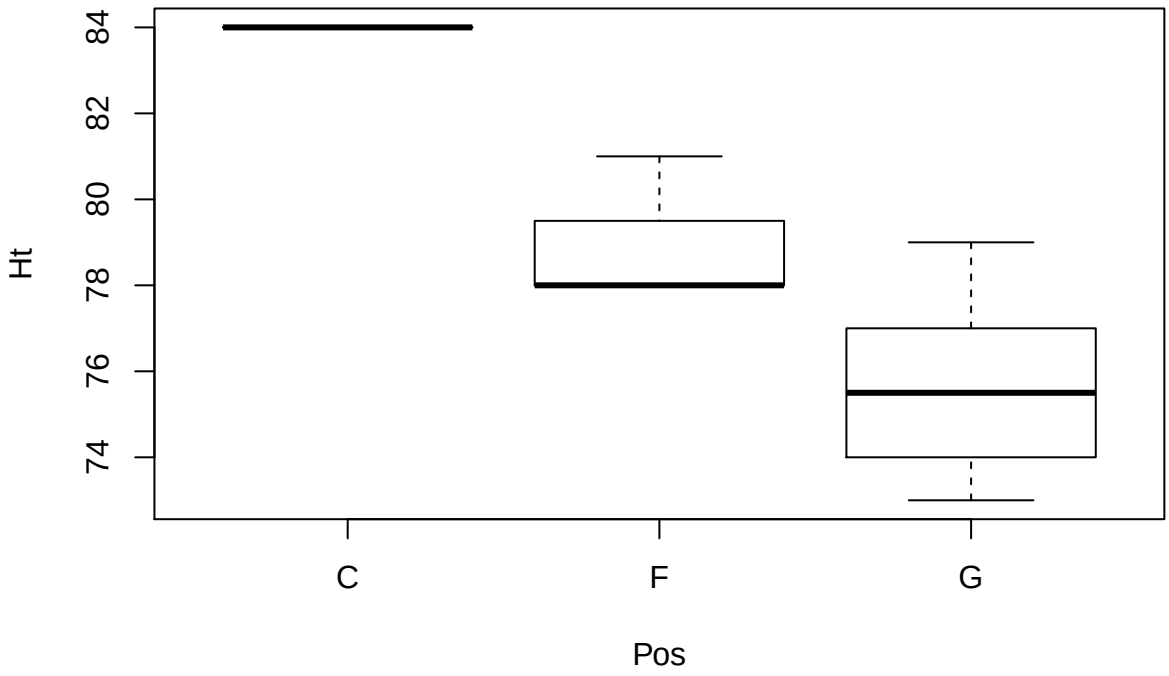
STAT 578 - Advanced Bayesian Modeling - Fall 2019
Assignment 6

Xiaoming Ji

Solution for Problem 1

```
perf_data = read.csv("illinimensbb.csv", header=TRUE)

plot(Ht ~ Pos, data= perf_data)
```



By checking the plot, we do see height and position are highly correlated. *center* has highest mean of height, *forward* has shortest mean of height and *forward* has in between these two. Their value ranges also don't seem to cross each other significantly.

