Censored observations

Week2-ex1, solution

Exercise instructions

Suppose you have a Gamma($\alpha = 1, \beta = 1$) prior distribution on the parameter λ which corresponds to the expected number of ship ice besetting events (=events where a ship gets stuck in ice) during 1000 nautical miles in ice infested waters. The number of besetting events, y per distance d (nm) is modeled with a Poisson distribution Poisson($\lambda \times d$). The hyper-parameter α is the shape and β is the inverse scale parameter. You are told that during winters 2013-2017 category A ice breakers traveled in total 6560 nautical miles in the Kara Sea (a sea area in the Arctic Sea). Within this distance they experienced in total more than 2 but less than 7 ice besetting events.

1) Write down the equation for the posterior probability density for λ.
2) Discretize interval $[0, 3]$ into 100 equally spaced intervals, calculate the unnormalized posterior probability density at each of the discrete bins, normalize the densities to sum up to one and draw the posterior density at the discrete cells.
3) Using the discretized values of λ and their corresponding posterior density values
   a) draw the posterior cumulative distribution function of λ and
   b) calculate the posterior probability that $\lambda > 1$.
4) Draw the posterior probability density for λ in case where you are told that the exact number of besetting events is 6. What are the differences between the posterior densities from 2) and 4)?
5) Calculate the posterior predictive probability mass function for number of besetting events, $\tilde{y}$, within a 500 nm distance by using the posterior predictive density from step 2) and draw it. Note, you can restrict $\tilde{y} \in [0, 4]$

Model answer

1-3)

The posterior probability density function in case of censored observation is

$$p(\lambda|2 < y < 7, d = 6.56) \propto (\text{Poisson}(3|\lambda \times d) + \text{Poisson}(4|\lambda \times d) + \text{Poisson}(5|\lambda \times d) + \text{Poisson}(6|\lambda \times d))\,\text{Gamma}(\lambda|1, 1)$$

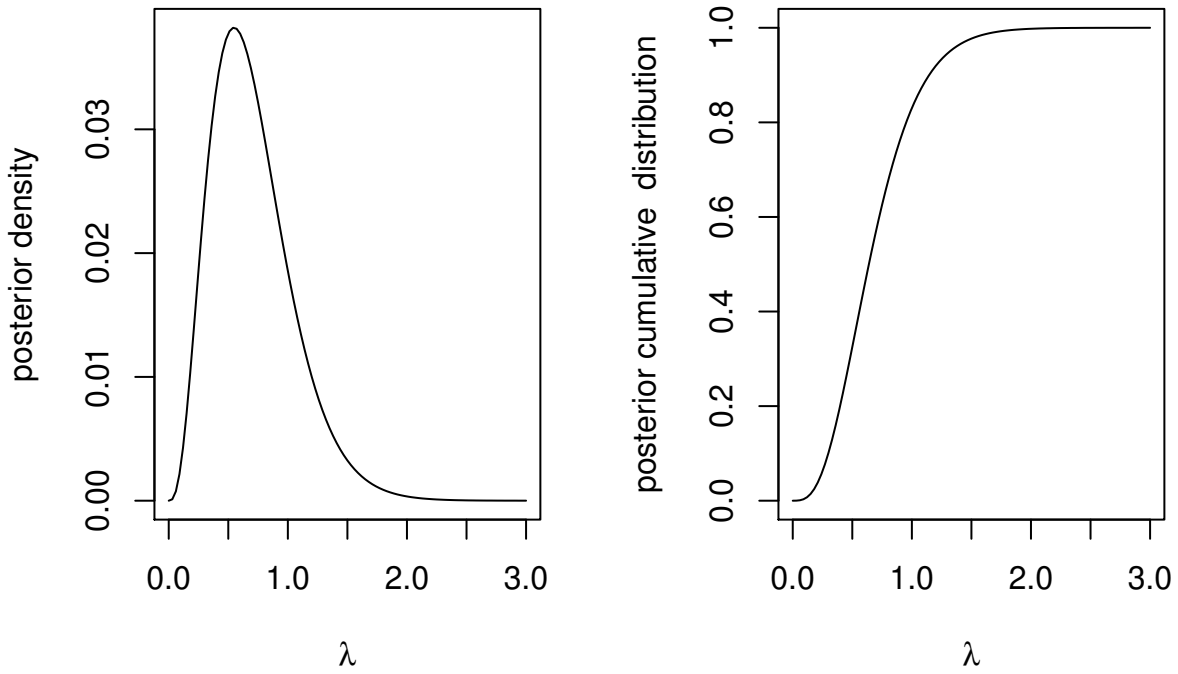
When discretizing λ we can normalize the distribution as done below

```
# vectorize th into 100 bins
lambda = seq(0, 3, length=101)
d = 6.56

# calculate the unnormalized density at each bin
dens = (dpois(3,lambda*d)+dpois(4,lambda*d)+dpois(5,lambda*d)+dpois(6,lambda*d))*dgamma(lambda,1,1)
# normalize the discretized probability densities
dens=dens/sum(dens)
```

```r
# calculate the cumulative distribution function
post_cdf = cumsum(dens)

# plot the posterior density
par(mfrow=c(1,2))           # divide plot into 2 subplots
plot (lambda, dens, type="l", xlab=expression(lambda), ylab="posterior density",cex=2)
# plot the posterior cumulative distribution function
plot (lambda, post_cdf, type="l", xlab=expression(lambda), ylab="posterior cumulative  distribution",ce
```




```r
# calculate the probability that lambda > 1
1-max(post_cdf[which(lambda<=1)])
```

```
## [1] 0.1759454
```

