STAT 2650 Final Project

Name: Kuan-Cheng Fu

Date: 12/01/2020

Problem 1

After excluding the players whose at bats, AB, either in the first or second period, are no larger than 10 ($N_{i1} \leq 10$ or $N_{i2} \leq 10$), 491 players are left in the analysis.

```
raw_data <- read.table("Bat.dat")
data <- raw_data[raw_data$AB.1>10 & raw_data$AB.2>10,]
data_shape <- as.data.frame(dim(data)[1])
colnames(data_shape) <- "numbers"
rownames(data_shape) <- "players"
data_shape
```

```
##         numbers
## players     491
```

Problem 2

1. Data: $H_{i1} \sim Bin(N_{i1}, p_i)$

2. Prior: $p_i \sim beta(a, b)$

3. Posterior: $p_i|H_{i1} \sim beta(H_{i1} + a, N_{i1} - H_{i1} + b)$

4. Regarding the parameters a and b of our prior distribution, a and b will be set to equal 1 and 4 respectively since we believe that the mean of batting average is approximately equal to 0.2 (i.e., $E(p_i) = \frac{a}{a+b} = 0.2$). After that, by using Monte Carlo technique, we will sample 5000 values of each p_i from our posterior distributions. Finally, the estimate of of each p_i will be obtained by averaging its 5000 sample values.

```r
# data information
data_hitting_ability.1 <- matrix(0,dim(data)[1],1)
colnames(data_hitting_ability.1) <- c("Batting Average")

for (i in 1:dim(data)[1]) {
  data_hitting_ability.1[i,1] <- data$H.1[i]/data$AB.1[i]
}
summary(data_hitting_ability.1)
```

```
##  Batting Average
##  Min.   :0.0000
##  1st Qu.:0.1345
##  Median :0.1691
##  Mean   :0.1638
##  3rd Qu.:0.1972
##  Max.   :0.3846
```

```r
# monte carlo
posterior_hitting_ability_p2 <- matrix(0,dim(data)[1],1)
colnames(posterior_hitting_ability_p2) <- c("Batting Average")

set.seed(1)
for (i in 1:dim(data)[1]) {
  a <- 1
  b <- 4
  p.mc5000 <- rbeta(5000,data$H.1[i]+a,data$AB.1[i]-data$H.1[i]+b)
  posterior_hitting_ability_p2[i,1] <- mean(p.mc5000)
}
```

Problem 3

1. Data: $X_{i1} \sim N(\theta_i, \sigma_{i1}^2)$, $\sigma_{i1}^2 = \frac{1}{4N_{i1}}$

2. Prior: $\theta_i \sim N(\mu, \tau^2)$; $\mu \sim N(\mu_0, \gamma_0^2)$; $\tau^2 \sim Inverse - gamma(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2})$

3. $p(\theta_1, ..., \theta_n, \mu, \tau^2 | X_{11}, ..., X_{n1})$
   $\propto p(X_{11}, ..., X_{n1} | \theta_1, ..., \theta_n, \mu, \tau^2) \times p(\theta_1, ..., \theta_n | \mu, \tau^2) \times p(\mu) \times p(\tau^2)$
   $\propto \prod_{k=1}^{n} p(X_{k1} | \theta_k, \sigma_{k1}^2) \times \prod_{k=1}^{n} p(\theta_k | \mu, \tau^2) \times p(\mu) \times p(\tau^2)$

4. Posterior of θ_i: $p(\theta_i | \mu, \tau^2, X_{11}, ..., X_{n1}) \propto p(X_{i1} | \theta_i, \sigma_{i1}^2) p(\theta_i | \mu, \tau^2)$
   $\Rightarrow \theta_i | \mu, \tau^2, X_{11}, ..., X_{n1} \sim N(\frac{\frac{X_{i1}}{\sigma_{i1}^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_{i1}^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma_{i1}^2} + \frac{1}{\tau^2}})$

5. Posterior of μ: $p(\mu | \theta_1, ..., \theta_n, \tau^2, X_{11}, ..., X_{n1}) \propto \prod_{k=1}^{n} p(\theta_k | \mu, \tau^2) \times p(\mu)$
   $\Rightarrow \mu | \theta_1, ..., \theta_n, \tau^2, X_{11}, ..., X_{n1} \sim N(\frac{\frac{n\bar{\theta}}{\tau^2} + \frac{\mu_0}{\gamma_0^2}}{\frac{n}{\tau^2} + \frac{1}{\gamma_0^2}}, \frac{1}{\frac{n}{\tau^2} + \frac{1}{\gamma_0^2}})$

6. Posterior of τ^2: $p(\tau^2 | \theta_1, ..., \theta_n, \mu, X_{11}, ..., X_{n1}) \propto \prod_{k=1}^{n} p(\theta_k | \mu, \tau^2) \times p(\tau^2)$
   $\Rightarrow \tau^2 | \theta_1, ..., \theta_n, \mu, X_{11}, ..., X_{n1} \sim Inverse - gamma(\frac{\eta_0 + n}{2}, , \frac{\eta_0 \tau_0^2 + \sum_{k=1}^{n} (\theta_k - \mu)^2}{2})$

