1.1 The likelihood function

1.1.1 Two simple mark-recapture models

Sampling with replacement:
Suppose that we are studying a closed population of desert mice. In a first visit to the desert, we trap 49 mice, mark them with a red tag and then release them. After some time, we come back to the study area and trap mice again. Each time we capture a mouse, we record whether it is marked or not and release it. That is, we sample mice *with replacement*. With the recorded data, we seek to estimate the total number of individuals in the population. How do we go about writing a probability model for this experiment? Can we build a statistical model to explain how the data arose? Let

- X be the r.v. that counts the number of marked mice recaptures in the second visit.
- x denote the realized value of X.
- m be the number of marked mice in the population.
- t be the total number of mice in the population.
- n be the total number of mice captured in the second visit (23).

Suppose that the experimental data consist of the following results: $x = 5$, $m = 49$, $n = 23$. Here, t is the only unknown quantity. In what follows, after building a probabilistic model for this experiment we derive the Maximum Likelihood (ML) estimate of t.

In order to build a probabilistic model, first note that the experiment "*recording the number of marked mice among the n captured mice*" can be viewed as a sequence of n trials with binary outcome (marked/not marked or "Success"/"Failure"). Let's assume for now that each of these n trials is independent from each other. Then, the probability of observing a marked mouse (*i.e.* the probability of a success) in one of these trials is $\frac{m}{t}$. Likewise, the probability of observing an unmarked mouse is $\left(1 - \frac{m}{t}\right)$. Hence, the probability of a particular sequence of x successes and $n - x$ failures is $\left(\frac{m}{t}\right)^x \left(1 - \frac{m}{t}\right)^{n-x}$. Noting that the total number of such sequences is equal to

$$\begin{aligned} \frac{\#\text{ of ways of assigning } x \text{ marked mice in } n \text{ trials}}{\#\text{ of ways that } x \text{ marked mice can be ordered}} &= \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!} \\ &= \frac{n!}{x!(n-x)!} = \binom{n}{x}, \end{aligned}$$

we get that

$$P(X = x) = \binom{n}{x} \left(\frac{m}{t}\right)^x \left(1 - \frac{m}{t}\right)^{n-x}, \quad x \in \{0, 1, 2, \ldots, n\}.$$

This the binomial distribution with parameters n and m/t and from here on we will write $X \sim \text{Bin}\left(n, \frac{m}{t}\right)$. The probability of drawing 5 marked mice in 23 trials is then:

$$P(X = 5) = \binom{23}{5} \left(\frac{49}{t}\right)^5 \left(1 - \frac{49}{t}\right)^{23-5}.$$

Since t is an unknown quantity, we can view the right hand side (RHS) of the above equation as function of plausible values of t. This function is plotted in Figure 1.

