Estimating population size by mark-and-recapture

Ege Rubak and Jesper Møller
Based on Material by Søren Højsgaard

Introduction

Mark and recapture is a method commonly used in ecology to estimate the size of an animal population. A portion of the population is captured, marked, and released. Later, another portion is captured and the number of marked individuals within the sample is counted. Since the number of marked individuals within the second sample should be proportional to the number of marked individuals in the whole population, an estimate of the total population size can be obtained by dividing the number of marked individuals by the proportion of marked individuals in the second sample. The method is most useful when it is not practical to count all the individuals in the population

This method assumes that the study population is "closed". In other words, the two visits to the study area are close enough in time so that no individuals die, are born, move into or out of the study area. The model also assumes that no marks fall off animals between visits to the field site by the researcher, and that the researcher correctly records all marks.

In this exercise we will take a frequentist approach to the estimation problem, while at a later stage in the course we will consider a full Bayesian analysis.

The setting

We shall use this notation:

- Fixed positive integers:
  - U: The unknown number of unmarked individuals (in the population).
  - M: The number of marked individuals (in first catch/sample). (M is predetermined in the experimental design.)
- Random non-negative integers coming from the main experiment (second sample):
  - Z: The number of unmarked individuals captured in the second catch/sample.
  - R: The number of recaptures (i.e. number of catches in second sample with a mark).

Note that knowing U is equivalent to knowing the total population size $N = M + U$, which is the main parameter of interest.

We organize the quantities in a table where we introduce $K = Z + R$ for the total number of captures in the main experiment (second sample), and recall that only M, Z and R are observed.

	captured	not captured	total
umarked	Z	U-Z	U
marked	R	M-R	M
total	K=Z+R	N-K	N=M+U

Notice that for given data M, Z and R we must have the following lower limit on U:

$$U \geq Z + R - M$$

Exercise 1: Intuitive estimate

The formula for estimating population size N is

$$N \approx \frac{MK}{R}.$$

- Use intuition (and a few simple manipulations of equations) to argue why this is a sensible estimate.

Estimate the population size for the following examples:

1. Catch and mark $M = 9$ squirrels in a tree; later catch $K = 7$ squirrels and notice that $R = 6$ of these are marked.
$$N \approx \frac{M}{R/K} = \frac{9}{6/7} = 10.5$$
2. Catch and mark $M = 100$ mice in the field. Next night, catch $K = 120$ mice of which only $R = 10$ are recaptures.
$$N \approx \frac{M}{R/K} = \frac{100}{10/120} = 1200$$
3. Catch and mark $M = 40$ snails in a garden. Next night, catch $K = 60$ snails of which only $R = 2$ are recaptures.
$$N \approx \frac{M}{R/K} = \frac{40}{2/60} = 1200$$

Notice this:

$$N = \frac{MK}{R} = \frac{M}{R/K}$$

so what controls N is the number M of individuals marked in first catch/sample and the fraction R/K of recaptures in the second sample.

Exercise 2: Statistical model

For the statistical model we shall assume as follows:

- M and U are fixed numbers; M is determined by us; U is unknown to us but it is simply the rest of the population.
- Each individual has the same probability θ of being recaptured independently of whether the individual was marked or not in the first catch. Furthermore, we ignore the effect of sampling without replacement.

1. Specify a sensible distribution for R using M and θ.
$$R \sim B(M, \theta)$$
2. Specify a sensible distribution for Z using U and θ.
$$Z \sim B(U, \theta)$$
3. What is the MLE $\hat{\theta}$ for θ in the first distribution above?
$$\hat{\theta} = R/M$$

4. What is $\mathrm{E}(Z)$ in the second distribution?

$$\mathrm{E}(Z) = U\theta$$

5. Approximate $\mathrm{E}(Z)$ by the observation Z and plug-in $\hat{\theta}$ to obtain an estimate of U in terms of Z, M and R.

$$Z \approx U\hat{\theta} = U\frac{R}{M} \Leftrightarrow U \approx Z\frac{M}{R}$$

6. Use the relations $K = Z + R$ and $N = U + M$ to find an estimate of N in terms of K, M and R.

$$N = U + M \approx \frac{Z}{R}M + M = M(\frac{K-R}{R} + 1) = M\frac{K}{R}$$

The same estimate as before.

Exercise 3: The likelihood for mark and recapture

- Given U and M we have independence between Z and R. Use this to derive the following likelihood:

$$L(\theta, U|M, Z, R) \propto \binom{U}{Z}\theta^{Z+R}(1-\theta)^{U-Z+M-R}$$

which is to be maximized over (θ, U).

$$L(\theta, U|M, Z, R) \propto \binom{M}{R}\theta^{R}(1-\theta)^{M-R}\binom{U}{Z}\theta^{Z}(1-\theta)^{U-Z} \propto \binom{U}{Z}\theta^{R+Z}(1-\theta)^{M-R+U-Z}$$

- For any known value of U show that, the maximum likelihood estimate of θ is

$$\hat{\theta} = \hat{\theta}(U) = \frac{Z+R}{U+M} = \frac{K}{N}.$$

For fixed U the likelihood is proportional to a binomial density with index parameter $N = U + M$ and success probability θ, where we have observed $K = Z + R$ successes:

$$L(\theta|U, M, Z, R) \propto \theta^{Z+R}(1-\theta)^{M+U-(Z+R)}$$

The MLE is then as given above.