Solution for Problem 2

(a)

```
model {
    for (i in 1:length(FGM)) {
        FGM[i] ~ dbin(prob[i], FGA[i])
```

```
        logit(prob[i]) <- beta_pos[Pos[i]] + beta_ht * Ht_Scaled[i]
        FGM_rep[i] ~ dbin(prob[i], FGA[i])
    }
    for (j in 1:max(Pos)) {
        beta_pos[j] ~ dt(0, 0.01, 1)
    }

    beta_ht ~ dt(0, 0.16, 1)
}
```

```
library(rjags)

df_jags_1 <- list( FGM = perf_data$FGM, FGA = perf_data$FGA,
                   Pos = unclass(perf_data$Pos),
                   Ht_Scaled = as.vector(scale(perf_data$Ht, scale=2*sd(perf_data$Ht))))

initial_vals_1 <- list(list(beta_pos = c(10,10,10), beta_ht=10),
                       list(beta_pos = c(10,10,-10), beta_ht=-10),
                       list(beta_pos = c(10,-10,10), beta_ht=-10),
                       list(beta_pos = c(10,-10,-10), beta_ht=10))

model_1 <- jags.model("perf_1.bug", df_jags_1, initial_vals_1, n.chains = 4,
                      n.adapt = 1000)
update(model_1, 1000)

#Need only check top-level parameters (in the DAG) for convergence.
x1 <- coda.samples(model_1, c("beta_pos","beta_ht"), n.iter = 2000)
```

```
gelman.diag(x1, autoburnin=FALSE)
```

```
## Potential scale reduction factors:
##
##             Point est. Upper C.I.
## beta_ht              1       1.01
## beta_pos[1]          1       1.01
## beta_pos[2]          1       1.01
## beta_pos[3]          1       1.01
##
## Multivariate psrf
##
## 1
```

```
coef_sample_1 <- coda.samples(model_1, c("beta_pos","beta_ht","prob","FGM_rep"),
                              n.iter = 10000, thin = 5)
effectiveSize(coef_sample_1[,c("beta_pos[1]", "beta_pos[2]", "beta_pos[3]", "beta_ht")])
```

```
## beta_pos[1] beta_pos[2] beta_pos[3]     beta_ht
##    6051.941    6250.811    5725.131    4706.478
```

(b)

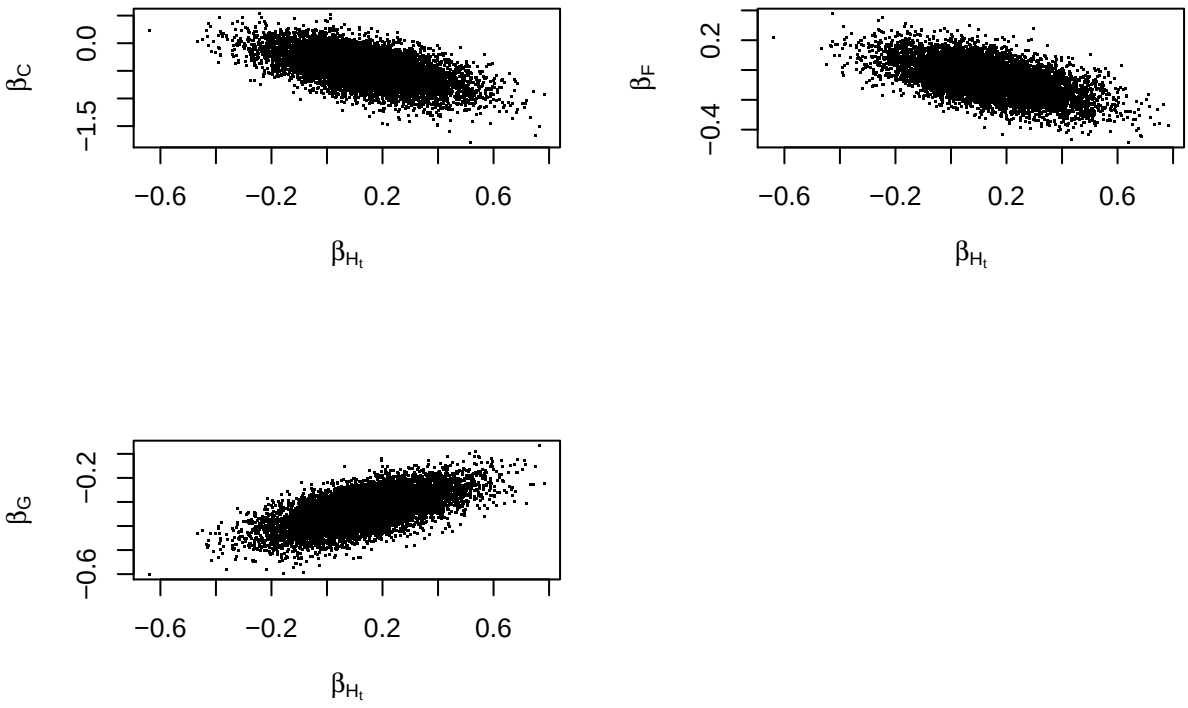
```
summary(coef_sample_1[, c("beta_pos[1]", "beta_pos[2]", "beta_pos[3]", "beta_ht")])
```

```
##
```

```
## Iterations = 4005:14000
## Thinning interval = 5
## Number of chains = 4
## Sample size per chain = 2000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                 Mean     SD  Naive SE Time-series SE
## beta_pos[1] -0.45744 0.2923 0.0032677      0.0037629
## beta_pos[2] -0.06157 0.1115 0.0012462      0.0014130
## beta_pos[3] -0.33432 0.0708 0.0007915      0.0009416
## beta_ht      0.13828 0.1804 0.0020175      0.0026321
##
## 2. Quantiles for each variable:
##
##                2.5%      25%     50%      75%   97.5%
## beta_pos[1] -1.0332 -0.64850 -0.4573 -0.25674  0.1085
## beta_pos[2] -0.2787 -0.13610 -0.0629  0.01292  0.1577
## beta_pos[3] -0.4750 -0.38122 -0.3334 -0.28628 -0.1981
## beta_ht     -0.2095  0.01519  0.1368  0.26105  0.4919
```

