This exercise is about estimating the size of a population of wild animals. The method described below is called the capture-recapture method. As we shall see, it does not yield a very good estimator, but that is partly because the problem is inherently difficult. Wildlife biologists use the capture-recapture method in combination with other methods.

Let $N$ denote the unknown size of the population of animals. In the capture-recapture method, some number $t$ of animals in the population is captured and tagged. After the captured animals are tagged, they are released and allowed to mingle with other members of the population. Then a sample of $n$ animals is captured and the number of tagged animals in the sample is observed. We will assume that both $t$ and $n$ are predetermined numbers. Let $X$ denote the number of tagged animals in the sample. $X$ is a random variable. The objective of the experiment is to use the observed value of $X$ to estimate the unknown value of $N$.

The basic idea is simple. If we equate the proportion of tagged animals in the sample to the proportion of tagged animals in the population we can solve the resulting equation for the population size. The sample proportion of tagged animals is $X/n$ and the population proportion is $t/N$. If these two quantities are equal, then $N = tn/X$. Of course, the two proportions are not likely to be exactly equal. Therefore, we must regard the random variable $tn/X$ as an estimator of the parameter $N$.

If the sample is taken with replacement, then an animal could be included more than once in the sample. In this case, $tn/X$ is the maximum likelihood estimator of $N$. (Actually, it is within one unit of the maximum likelihood estimator of $N$, since we want an integer estimate of $N$.) Here is the reason: the random variable $X$ has a binomial distribution based on $n$ trials with success probability $p = t/N$. The maximum likelihood estimator of $p$ is $X/n$. Since $N = t/p$, it follows that the maximum likelihood estimator of $N$ is $t / (X/n) = tn/X$.

It can be shown that even if sampling is done without replacement, the maximum likelihood estimator of $N$ is $tn/X$. Thus, there are several reasons for considering the estimator $tn/X$ of the population size.

Maximum likelihood estimators usually have nearly optimal statistical properties. In other words, the maximum likelihood estimator is about as good as any estimator can be in a given statistical model of an experiment. Thus, if the estimator $tn/X$ is not very good it may indicate that a simple random sampling procedure is not adequate to solve the problem.

If we are sampling with replacement, there is no restriction on the size of $n$. We are going to simulate the experiment above with various values of $n$. First, we establish realistic values of $N$ and $t$.

> $N = 5000$
> $t = 200$

Suppose the recapture sample size is the same as the number tagged.
\[ n = 200 \]

We will do 100 simulations of the experiment, leading to 100 observed values of \( X \) and 100 estimates of \( N \). The vector of 100 \( X \) values is produced by the command

\[ Xs = \text{rbinom}(100, \text{size} = n, \text{prob} = t/N) \]

and the estimated values of \( N \) by

\[ \text{Nhats} = t*n/Xs \]

Notice that the expression above tells R to divide a number by a vector. R understands an expression like this and gives back a vector of 100 estimated values of \( N \).

Make a stemplot of \( \text{Nhats} \) with the command

\[ \text{stem}(\text{Nhats}). \]

Make a histogram and a normal quantile plot of \( \text{Nhats} \) with the commands

\[ \text{hist}(\text{Nhats}, \text{main} = "\text{your own last name}") \]

\[ \text{qqnorm}(\text{Nhats}, \text{main} = "\text{your name}") \]

\[ \text{qqline}(\text{Nhats}) \]

\textbf{TURN IN THE HISTOGRAM AND THE NORMAL QUANTILE PLOT.}

Does it look like this estimator is approximately normally distributed when \( n=200 \)?

You can also get a numerical summary of the data by typing

\[ \text{summary}(\text{Nhats}) \]

One peculiarity of R is that this summary does not include the standard deviation. To see the variance of the data, type

\[ \text{var}(\text{Nhats}) \]

Call the help function for \text{var} and notice that by default it gives the unbiased sample variance instead of the so-called population variance. If you want the standard deviation, type either

\[ \text{sqrt(var(\text{Nhats}))} \]

or
Repeat the steps above with \( n = 1000 \) and \( n = 5000 \) for the same values of \( t \) and \( N \). Print the histograms and normal quantile plots.

TURN IN THE HISTOGRAMS AND QQNORM PLOTS FOR \( n = 1000 \) and \( n = 5000 \).

Does it appear that the estimator is approximately normal for larger values of \( n \)? For each value of \( n \), make a note of the mean, median and variance of the 100 estimated values of \( N \).

It can be shown theoretically that for large enough values of \( n \) the estimator \( \frac{tn}{X} \) is approximately normal with mean \( N \) and variance \( \frac{N^2(N-t)}{tn} \). Calculate the value of this expression for \( n=200, 1000, \) and \( 5000 \). Compare the values to the sample variances of the \( Nhats \) vectors. You can do the calculations from the command line of \( R \), using it as if it were a desk calculator, e.g.,

\[
> \text{val} = N^2*(N-t)/t
> \text{val}/200
> \text{val}/1000
> \text{val}/5000
\]

Whenever an estimator is approximately normal there is a possibility that we can use the normal distribution to construct a confidence interval for the value of the parameter. In this case, there is a problem. The variance of the estimator depends on the unknown parameter \( N \). A straightforward remedy is to replace the unknown parameter in this expression for the variance by its estimated value. This generally works well, although the reliability of the conclusions requires a somewhat larger sample size than would otherwise be necessary. You have already seen an example of this approach in class, where we derived a large sample confidence interval of the form \( \hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \) for a binomial success probability.

For the estimated values of \( N \) just computed we can produce a vector of corresponding estimated standard errors by

\[
> \text{std.errors} = \text{Nhats} * \sqrt{((\text{Nhats} - t)/(t*n))}
\]

Pick a standard normal \( z \) value corresponding to whatever confidence level you want, e.g., \( z = 1.96 \) for 95% confidence. You can use the normal table or you can use the normal quantile function.

\[
> z = \text{qnorm}(0.975) \quad \text{(For 95% confidence. Adjust as necessary)}
\]
Then calculate 100 lower confidence limits and 100 upper confidence limits:

\[ lcl = \text{Nhats} - z \times \text{std.errors} \]

\[ ucl = \text{Nhats} + z \times \text{std.errors} \]

If our procedure works as it is supposed to, then about 95% (or whatever) of our simulations should lead to lower and upper confidence limits that enclose the true value of \( N = 5000 \). To count the number of times this favorable outcome occurred, type

\[ \text{sum(lcl < N & N < ucl)} \]

See if the result is close to the confidence level you picked.

The preceding command should be examined carefully. When the argument of the \textit{sum} function is a numeric object, such as a vector or matrix, the function simply returns the sum of its elements. In this case the argument “lcl < N & N < ucl” is a logical vector whose components are either T or F. \textit{sum} treats a T as the number 1 and an F as the number 0.

Next, we want to put the confidence intervals in readable matrix form. Type

\[ \text{intervals = matrix(c(lcl,ucl), ncol=2)} \]

Then type

\[ \text{intervals} \]

to see the results. Highlight the displayed intervals and print them.

\textbf{TURN IN THE MATRIX OF CONFIDENCE INTERVALS.}

As you see, we had to have \( n \geq 1000 \) to obtain an approximately normally distributed estimator. In real life, this probably would be an unacceptably large sample size. Also, in most surveys the recapture sample could not be obtained with replacement and the size of the recapture sample could not be determined in advance. A good reason for doing simulations like this is to reveal the difficulties in an experimental plan before the experiment is actually carried out.