7. Regarding the parameters μ_0 and γ_0^2 of our prior distribution of μ, μ_0 and γ_0^2 will be set to equal 0.4 and 0.01 respectively since the mean of X_{i1} is approximately equal to 0.4 and the prior probability that μ is in the interval $(0.2, 0.6)$ is about 95%. Besides, regarding the parameters η_0 and τ_0^2 of our prior distribution of τ, η_0 and τ_0^2 will be both set to equal 1 which represents weak prior information. Similarly, we will sample 5000 values of each θ_i from our posterior distributions. Finally, after transforming, the estimate of of each p_i will be obtained by averaging its 5000 sample values.

```
data$X <- 0
data$var <- 0
for (i in 1:dim(data)[1]) {
  value <- (data$H.1[i]+0.25)/(data$AB.1[i]+0.5)
  data$X[i] <- asin(sqrt(value))
  data$var[i] <- 1/(4*data$AB.1[i])
}
```

```
# weakly informative priors
eta0 <- 1 ; t20 <- 1
mu0 <- 0.4 ; g20 <- 0.01

# starting values
m <- dim(data)[1]
n <- 1
theta <- ybar <- data$X
sigma2 <- data$var
mu <- mean(theta)
tau2 <- var(theta)

# setup MCMC
set.seed(1)
S <- 5000
THETA <- matrix(nrow=S,ncol=m)
MST <- matrix(nrow=S,ncol=2)
```

```r
# MCMC algorithm
for(s in 1:S)
{
  # sample new values of the thetas
  for(j in 1:m)
  {
    vtheta <- 1/(n/sigma2[j]+1/tau2)
    etheta <- vtheta*(ybar[j]*n/sigma2[j]+mu/tau2)
    theta[j] <- rnorm(1,etheta,sqrt(vtheta))
   }

  # sample a new value of mu
  vmu <- 1/(m/tau2+1/g20)
  emu <- vmu*(m*mean(theta)/tau2+mu0/g20)
  mu <- rnorm(1,emu,sqrt(vmu))

  # sample a new value of tau2
  etam <- eta0+m
  ss <- eta0*t20+sum((theta-mu)^2)
  tau2 <- 1/rgamma(1,etam/2,ss/2)

  # store results
  THETA[s,] <- theta
  MST[s,] <- c(mu,tau2)
}

mcmc <- list(THETA=THETA,MST=MST)
theta.mc5000 <- apply(THETA, 2, mean)
posterior_hitting_ability_p3 <- (sin(theta.mc5000))^2
```

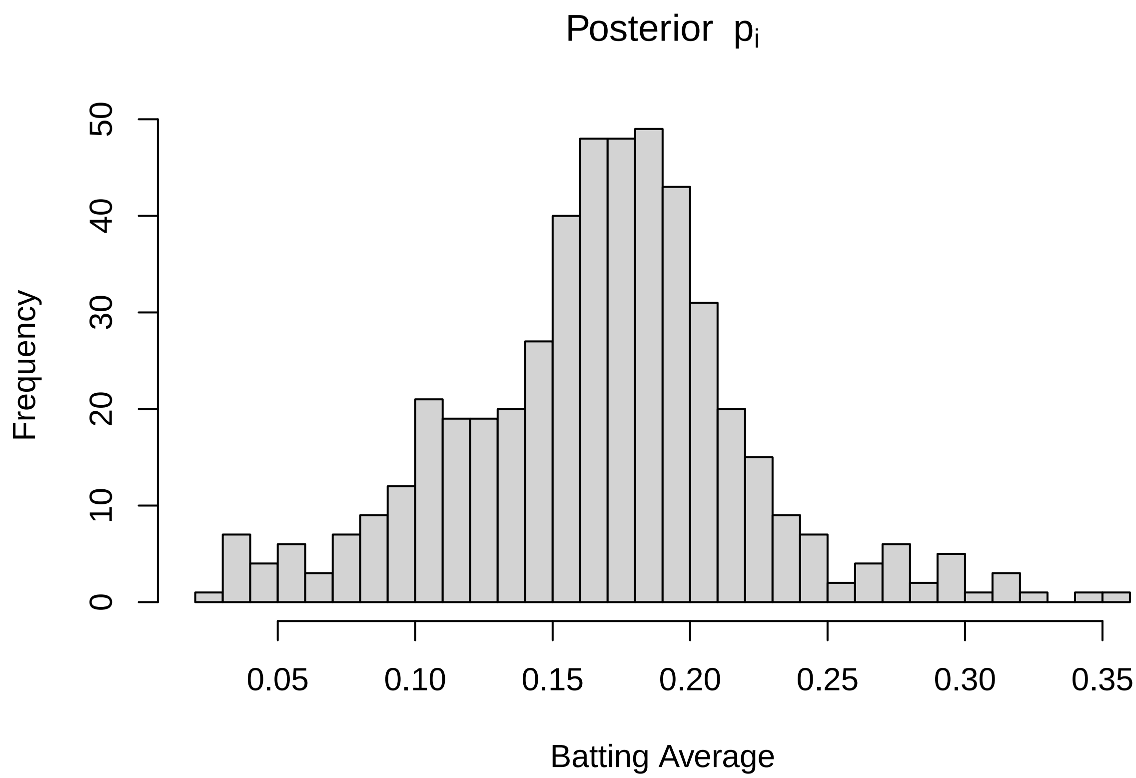
Problem 4

The information and histograms of our estimates $\frac{H_{i2}}{N_{i2}}$ from Problem 2 and Problem 3 are presented below.

```r
# estimates (Problem 2)
summary(posterior_hitting_ability_p2)
```

```
##  Batting Average
##  Min.   :0.02968
##  1st Qu.:0.13649
##  Median :0.17047
##  Mean   :0.16707
##  3rd Qu.:0.19700
##  Max.   :0.35481
```

