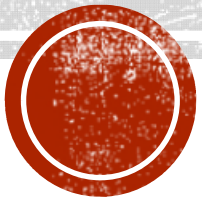
Chapter 6

MH4511: SAMPLING AND SURVEY



Estimation of Population Size

WHY ESTIMATE POPULATION SIZE?

- Knowing the size of a population of animals is important in making environmental decisions that would affect the population, but estimating the size of wild populations is extremely difficult.
- In the case of ocean dwellers, such as whales, the task is especially challenging. Estimates of the number of minke whales, for example, have differed by as much as a factor of 10.
  - Deciding whether to allow hunting of minke whales, based on population estimates that are too high, could lead to extinction of the species.
  - On the other hand, basing a decision on an estimate that is too low could unnecessarily ban hunting of minkes by people that depend on whales for food.

WHY ESTIMATE POPULATION SIZE?

- The best way to measure the size of a population is to count all the individuals in that population.
- When determining the population sizes of trees or other relatively immobile organisms, this method is practical.
- If the organism is mobile, however, such as a fish, counting every individual would be difficult. Some individuals might be counted twice or not at all, since the experimenter would not know which fish had been counted and which had not.

EXAMPLE (CAPTURE-RECAPTURE)

Suppose we want to estimate N, the number of fish in a lake. One method is as follows:

- Catch and mark 200 fish (first sample) in the lake, then release them.
- Allow the marked and released fish to mix with the other fish in the lake.
- Then, take a second, independent sample of 100 fish.
- Suppose that 20 of the fish in the second sample are marked.
- Assuming that the population of has not changed between the two samples and that each catch gives a SRS of fish in the lake,
- Then we estimate that 20% of the fish in the lake are marked.
- And, that the 200 fish tagged in the first sample represent approximately 20% of the population of fish.
- The population size N is then estimated to be approximately 1000.

CAPTURE-RECAPTURE ESTIMATION

- This is the basic (two-sample) capture-recapture method.
- This method relies on some assumptions:
  - The population is closed –no fish enter or leave the lake between the samples. That is, N is the same for each sample.
  - Each sample of fish is an SRS from the population. That is, each fish is equally likely to be chosen in a sample.
  - The two samples are independent. The marked fish from the first sample become re-mixed in the population, so that the marking status of a fish is unrelated to the probability that the fish is selected in the second sample.
  - Fish do not lose their markings, and marked fish can be identified as such.

MARKING METHODS

THE THEORY

- Let n_1 be the size of the first sample, n_2 be the size of the second sample, and m the number of marked fish caught in the second sample.
- The estimator of N is then

$$\widehat{N} = n_1 \times \frac{1}{\hat{p}} = n_1 \times \frac{1}{m/n_2} = \frac{n_1 n_2}{m}$$

  where $\hat{p}$ is the proportion estimate of the marked fish in the lake based on second sample.
- This is a special case of ratio estimation.
- Let

  $y_i = 1$ for every fish i in the lake and

  $x_i = 1$ if the fish is marked, 0 otherwise.g

THE RATIO ESTIMATOR

Then

$N = t_y = \sum_{i=1}^{N} y_i$ can be estimated by the ratio estimator

$$\widehat{N} = \hat{t}_{yr} = t_x \times \hat{B}$$

where

$$t_x = \sum_{i=1}^{N} x_i = n_1 \text{ and}$$

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{n_2 \times \bar{y}}{n_2 \times \bar{x}} = \frac{\text{sum of fish in 2nd sample}}{\text{sum of marked fish in 2nd sample}} = \frac{n_2}{m}$$

Hence,

$$\widehat{N} = \hat{t}_{yr} = t_x \times \hat{B} = \frac{n_1 n_2}{m}$$

We can show (in tutorial) that $\widehat{N}$ is an MLE of N.

