Additional Notes for Hypergeometric Random Variables

Capture – Recapture Sampling

This is similar to exercise 18.18 (p. 263)

This is a method that is used to estimate the size of wildlife populations

Suppose that an unknown number, N, of bluegills inhabit a small lake and that we want to estimate that number. One procedure for doing so, often referred to as the capture-recapture method, is to proceed as follows:

1. Capture and tag some of the fish, say 250 and then release the fish back into the lake and give them time to disperse.
2. Capture some more of the animals, say 150, and determine the number that are tagged, say 16. These are the recaptures.
3. Use the data to estimate N.

Note: we are trying to find N here

Step 1:
We are interested in the marked animals, M. The percentage of marked animals is $\frac{M}{N}$.

Step 2:
Our sample size is n = 150, The number that are tagged is x = 16. This proportion is $\frac{x}{n}$.

Step 3:
assuming that n is large enough, we would expect that
$\frac{x}{n} \approx \frac{M}{N} \Longrightarrow N \approx \frac{Mn}{x} = \frac{250 \cdot 150}{16} = 2344$
where I rounded the last value to an integer.

This is an approximation, how can we do this more rigorously?

The method of choice is using the method of maximum likelihood. That is, we want to choose the value of N for which the observed value of X has the highest probability.

The distribution is a hypergeometric because we are choosing items (in this case fish) without replacement.

The parameters:
N: the total number of fish in the lake
M: the number of tagged fish (Step 1)
n: the number of fish that we capture the second time (Step 2)
X: the number of fish that are tagged when we capture them the second time (Step 2)

The distribution for a hypergeometric is:
$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, N \in \mathbb{N}$

We want the value of N which will maximize $\frac{P(N)}{P(N-1)}$ which is called $\widehat{N}$.

$$\frac{P(N)}{P(N-1)} = \frac{\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}}{\frac{\binom{M}{x}\binom{N-1-M}{n-x}}{\binom{N-1}{n}}} = \frac{\frac{\binom{N-M}{n-x}}{\binom{N}{n}}}{\frac{\binom{N-1-M}{n-x}}{\binom{N-1}{n}}} = \frac{\frac{\frac{(N-M)!}{(N-M-n+x)!\,(n-x)!}}{\frac{N!}{(N-n)!\,n!}}}{\frac{\frac{(N-1-M)!}{(N-1-M-n+x)!\,(n-x)!}}{\frac{(N-1)!}{(N-1-n)!\,n!}}} = \frac{\frac{\frac{(N-M)!}{(N-M-n+x)!}}{\frac{N!}{(N-n)!}}}{\frac{\frac{(N-1-M)!}{(N-1-M-n+x)!}}{\frac{(N-1)!}{(N-1-n)!}}}$$

$$= \frac{(N-M)(N-n)}{(N-M-n+x)N} = \frac{N^2 - MN - nN + Mn - xN + xN}{(N^2 - MN - nN + xN)}$$

$$= \frac{N^2 - MN - nN + xN + Mn - xN}{(N^2 - MN - nN + xN)} = 1 + \frac{Mn - xN}{(N-M-n+x)N}$$

therefore,

P(N) > P(N-1) if Mn > xN or N < $\frac{Mn}{x}$

Therefore P(N) is a maximum when $\widehat{N} = \left\lfloor \frac{Mn}{x} \right\rfloor$= 2343

The only difference between the rigorous derivation and the heuristic one is that we use the 'floor' for the latter.

Binomial Approximation to the Hypergeometric?

The difference between these two random variables is that the Binomial random variable is with replacement and the Hypergeometric is without replacement.

Therefore this question can be rephrased to ask when is without replacement approximately the same as with replacement?

Answer: when the sample size is small relative to the population size.
In practice, if $\frac{n}{N} < 0.05$ then the distributions are practically identical.

For the following comparisons, let X be the Binomial random variable and Y be the Hypergeometric random variable

Mass: It can be shown that the mass are the same in this situation.

Expected Values:
E(X) = np
$E(Y) = n\frac{M}{N} \approx np$

Variance:

Var(X) = np(1 – p)

$Var(Y) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1} \approx np(1-p)$

When N is large n ≈ 1 so $\frac{N-n}{N-1} \approx 1$

The following is a comparison of the hypergeometric probability to the binomial probability:

Table 5.15 Comparison of the hypergeometric and binomial distributions: $N = 1000$, $n = 8$, and $p = 0.2$.

Successes x	Hypergeometric probability	Binomial probability
0	0.1666	0.1678
1	0.3361	0.3355
2	0.2949	0.2936
3	0.1469	0.1468
4	0.0454	0.0459
5	0.0089	0.0092
6	0.0011	0.0011
7	0.0001	0.0001
8	0.0000	0.0000

In this case, $\frac{n}{N} = \frac{8}{1000} = 0.008 < 0.05$. Note that the two probabilities are the same to 3 decimal places.