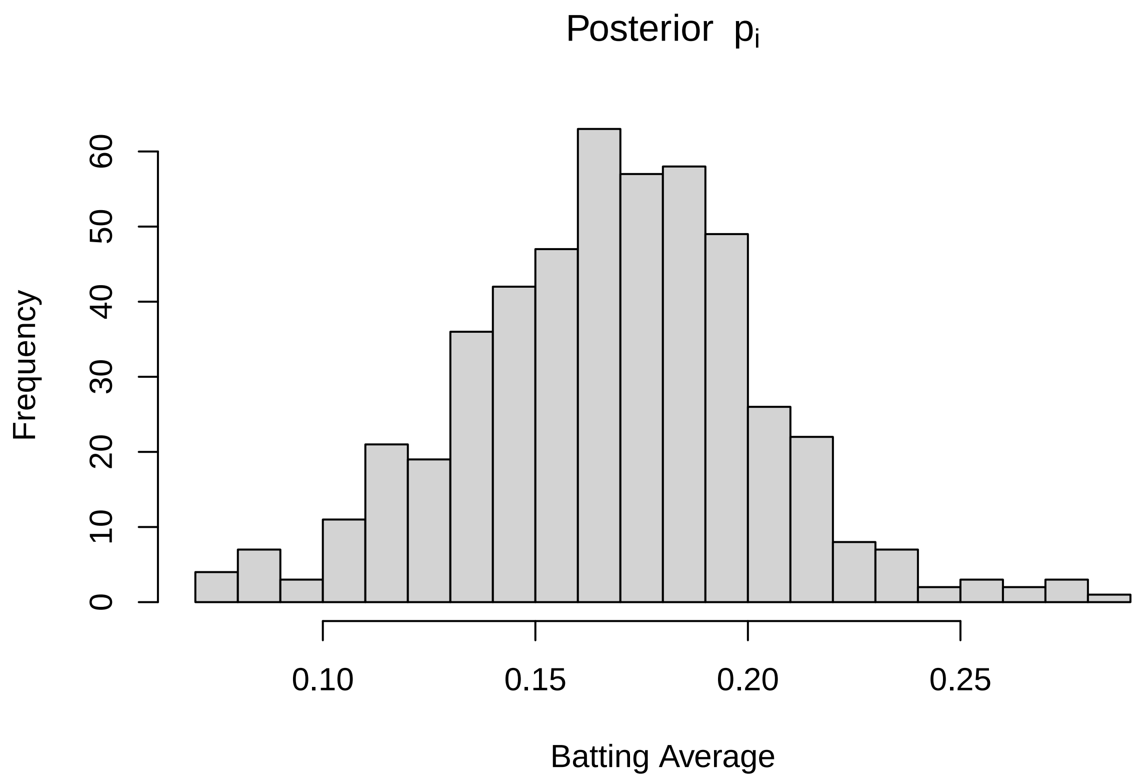
```r
# histogram (Problem 2)
hist(posterior_hitting_ability_p2,main=expression("Posterior "~p[i]),
     xlab="Batting Average",breaks=25)
```




```r
# estimates (Problem 3)
posterior_hitting_ability_p3 <- matrix(posterior_hitting_ability_p3,dim(data)[1],1)
colnames(posterior_hitting_ability_p3) <- c("Batting Average")
summary(posterior_hitting_ability_p3)
```

```
##  Batting Average
##  Min.   :0.07102
##  1st Qu.:0.14612
##  Median :0.16926
##  Mean   :0.16813
##  3rd Qu.:0.19003
##  Max.   :0.28837
```

```r
# histogram (Problem 3)
hist(posterior_hitting_ability_p3,main=expression("Posterior "~p[i]),
     xlab="Batting Average",breaks=25)
```



Problem 5

The MSE of the estimates from Problem 2 is slightly larger than the MSE of the estimates from Problem 3. We believe that the reason for this is because the variance of the estimates from Problem 3 is smaller than the variance of the estimates from Problem 2 and the biases of the estimates from Problem 2 and Problem 3 are similar based on the information and histograms in Problem 4.

```
data_hitting_ability.2 <- matrix(0,dim(data)[1],1)
colnames(data_hitting_ability.2) <- c("Batting Average")

for (i in 1:dim(data)[1]) {
  data_hitting_ability.2[i,1] <- data$H.2[i]/data$AB.2[i]
}

#MSE
mse2 <- sum((posterior_hitting_ability_p2-data_hitting_ability.2)^2)/nrow(data)
mse3 <- sum((posterior_hitting_ability_p3-data_hitting_ability.2)^2)/nrow(data)
MSE <- matrix(c(mse2,mse3),1,2)
rownames(MSE) <- c("MSE")
colnames(MSE) <- c("Problem 2","Problem 3")
MSE
```

```
##       Problem 2  Problem 3
## MSE 0.01176792 0.01073689
```