VARIANCE OF THE RATIO ESTIMATOR

Recall that

$$\widehat{Var}(\hat{B}) = \left(1 - \frac{n}{N}\right) \times \left(\frac{s_e^2}{n\bar{x}^2}\right)$$

with

$$s_e^2 = \frac{1}{n-1}\sum_{S} e_i^2 = \frac{1}{n-1}\sum_{S} \left(y_i - \hat{B}\, x_i\right)^2$$

In our case,

$$\hat{B} = \frac{n_2}{m}$$

$$s_e^2 = \frac{1}{n_2-1}\sum_{S} \left(y_i - \hat{B}\, x_i\right)^2 = \frac{1}{n_2-1}\sum_{S_2} \left(y_i - \frac{n_2}{m} x_i\right)^2$$

$$= \frac{1}{n_2-1}\sum_{S} \left[y_i^2 - 2\frac{n_2}{m} x_i y_i + \left(\frac{n_2}{m} x_i\right)^2\right]$$

$$= \frac{1}{n_2-1}\left[n_2 - 2\frac{n_2}{m} m + \frac{n_2^2}{m^2} m\right] = \frac{1}{n_2-1}\left[-n_2 + \frac{n_2^2}{m}\right] = \frac{n_2(n_2-m)}{m(n_2-1)}$$

VARIANCE OF THE RATIO ESTIMATOR

$$\widehat{Var}(\hat{N}) = t_x^2 \times \widehat{Var}(\hat{B}) \approx t_x^2 \times \left(\frac{s_e^2}{n_2 \bar{x}^2}\right), \text{ ignoring fpc}$$

$$\widehat{Var}(\hat{N}) \approx n_1^2 \times \frac{\frac{n_2(n_2-m)}{m(n_2-1)}}{n_2\left(\frac{m}{n_2}\right)^2} = \frac{n_1^2 n_2^2}{m^2} \frac{(n_2-m)}{m(n_2-1)}$$

$$\approx \frac{n_1^2\, n_2\, (n_2-m)}{m^3}$$

VARIANCE OF THE RATIO ESTIMATOR

$$\widehat{Var}(\hat{N}) = t_x^2 \times \widehat{Var}(\hat{B}) \approx t_x^2 \times \left(1 - \frac{n_2}{\hat{N}}\right) \times \left(\frac{s_e^2}{n_2 \bar{x}^2}\right) \text{ with fpc}$$

$$\widehat{Var}(\hat{N}) \approx n_1^2 \times \left(1 - \frac{n_2}{\frac{n_1 n_2}{m}}\right) \times \frac{\frac{n_2(n_2 - m)}{m(n_2 - 1)}}{n_2 \left(\frac{m}{n_2}\right)^2} = \frac{n_1^2 n_2^2}{m^2} \frac{(n_2 - m)}{m(n_2 - 1)}$$

$$\approx \left(1 - \frac{m}{n_1}\right) \frac{n_1^2 \, n_2 \, (n_2 - m)}{m^3}$$

EXAMPLE

In our fish example:

$n_1 = 200$ is the size of the first sample,

$n_2 = 100$ is the size of the second sample, and

$m = 20$ is the number of marked fish caught in the second sample.

Hence, $\widehat{N} = \frac{n_1 n_2}{m} = \frac{200 \times 100}{20} = 1000$ and

$$\widehat{Var}(\widehat{N}) \approx \frac{n_1^2 \, n_2 \, (n_2 - m)}{m^3}$$

$$= \frac{200^2 \times 100 \times (100 - 20)}{20^3} = 40000$$

BIASNESS OF RATIO ESTIMATOR

Note that being a ratio estimator, $\widehat{N}$ is biased, and the bias can be large in wildlife applications with small sample size.

Indeed, it is possible for the second sample to consist entire of unmarked fish, making the estimate to be infinite.

An alternative estimator was proposed by Chapman (1951) and Seber (1970):

$$\tilde{N} = \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1$$

$$\widehat{Var}(\tilde{N}) \approx \frac{(n_1 + 1)(n_2 + 1)(n_1 - m)(n_2 - m)}{(m + 1)^2(m + 2)}$$

CONTINGENCY TABLES FOR CAPTURE-RECAPTURE METHOD

In general, the observed counts (number of fish) can be considered as

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	$x_{11}(=m)$	x_{12}	$x_{1+}(=n_1)$
	No	x_{21}	x^*_{22}	x^*_{2+}
	Total	$x_{+1}(=n_2)$	x^*_{+2}	$x^*_{++}(=N)$

An asterisk indicates that we do not observe that value.

CONTINGENCY TABLES FOR CAPTURE-RECAPTURE METHOD

The expected counts (number of fish) are as follows

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	m_{11}	m_{12}	m_{1+}
	No	m_{21}	m_{22}^*	x_{2+}^*
	Total	m_{+1}	m_{+2}^*	m_{++}^*

An asterisk indicates that we do not observe that value.

