and the RMarkdown examples on Blackboard. **Show all of your code in the knitted Word document.**

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.

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 es-

[1]Interestingly, this estimator is also the maximum likelihood estimate. As you probably guessed, more complex versions of this idea have been developed since the 1600s.

timators 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 8 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:

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

   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 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 an estimator $\hat{\theta}$ is

[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).

an unbiased estimate of the true value θ, we would need to mathematically prove that $\mathbb{E}[\hat{\theta}] - \theta = 0$ where $\mathbb{E}[\cdot]$ is the expectation (i.e., theoretical average)[3]. Instead, we will investigate this property empirically by replacing the theoretical average $\mathbb{E}[\hat{\theta}]$ 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)[4].

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, is the Lincoln-Peterson or Chapman estimator better? Explain your answer.

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

[3]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.

[4]Note: This procedure is not a replacement for a mathematical proof, but it's a good way to explore statistical properties.