Problem 6.1 (nonpitchers)

In this case, there are 431 nonpitchers.

```
raw_data <- read.table("Bat.dat")
data <- subset(raw_data[raw_data$AB.1>10 & raw_data$AB.2>10,],Pitcher==0)
```

Problem 6.2 (nonpitchers)

1. The Bayesian model in this case is the same as the one in Problem 2.
2. In this case, a and b will also be set to equal 1 and 4 respectively. Besides, the process of how we estimate p_i is also the same as the one in Problem 2.

```
# data information
data_hitting_ability.1 <- matrix(0,dim(data)[1],1)
colnames(data_hitting_ability.1) <- c("Batting Average")

for (i in 1:dim(data)[1]) {
  data_hitting_ability.1[i,1] <- data$H.1[i]/data$AB.1[i]
}
summary(data_hitting_ability.1)
```

```
##  Batting Average
##  Min.   :0.0000
##  1st Qu.:0.1480
##  Median :0.1744
##  Mean   :0.1737
##  3rd Qu.:0.1996
##  Max.   :0.3846
```

```
# monte carlo
posterior_hitting_ability_p2 <- matrix(0,dim(data)[1],1)
colnames(posterior_hitting_ability_p2) <- c("Batting Average")

set.seed(1)
for (i in 1:dim(data)[1]) {
  a <- 1
  b <- 4
  p.mc5000 <- rbeta(5000,data$H.1[i]+a,data$AB.1[i]-data$H.1[i]+b)
  posterior_hitting_ability_p2[i,1] <- mean(p.mc5000)
}
```

Problem 6.3 (nonpitchers)

1. The Bayesian hierarchical model in this case is the same as the one in Problem 3.
2. μ_0 and γ_0^2 will also be set to equal 0.4 and 0.01 respectively while η_0 and τ_0^2 will also be both set to equal 1. Besides, the process of how we estimate p_i is also the same as the one in Problem 3.

```r
data$X <- 0
data$var <- 0
for (i in 1:dim(data)[1]) {
  value <- (data$H.1[i]+0.25)/(data$AB.1[i]+0.5)
  data$X[i] <- asin(sqrt(value))
  data$var[i] <- 1/(4*data$AB.1[i])
}
```

```r
# weakly informative priors
eta0 <- 1 ; t20 <- 1
mu0 <- 0.4 ; g20 <- 0.01

# starting values
m <- dim(data)[1]
n <- 1
theta <- ybar <- data$X
sigma2 <- data$var
mu <- mean(theta)
tau2 <- var(theta)

# setup MCMC
set.seed(1)
S <- 5000
THETA <- matrix(nrow=S,ncol=m)
MST <- matrix(nrow=S,ncol=2)
```

```r
# MCMC algorithm
for(s in 1:S)
{
  # sample new values of the thetas
  for(j in 1:m)
  {
    vtheta <- 1/(n/sigma2[j]+1/tau2)
    etheta <- vtheta*(ybar[j]*n/sigma2[j]+mu/tau2)
    theta[j] <- rnorm(1,etheta,sqrt(vtheta))
   }

  # sample a new value of mu
  vmu <- 1/(m/tau2+1/g20)
  emu <- vmu*(m*mean(theta)/tau2+mu0/g20)
  mu <- rnorm(1,emu,sqrt(vmu))

  # sample a new value of tau2
  etam <- eta0+m
  ss <- eta0*t20+sum((theta-mu)^2)
  tau2 <- 1/rgamma(1,etam/2,ss/2)

  # store results
  THETA[s,] <- theta
  MST[s,] <- c(mu,tau2)
}

mcmc <- list(THETA=THETA,MST=MST)
theta.mc5000 <- apply(THETA, 2, mean)
posterior_hitting_ability_p3 <- (sin(theta.mc5000))^2
```