To estimate the expected counts, we use $\hat{m}_{11} = x_{11}$, $\hat{m}_{12} = x_{12}$, and $\hat{m}_{21} = x_{21}$

CONTINGENCY TABLES FOR CAPTURE-RECAPTURE METHOD

If sample 1 and sample 2 are independent, then the odds of being in sample 2 are the same for marked and unmarked fish.

That is, $\frac{m_{11}}{m_{12}} = \frac{m_{21}}{m_{22}}$, and consequently,

$$\hat{m}_{22} = \frac{\hat{m}_{12}\hat{m}_{21}}{\hat{m}_{11}} = \frac{x_{12}x_{21}}{x_{11}}$$

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	m_{11}	m_{12}	m_{1+}
	No	m_{21}	m^*_{22}	x^*_{2+}
	Total	m_{+1}	m^*_{+2}	m^*_{++}

CONTINGENCY TABLES FOR CAPTURE-RECAPTURE METHOD

Finally,

$$\widehat{N} = \hat{m}_{11} + \hat{m}_{12} + \hat{m}_{21} + \hat{m}_{22} = x_{11} + x_{12} + x_{21} + \frac{x_{12}x_{21}}{x_{11}}$$

$$= \frac{x_{11}(x_{11} + x_{12} + x_{21}) + x_{12}x_{21}}{x_{11}}$$

$$= \frac{x_{11}(x_{11} + x_{12}) + x_{21}(x_{11} + x_{12})}{x_{11}}$$

$$= \frac{(x_{11} + x_{12})(x_{11} + x_{21})}{x_{11}} = \frac{x_{1+}x_{+1}}{x_{11}} = \frac{n_1 n_2}{m}$$

CONFIDENCE INTERVAL FOR N

In many application of capture-recapture, (95%) confidence intervals can be constructed using:

$$\hat{N} \pm 1.96\sqrt{\widehat{Var}(\hat{N})} \quad \text{or} \quad \tilde{N} \pm 1.96\sqrt{\widehat{Var}(\tilde{N})}$$

For example, in our capture-recapture of fish in the lake. A 95% CI for N can be

$$1000 \pm 1.96\sqrt{40000} = 1000 \pm 392 = (608, 1392)$$

Unfortunately, these CIs based on the assumption that $\hat{N}$ or $\tilde{N}$ follows a normal distribution, which is hard to verify from in practice, or when the sample is small.

OTHER PROBLEM WITH NORMAL APPROXIMATION

Consider an example with, $n_1 = 30, n_2 = 20, m = 15$.

Then,

$$\widehat{N} = \frac{n_1 n_2}{m} = \frac{30 \times 20}{15} = 40 \text{ and}$$

$$\widehat{Var}(\widehat{N}) \approx \frac{n_1^2\, n_2\, (n_2 - m)}{m^3} = \frac{30^2 \times 20 \times (20 - 15)}{15^3} = 26.67$$

Using normal approximation, a 95% CI is roughly (30, 50).

The lower bound of 30 is not realistic, because a total of 35 distinct fish were observed.

USING CHI-SQ TEST TO CONSTRUCT CI

We recall the contingency table.

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	$x_{11}(=m)$	x_{12}	$x_{1+}(=n_1)$
	No	x_{21}	x_{22}^*	x_{2+}^*
	Total	$x_{+1}(=n_2)$	x_{+2}^*	$x_{++}^*(=N)$

With the observed data, we may perform a chi-square (goodness-of-fit) test by filling in the missing observation x_{22} with some arbitrary value u.

The 95% CI for x_{22} is then all the values of u for which the null hypothesis of independence for the two samples would not be rejected at 5% level.

EXAMPLE

With our fish in the lake example, when we use $u = 600$,

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	20	180	200
	No	80	$u = 600$	680
	Total	100	780	880

A chi-square test would give a p-value of 0.49, which is greater than 0.05. We then know that the value 600 would be inside the 95% CI for x_{22}, and the value 880 would be inside the 95% CI for N.