(c)

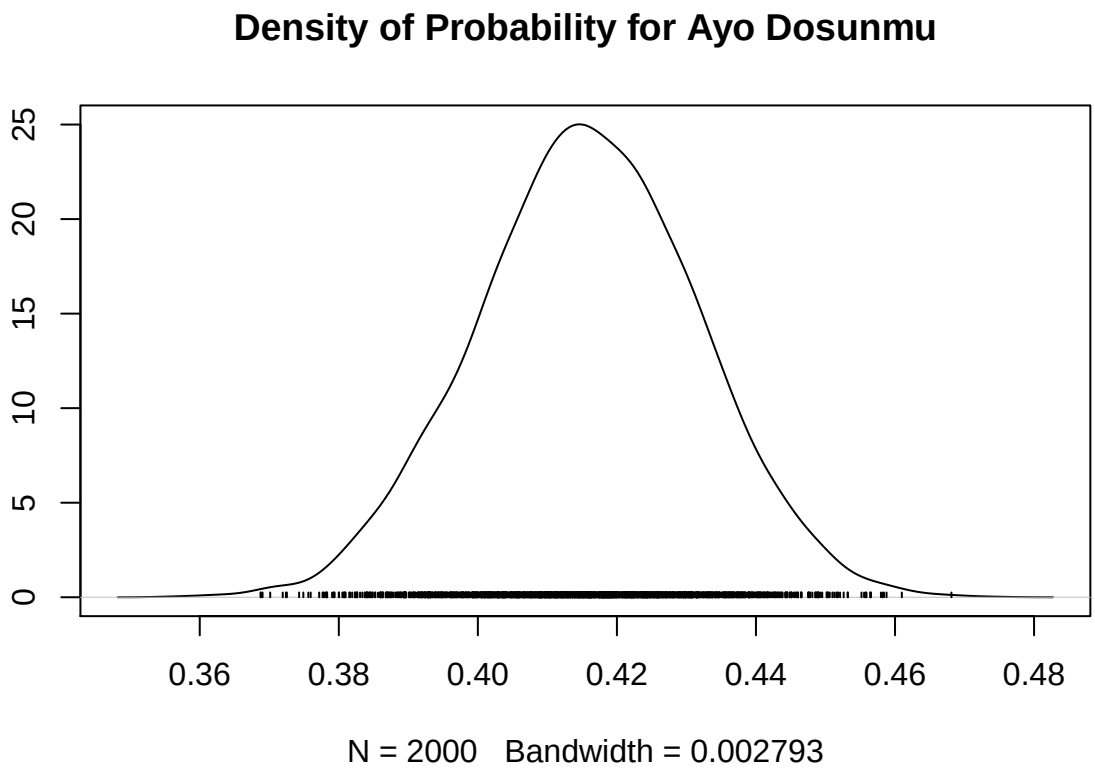
```
par(mfrow=c(2, 2))

plot(as.matrix(coef_sample_1)[,"beta_pos[1]"] ~ as.matrix(coef_sample_1)[,"beta_ht"],
     xlab = expression(paste(beta[H[t]])), ylab = expression(paste(beta[C])), pch='.')
plot(as.matrix(coef_sample_1)[,"beta_pos[2]"] ~ as.matrix(coef_sample_1)[,"beta_ht"],
     xlab = expression(paste(beta[H[t]])), ylab = expression(paste(beta[F])), pch='.')
plot(as.matrix(coef_sample_1)[,"beta_pos[3]"] ~ as.matrix(coef_sample_1)[,"beta_ht"],
     xlab = expression(paste(beta[H[t]])), ylab = expression(paste(beta[G])), pch='.')
```