- Plug this estimate for θ into $L()$ to obtain a function $\tilde{L}(U) = L(\hat{\theta}(U), U)$ that depends on U alone; this is called the *profile likelihood*.

$$\tilde{L}(U) = L(\hat{\theta}(U), U) = \binom{U}{Z}\hat{\theta}(U)^{Z+R}(1-\hat{\theta}(U))^{U-Z+M-R}$$

- Write an R function `logpl()` which computes the log profile likelihood as as a function of a a single input `U` assuming global variables `M`, `Z` and `R` have already been defined (Hint: use `lchoose()` to get the logarithm of the binomial coefficient):

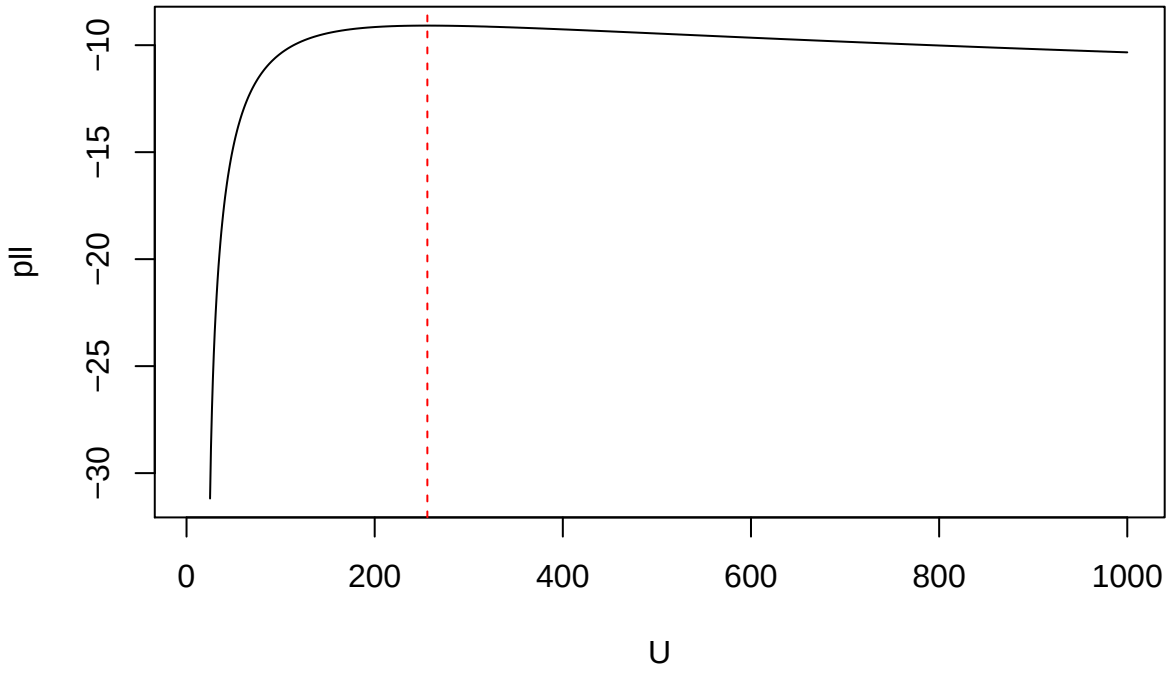
```
logpl <- function(U){
    hat_theta <- (Z+R) / (U + M)
    rslt <- lchoose(U, Z) +
            (Z+R) * log(hat_theta) +
            (U - Z + M - R) * log(1 - hat_theta)
    return(rslt)
}
```

- For the data $M = 21$, $Z = 25$ and $R = 2$, find the value of U that maximizes the log profile likelihood:
    1. by evaluating `logpl()` for all values of U from the lower limit $Z + R - M = 6$ to 1000 and finding the maximal value (Hint: use `which.max()` to find the index of the maximal value). Also plot `logpl()` as a function of `u` from `6` to `1000`.

```
M <- 21 # Marked individuals
Z <- 25 # Unmarked captures
R <-  2 # Recaptured
U <- 6:1000 # Potential values of U
pll <- logpl(U)  # profile log-likelihood values for each U
index <- which.max(pll)
U[index]
```

```
## [1] 256
```

```
plot(U, pll, type = "l")
abline(v=U[index], col = 2, lty = 2)
```



2. by using the built in function minimizer `optim()` in R as follows:

```
M <- 21 # Marked individuals
Z <- 25 # Unmarked captures
R <-  2 # Recaptures
start_value <- Z * M / R
opt <- optim(start_value, logpl, method = "Brent", lower = Z+R-M,
             upper = 10*start_value, control = list(fnscale=-1), hessian = TRUE)
opt$par ## Maximum of `logpl()` (actually minimum of `-logpl()`).
```

```
## [1] 255.7454
```

- Compare the MLE with the intuitive estimate. **Intiutive estimate is 262.5**
- Recall from module 3 that the asymptotic variance of the MLE is $-1/\ell''(\hat{\theta})$, where $\ell''(\hat{\theta})$ is the Hessian of the log-likelihood evaluated at $\hat{\theta}$. Based on this, use `par$hessian` to construct an approximate 95% confidence interval for U.

```
v <- as.numeric(-1/opt$hessian)
v
```

```
## [1] 31996.7
```

```
opt$par
```

```
## [1] 255.7454
```

```
opt$par + c(-2, 2) * sqrt(v)
```

```
## [1] -102.0071  613.4978
```