Effect of bottom coverage to larval presence

Week2-ex3, solution

In this exercise you need the following special result. Assume we have a Binomial observation model

$$p(y|\theta, N) = \text{Bin}(y|N, \theta) \propto \theta^y (1-\theta)^{N-y}. \tag{1}$$

The number of trials, N, is considered to be fixed and the parameter θ is given a Beta prior

$$p(\theta) = \text{Beta}(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{2}$$

Comparing the observation model and prior gives insight that the Beta prior corresponds to $\alpha - 1$ prior successes and $\beta - 1$ prior failures. A special case is $\alpha = \beta = 1$ when the prior is uniform on the interval $[0, 1]$. The posterior distribution of θ is now also Beta distribution (see BDA 3, Chapter 2 for more details)

$$\theta|N, y \sim \text{Beta}(\alpha + y, \beta + N - y). \tag{3}$$

Note! The Beta posterior arizes only if we use Beta prior. If we use some other prior than Beta distribution the posterior will not be Beta distribution.

Using the above result solve the below problem.

Problem statement

We analyze the data presented by Veneranta et al. (2013). White fish is a fresh water origin fish species that is found also from the northern parts of the brackish water Gulf of Bothnia. The species is important for both commercial and recreational fisheries. White fish spawn in shallow coastal areas and former studies suggests that the survival of white fish larvae is decreased by algal or other bottom vegetation which have been increasing throughout Finnish and Swedish coastal region due to eutrophication. Hence, we want to study whether there is difference in the probability of presence of white fish in clear and vegetated areas.

A number of sites (n=653) along the Finnish and Swedish coastal region in the Gulf of Bothnia were sampled during 2009-2011. At each site, researchers sampled a volume of water using a fine meshed net and recorded whether or not white fish larvae were detected or not. Each site was classified with COVERAGE: 0 = clear and 1 = covered with vegetation. The data contains also other covariates and spatial information but these will be considered in later weeks.

The data (contingency table) for this exercise can be downloaded and formed as follows

```
data = read.csv("white_fishes_data.csv")

y = table(data$WHIBIN, data$BOTTOMCOV)
colnames(y) <- c("COV=0","COV=1")
rownames(y) <- c("y=0","y=1")
print(y)
```

```
##
##       COV=0 COV=1
##   y=0    65   104
##   y=1   212   121
```

The variable y groups the sampling sites into locations with respect to the vegetation cover (COV=0 vs. COV=1) and occurrence of white fish larvae (y=0 vs. y=1).

Let's assume that the outcomes of sampling occasions (presence/absence of whitefish) are independent Bernoulli (Binomial with sample size 1) distributed random variables with success probabilities θ_0 for sites with no vegetation cover and θ_1 for sites with vegetation cover.
Let's further assume that there is no prior information on θ_c, $c \in \{0, 1\}$ so that their prior is uniform between 0 and 1, and that the parameters θ_c are mutually independent.

1. Write down the equation for the posterior distribution for both θ_0 and θ_1
2. Sample random draws from both posterior distributions, draw a histogram of the samples and report the posterior mean and standard deviation.
3. Visualize the posterior distribution of $\phi = \theta_0 - \theta_1$ and calculate the posterior probability that $\theta_1 < \theta_0$
4. Analyze and discuss the sensitivity of the results to the choice of the prior distribution

Solution

1.

The uniform prior over $[0, 1]$ corresponds to $\theta_c \sim \text{Beta}(1, 1)$. Let's denote by N_c the total number of sampling sites with bottom coverage status c and by y_c the number of sampling sites where white fish larvae were detected and bottom coverage status was c. Since the outcomes at sampling sites are independent given θ_0 and θ_1 the observation model is $y_c \sim \text{Bin}(\theta_c, N_c)$. Hence, using the result given at the start of the exercise, we can write the posterior probability distribution as

