Counting Fish: Exploring Estimation by Example

Darryl Nester

Bluffton University, Bluffton, Ohio

Mathematics & Statistics Conference
Miami University
September 29, 2006

www.bluffton.edu/mat/seminar/

A pond contains t tagged fish and x untagged fish. (Or, a bucket contains t red and x black marbles.)

We know t, and we want to estimate x.

How should we do this?

(1) catch-and-release n fish, and observe T_1, the number of tagged fish caught.

(2) collect n fish, and observe T_2, the number of tagged fish caught.

(3) count N_3, the number of fish we must catch-and-release in order to find k tagged fish.

(4) count N_4, the number of fish we must collect in order to find k tagged fish.

How should we compute the estimates for each method? For all four methods, if k out of n fish were tagged, it seems reasonable that:

$$\begin{aligned}
\text{lake ratio of untagged to tagged fish} &\doteq \text{sample ratio of untagged to tagged fish} \\
\frac{x}{t} &\doteq \frac{n-k}{k} \\
x &\doteq \left(\frac{n-k}{k}\right) t
\end{aligned}$$

“Reasonable” is nice, but what is “best” (according to usual statistical standards)?

- We like *unbiased* estimates—we want them to be correct "on the average."

So, if many people used our estimation method (independently, on the same lake), the average of their estimates for x would be (very close to) x.

- We like *efficient* estimates—little variation from one estimate to the next.

If one person estimates 5 and another estimates 2000, that's a bad sign. (The standard deviation should be small.)

A common way of choosing a “good” estimate is to compute the probability (likelihood) of the observed outcome as a function of x. Call this function $L(x)$.

We then choose as our estimate the value of x that maximizes $L(x)$—typically, this amounts to solving a first-semester calculus problem. This is called ...

... the Maximum Likelihood Estimate (MLE).

(First proposed by Gauss in 1821, and rediscovered by Fisher in 1922.)

Method (1): T_1 has a binomial distribution with sample size n and success probability $p = \frac{t}{t+x}$ ("success" = catching a tagged fish).

Then if we catch k tagged fish,

$$L(x) = \binom{n}{k} p^k (1-p)^{n-k} \propto x^{n-k}(t+x)^{-n}.$$

Solving $L'(x) = 0$ leads to $x = \left(\frac{n-k}{k}\right) t$. (Trust me.)

So the MLE agrees with our "reasonable" estimate.

But it's not unbiased or efficient. :-(

Method (2): T_2 has a hypergeometric distribution with population size $t + x$, sample size n, and t special (tagged) fish.

Then if we catch k tagged fish,

$$L(x) = \frac{\binom{t}{k}\binom{x}{n-k}}{\binom{t+x}{n}} \propto \frac{x!\,(t+x-n)!}{(t+x)!\,(k+x-n)!}.$$

Solving $L'(x) = 0$ leads to $x \doteq \left(\frac{n-k}{k}\right) t - \frac{1}{2}$ (for most values of t, n, and k).

Again, this is not unbiased, but it's better ...

Method (3): N_3 has a negative binomial distribution, seeking k successes (tagged fish) with success probability $p = \frac{t}{t+x}$.

Then if we catch n total fish,

$$L(x) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \ \propto \ x^{n-k}(t+x)^{-n}.$$

Déjà vu! This is basically the same as the likelihood function for method (1), so the MLE is $x = \left(\frac{n-k}{k}\right) t$.

And this time, it is unbiased!

The standard deviation is $\sqrt{\frac{x(x+t)}{k}}$.

Method (4): N_4 has a Pòlya distribution with population size $t + x$ and t special (tagged) fish, seeking k successes.

Then if we catch n total fish,

$$L(x) = \frac{\binom{n-1}{k-1}\binom{t+x-n}{t-k}}{\binom{t+x}{t}} \;\propto\; \frac{x!\,(t+x-n)!}{(t+x)!\,(k+x-n)!}$$

(Déjà vu)^2! The MLE is again $x \doteq \left(\frac{n-k}{k}\right) t - \frac{1}{2}$.

Not quite unbiased: the mean is $\left(\frac{t}{t+1}\right) x - \frac{1}{2}$.

The standard deviation is $\left(\frac{t}{t+1}\right) \sqrt{\frac{x(x+t+1)(t-k+1)}{k(t+2)}}$.

Method (4a): Unbiased, but not the MLE.

The mean of N_4 is $\left(\frac{t+x+1}{t+1}\right) k$, so we can produce an unbiased estimate by solving for x:

$$\left(\frac{t+x+1}{t+1}\right) k = n \quad \Longrightarrow \quad x = \left(\frac{n-k}{k}\right)(t+1)$$

(Compare to our "reasonable" equation: $\frac{x}{t} \doteq \frac{n-k}{k}$.)

The standard deviation is larger by a factor of $\frac{t+1}{t}$, but the benefit of an unbiased estimate outweighs the small efficiency penalty.

Appendix: *Finding MLEs*

For methods (1) and (3): If $L(x) \propto x^{n-k}(t+x)^{-n}$, let

$$\mathcal{L}(x) = \ln L(x) = C + (n-k)\ln(x) - n\ln(t+x).$$

This is the "log-likelihood function"; $L(x)$ is maximized when $\mathcal{L}(x)$ is maximized. We find

$$\mathcal{L}'(x) = \frac{n-k}{x} - \frac{n}{t+x},$$

which equals zero when $x = \left(\frac{n-k}{k}\right)t$.

For methods (2) and (4), $L(x) \propto \frac{x!\,(t+x-n)!}{(t+x)!\,(k+x-n)!}$. Expand and cancel the factorials to obtain

$$\frac{x(x-1)(x-2)\cdots(k+x-n+1)}{(t+x)(t+x-1)(t+x-2)\cdots(t+x-n+1)}.$$

Then

$$\mathcal{L}'(x) = \sum_{j=1}^{n-k}\frac{1}{k+x-n+j} - \sum_{j=1}^{n}\frac{1}{t+x-n+j} = S(k+x-n, n-k) - S(t+x-n, n),$$

where $S(a,b) = \sum_{j=1}^{b} \frac{1}{a+j}$. Using the midpoint approximation for a definite integral,

$$S(a,b) = \frac{1}{a+1} + \cdots \frac{1}{a+b} \doteq \int_{a+1/2}^{a+b+1/2} \frac{dx}{x} = \ln\left(\frac{a+b+1/2}{a+1/2}\right),$$

so the MLE is (approximately) the value of x such that

$$\ln\left(\frac{x+1/2}{k+x-n+1/2}\right) = \ln\left(\frac{t+x+1/2}{t+x-n+1/2}\right).$$

The solution to this equation is $x = \left(\frac{n-k}{k}\right)t - \frac{1}{2}$.

The approximation $S(a,b) \doteq \ln\left(\frac{a+b+1/2}{a+1/2}\right)$ is good except when a is small. This leads to slight innaccuracies for extreme values of n or k, but numerical solution to $\mathcal{L}'(x) = 0$ suggests that we shouldn't lose any sleep over this issue.