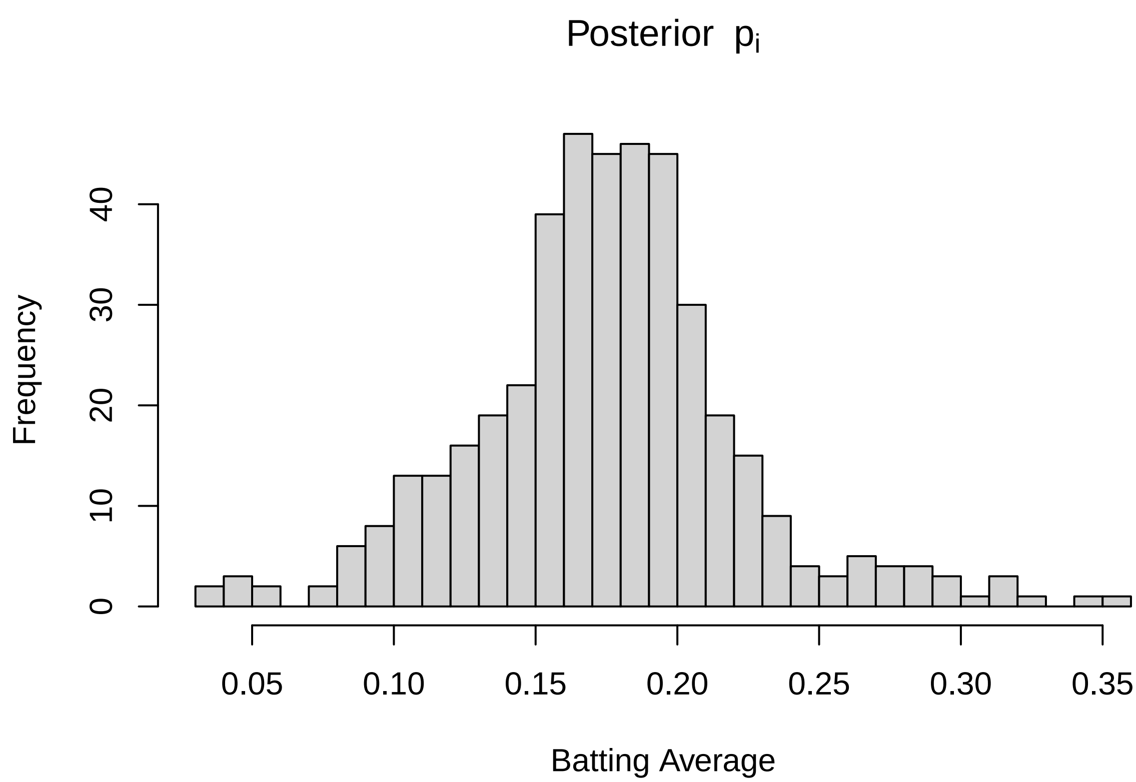
Problem 6.4 (nonpitchers)

The information and histograms of our estimates $\frac{H_{i2}}{N_{i2}}$ from Problem 6.2 (nonpitchers) and Problem 6.3 (nonpitchers) are presented below.

```r
# estimates (Problem 2)
summary(posterior_hitting_ability_p2)
```

```
##  Batting Average
##  Min.   :0.03119
##  1st Qu.:0.15042
##  Median :0.17545
##  Mean   :0.17506
##  3rd Qu.:0.19925
##  Max.   :0.35549
```

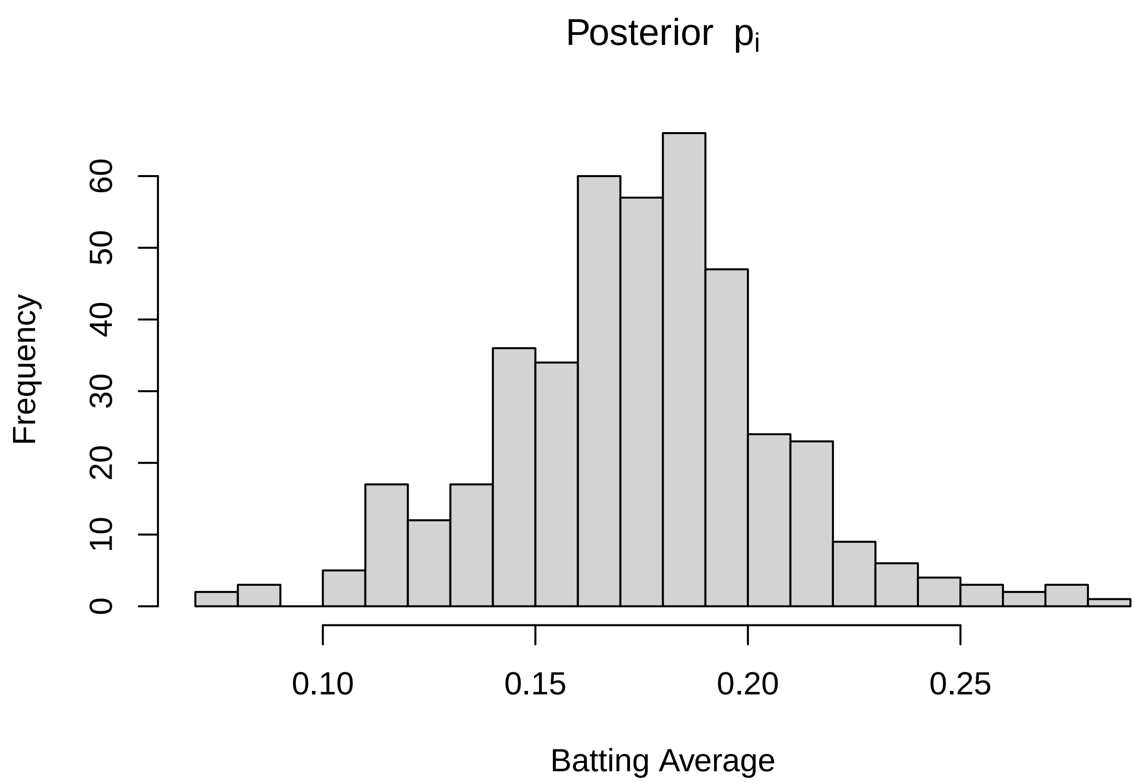
```r
# histogram (Problem 2)
hist(posterior_hitting_ability_p2,main=expression("Posterior "~p[i]),
     xlab="Batting Average",breaks=25)
```



```r
# estimates (Problem 3)
posterior_hitting_ability_p3 <- matrix(posterior_hitting_ability_p3,dim(data)[1],1)
colnames(posterior_hitting_ability_p3) <- c("Batting Average")
summary(posterior_hitting_ability_p3)
```

```
##  Batting Average
##  Min.   :0.07462
##  1st Qu.:0.15484
##  Median :0.17487
##  Mean   :0.17422
##  3rd Qu.:0.19386
##  Max.   :0.28815
```

```r
# histogram (Problem 3)
hist(posterior_hitting_ability_p3,main=expression("Posterior "~p[i]),
     xlab="Batting Average",breaks=25)
```



