How Many Fish are in the Pond?

KEYWORDS:
Teaching;
Capture-Recapture Method;
Minitab;
Simulation.

Roger W. Johnson
Carleton College, Northfield, Minnesota, USA.

Summary
This article describes the capture-recapture method of estimating the size of an animal population and illustrates it with a "hands-on" classroom activity. Properties of the estimate, such as its variability, may be explored with the Minitab macro provided.

◆ THE CAPTURE-RECAPTURE ◆ ESTIMATE

THE "capture-recapture method" or "Petersen's method", as it is sometimes referred to by fishery biologists, is a simple method of estimating the size of an animal or human population. A number of applications to the estimation of animal population size are given in Seber (1982). Lock and Moore in Gordon and Gordon (1992) note that the capture-recapture method is "at the heart of proposals to adjust for undercount in the U.S. Census" (c.f. the November 1994 issue of *Statistical Science*). McKeganey (1993) describes its application to estimating the number of prostitutes in Glasgow in an attempt to monitor the spread of HIV. A description of the method now follows. For concreteness, suppose we are interested in estimating the number of fish, N, in a pond. We begin by capturing a sample of size c of the fish, tagging them, and then returning them to the pond. After allowing some time for mixing, a second recaptured sample of size r is taken from all the fish in the pond and the number t of them which are tagged is recorded. An estimate of the number of fish in the pond is

$$\hat{N} = \frac{cr}{t}$$

(Ornithologists and mammalogists sometimes refer to this capture-recapture method estimate as the "Lincoln Index"). In this paper I briefly describe two ways of deriving the above estimate. A "hands-on" classroom activity is also suggested that helps students think about implicit assumptions in using it. I conclude by giving a Minitab macro which can be used to investigate the performance of the estimate.

◆ DERIVATION OF THE ESTIMATE ◆

I will now describe a simple, common sense way of obtaining $\hat{N}$ as an estimate of N. Consider the situation just before the recaptured sample is taken. At this time the population proportion tagged is c/N. When we then take the recaptured sample of size r from this population we observe the sample proportion tagged to be t/r. Setting

Population proportion = Sample proportion

or

$$\frac{c}{N} = \frac{t}{r}$$

we come up with the estimate $\hat{N}$. Implicit in this procedure is that the sample is representative of the population.

A second, more complicated way of coming up with the estimate $\hat{N}$ (not needed in the sequel) involves a maximum likelihood argument. If $T=t$ is the event of obtaining t tagged and r-t untagged fish in the recaptured sample of size r, then

$$P(T = t) = \frac{\binom{c}{t}\binom{N-c}{r-t}}{\binom{N}{r}}.$$

This assumes that the fish mix well so that each selection of r of the N fish in the lake is equally likely. The above, of course, indicates the chances of seeing various t prior to our conducting the recapture. Think about the situation upon completing the capture-recapture. Now the value of t as well as the values of c and r are known and, from this perspective, $P(T=t) = L(N)$ may be interpreted as the "likelihood" of a particular N. It is reasonable to estimate N by choosing that N which maximizes $L(N)$. Going through some algebra (see Rice (1995), for example, for further details) one can show that $L(N)/L(N-1)>1$ precisely when $N<cr/t$. It follows that the value of N that maximizes $L(N)$ is the greatest integer not exceeding cr/t. This so-called maximum likelihood estimate is essentially the estimate $\hat{N}$ given above. Because our population size estimate should be an integer, in what follows redefine $\hat{N}$ to be the maximum likelihood estimate. That is, in what follows take

$$\hat{N} = \left[\frac{cr}{t}\right]$$

where $[x]$ indicates the greatest integer value of x.

◆ A CLASSROOM ACTIVITY ◆

To illustrate the capture-recapture method in the classroom one can use, as suggested by Jeff Witmer of Oberlin College, two different varieties of Pepperidge Farm™ Goldfish crackers (see Figure 1).

