SDGB 7844 HW 3: Capture-Recapture Method

Jiayin Hu

2018/11/1

The two estimators in *capture-recapture method* are Lincoln-Peterson estimator and Chapman estimator.

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

$$\hat{N}_{C} = \frac{(n_1 + 1)(n_2 + 1)}{m_2 + 1} - 1$$

The package we use to plot is `ggplot2`.

```
require(ggplot2)
```

Q1

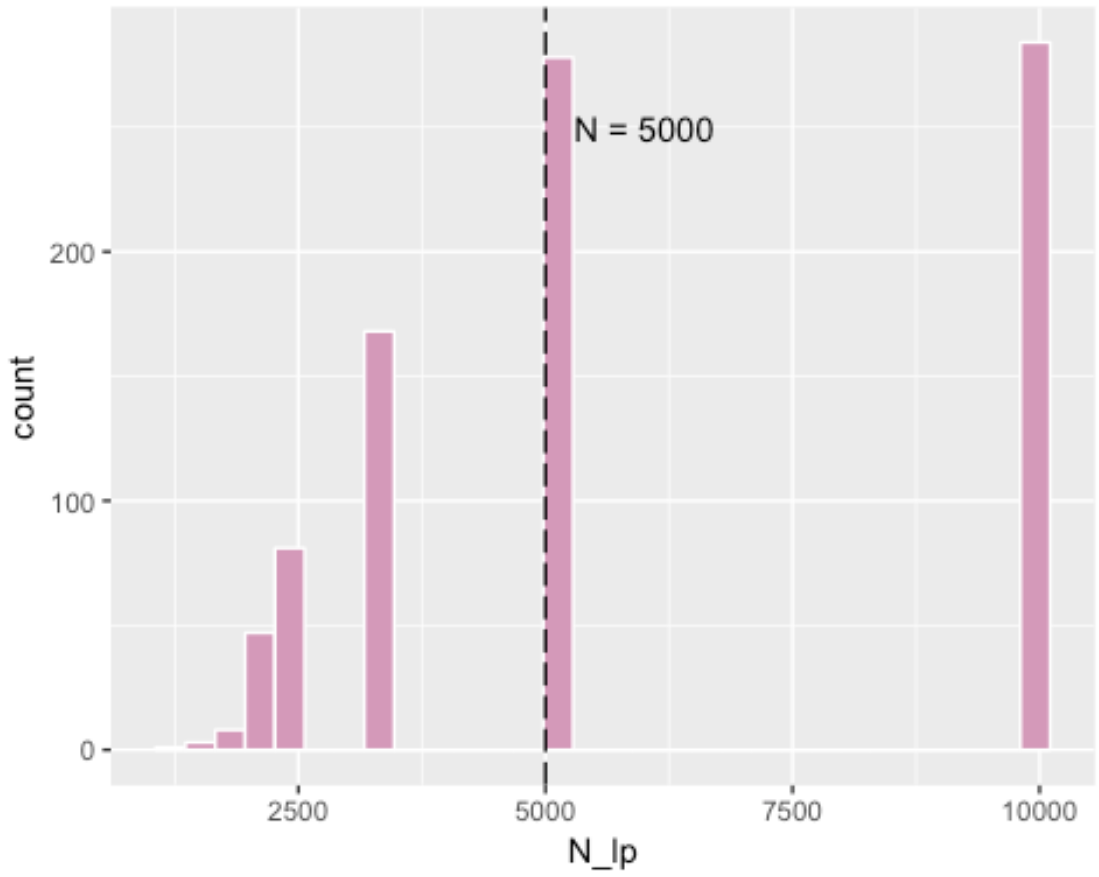
Simulate the capture-recapture method for a population of size $N = 5000$ when $n_1 = 100$ and $n_2 = 100$ using the `sample()` function (we assume that each individual is equally likely to be "captured"). Determine m_2 and calculate $\hat{N}_{LP}$ using Eq.1.

```
# simulation
n <- 5000
n1 <- 100
n2 <- 100

population <- c(1:n)
s1 <- sample(population, size = n1, replace = FALSE)
s2 <- sample(population, size = n2, replace = FALSE)
s2in1 <- intersect(s1, s2)
m2 <- length(s2in1)
n_lp <- n1 * n2 / m2

m2

## [1] 2

n_lp

## [1] 5000
```

Q2

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. Indicate N on your plot.

```r
# define a function
SimulateCapRecapNew <- function(n = 5000, n1 = 100, n2 = 100, iter = 100){
  # n: size of population
  # n1: number of individuals “captured,” “tagged”, and released
  # n2: the number of individuals "recaptured"
  # m2: the number of individuals "tagged"
  # n_lp: the estimator of the population size

  df <- data.frame()
  population <- c(1:n)
  i <- 1

  while (i <= iter) {
    s1 <- sample(population, size = n1, replace = FALSE)
    s2 <- sample(population, size = n2, replace = FALSE)
    s2in1 <- intersect(s1, s2)
    m2 <- length(s2in1)
    n_lp <- n1 * n2 / m2
    df[i, 1] <- m2
    df[i, 2] <- n_lp
    i <- i + 1
  } # end while loop

  N <- n
  return(list(df, N))
}

n <- 5000
n1 <- 100
n2 <- 100
iter <- 1000

simu <- SimulateCapRecapNew(n, n1, n2, iter)
simu_df <- simu[[1]]
names(simu_df) <- c("m2", "N_lp")

ggplot(simu_df, aes(N_lp)) +
  geom_histogram(na.rm = TRUE, color = "white", fill = "#D499B9") +
  geom_vline(xintercept = 5000, linetype = "longdash") +
  annotate("text", x = 6000, y = 250, label = "N = 5000")
```



Q3

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

```
num_inf = sum(is.infinite(simu_df[,2]))
percent_inf = num_inf/iter
percent_inf
```

```
## [1] 0.13
```

13% of the setimated population values were infinite.

