

## STAT 325 Introduction to Probability Models Spring 2012

### Investigation 1: Counting Goldfish (assigned on Thur Apr 5, due on Mon Apr 16)

*You may work with one partner, submitting one report with both names, provided that you both contribute substantially to the work. Word-processed reports are strongly preferred to hand-written ones. Please integrate computer output into your report.*

Suppose that you want to estimate the number of fish in a lake; call this number  $N$ . In a technique known as capture-recapture, you first capture a random sample of fish, let's say 20 fish. You mark these 20 fish so they can be identified later and then release them back into the lake. You allow them ample time to mingle thoroughly with the other fish, and then you capture another random sample of fish, let's say 10 fish. Then you count how many of the fish captured in the second sample are marked as having already been caught in the first sample. Let's say you observe that 3 of these 10 fish had already been caught the first time.

Let  $E$  denote the evidence/data that 3 of the 10 fish caught in the second sample had also been caught in the first sample.

- What are the possible values of  $N$ ? In other words, for what values of  $N$  is  $\Pr(E)$  positive?
- Determine  $\Pr(E)$ , as a function of  $N$ .
- Use R to graph this function, for all values of  $N \leq 200$  that have positive probability. Submit the graph and also the R code that you use to produce the relevant vectors. [R advice: You might use  $N$  and `prob` for the names of the vectors.]
- What value of  $N$  gives the largest probability of producing the observed evidence (that 3 of the 10 fish captured in the second sample were marked)? Also report that probability. [R tip: You can use the command `which.max(prob)` to determine the index number for the largest value in the vector of probabilities. Then use `N[which.max(prob)]` to give the value for the vector  $N$  at the index number.]

A simple, non-probabilistic approach to this problem of estimating the value of  $N$  might equate the proportion of marked fish in the second sample to the proportion of marked fish in the entire lake.

- Set these two proportions equal to each other and solve for  $N$ . Is the answer close to your answer to d)?

A Bayesian approach to estimating the value of  $N$  would begin with prior probabilities that  $N$  equals particular values. Then a Bayesian would update the probabilities of those values of  $N$ , based on the evidence/data  $E$ , using Bayes' Theorem.

Suppose for now that before taking the second sample we consider only 5 possible values of  $N$ : 25, 50, 75, 100, 125. Furthermore, before taking the second sample, we consider these 5 possibilities to be equally likely. Let the event  $N_i$  mean that  $N = i$ , for  $i = 25, 50, 75, 100, 125$ .

f) Use Bayes' Theorem to determine the updated probabilities of these 5 events  $N_{25}, N_{50}, N_{75}, N_{100}$ , and  $N_{125}$ , conditional on the evidence/data  $E$  that 3 of the fish captured in the second sample were marked. [R advice: You could do these calculations by hand. If you want to use R, you might use `prior` and `updated` for the names of the vectors.]

g) Which of these 5 possible values of  $N$  is most likely, given the evidence/data  $E$ ? Which is least likely?

h) Which of these 5 possible values of  $N$  have become more likely, given the data/evidence, than they were originally? Which have become less likely?

i) Repeat f), g), and h), but now starting with the following non-equally-likely prior probabilities:  $\Pr(N_{25}) = .1$ ,  $\Pr(N_{50}) = .2$ ,  $\Pr(N_{75}) = .3$ ,  $\Pr(N_{100}) = .3$ ,  $\Pr(N_{125}) = .1$ .

Now consider all 181 integer values of  $N$  from 20 to 200 as equally likely, prior to observing the second sample.

j) Use Bayes' Theorem (with R) to determine the updated probabilities, given the evidence/data  $E$ . Submit a graph of these updated probabilities vs.  $N$ , and also submit the R code that you use to produce the calculations and graph. [R advice: Use `prior` and `updated` for the names of the vectors. R tip: To create a vector  $x$  of length  $n$  that consists of only the value  $k$  repeated  $n$  times, use: `x = rep(k, n)`.]

k) Which value of  $N$  has the largest updated probability? What is that probability?