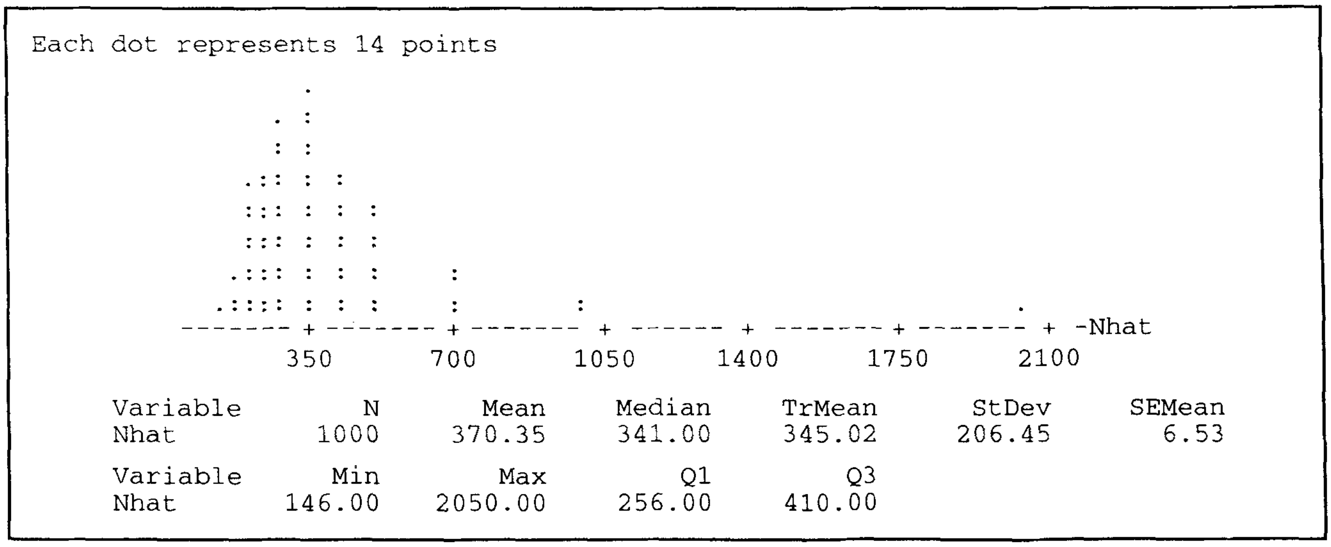
Fig. 1. Pepperidge Farm Goldfish crackers.

In particular, in one class I placed a bag of the original variety in a bowl to correspond to the initial state of the pond. Unknown to my students was the fact that N=323. Students then captured c=50 of these fish and, because they were hard to tag (!), we replaced these with 50 fish of a flavoured variety of a different colour. After mixing the contents of the bowl we found 6 'tagged' fish - fish of the flavoured variety, out of a recaptured sample size of 41 giving the estimate $\hat{N}$=[(50)(41)/6]=341. I then revealed the actual population size. One can, of course, use any objects of uniform shape (e.g. beads, M & M's) of two different colours to represent animals, if the Goldfish crackers are unobtainable.

Some classroom time should then be devoted to talking about the assumptions implicit in using $\hat{N}$. To initiate discussion about such assumptions I ask my students about what they think about the method in the event that the tagged fish were more likely than the untagged fish to be in the recaptured sample. This, of course, leads to an inflated value of t and, consequently, to a value of that is likely to underestimate N. McKeganey (1993) believes this happened in his study as the women "captured" for interviews tended to be the more gregarious. Occasionally I have used tagged fish which are somewhat larger than the untagged fish (Pepperidge Farm chocolate flavoured fish, for example, are somewhat larger than the original variety) to make this point. The discussion then leads into other potential difficulties in using the capture-recapture method such as the possibility of tags falling off.

◆ SIMULATION ◆

In the above classroom exercise the estimate of 341 was fairly close to the actual population size of 323. Does the capture-recapture method usually perform this well? A simulation can be used to find out. The above classroom experiment with N=323, c=50, and



```
Each dot represents 14 points

                 .
               . :
               : :
             .:: : :
             ::: : : :
             ::: : : :
            .::: : : :    :
           .:::: : : :    :       :                          .
         --------+---------+---------+---------+---------+---------+-Nhat
                350      700      1050     1400     1750     2100

     Variable         N      Mean    Median    TrMean      StDev    SEMean
     Nhat          1000    370.35    341.00    345.02     206.45      6.53

     Variable       Min       Max        Q1        Q3
     Nhat        146.00   2050.00    256.00    410.00
```

Fig. 2. Simulation of $\hat{N}$.

r=41 was repeated 1,000 times using the Minitab commands below taking about 2 minutes to run on a 486, 66MHz PC. The 1,000 values of $\hat{N}$ generated had an average of 370.35 with a standard deviation of 206.45 (see Figure 2).