4)

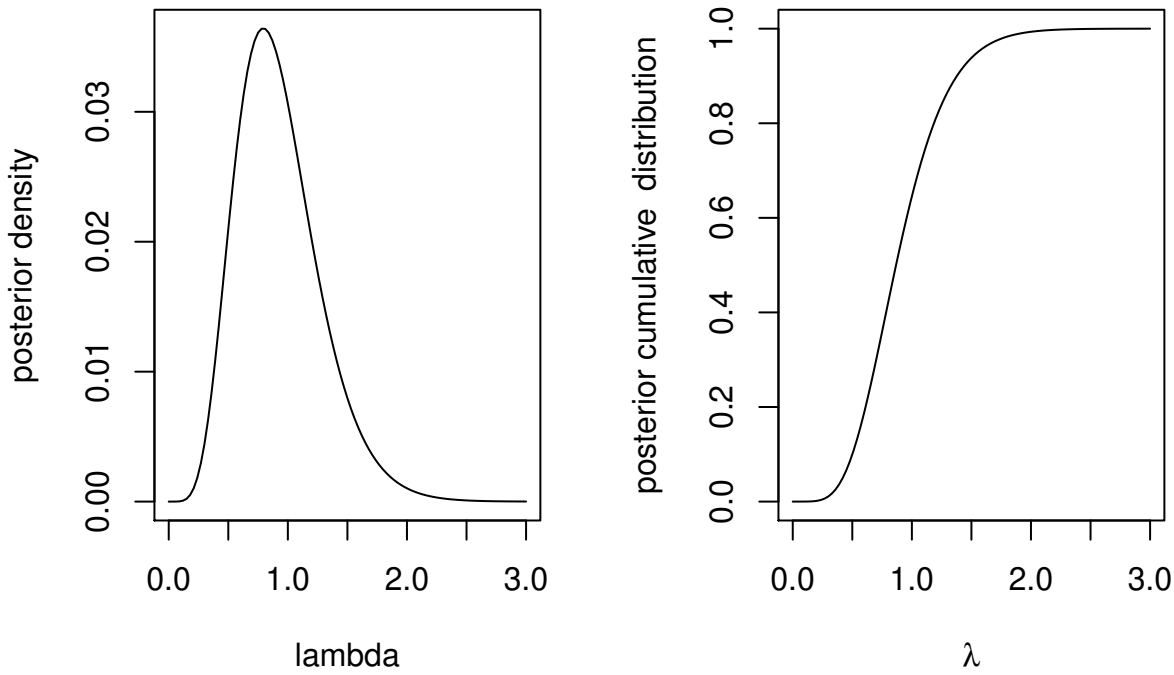
The posterior density function in case of $y = 6$ observation is

$p(\lambda|y = 6, d = 6.56) \propto \text{Poisson}(6|\lambda \times d)\text{Gamma}(\lambda, 1, 1)$

```r
# calculate the density at each bin
dens2 = dpois(6,lambda*d)*dgamma(lambda,1,1)
dens2 = dens2 /sum(dens2)

# calculate the cumulative distribution function
post_cdf2 = cumsum(dens2)

# plot the unnormalized posterior
par(mfrow=c(1,2))           # divide plot into 2 subplots
plot (lambda, dens2, type="l", xlab="lambda", ylab="posterior density", cex=2)
plot (lambda, post_cdf2, type="l", xlab=expression(lambda), ylab="posterior cumulative  distribution",c
```




```
# calculate the probability that lambda > 1
1-max(post_cdf2[which(lambda<=1)])
```

```
## [1] 0.3648346
```

The apparent differences are that the latter posterior probability density is centered more to the right than the first one and that it is narrower.