According to the plots, β_C, β_F, β_G are correlated with β_{H_t}.

(d)

```
Dosunmu_index = which(perf_data$X==11)
densplot(coef_sample_1[, paste("prob[",Dosunmu_index,"]",sep="")],
         main = "Density of Probability for Ayo Dosunmu")
```



(e)

Probability of $\beta_F > \beta_G$,

```
beta_F = as.matrix(coef_sample_1)[, "beta_pos[2]"]
beta_G = as.matrix(coef_sample_1)[, "beta_pos[3]"]
mean(beta_F > beta_G)
```

```
## [1] 0.966125
```

Bayes factor favoring $\beta_F > \beta_G$ versus $\beta_F < \beta_G$,

```
mean(beta_F > beta_G) / mean(beta_F < beta_G)
```

```
## [1] 28.5203
```

Given the Bayes factor is between 20 to 150, we can say that the data has **Strong** evidence that $\beta_F > \beta_G$.

(f)

```
probs <- as.matrix(coef_sample_1)[, paste("prob[",1:nrow(perf_data),"]", sep="")]
FGM_rep <- as.matrix(coef_sample_1)[, paste("FGM_rep[",1:nrow(perf_data),"]", sep="")]


Tchi <- numeric(nrow(FGM_rep))
Tchirep <- numeric(nrow(FGM_rep))
```

```
for(s in 1:nrow(FGM_rep)){
  Tchi[s] <- sum((perf_data$FGM - perf_data$FGA * probs[s,])^2 /
                   (perf_data$FGA * probs[s,] * (1 - probs[s,])))
  Tchirep[s] <- sum((FGM_rep[s,] - perf_data$FGA * probs[s,])^2 /
                      (perf_data$FGA * probs[s,] * (1 - probs[s,])))
 }

mean(Tchirep >= Tchi)
```

```
## [1] 0.045875
```

The posterior predictive p-value is small, although not exceedingly so. Given we don't find any outliers, we conclude that there is a problem of overdispersion.

(g)