The reason is comparing to 5000 numbers, 100 numbers are only small part among the numbers. The ways of combination are so much. So the occurance of no same number in two samples is not very infrequent. By using the knowledge of probability, we could calculate its possibility.

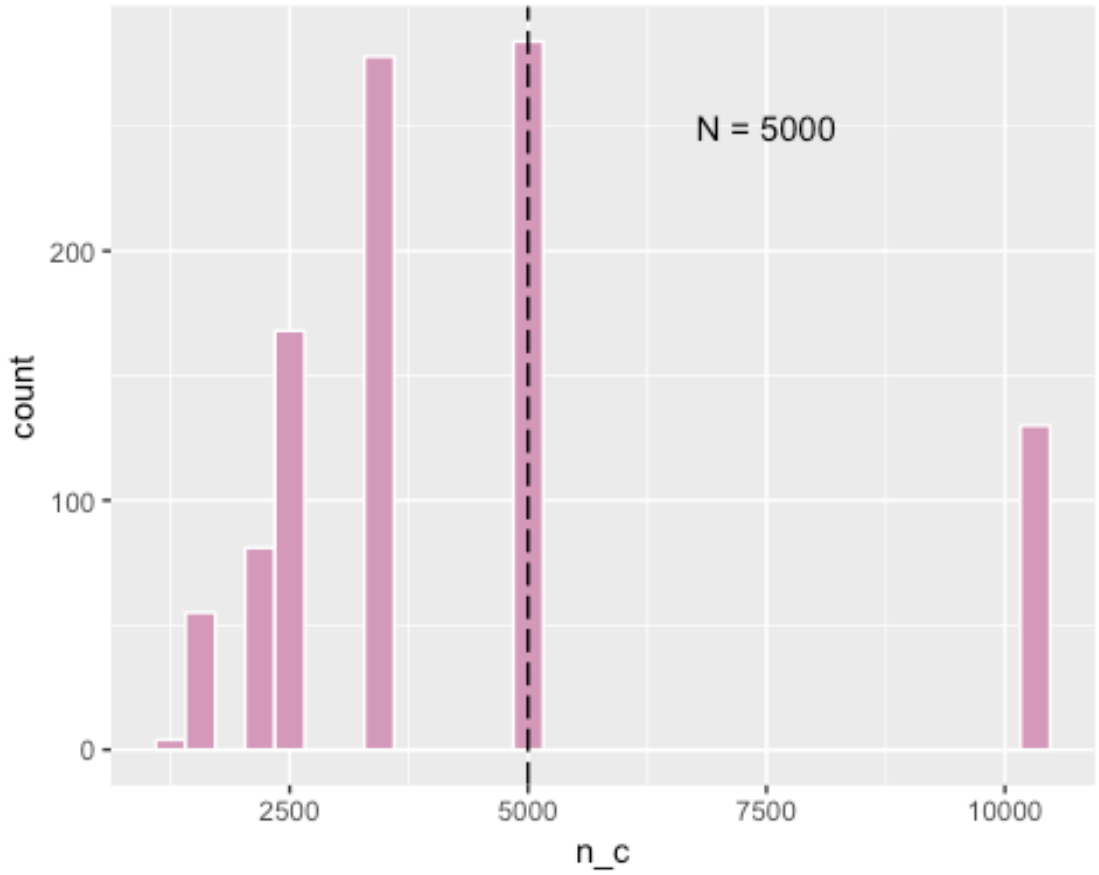
$$P = \frac{C_{100}^{4900}}{C_{100}^{5000}} = 0.1299$$

Our result is almost as much as the theoretical result.

Q4

Use the saved m2 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.

```
m_2 <- simu_df[,1]
n_c <- (n1 + 1) * (n2 + 1) / (m_2 + 1) - 1

df <- data.frame(simu_df[,1], n_c)
ggplot(df, aes(n_c)) +
  geom_histogram(na.rm = TRUE, color = "white", fill = "#D499B9") +
  geom_vline(xintercept = 5000, linetype = "longdash") +
  annotate("text", x = 7500, y = 250, label = "N = 5000")
```



Q5

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

```r
n_lp_mean <- mean(simu_df[,2][which(simu_df[,2] != Inf)])
n_c_mean <- mean(n_c)

n_lp_bias <- n_lp_mean - n
n_c_bias <- n_c_mean - n

n_lp_bias
```

```
## [1] 868.2403
```

```r
n_c_bias
```

```
## [1] -590.8122
```

The bias of the LP estimator is 868.2402846 and the bias of the Chapman estimator is -590.8121948 According to this result, they are not unbiased estimators.

Q6

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

Based on the result above, I think Chapman estimator is better.

First, Chapman estimator could handle the situation that no tagged individual is recaptured, while LP estimator gives infinite.

Second, the bias of Chapman estimator is smaller than LP estimator, which means Chapman estimator is nearer to the real number.

Q7

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

(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)

Each individual must has some connection with others. For some animals, maybe one is captured, its relative following it will be captured, or others will avoid being captured. So each individual is not independently captured.

Because we could only set several places to capture the sample, we could not determine whether the individuals with specific similarity tend to aggregate in some places. This may lead to unequal likelihood to be captured. Also for animals, the individuals captured at the first time may have experience to avoid being captured next time.

In a specific area, individuals are mobile and we could not control the birth, death and migration. Therefore, the replacement and change in the population is inevitable.