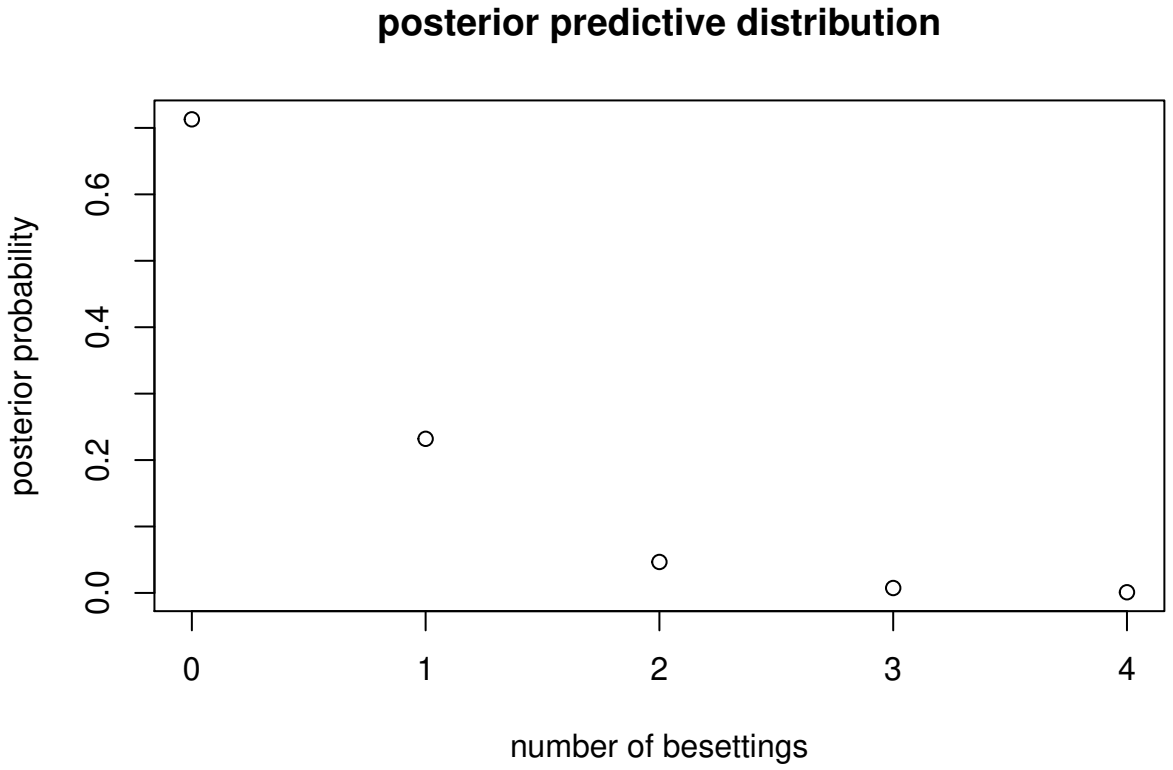
5

The posterior predictive probability for $\tilde{y}$ is

$$p(\tilde{y}|2 < y < 7, d = 6.56, \tilde{d} = 0.5) = \sum_{i=1}^{100} \text{Poisson}(\tilde{y} \times 0.5|\lambda) p(\lambda|2 < y < 7, d = 6.56)$$

We need to now calculate this sum for each $\tilde{y} \in [0, 4]$.

```
y_tilde = c(0,1,2,3,4)
p_pred = vector(length=length(y_tilde))
for (i in 1:length(y_tilde)){
  p_pred[i] = sum(dpois(y_tilde[i],lambda*0.5)*dens)
}
plot(y_tilde,p_pred, main="posterior predictive distribution",
     ylab="posterior probability", xlab="number of besettings")
```



Grading

Total 10 points: Two points from correct answer in each step. One point per step if the main idea is correct but the result is erroneous because of a minor bug or misthought.

Mark-recapture method for population size estimation

Week2-ex2, solution

Introduction

This exercise serves also as an example to Bayesian inference and summarizing posterior probability distribution of discrete variables using exact probabilities and Monte Carlo method. Hence, go through the whole file before starting to do the exercise.

Mark-recapture method

The basic idea in mark-recapture method for estimating animal populations is that a researcher visits a study area and captures a group of individuals. Each of these individuals is marked with a unique identifier (e.g., a numbered tag, ring or band), and is released back into the environment. Sufficient time is allowed to pass for the marked individuals to redistribute themselves among the unmarked population. After a while, the researcher returns and captures another sample of individuals.

Assumptions in the basic implementation of the method are, among others, that the time between consecutive captures is long enough for "perfect mixing", marks are not lost, the behavior and capture probability of an individual does not change due to marking and that the study population is "closed". In other words, the two visits to the study area are close enough in time so that no individuals die, are born, move into the study area (immigrate) or move out of the study area (emigrate) between visits. If these assumptions hold, we can reasonably assume that the *marked animals are randomly distributed in the total population" which then allows for inference on the total population size.

This method is illustrated during the lecture where we estimate the number of balls in a bag (the total *population* comprises of all the balls in the bag).

Let N denote the total population size, M the number of marked individuals at first visit, C the total number of animals captured at the second time and R the number of recaptured animals. By assuming that N is large compared to M and that the marked individuals are randomly distributed in the population, we can use Binomial distribution as our observation model for R as follows

$$p(R|C, M, N) = \text{Bin}(R|C, M/N) \tag{1}$$

We have to define a prior for N after which we can solve its posterior

$$p(N|M, C, R) \propto \text{Bin}(R|C, M/N)p(N) \tag{2}$$

The number of marked balls is

```
M=25
```

We will now analyze the total number of balls in the bag. This will be done first by exact calculations with discrete valued N and after that using Markov chain Monte Carlo.

Conduct the inference with discretization

Since there is only one, discrete, variable that we are interested in, we can easily discretize the problem and work with array(s) of probabilities

Let's define an array of values N that we think are a priori plausible at all. The below values are "hard" limits. Prior probability below the minimum and above the maximum is zero

```
abs_min <- M    # the number of balls cannot be negative
abs_max <- 500  # No way that bag can contain more than 1000 balls (a subjective assumption)


# Define the evaluation points so that all integers between
# abs_min and abs_max are included
Nseq <- seq(abs_min, abs_max, length=abs_max-abs_min+1)
```

Next we define prior for N and draw it.

Now that we have a discrete variable we have to give a prior probability for each of the elements in Nseq. You can do this in multiple ways. Here are few examples:

```
par(mfrow=c(2,3))              # Open figure for plotting the examples

# uniform prior
Nprior <- rep(1,length(Nseq))/length(Nseq)
sum(Nprior)              # check that prior probabilities sum up to to one
```

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

```
plot(Nseq,Nprior, main="Uniform prior", xlab="N", pch=16)
# "Gaussian" prior
Nprior <- dnorm(Nseq, mean=50, sd=20)
Nprior <- Nprior/sum(Nprior)     # Normalize the prior probabilities to sum to one
sum(Nprior)        # check that prior probabilities sum up to to one
```

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

```
plot(Nseq,Nprior, main="Gaussian prior", xlab="N", pch=16)
# log-Gaussian prior
Nprior <- dlnorm(Nseq, mean=5, sd=1)
Nprior <- Nprior/sum(Nprior)   # Normalize the prior probabilities to sum to one
sum(Nprior)        # check that prior probabilities sum up to to one
```