2-WAY CONTINGENCY TABLES

- It provides a method for testing the association between the row and column variables in a two-way table.
- The null hypothesis H_0 assumes that there is no association (Independent) between the variables (in other words, one variable does not vary according to the other variable).
- The alternative hypothesis H_1 claims that some association does exist (Dependent).
- **The chi-square test is based on a test statistic that measures the divergence of the observed data from the values that would be *expected* under the null hypothesis of no association.**
- **This requires calculation of the expected values based on the data.**
- **The expected value for each cell in a two-way table is equal to *(row total*column total)/n*, where *n* is the total number of observations included in the table.**

Independent variable

Dependent variable	+ (Control)	
+	E_1	E_2
-	E_3	E_4

Fill in theoretical expected values if no association

$$E_i = n * \text{row total}/n * \text{column total}/n$$
$$= (\text{row total} * \text{column total})/n$$

Independent variable

Dependent variable	+ (Control)	
+	O_1	O_2
-	O_3	O_4

Fill in actual values found in your study

$$x^2 = \sum \frac{[O_i - E_i]^2}{E_i}$$

2-WAY CONTINGENCY TABLES

$\frac{780 * 200}{880} \approx 177.27$

› The expected values in our example are:

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	22.73	177.27	200
	No	77.27	602.73	680
	Total	100	780	880

› The chi-square statistic is

$$\chi^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$$

› This chi-square statistic has degrees of freedom v = (number of rows -1)*(number of columns -1).

› The p-value for the chi-square test is $\Pr(\chi_v^2 > \chi^2)$

2-WAY CONTINGENCY TABLES

› The expected values in our example are:

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	22.73 (20)	177.27 (180)	200
	No	77.27 (80)	602.73 (600)	680
	Total	100	780	880

› The chi-square statistic is

$$\chi^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$$

$$= \frac{(20 - 22.73)^2}{22.73} + \cdots + \frac{(600 - 602.73)^2}{602.73} = 0.478$$

› This chi-square statistic has degrees of freedom $v = (2 - 1) * (2 - 1) = 1$.

› The p-value for the chi-square test is $\Pr(\chi_v^2 > \chi^2) = \Pr(\chi_1^2 > 0.478) = 0.489$

EXAMPLE

On the other hand, when we use $u = 1500$,

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	20	180	200
	No	80	$u = 1500$	1580
	Total	100	1680	1780

A chi-square test would give a p-value of 0.0043, which is less than 0.05. We then know that the value 1500 would be outside the 95% CI for x_{22}, and the value 1780 would be outside the 95% CI for N.

2-WAY CONTINGENCY TABLES

› The expected values in our example are:

		In Sample 2?		
		Yes	No	Total
In Sample 1?	Yes	11.24 (20)	188.76 (180)	200
	No	88.76 (80)	1491.24 (1500)	1580
	Total	100	1680	1780

› The chi-square statistic is

$$\chi^2 = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$$

$$= \frac{(20 - 11.24)^2}{11.24} + \cdots + \frac{(1500 - 1491.24)^2}{1491.24} = 8.160$$

› This chi-square statistic has degrees of freedom $v = (2 - 1) * (2 - 1) = 1$.

› The p-value for the chi-square test is $\Pr(\chi_v^2 > \chi^2) = \Pr(\chi_1^2 > 8.160) = 0.0043$

EXAMPLE

Continuing in this manner, probably using algebra or a computer program, we would find that values of u between 430 and 1198 would result in p-value>0.05.

So, a 95% CI for x_{22} would be (430, 1198).

The corresponding 95% CI for N is obtained by adding the x_{ij} observed , resulting in the interval (710,1478).

MULTIPLE RECAPTURE METHOD

The 2-sample capture-recapture method can be generalized to multiple samples.

For fish example, we may take $k > 2$ random samples, and for each recapture sample, we mark the fish with a different marking (colour or marking position).

For example,

First sample: we mark the left pectoral fin

Second sample: we mark the right pectoral fin

Third sample: we mark the dorsal fin

Then, a fish caught in Sample 4 that has markings on the left pectoral fin and dorsal fin , will be known to have been caught in Sample 1 and Sample 3, but not in Sample 2.

MULTIPLE RECAPTURE METHOD

To estimate N with k samples, the maximum likelihood estimator is the solution $\widehat{N}$ to the following equation (Schnable, 1983):

$$\sum_{i=1}^{k} \frac{(n_i - r_i)M_i}{\widehat{N} - M_i} = \sum_{i=1}^{k} r_i$$

where

n_i is the sample size of sample i,

r_i is the number of recaptured fish in sample i,

M_i is the number of tagged fish in the lake when sample i is drawn.

Further analysis of the data can be performed using more sophisticated techniques.