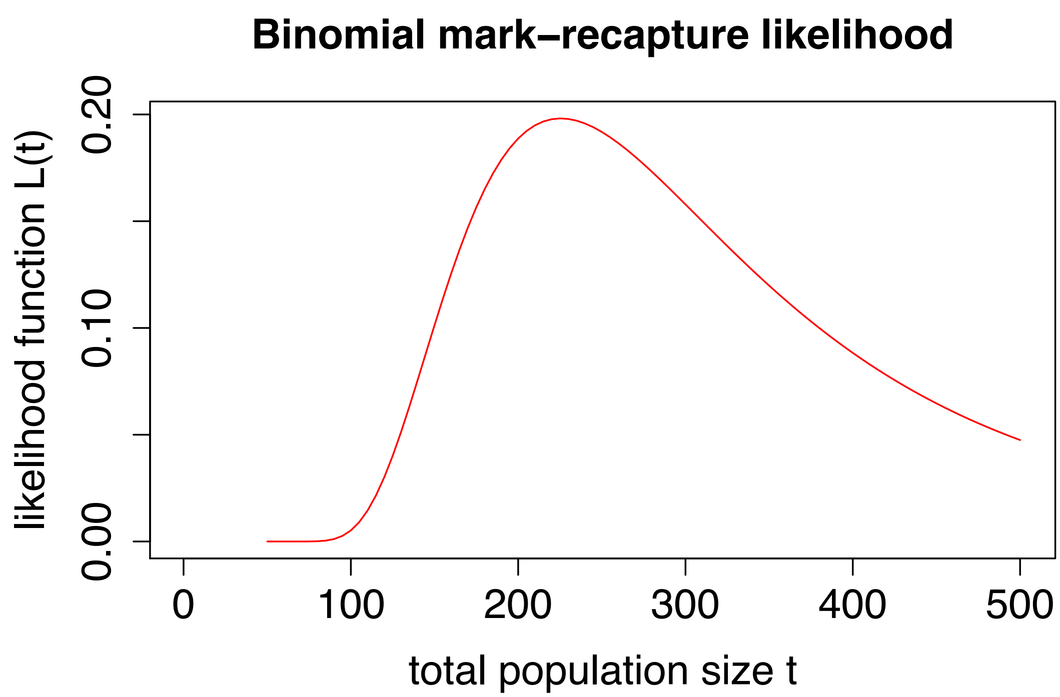

Figure 1: Plot of $P(X = 5)$ as a function of the unknown quantity t.

By doing this exercise, we find that one value around 230 of the unknown quantity t would yield the observed result ($x = 5$) more frequently than any other value. Noting that 'probability' implies a ratio of frequencies and "*about the frequencies of such values we can know nothing whatever*", Fisher (1922) suggested to talk instead of the *likelihood of one value of the unknown parameter being a number of times bigger than the likelihood of another value*. Thus, following Fisher, we refer to the function

$$\ell(t) = \binom{n}{x} \left(\frac{m}{t}\right)^x \left(1 - \frac{m}{t}\right)^{n-x}$$

as the *likelihood function* of t and use it to quantify the relative frequencies with which the values of the hypothetical quantity t would in fact yield the observed sample

(Fisher 1922). The value $\hat{t}$ that maximizes this function is called the Maximum Likelihood (ML) estimate of t. Finding this value analytically is straightforward in this case. To do that, we 1) compute $\ln \ell(t)$, 2) find its derivative with respect to t and 3) set it equal to 0 and solve for t:

1)

$$\ln \ell(t) = \ln \binom{n}{x} + x\ln m - x\ln t + (n-x)\ln(t-m) - (n-x)\ln t$$

2)

$$\frac{d\ln \ell(t)}{dt} = -\frac{x}{t} + \frac{(n-x)}{(t-m)} - \frac{n-x}{t},$$

and **3)**

$$\frac{d\ln \ell(t)}{dt} = \frac{n-x}{t-m} - \frac{n}{t} = 0 \Rightarrow \hat{t} = \frac{nm}{x} = 225.4$$

This estimator of t is known as the "Lincoln-Petersen" index in the scientific literature. Finding $\hat{t}$ using `R` is also straightforward. Instead of doing the above calculations in `R`, we will find the integer ML estimate "by hand": First, let's define a function that computes $\ell(t)$ for various values of t, given the (known) values of x, m and n. We can do that using the function `dbinom` that computes the pmf of the Binomial random variable:

```
binom.like<- function(t,n,m,x){
like<- dbinom(x=x,size=n,prob=(m/t),log=FALSE);
return(like)
}
```

Alternatively, instead of using function `dbinom` we could have used the function `lgamma(x)` that computes $\ln\,(\Gamma(x))$[1] :

```
binom.like<- function(t,n,m,x){
like <- exp(lgamma(n+1)-lgamma(x+1)-lgamma(n-x+1)+x*log(m/t)+(n-x)*log(1-(m/t)));
return(like)
}
```