Problem 6.5 (nonpitchers)

The MSE of the estimates from Problem 6.2 (nonpitchers) is slightly larger than the MSE of the estimates from Problem 6.3 (nonpitchers). We believe that the reason for this is because the variance of the estimates from Problem 6.3 (nonpitchers) is smaller than the variance of the estimates from Problem 6.2 (nonpitchers) and the biases of the estimates from Problem 6.2 (nonpitchers) and 6.3 (nonpitchers) are similar based on the information and histograms in Problem 6.4 (nonpitchers).

```
data_hitting_ability.2 <- matrix(0,dim(data)[1],1)
colnames(data_hitting_ability.2) <- c("Batting Average")

for (i in 1:dim(data)[1]) {
  data_hitting_ability.2[i,1] <- data$H.2[i]/data$AB.2[i]
}

#MSE
mse2 <- sum((posterior_hitting_ability_p2-data_hitting_ability.2)^2)/nrow(data)
mse3 <- sum((posterior_hitting_ability_p3-data_hitting_ability.2)^2)/nrow(data)
MSE <- matrix(c(mse2,mse3),1,2)
rownames(MSE) <- c("MSE")
colnames(MSE) <- c("Problem 6.2","Problem 6.3")
MSE
```

```
##     Problem 6.2 Problem 6.3
## MSE  0.01169603  0.01074855
```

Problem 6.1 (pitchers)

In this case, there are 60 pitchers.

```
raw_data <- read.table("Bat.dat")
data <- subset(raw_data[raw_data$AB.1>10 & raw_data$AB.2>10,],Pitcher==1)
```

Problem 6.2 (pitchers)

1. The Bayesian model in this case is the same as the one in Problem 2.

2. In this case, a and b will be set to equal 1 and 9 respectively since we believe that the mean of batting average of the pitchers is approximately equal to 0.1 (i.e., $E(p_i) = \frac{a}{a+b} = 0.1$). Besides, the process of how we estimate p_i is also the same as the one in Problem 2.

```
# data information
data_hitting_ability.1 <- matrix(0,dim(data)[1],1)
colnames(data_hitting_ability.1) <- c("Batting Average")

for (i in 1:dim(data)[1]) {
  data_hitting_ability.1[i,1] <- data$H.1[i]/data$AB.1[i]
}
summary(data_hitting_ability.1)
```

```
##  Batting Average
##  Min.   :0.00000
##  1st Qu.:0.04708
##  Median :0.08957
##  Mean   :0.09238
##  3rd Qu.:0.12125
##  Max.   :0.31579
```

```
# monte carlo
posterior_hitting_ability_p2 <- matrix(0,dim(data)[1],1)
colnames(posterior_hitting_ability_p2) <- c("Batting Average")

set.seed(1)
for (i in 1:dim(data)[1]) {
  a <- 1
  b <- 9
  p.mc5000 <- rbeta(5000,data$H.1[i]+a,data$AB.1[i]-data$H.1[i]+b)
  posterior_hitting_ability_p2[i,1] <- mean(p.mc5000)
}
```

Problem 6.3 (pitchers)