$\theta_c \sim \text{Beta}(y_c + 1, N_c - y_c + 1)$

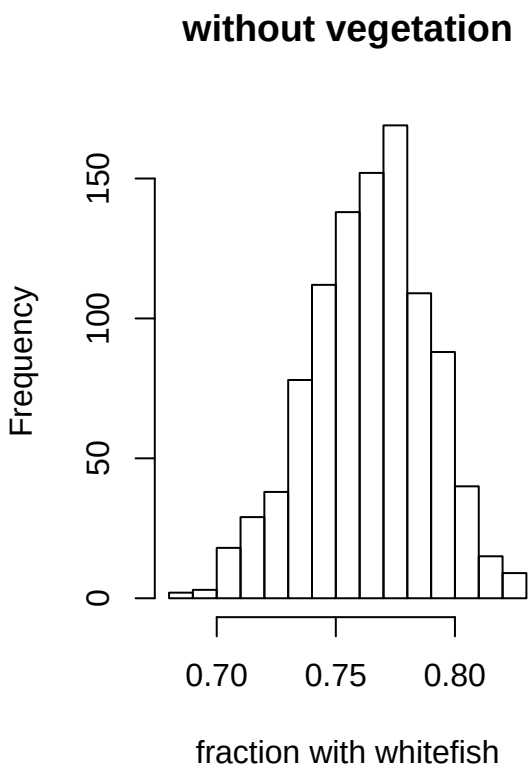
2.

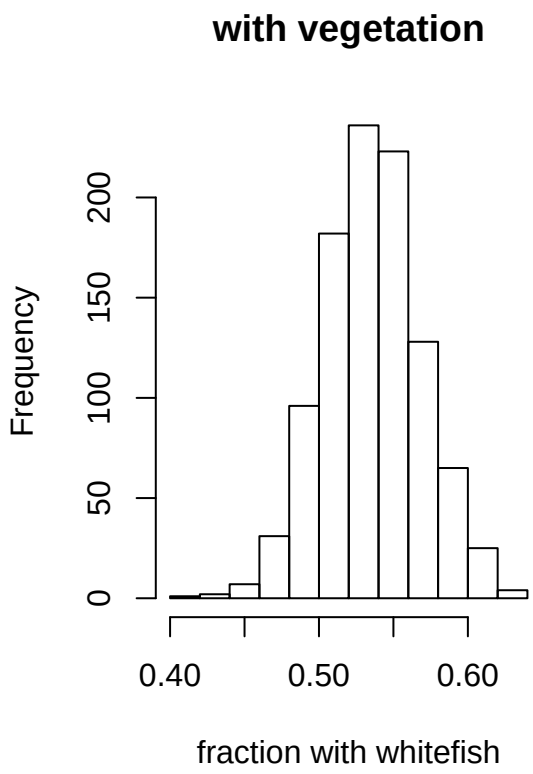
Let's sample from the posterior distributions

```
# calculate N_c and y_c
N_0 = sum(y[,1])
N_1 = sum(y[,2])
y_0 = y[2,1]
y_1 = y[2,2]
# sample from the posterior
theta_0 = rbeta(1000, y_0+1, N_0-y_0+1)
theta_1 = rbeta(1000, y_1+1, N_1-y_1+1)
# # Note this is the same as
# theta_0 = rbeta(1000, y[2,1]+1, y[1,1]+1)
# theta_1 = rbeta(1000, y[2,2]+1, y[1,2]+1)
```

Let's then draw the histogram of the posterior samples and calculate the posterior mean and standard deviation

```
par(mfrow=c(1,2))
hist(theta_0, main="without vegetation", xlab="fraction with whitefish")
hist(theta_1, main="with vegetation", xlab="fraction with whitefish")
```






```
# Posterior mean and standard deviation
cbind(mean(theta_0),mean(theta_1))
```

```
##           [,1]      [,2]
## [1,] 0.7639945 0.5357722
```