(i)

```
model {
    for (i in 1:length(FGM)) {
        FGM[i] ~ dbin(prob[i], FGA[i])
        logit(prob[i]) <- beta_pos[Pos[i]] + beta_ht * Ht_Scaled[i] + epsilon[i]
        epsilon[i] ~ dnorm(0, 1 / sigma_epsilon^2)
        FGM_rep[i] ~ dbin(prob[i], FGA[i])
    }
    for (j in 1:max(Pos)) {
        beta_pos[j] ~ dt(0, 0.01, 1)
    }

    beta_ht ~ dt(0, 0.16, 1)
    sigma_epsilon ~ dunif(0,10)
}
```

```
df_jags_2 <- list( FGM = perf_data$FGM, FGA = perf_data$FGA,
                   Pos = unclass(perf_data$Pos),
                   Ht_Scaled = as.vector(scale(perf_data$Ht, scale=2*sd(perf_data$Ht))))

initial_vals_2 <- list(list(beta_pos = c(10,10,10), beta_ht=10, sigma_epsilon = 0.01),
                       list(beta_pos = c(10,10,-10), beta_ht=-10, sigma_epsilon = 9),
                       list(beta_pos = c(10,-10,10), beta_ht=-10, sigma_epsilon = 0.01),
                       list(beta_pos = c(10,-10,-10), beta_ht=10, sigma_epsilon = 9))

model_2 <- jags.model("perf_2.bug", df_jags_2, initial_vals_2, n.chains = 4,
                      n.adapt = 1000)
update(model_2, 1000)
x2 <- coda.samples(model_2, c("beta_pos","beta_ht", "sigma_epsilon"), n.iter = 10000)
```

```
gelman.diag(x2, autoburnin=FALSE)
```

```
## Potential scale reduction factors:
##
##               Point est. Upper C.I.
## beta_ht             1.02       1.04
## beta_pos[1]         1.01       1.02
```

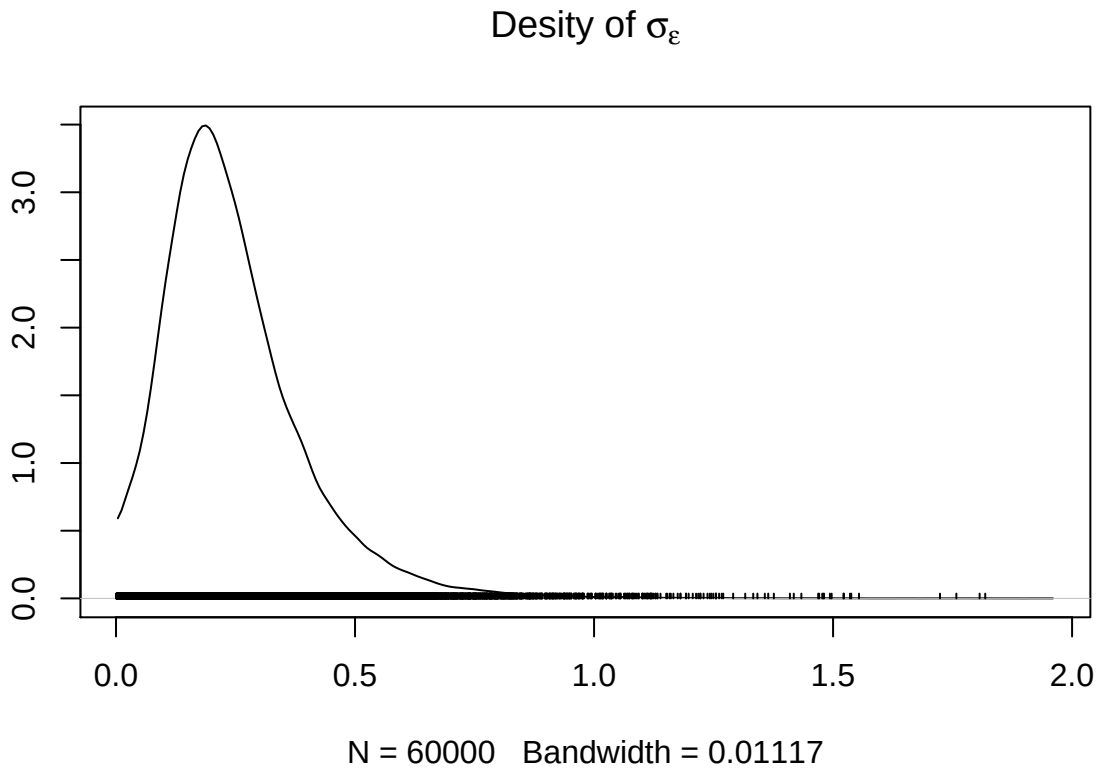
```
## beta_pos[2]          1.01       1.02
## beta_pos[3]          1.01       1.01
## sigma_epsilon        1.01       1.03
##
## Multivariate psrf
##
## 1.01
```

```
coef_sample_2 <- coda.samples(model_2, c("beta_pos","beta_ht","prob","FGM_rep",
                                         "sigma_epsilon"), n.iter = 60000)
effectiveSize(coef_sample_2[,c("beta_pos[1]", "beta_pos[2]", "beta_pos[3]", "beta_ht",
                               "sigma_epsilon")])
```

```
##   beta_pos[1]   beta_pos[2]   beta_pos[3]       beta_ht sigma_epsilon
##      8335.420      6418.475      7133.235      5479.946      5458.723
```