1. The Bayesian hierarchical model in this case is the same as the one in Problem 3.
2. In this case, μ_0 and γ_0^2 will be set to equal 0.3 and 0.01 respectively since the mean of X_{i1} is approximately equal to 0.3 and the prior probability that μ is in the interval $(0.1, 0.5)$ is about 95%. Meanwhile, η_0 and τ_0^2 will also be both set to equal 1. Besides, the process of how we estimate p_i is also the same as the one in Problem 3.

```
data$X <- 0
data$var <- 0
for (i in 1:dim(data)[1]) {
  value <- (data$H.1[i]+0.25)/(data$AB.1[i]+0.5)
  data$X[i] <- asin(sqrt(value))
  data$var[i] <- 1/(4*data$AB.1[i])
}
```

```
# weakly informative priors
eta0 <- 1 ; t20 <- 1
mu0 <- 0.3 ; g20 <- 0.01

# starting values
m <- dim(data)[1]
n <- 1
theta <- ybar <- data$X
sigma2 <- data$var
mu <- mean(theta)
tau2 <- var(theta)

# setup MCMC
set.seed(1)
S <- 5000
THETA <- matrix(nrow=S,ncol=m)
MST <- matrix(nrow=S,ncol=2)
```

```r
# MCMC algorithm
for(s in 1:S)
{
  # sample new values of the thetas
  for(j in 1:m)
  {
    vtheta <- 1/(n/sigma2[j]+1/tau2)
    etheta <- vtheta*(ybar[j]*n/sigma2[j]+mu/tau2)
    theta[j] <- rnorm(1,etheta,sqrt(vtheta))
   }

  # sample a new value of mu
  vmu <- 1/(m/tau2+1/g20)
  emu <- vmu*(m*mean(theta)/tau2+mu0/g20)
  mu <- rnorm(1,emu,sqrt(vmu))

  # sample a new value of tau2
  etam <- eta0+m
  ss <- eta0*t20+sum((theta-mu)^2)
  tau2 <- 1/rgamma(1,etam/2,ss/2)

  # store results
  THETA[s,] <- theta
  MST[s,] <- c(mu,tau2)
}

mcmc <- list(THETA=THETA,MST=MST)
theta.mc5000 <- apply(THETA, 2, mean)
posterior_hitting_ability_p3 <- (sin(theta.mc5000))^2
```

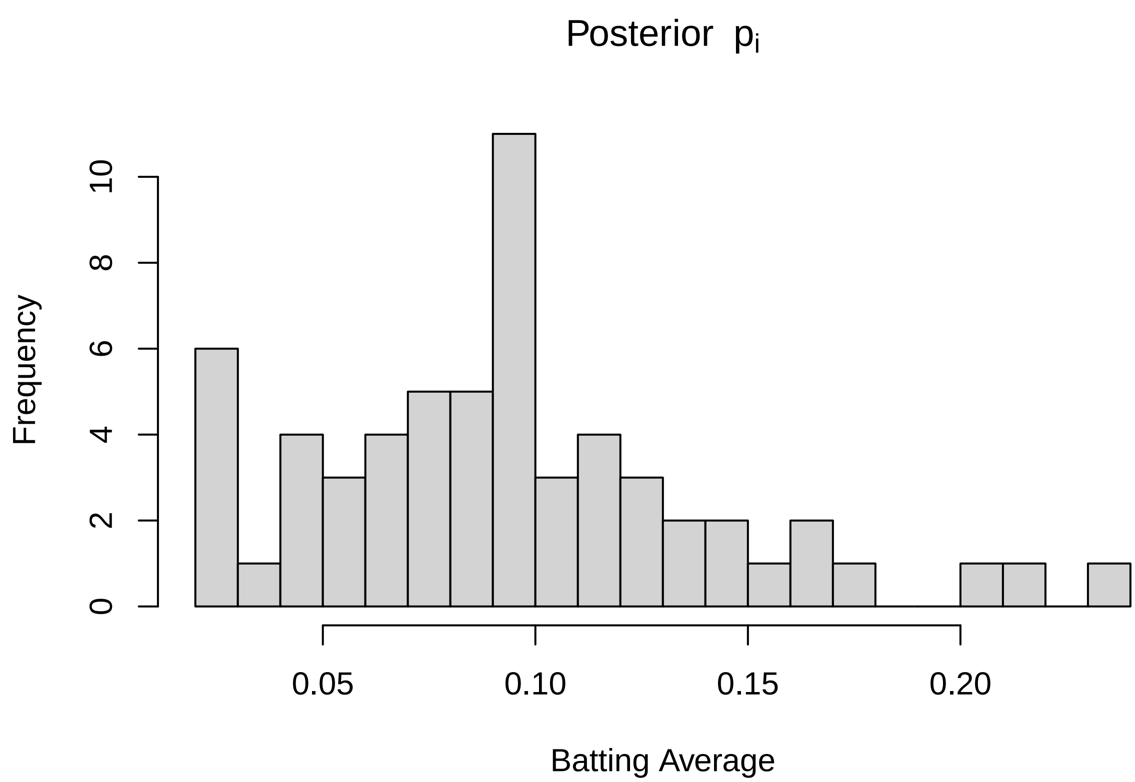
Problem 6.4 (pitchers)

The information and histograms of our estimates $\frac{H_{i2}}{N_{i2}}$ from Problem 6.2 (pitchers) and Problem 6.3 (pitchers) are presented below.

```r
# estimates (Problem 2)
summary(posterior_hitting_ability_p2)
```

```
##  Batting Average
##  Min.   :0.02548
##  1st Qu.:0.06394
##  Median :0.09147
##  Mean   :0.09473
##  3rd Qu.:0.11618
##  Max.   :0.23898
```