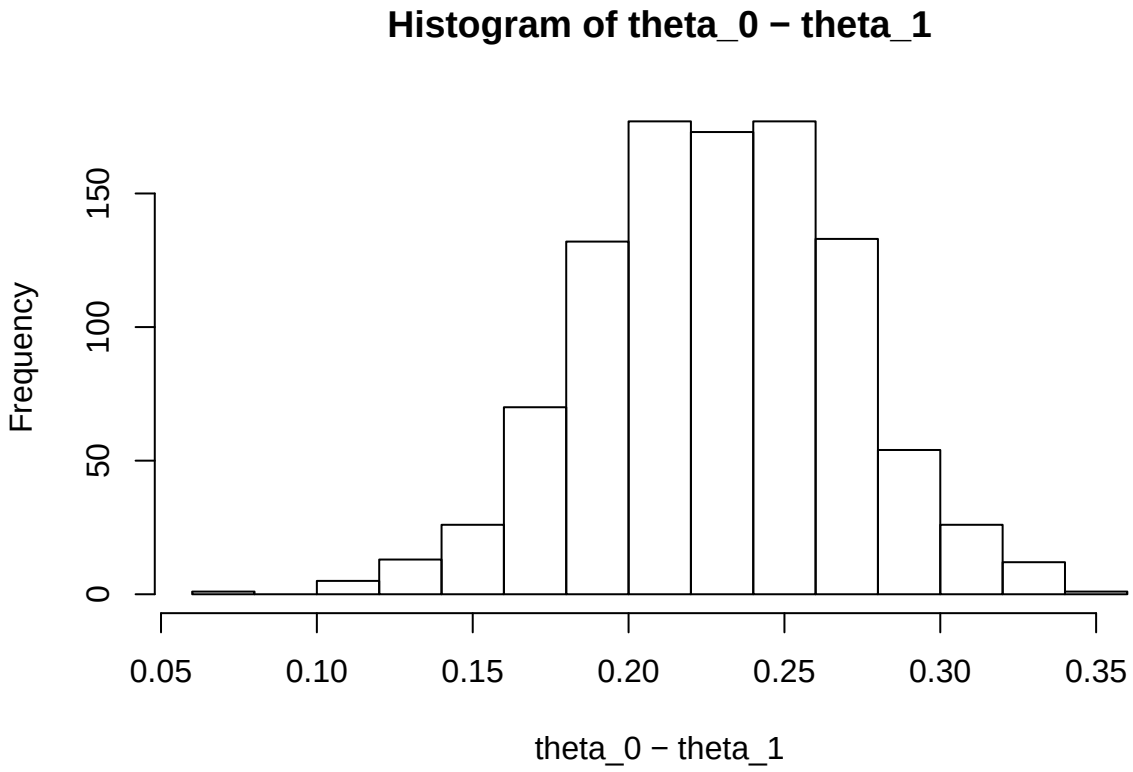
```
cbind(sd(theta_0),sd(theta_1))
```

```
##            [,1]      [,2]
## [1,] 0.02476903 0.0329427
```

3.

The posterior distribution of $\beta_0 - \beta_1$ and the probability that $\beta_0 > \beta_1$ are

```
hist(theta_0-theta_1)
```




```
mean(theta_0-theta_1>0)
```

```
## [1] 1
```