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

```
plot(Nseq,Nprior, main="log-Gaussian prior", xlab="N", pch=16)
# Step wise prior by giving different relative weights for different values
Nprior <- rep(1,length(Nseq))
Nprior[Nseq>50 & Nseq<600] <- 2
Nprior[Nseq>70 & Nseq<400] <- 4
Nprior[Nseq>200 & Nseq<300] <- 6
Nprior <- Nprior/sum(Nprior)    # Normalize the prior probabilities to sum to one
sum(Nprior)        # check that prior probabilities sum up to to one
```

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

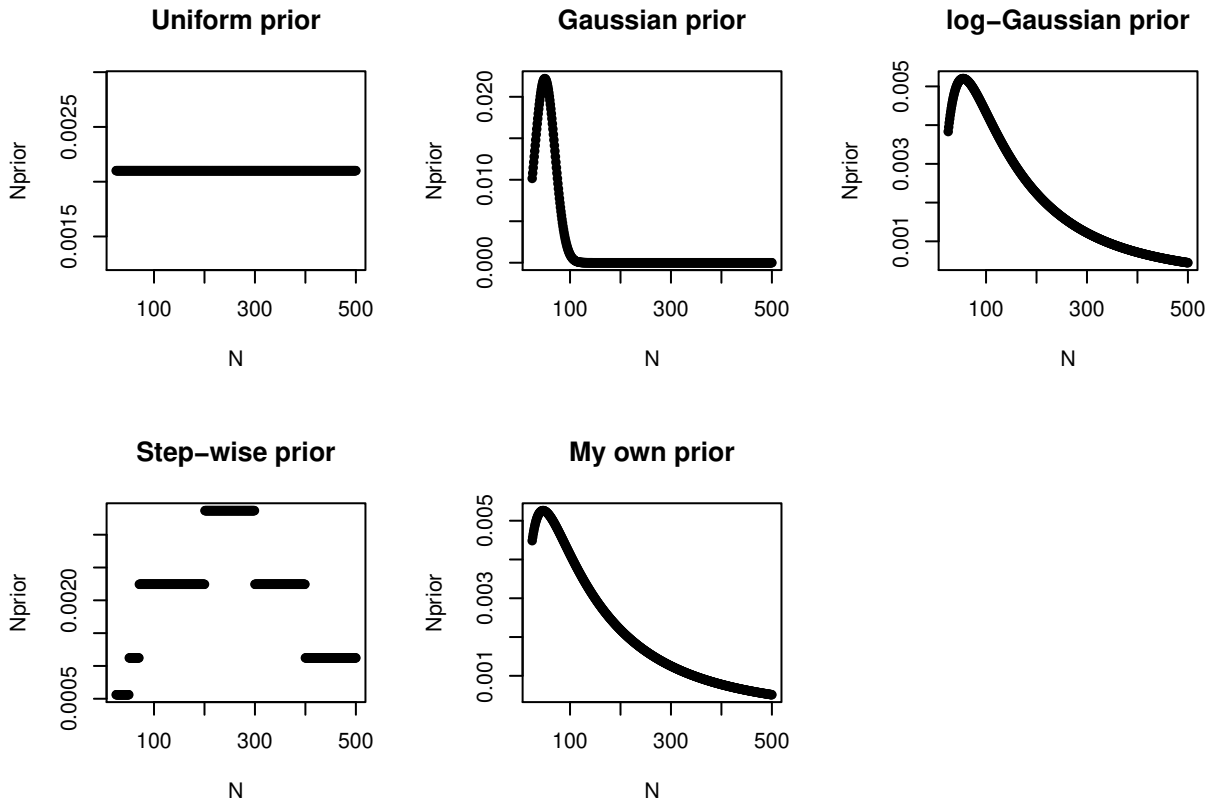
```
plot(Nseq,Nprior, main="Step-wise prior", xlab="N", pch=16)

# --- Here we will fill in the prior defined during the lecture ---
```

```
Nprior <- dlnorm(Nseq, mean=5.05, sd=1.1)
Nprior <- Nprior/sum(Nprior)    # Normalize to sum to one
sum(Nprior)                     # check that prior probabilities sum up to to one
```