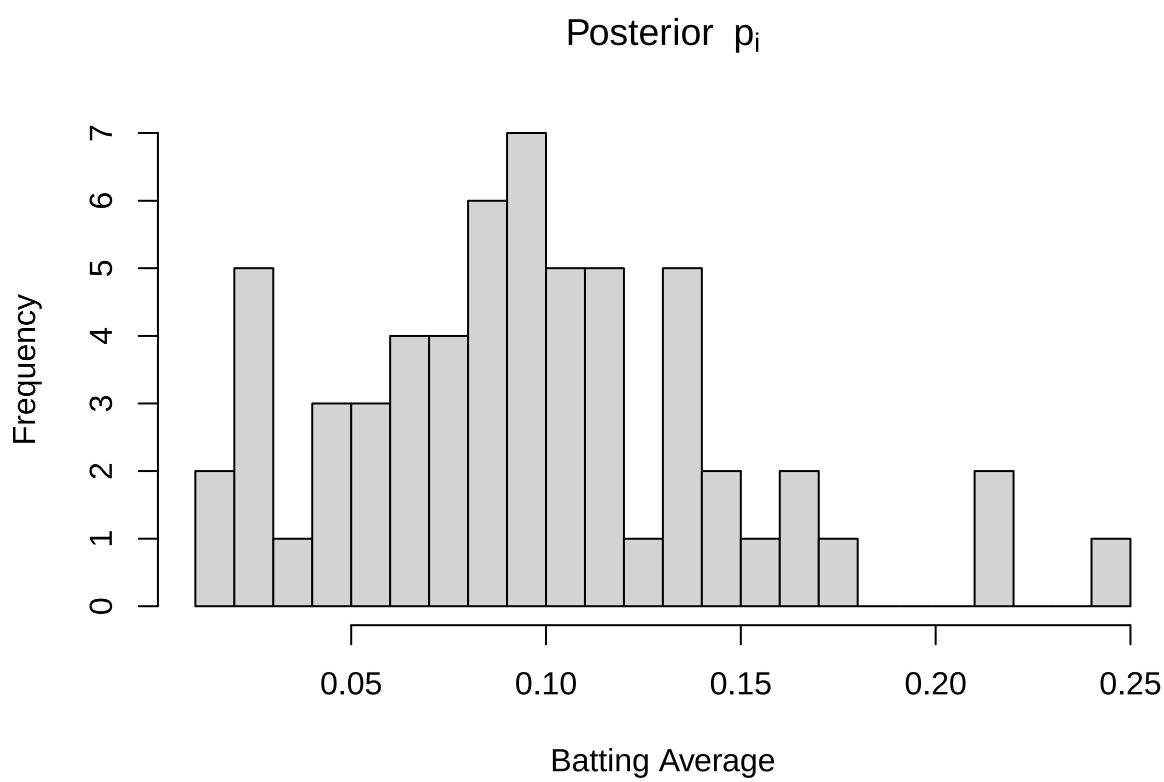
```r
# histogram (Problem 2)
hist(posterior_hitting_ability_p2,main=expression("Posterior "~p[i]),
     xlab="Batting Average",breaks=25)
```



```r
# estimates (Problem 3)
posterior_hitting_ability_p3 <- matrix(posterior_hitting_ability_p3,dim(data)[1],1)
colnames(posterior_hitting_ability_p3) <- c("Batting Average")
summary(posterior_hitting_ability_p3)
```

```
##  Batting Average
##  Min.   :0.01918
##  1st Qu.:0.06588
##  Median :0.09491
##  Mean   :0.09594
##  3rd Qu.:0.11917
##  Max.   :0.24213
```

```r
# histogram (Problem 3)
hist(posterior_hitting_ability_p3,main=expression("Posterior "~p[i]),
     xlab="Batting Average",breaks=25)
```



Problem 6.5 (pitchers)

The MSE of the estimates from Problem 6.2 (pitchers) is slightly larger than the MSE of the estimates from Problem 6.3 (pitchers); however, they are really similar. We believe that the reason for this is because the variances of the estimates from Problem 6.2 (pitchers) and 6.3 (pitchers) are similar and the biases of the estimates from Problem 6.2 (pitchers) and 6.3 (pitchers) are also similar based on the information and histograms in Problem 6.4 (pitchers).
On the other hand, the MSE from the model in Problem 3 is generally smaller than the MSE from the model in Problem 2. Besides, the MSE for the pitchers is largest while the MSE for nonpitchers is similar to the MSE for all players. We believe that the reason for this is because the dataset of the pitchers is relatively small while the batting averages of the pitchers have a larger variance and their distribution is skewed.

```r
data_hitting_ability.2 <- matrix(0,dim(data)[1],1)
colnames(data_hitting_ability.2) <- c("Batting Average")

for (i in 1:dim(data)[1]) {
  data_hitting_ability.2[i,1] <- data$H.2[i]/data$AB.2[i]
}

#MSE
mse2 <- sum((posterior_hitting_ability_p2-data_hitting_ability.2)^2)/nrow(data)
mse3 <- sum((posterior_hitting_ability_p3-data_hitting_ability.2)^2)/nrow(data)
MSE <- matrix(c(mse2,mse3),1,2)
rownames(MSE) <- c("MSE")
colnames(MSE) <- c("Problem 6.2","Problem 6.3")
MSE
```

```
##     Problem 6.2 Problem 6.3
## MSE  0.01287095  0.01283949
```