(ii)

```
densplot(coef_sample_2[, "sigma_epsilon"],
         main = expression(paste("Desity of ", sigma[epsilon])))
```



(iii)

```
beta_F = as.matrix(coef_sample_2)[, "beta_pos[2]"]
beta_G = as.matrix(coef_sample_2)[, "beta_pos[3]"]
mean(beta_F > beta_G)
```

```
## [1] 0.78665
```

This posterior probability is smaller than previous model.

```
mean(beta_F > beta_G) / mean(beta_F < beta_G)
```

```
## [1] 3.687134
```

This Bayes factor favoring $\beta_F > \beta_G$ versus $\beta_F < \beta_G$ is much smaller than previous model, and we can only say the data has **Positive** (between 3 to 30) evidence that $\beta_F > \beta_G$.

Also Chi-square discrepancy,

```
## [1] 0.3826792
```

Thus we says no overdispersion problems for this model.

Solution for Problem 3

(a)

```
model {
    for (i in 1:length(BLK)) {
        BLK[i] ~ dpois(lambda[i])
        log(lambda[i]) <- log_MIN[i] + beta_pos[Pos[i]] + beta_ht * Ht_Scaled[i]
        BLK_rep[i] ~ dpois(lambda[i])
    }

    for (j in 1:max(Pos)) {
        beta_pos[j] ~ dnorm(0, 0.0001)
    }

    beta_ht ~ dnorm(0, 0.0001)
}
```

```
df_jags_3 <- list( BLK = perf_data$BLK,
                   Pos = unclass(perf_data$Pos),
                   log_MIN = log(perf_data$MIN),
                   Ht_Scaled = as.vector(scale(perf_data$Ht, scale=sd(perf_data$Ht))))

initial_vals_3 <- list(list(beta_pos = c(100,100,100), beta_ht=100),
                       list(beta_pos = c(100,100,-100), beta_ht=-100),
                       list(beta_pos = c(100,-100,100), beta_ht=-100),
                       list(beta_pos = c(100,-100,-100), beta_ht=100))

model_3 <- jags.model("perf_3.bug", df_jags_3, initial_vals_3, n.chains = 4,
                      n.adapt = 1000)
update(model_3, 1000)
x3 <- coda.samples(model_3, c("beta_pos","beta_ht"), n.iter = 2000)
```

```
gelman.diag(x3, autoburnin=FALSE)
```

```
## Potential scale reduction factors:
##
##             Point est. Upper C.I.
## beta_ht              1          1
## beta_pos[1]          1          1
```

```
## beta_pos[2]          1          1
## beta_pos[3]          1          1
##
## Multivariate psrf
##
## 1
```

```
coef_sample_3 <- coda.samples(model_3, c("beta_pos","beta_ht", "lambda","BLK_rep"),
                              n.iter = 20000, thin = 5)
effectiveSize(coef_sample_3[,c("beta_pos[1]", "beta_pos[2]", "beta_pos[3]", "beta_ht")])
```

```
## beta_pos[1] beta_pos[2] beta_pos[3]     beta_ht
##    4881.227    5760.599    9429.966    4620.907
```