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

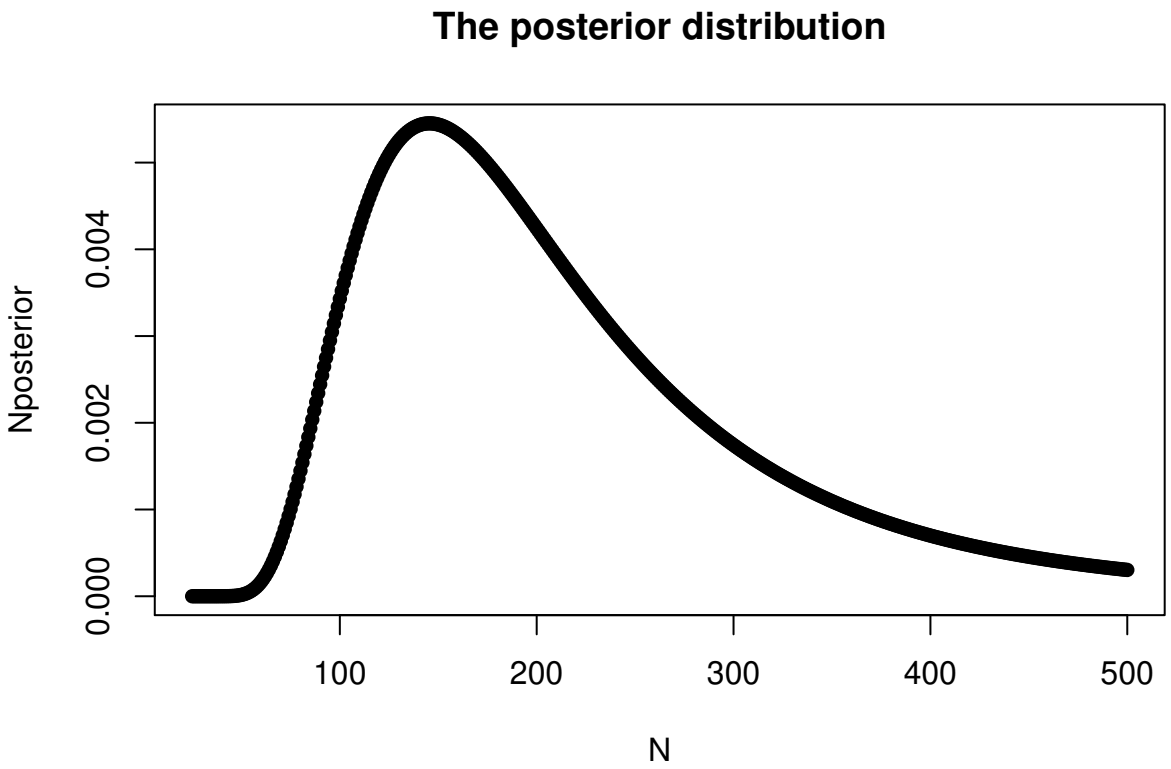
```
plot(Nseq,Nprior, main="My own prior", xlab="N", pch=16)
```



Now that we have defined the vector of prior probabilities for different values of N we can can conduct the second sampling round, to obtain data C and R, and after that calculate the posterior distribution for it by using the Bayes Theorem explicitly

```
# The result from the other sampling time
C=22
R=3

Nposterior <- Nprior*dbinom(R,C,M/Nseq)  # numerator of Bayes theorem
Nposterior <- Nposterior/sum(Nposterior) # divide by marginal likelihood
plot(Nseq,Nposterior, main="The posterior distribution", xlab="N", pch=16)
```



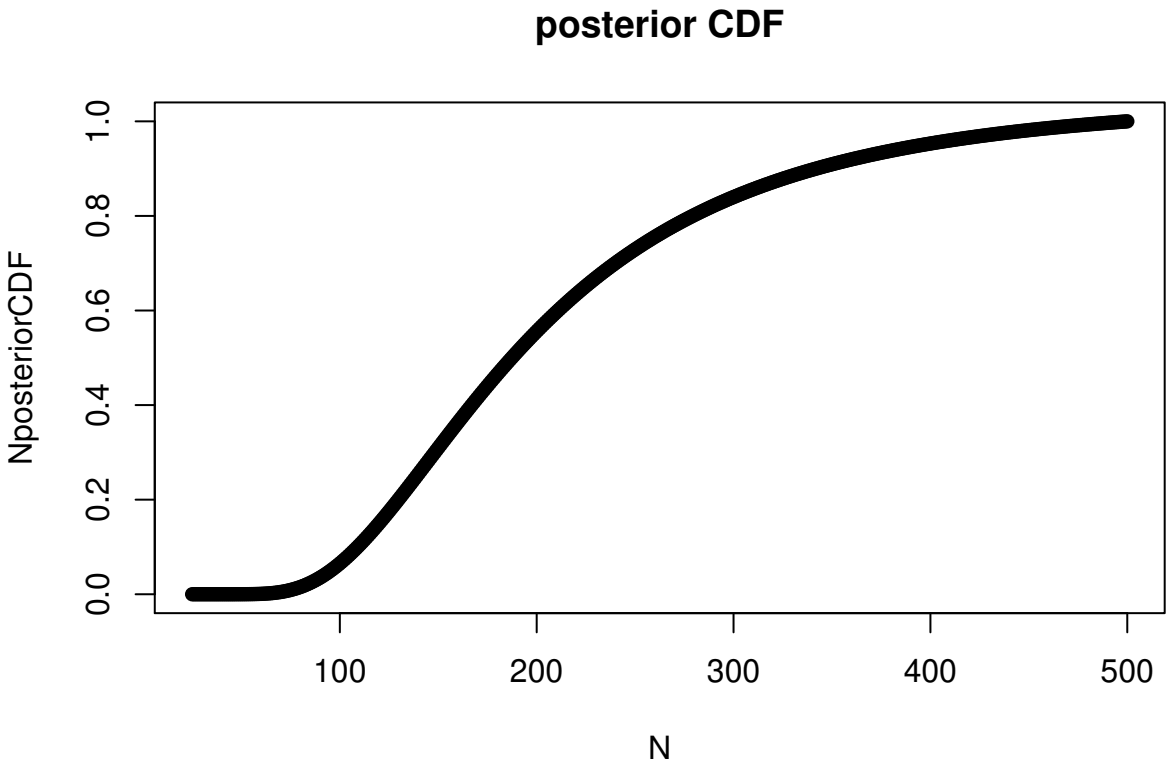
Given the vector of posterior probabilities for different values of N we can calculate various summaries for the posterior distribution. Such as the posterior mean.

```
posteriorMean = sum(Nposterior*Nseq)
print(posteriorMean)
```

```
## [1] 208.4088
```

In order to calculate quantiles, we need to first calculate the cumulative distribution function

```
NposteriorCDF <- cumsum(Nposterior)
# Plot CDF
plot(Nseq,NposteriorCDF, main="posterior CDF", xlab="N", pch=16)
```



