Homework 4

Megan Robertson

Monday, September 21, 2015

Suppose we had the following hypotheses:

- H_0: Rate of fish biting less than one per hour
- H_1: Rate of fish biting between one and five per hour
- H_2: Rate greater than five per hour

Describe Bayesian approach assessing these hypotheses based on data collected for n fishermen with fisherman i fishing for t_i hours.

The first step is the set a prior over the hypotheses (models). In this example, the probability of each model is 1/3. Then, each parameter in the models needs to be assigned a prior. Each fisherman has n_i bites in the time t_i that they spend fishing. Each catch occurs at a rate per hour that follows a Poisson distribution. Thus, each of the j fisherman catches a total number of fish that is $\sum_{i=1}^{j}$ Poi(θ) = Poisson(θt_i). The next step is to find marginal likelihoods under each hypothesis and use these to calculate the Bayes factors, BF_{10}, BF_{20}, BF_{01}, BF_{21}, BF_{02}, and BF_{12}. These Bayes factors would then be used to find P(M=0|data) = 1/(1+BF_{10}, BF_{20}), P(M=1|data) = 1/(1+BF_{01}, BF_{21}), and P(M=2|data) = 1/(1+BF_{02}, BF_{12}). Then, the conclusion of the test would be the hypothesis that has the highest posterior probability.

Simulate data under one hypothesis and see how approach does.

Suppose that H_1 is true. (see derivations and simulation code below)

The test performed very well for this simulation. The probability of the null being true was found to be 1.75 e-27 and the probability of the second hypothesis was 0.15. The probability of the first hypothesis was 1. Thus the approach performed very well since the highest probability of any hypothesis was the one that was true.

What happens to performance as we had more narrow hypothesis?

As there are more narrow hypotheses, the performance of the test will worsen. If the intervals of possible rates are smaller, then it is more likely that the data will indicate that the true rate is in an incorrect interval. Suppose the possible hypotheses had the following intervals for the rates of fish per hour: 0-1, 1-1.5, 1.5-3, 3-4, 4-5, and 5+. If the true rate were 1.6, there is a higher probability that the data could indicate that the true interval was in the 1-1.5 interval. This would be less likley to occur if the intervals were wider.

Describe how to test whether fishing better in one pond than another (no simulation, just describe)

In order to test this, I would establish two hypotheses. The null hypothesis would be that neither pond is better for fishing (the rates of fish caught per hour would be the same). The alternative hypothesis would be that one pond is better than the other. I would assign prior probabilities to the models, each would have a probability of 1/2 of occurring. The next step would be to assign priors to θ_A (the overall rate of bites per hour in pond A) and θ_B (the overall rate of bites per hour in pond B). These priors would be based on historical data or chosen to be uninformative. Then, the marginal likelihoods for each hypothesis would be calculated and used to find the Bayes Factors. The Bayes Factors would then be used to find the probability of each hypotheses given the data. Then the hypothesis with the highest probability would be concluded to be the correct hypothesis.

Code Section

Problem 2

```r
set.seed(100)
time.vec = rep(1:5, 2)  #vector of number of hours fished
true.theta = 3.5  #this is because hypothesis one is true
total.fish = true.theta * time.vec
c.fun = function(a, b) {
    return(1/beta(a, b))
}

a = 0.01
b = 0.01/0.5
likelihood.null = (pgamma(1, sum(total.fish) + a - 1, sum(time.vec) +
    b) - pgamma(0, sum(total.fish + a - 1), sum(time.vec) + b))/(pgamma(1,
    a, b) - pgamma(0, a, b)) * (c.fun(a, b)/c.fun(sum(total.fish) +
    a - 1, sum(time.vec + b))) * ((b^a)/gamma(a))

c = 0.01
d = 0.01/3.5
likelihood.one = (pgamma(5, sum(total.fish) + c - 1, sum(time.vec) +
    d) - pgamma(1, sum(total.fish + c - 1), sum(time.vec) + d))/(pgamma(5,
    c, d) - pgamma(1, c, d)) * (c.fun(c, d)/c.fun(sum(total.fish) +
    c - 1, sum(time.vec + d))) * ((d^c)/gamma(c))

e = 0.01
f = 0.01/6
likelihood.two = (1 - pgamma(1, sum(total.fish) + e - 1, sum(time.vec) +
    f))/(1 - pgamma(1, e, f)) * (c.fun(e, f)/c.fun(sum(total.fish) +
    e - 1, sum(time.vec + f))) * ((f^e)/gamma(e))

# calculating the Bayes factor
BF.01 = likelihood.null/likelihood.one
BF.21 = likelihood.null/likelihood.two
BF.10 = likelihood.one/likelihood.null
BF.20 = likelihood.two/likelihood.null
BF.02 = likelihood.null/likelihood.two
BF.12 = likelihood.one/likelihood.two

# posterior probabilities
prob.null = 1/(1 + BF.10 + BF.20)  #PR(M=0|data)
prob.one = 1/(1 + BF.01 + BF.21)  #PR(M=1|data)
prob.two = 1/(1 + BF.02 + BF.12)  #PR(M=2|data)
```