Apparently then, there is substantial bias and variability in using $\hat{N}$ to estimate N. (Note that only 41 values of $\hat{N}$ are possible in our example: If t=1 then $\hat{N}$= 2050, if t=2 then $\hat{N}$= 1025, . . ., if t=41 then $\hat{N}$= 50). Figure 3 shows the Minitab script for producing $\hat{N}$.

```
mtb > let k1=323                  population size
mtb > let k2=50                   number of captured animals tagged
mtb > let k3=k1-k2                untagged animals in population
mtb > let k4=41                   number of animals in recaptured sample
mtb > set c1
data > k3(0) k2(1)                untagged animals 0s,
                                  tagged animals 1s
data > end
mtb > erase c3                    clear column to contain the estimates
mtb > name c3 'Nhat'
mtb > noecho
mtb > exec 'pop.mtb' 1000         produce 1000 estimates of pop. size
mtb > dotplot 'Nhat'              examine the estimates graphically, and
mtb > describe 'Nhat'             numerically (compare mean/median)

where pop.mtb is the text file:
sample k4 c1 c2                   recapture k4 animals
let k5 = sum(c2)                  count recaptured animals tagged
let k6 = round(k2*k4/k5 - 0.5)    estimate the population size
stack k6 'Nhat' 'Nhat'            store the estimate along with the others
```

Fig. 3. Minitab script for producing estimates $\hat{N}$.

The observant reader will have noticed that the estimate $\hat{N}$ is not always well-defined. In particular, if there are no tagged animals in the recaptured sample, then the calculation of $\hat{N}$ involves division by zero. (No problem is encountered in the Minitab simulation in such an event as each such instance results in a missing value being stacked in the estimate column c3. The descriptive statistics given by Minitab exclude any such missing values.) Seber (1982, p. 60) suggests that a slight modification of $\hat{N}$ due to Chapman (1951), namely

$$\tilde{N} = \frac{(c+1)(r+1)}{t+1} - 1,$$

may be preferred to $\hat{N}$. The reasons for using $\tilde{N}$ instead of $\hat{N}$ extend beyond the simple avoidance of a divide by zero problem, see Seber (1982) and the references therein for details regarding the theoretical properties of $\hat{N}$ and $\tilde{N}$. Using $\tilde{N}$ instead of $\hat{N}$ in the above classroom experiment gives an estimate for N of 305. By appropriately modifying the line involving k6 in the Minitab commands above one can generate simulated values of $\tilde{N}$. In fact, for the same 1,000 simulated capture-recaptures used to generate the 1,000 values of $\hat{N}$ above, the corresponding 1,000 values of $\tilde{N}$ had an average of 317.09 and a standard deviation of 120.59 (see Figure 4).

Apparently $\tilde{N}$ has little or no bias and is much less variable than $\hat{N}$. Again, see Seber (1982) for further details.

```
Each dot represents 14 points

                      .
                    . :
                    : :
                   :: :   :
                  .:: :   :   :
                  ::: :   :   :
                 .::: :   :   :   :
               ..:::: :   :   :   :        :                  .
          -------+---------+---------+---------+---------+--------Ntilde
                200       400       600       800       1000


     Variable        N      Mean    Median    TrMean     StDev    SEMean
     Ntilde       1000    317.09    305.00    304.67    120.59      3.81

     Variable      Min       Max        Q1        Q3
     Ntilde     141.80   1070.00    237.00    356.00
```

Fig. 4. Simulation of $\tilde{N}$.

References

Chapman, D. (1951), "Some properties of the hypergeometric distribution with applications to zoological censuses", *University of California Publications in Statistics*, **1**, 131-160.

Gordon, F. and Gordon, S., editors, (1992), *Statistics for the Twenty-First Century*, MAA Notes Number 26, The Mathematical Association of America, 102.

McKeganey, N. (1993), Stalking HIV in the red light district, *New Scientist*, June 12, 22-23.

Rice, J. (1995), *Mathematical Statistics and Data Analysis*, second edition, Duxbury Press, Belmont, CA, 13-14.

Seber, G. (1982), *The Estimation of Animal Abundance and Related Parameters*, second edition, Charles Griffin, London.

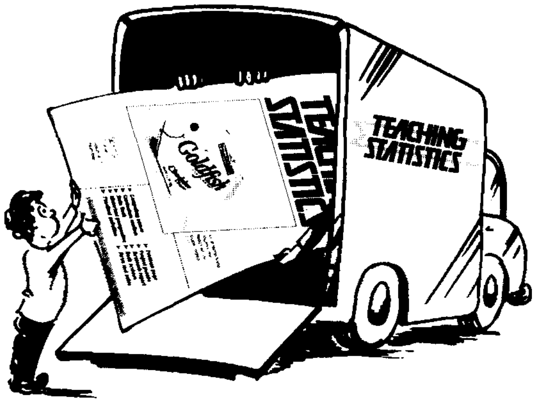
TEACHING STATISTICS

On the move...

The General Office for all administration for the journal has now moved from Sheffield to:

RSS Centre for Statistical Education
University of Nottingham
Nottingham
NG7 2RD

Telephone: 0115-951-4911
Fax: 0115-951-4951



- The editorial office continues at Coventry University.
- The Centre for Statistical Education at Sheffield has closed and all publications formerly available from there are now available from the RSS Centre at Nottingham.