Now we can calculate, for example, the 90% posterior quantile

```
# 10% quantile is the last N at which CDF is under 10%
Nseq[which(NposteriorCDF>=0.1)[1]-1]
```

```
## [1] 109
```

```
# 90% quantile is the first N at which CDF is over 90%
Nseq[which(NposteriorCDF>=0.9)[1]]
```

```
## [1] 342
```

Analysis using Monte Carlo

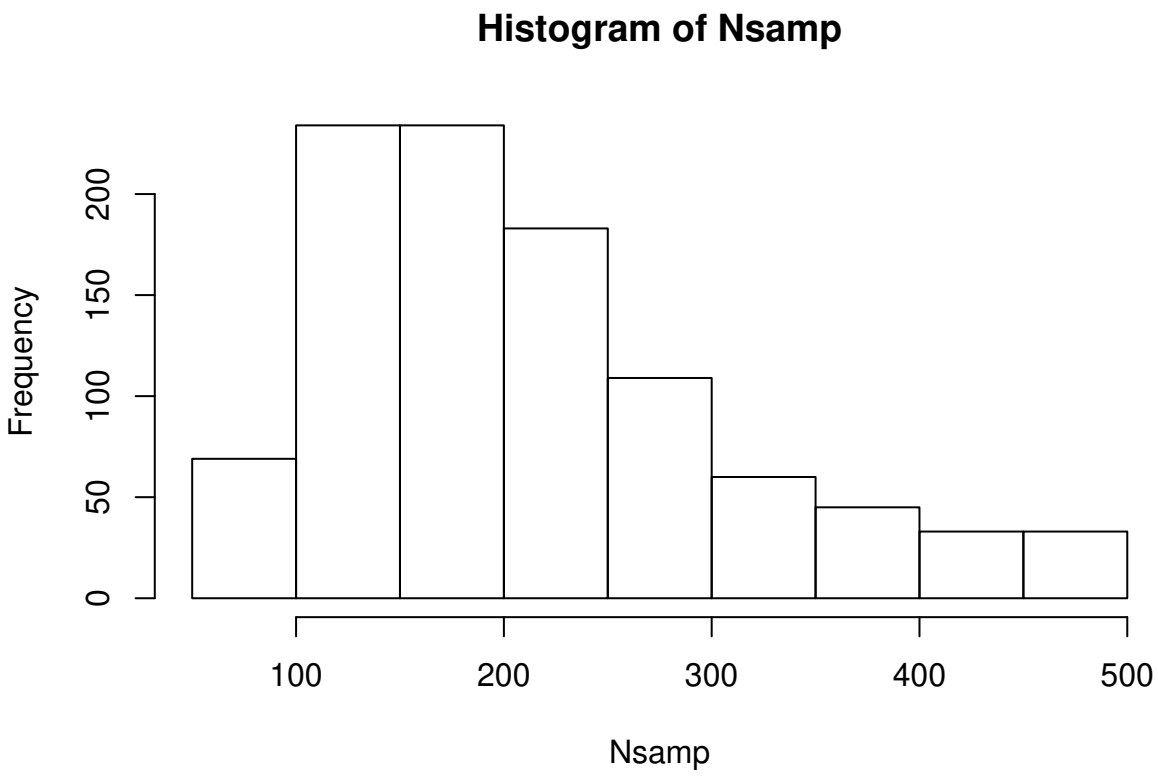
Next, we conduct the analysis using Monte Carlo technique. Here the idea is to draw random samples from the posterior distribution of N and then use these to calculate summaries of the posterior distribution.

In this example, we will draw the samples using a technique that utilizes the inverse cumulative distribution function of the posterior. Note, this technique is not part of the learning goals of the course so you can jump over the next code block. However, if you are interested in the technique see page 23 in BDA3.

```
Nsamp = Nseq[sapply(runif(1000,min=0,max=1),function(temp){ which(NposteriorCDF>=temp)[1] })]
```

We can now use the vector Nsamp as a collection of random samples from the posterior distribution. Hence, let's draw the histogram of them and calculate the mean and 10 % and 90 % quantiles

```
hist(Nsamp)
```




```
mean(Nsamp)
```

```
## [1] 213.431
```

```
quantile(Nsamp,probs=c(0.1,0.9))
```

```
##   10%   90%
## 107.9 369.0
```

Exercise for week 2

Calculate the posterior median, variance, standard deviation, and 80% central posterior interval using:

1. the vectorized form of the posterior probablity distribution (that is, the vectors Nseq, Nposterior, NposteriorCDF)
2. the Monte Carlo approximation using the random samples from the posterior that are stored in Nsamp
3. Additionally, why aren't the results from the above two approaches identical? How could you get them to match in theory?

Model solution

1 vectorized form of the posterior distribution

```
# posterior median
posteriorMedian = Nseq[which(NposteriorCDF>=0.5)[1]-1]
print(posteriorMedian)
```

```
## [1] 187
```

```
# Posterior variance
posteriorVariance = sum(Nposterior*(Nseq-posteriorMean)^2)
print(posteriorVariance)
```

```
## [1] 8300.412
```

```
# posterior standard deviation
posteriorSd = sqrt(posteriorVariance)
print(posteriorSd)
```

```
## [1] 91.1066
```

```
# 80% central posterior interval
interval80 = cbind(Nseq[which(NposteriorCDF>=0.1)[1]] , Nseq[which(NposteriorCDF>=0.9)[1]-1])
print(interval80)
```

```
##      [,1] [,2]
## [1,]  110  341
```