To do the plot in Figure 1 we type in `R` 's command line :

```
>tvec <- seq(50,500,by=5);
>like.caprecap<- binom.like(t=tvec,n=23,m=49,x=5);
>par(oma=c(1,2,1,1));
>plot(tvec,like.caprecap, col="red",type="l",main="Binomial mark-recapture likelihood",
xlab="total population size t", ylab="likelihood function L(t)",xlim=c(0,501),
cex.main=1.5,cex.lab=1.5,cex.axis=1.5);
```

Finally, the integer ML estimate of t is found by typing

[1] Why exponentiate and then take the log? Because when dealing with very big and very small numbers, it is numerically more stable to compute sums than multiplications.

```
> that<- tvec[which(like.caprecap==max(like.caprecap),arr.ind=T)]
> that
[1] 225
```

Sampling without replacement:

Suppose now that in the second visit we sample n mice *without replacement*. Here again, we let X be the r.v. that counts the number of marked mice recaptures in the second visit. Under this setting we have that

$\binom{t}{n}$ = # of samples of size n from t mice

$1/\binom{t}{n}$ = probability of a particular batch of n mice captured from t mice

$\binom{m}{x}$ = # of ways of choosing x marked mice from m marked mice,

$\binom{t-m}{n-x}$ = # of ways of choosing $n-x$ unmarked mice from $t-m$ unmarked mice and

$\binom{m}{x}\binom{t-m}{n-x}$ = # of ways of choosing x marked and $n-x$ unmarked mice.

Then,

$$P(X=x)=f(x)=\frac{\binom{m}{x}\binom{t-m}{n-x}}{\binom{t}{n}}$$

Hence, X follows the hypergeometric distribution. Note two things: first, if n exceeds $(t-m)$ then some marked animals must appear in the sample. Second, the number of marked animals in the sample cannot exceed m or n. In other words

$$\max(0, m+n-t) \le x \le \min(m,n).$$

The ML estimate of t for this setting may be found using four different methods. The first method consists of drawing a picture of the likelihood function and finding graphically $\hat{t}$. The second approach is to take the derivative of $\ln \ell(t)$, set it equal to 0 and solve for t. However no closed form of $\hat{t}$ can be found in this case, and we have to resort to the third approach: numerical maximization of $\ln \ell(t)$. However, before giving up, we can try to find the integer ML estimate analytically. This last approach consists of finding an integer value of t such that $\ell(t)=\ell(t-1)$. Let $[a]$ denote the greatest integer $\le$ a. Then, first we set $\ell(t)=\ell(t-1)$, solve for t and and take $\hat{t}$ to be $[t]$:

$$\frac{\ell(t-1)}{\ell(t)}-1=0 \Rightarrow \frac{\binom{t-1-m}{n-x}\binom{t}{n}}{\binom{t-m}{n-x}\binom{t-1}{n}}-1=0,$$

and after simplifying (in fact, after some messy algebra) we get

$$(t-m-n+x)t = (t-n)(t-m) \Rightarrow t = \frac{nm}{x}.$$

Rounding to the nearest integer we get $\hat{t} = \left[\frac{nm}{x}\right]$, which is the Petersen index.
It is often the case that multiple independent samples are taken, in which case the setting is:

- $k = \#$ of independent samples taken,
- $t =$ total population size,
- $m_i = \#$ in population that are marked at time of the i^{th} sample,
- $n_i = \#$ captured in the i^{th} sample,
- $x_i = \#$ marked and captured in the i^{th} sample,

and the likelihood function is:

$$\ell(t) = \prod_{i=1}^{k} \frac{\binom{m_i}{x_i}\binom{t-m_i}{n_i-x_i}}{\binom{t}{n_i}}.$$

As an example, consider the following data set: In Alaska, 13 wild goats where captured and marked. Then 3 aerial surveys were done. The results are

flight	**Total # of goats seen**	**Total # of marked goats seen**
1	74	6
2	72	6
3	51	6