4

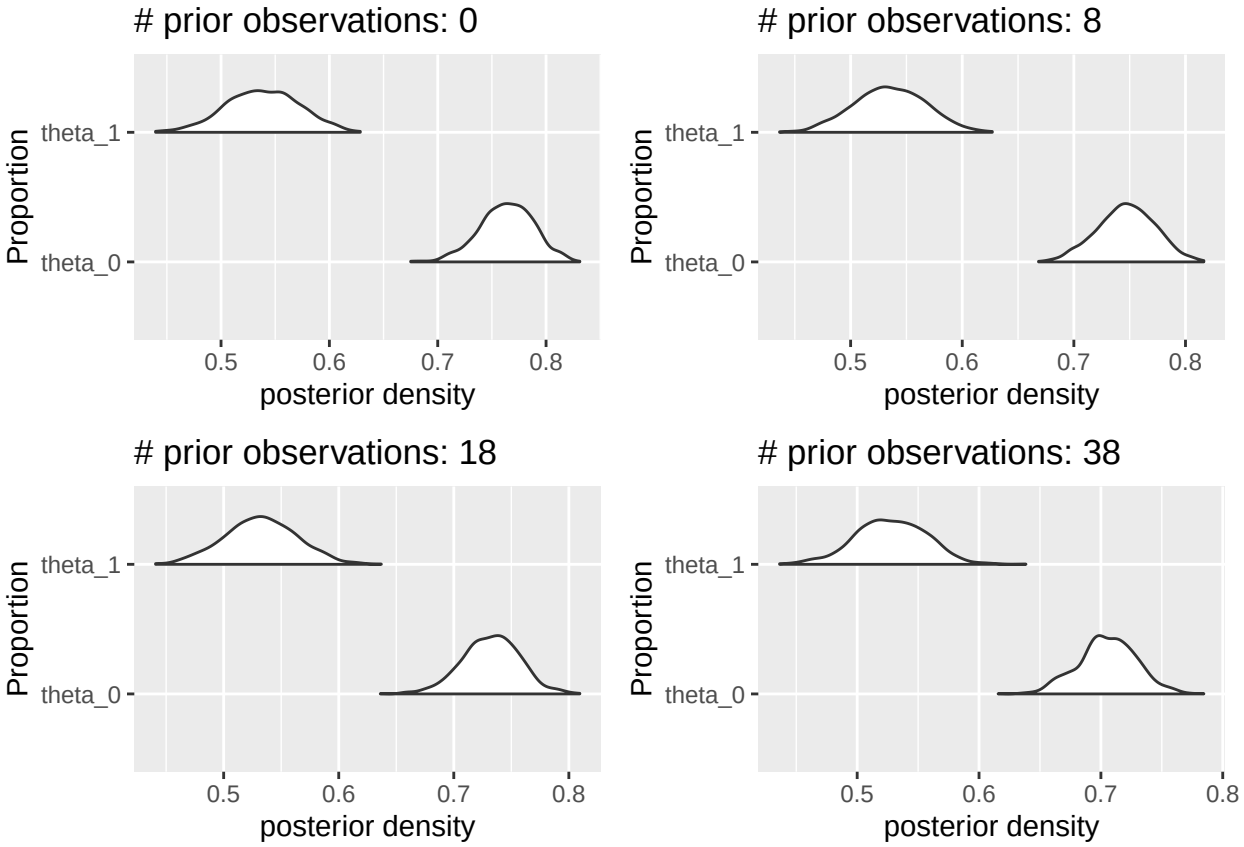
We can test the sensitivity of the posterior distribution to the choice of prior distribution by calculating the posterior distribution and its summaries with alternative prior distributions. Since we assumed that we don't have any prior knowledge on θ_0 or θ_1, we will test priors that have the same mean (0.5) but different amount of prior information. Below, we plot the posterior densities with different amount of prior data; that is, if $\theta \sim \text{Beta}(\alpha, \beta)$ then $\alpha + \beta - 2$ can be interpreted as the number of prior observations (compare to the equation of the posteior distribution above).

```
library(ggplot2)
library(gridExtra)
library(see)

#par(mfrow=c(2,2))
nsamp = 1000
i=1
p=c()
count = 1
theta.fac=list()
for (i in c(1,10,20,40)){
   # sample from the posterior
   theta_0 = rbeta(nsamp, y_0+i, N_0-y_0+i)
   theta_1 = rbeta(nsamp, y_1+i, N_1-y_1+i)
   # put samples into data frame in order to allow ggplotting
   theta.fac[[count]] <- data.frame(
```

```
      name=c( rep("theta_0",nsamp), rep("theta_1",nsamp)  ),
      value=c( theta_0, theta_1)
   )
   count = count+1
}
# Make a ggplot
p1 <- ggplot(theta.fac[[1]], aes(x=name,y=value)) + geom_violinhalf() +
   coord_flip() + labs(title="# prior observations: 0",
                       x="Proportion", y = "posterior density")
p2 <- ggplot(theta.fac[[2]], aes(x=name,y=value)) + geom_violinhalf() +
   coord_flip() + labs(title="# prior observations: 8",
                       x="Proportion", y = "posterior density")
p3 <- ggplot(theta.fac[[3]], aes(x=name,y=value)) + geom_violinhalf() +
   coord_flip() + labs(title="# prior observations: 18",
                       x="Proportion", y = "posterior density")
p4 <- ggplot(theta.fac[[4]], aes(x=name,y=value)) + geom_violinhalf() +
   coord_flip() + labs(title="# prior observations: 38",
                       x="Proportion", y = "posterior density")

# arrange ggplots into grid
grid.arrange(p1, p2, p3, p4, nrow = 2)
```



The posterior distributions get slightly narrower and move towards 0.5 as we increase the informativeness prior (that is, the size of prior data.). However, the analysis is not too sensitive to moderate amount of prior information (that is $\alpha + \beta \leq 38$) since the overall conclusions of the analysis do not change. **Note on grading**. One point from understanding the idea of how to test prior sensitivity; that is, you need to calculate your posterior distribution with alternative priors and compare them - either visually or, for example, with

mean, variance etc. Two points from actually executing this test.

Grading

Total 10 points Three points from both part 1 and 4. Two points from both part 2 and 3. In all sub-questions you should give point if the idea of the solution is correct. One extra point if the solution is in principle correct but contains minor typo/bug. Full points from totally correct answer.

References

Lari Veneranta, Richard Hudd and Jarno Vanhatalo (2013). Reproduction areas of sea-spawning Coregonids reflect the environment in shallow coastal waters. Marine Ecology Progress Series, 477:231-250. http://www.int-res.com/abstracts/meps/v477/p231-250/