2 Monte Carlo approach

```
# posterior median
posteriorMedianMC=median(Nsamp)
print(posteriorMedianMC)
```

```
## [1] 193
```

```
# Posterior variance
posteriorVarianceMC = var(Nsamp)
print(posteriorVarianceMC)
```

```
## [1] 9349.725
```

```
# posterior standard deviation
posteriorSdMC = sqrt(posteriorVarianceMC)
print(posteriorSdMC)
```

```
## [1] 96.69398
```

```
# 80% central posterior interval
interval80MC = quantile(Nsamp,c(0.1,0.9))
print(interval80MC)
```

```
##   10%   90%
## 107.9 369.0
```

3 Compare the results

```
# medians
cbind(posteriorMedian,posteriorMedianMC)
```

```
##      posteriorMedian posteriorMedianMC
## [1,]             187               193
```

```r
# variances
cbind(posteriorVariance,posteriorVarianceMC)
```

```
##      posteriorVariance posteriorVarianceMC
## [1,]          8300.412            9349.725
```

```r
# standard deviations
cbind(posteriorSd,posteriorSdMC)
```

```
##      posteriorSd posteriorSdMC
## [1,]     91.1066      96.69398
```

```r
# 80% central posterior intervals
rbind(interval80,interval80MC)
```

```
##                10% 90%
##              110.0 341
## interval80MC 107.9 369
```

Clearly each of the summary statistics is similar with Monte Carlo and the discretized approach so that the actual numbers are close to each others. However, since we use only 1000 samples in Monte Carlo there is random variation in it. In theory we would get the true result with Monte Carlo if we increased the sample size to infinity.

Grading

Total points 10: Four points from question 1 so that 1 point from correct implementation and result from each of the four summary statistics. Similarly, four points from question 2. Two points from, question 3 so that correct answer to each of the two sub-questions gives 1 point.

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

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

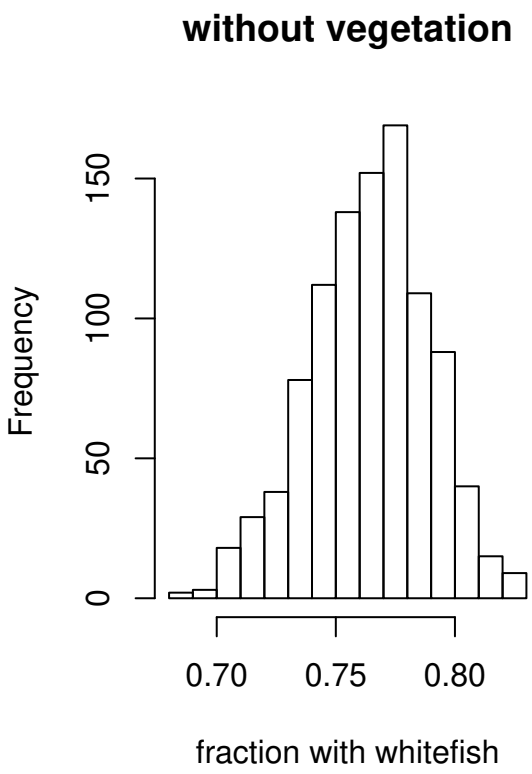
2.

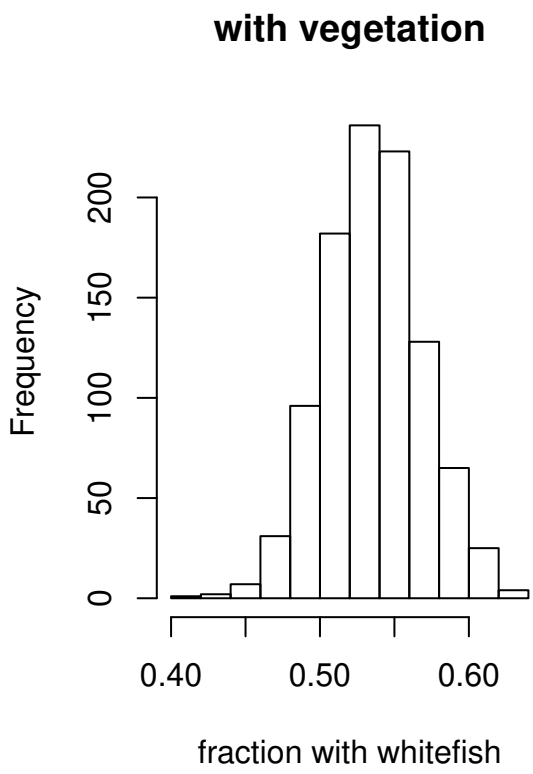
Let's sample from the posterior distributions

```
# 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
```