(b)

```
summary(coef_sample_3[, c("beta_pos[1]", "beta_pos[2]", "beta_pos[3]", "beta_ht")])
```

```
##
## Iterations = 4005:24000
## Thinning interval = 5
## Number of chains = 4
## Sample size per chain = 4000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                Mean     SD Naive SE Time-series SE
## beta_pos[1] -5.2725 0.6009 0.004750       0.008628
## beta_pos[2] -4.5005 0.2844 0.002248       0.003795
## beta_pos[3] -4.4522 0.1774 0.001402       0.001831
## beta_ht      0.9969 0.2727 0.002156       0.004027
##
## 2. Quantiles for each variable:
##
##                2.5%     25%     50%    75%  97.5%
## beta_pos[1] -6.4756 -5.6724 -5.2571 -4.857 -4.138
## beta_pos[2] -5.0832 -4.6864 -4.4929 -4.304 -3.972
## beta_pos[3] -4.8080 -4.5709 -4.4499 -4.329 -4.116
## beta_ht      0.4797  0.8097  0.9896  1.179  1.543
```

(c)

```
beta_ht = as.matrix(coef_sample_3)[, "beta_ht"]
quantile(exp(beta_ht), c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 1.615524 4.679455
```

The values within 95% central posterior credible interval are all greater than 1 and thus we can conclude that greater height is associated with a higher rate of blocking shots.

(d)

```
lambdas <- as.matrix(coef_sample_3)[, paste("lambda[",1:nrow(perf_data),"]", sep="")]
BLK_rep <- as.matrix(coef_sample_3)[, paste("BLK_rep[",1:nrow(perf_data),"]", sep="")]

Tchi <- numeric(nrow(BLK_rep))
Tchirep <- numeric(nrow(BLK_rep))

for(s in 1:nrow(BLK_rep)){
  Tchi[s] <- sum((perf_data$BLK - lambdas[s,])^2 / lambdas[s,])
  Tchirep[s] <- sum((BLK_rep[s,] - lambdas[s,])^2 / lambdas[s,])
 }

mean(Tchirep >= Tchi)
```

```
## [1] 0.0069375
```

The posterior predictive p-value is extremely small. Thus this could indicate a problem of overdispersion.

(e)

(i)

```
p_sample <- matrix(FALSE, nrow = nrow(BLK_rep), ncol = nrow(perf_data))
for(s in 1:nrow(BLK_rep)){
  p_sample[s,] <- BLK_rep[s,] > perf_data$BLK
}

p = apply(p_sample, 2, mean)
p_df = data.frame(name=perf_data$Player, p_value=p)
p_df
```

```
##                     name   p_value
## 1  Bezhanishvili, Giorgi 0.5311250
## 2            Cayce, Drew 0.0592500
## 3     De La Rosa, Adonis 0.9865625
## 4           Dosunmu, Ayo 0.7095000
## 5          Feliz, Andres 0.8339375
## 6         Frazier, Trent 0.8236250
## 7          Griffin, Alan 0.0075000
## 8         Griffith, Zach 0.1917500
## 9          Jones, Tevian 0.8935625
## 10         Jordan, Aaron 0.1326875
## 11           Kane, Samba 0.0022500
## 12       Nichols, Kipper 0.2376250
## 13     Oladimeji, Samson 0.1830625
## 14      Underwood, Tyler 0.2636250
## 15    Williams, Da'Monte 0.0436875
```

(ii)

```
p_df[p_df$p_value < 0.05,]
```

```
##                 name   p_value
```

```
## 7       Griffin, Alan 0.0075000
## 11        Kane, Samba 0.0022500
## 15 Williams, Da'Monte 0.0436875
```

(iii)

```
p_df[p_df$p_value > 0.95,]
```

```
##                 name   p_value
## 3 De La Rosa, Adonis 0.9865625
```

By looking at the original data, **Adonis** played in center position and was 84 height. He played 225 minutes but blocked only 1 shot. Samba in another hand, also played in center position and was also 84 height. For 86 minutes he played, blocked 10 shots. This makes the model always overestimate the blocks by Adonis. Thus the p-value is very high.