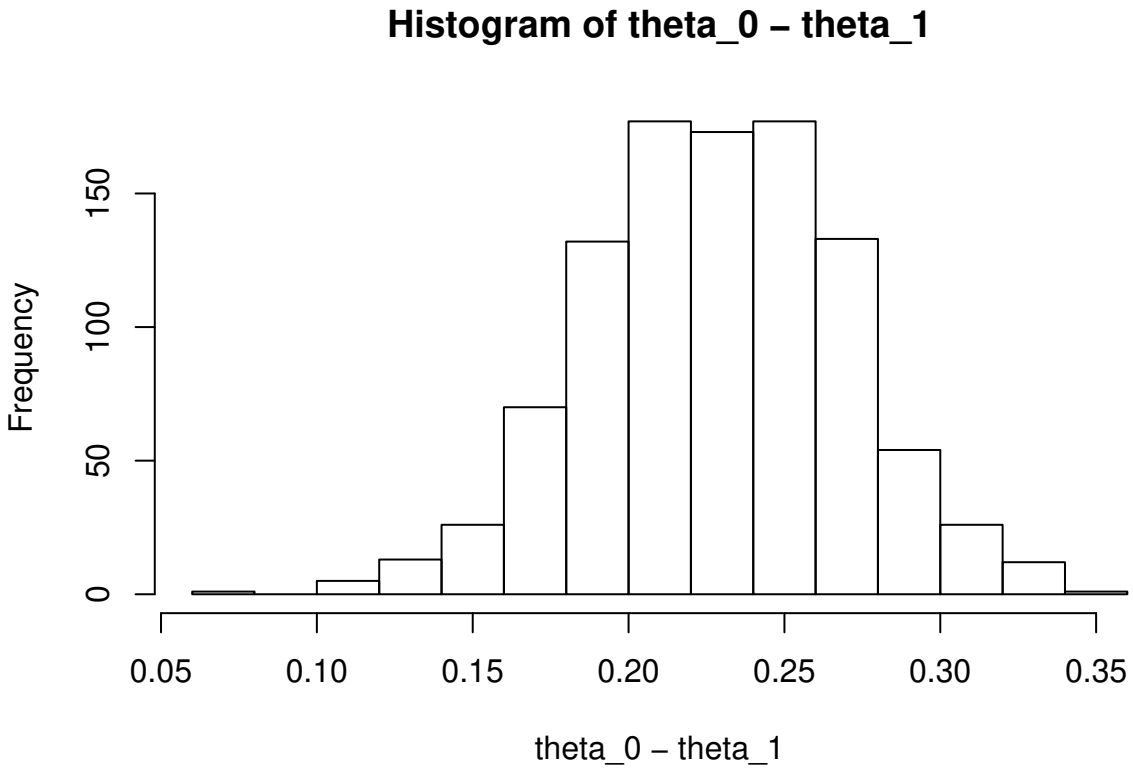
```
cbind(sd(theta_0),sd(theta_1))
```

```
##            [,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

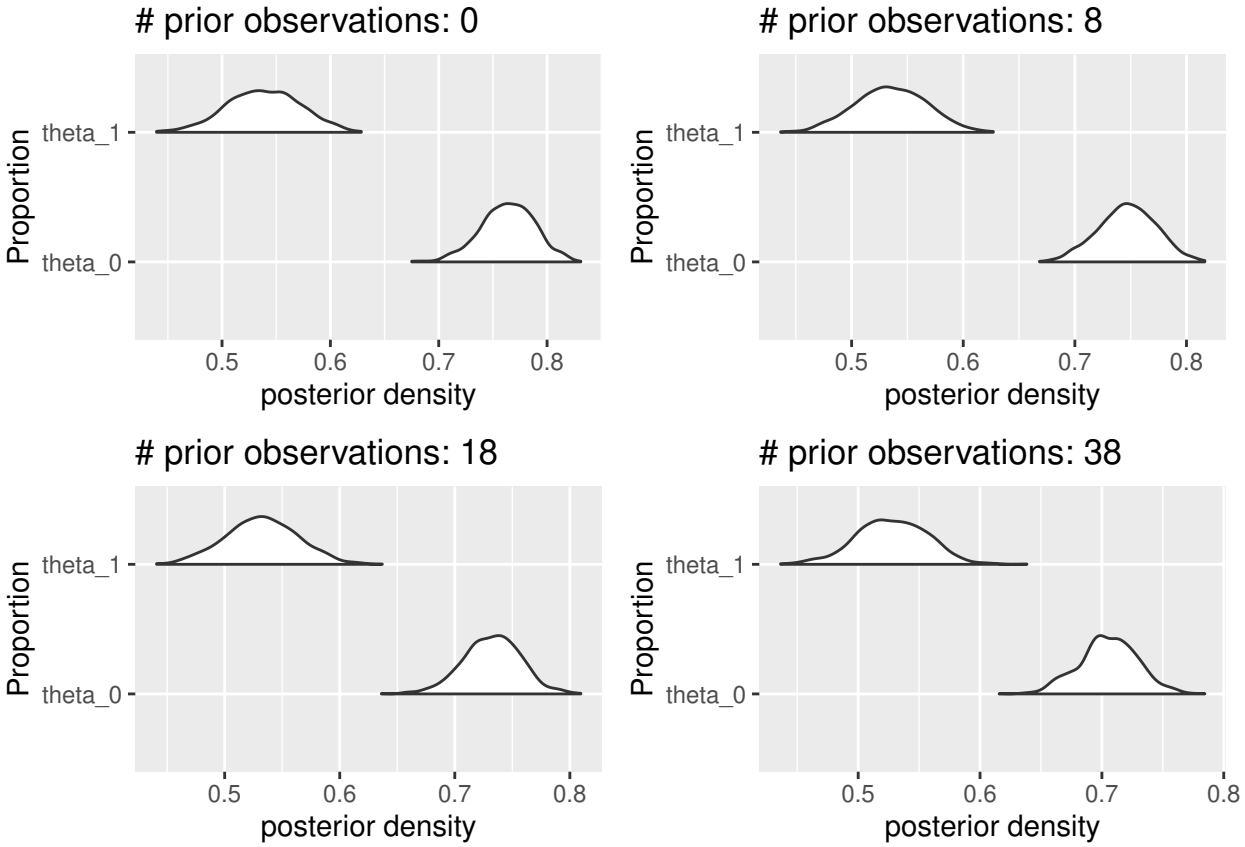
We can test the sensitivity of the posterior distribution to the choice of prior distribution by calculating the posterior distribution and its summaries with alternative prior distributions. Since we assumed that we don't have any prior knowledge on θ_0 or θ_1, we will test priors that have the same 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/