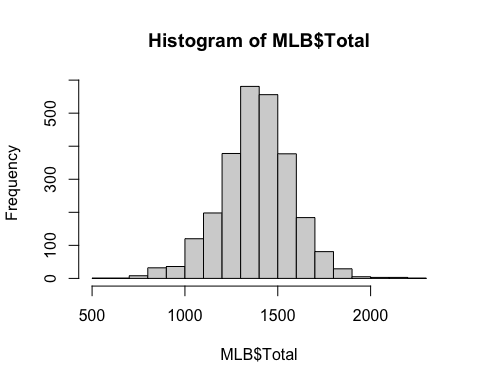
MLB Steroid Era Project

library(invgamma)  
# Reading and Cleaning data  
MLB <- read.csv(file = "baseballdata.csv", header = TRUE)  
MLB\_copy <- MLB  
MLB <- MLB[,c(3, 15, 16, 24)]  
MLB$Total <- MLB$R + MLB$RA  
hist(MLB$Total)



# Standardize 1995  
year1995 <- which(MLB\_copy$Year == 1995)  
(MLB\_copy$G[year1995])

## [1] 144 144 144 144 145 144 144 144 144 144 144 145 144 143 144 144 144 145 144  
## [20] 144 144 144 144 145 143 144 144 144

MLB$Total[year1995] <- (MLB$Total[year1995]/144) \* 162

# Standardize 1994  
year1994 <- which(MLB\_copy$Year == 1994)  
mean((MLB\_copy$G[year1994]))

## [1] 114.2857

MLB$Total[year1994] <- (MLB$Total[year1994]/114) \* 162

# Steroid data  
steroidyear <- c(which(MLB$Year == 1994), which(MLB$Year == 1995), which(MLB$Year == 1996), which(MLB$Year == 1997),   
 which(MLB$Year == 1998), which(MLB$Year == 1999), which(MLB$Year == 2000), which(MLB$Year == 2001),  
 which(MLB$Year == 2002), which(MLB$Year == 2003))  
steroid <- MLB[steroidyear, 5]  
# Range  
min(steroid)

## [1] 1130

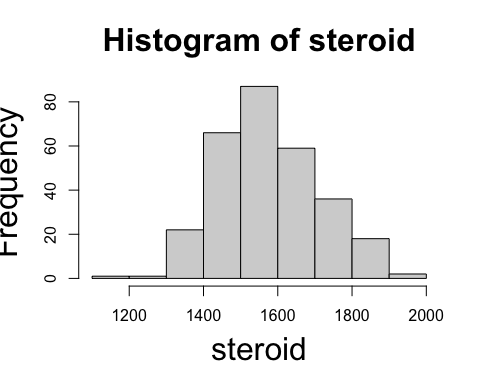
mean(steroid)

## [1] 1577.82

max(steroid)

## [1] 1934

# Histogram  
hist(cex.lab = 2, cex.main = 2, steroid)



# Clean data  
cleanyear <- c(which(MLB$Year == 2004), which(MLB$Year == 2005), which(MLB$Year == 2006), which(MLB$Year == 2007),   
 which(MLB$Year == 2008), which(MLB$Year == 2009), which(MLB$Year == 2010), which(MLB$Year == 2011),  
 which(MLB$Year == 2012), which(MLB$Year == 2013))  
clean <- MLB[cleanyear, 5]  
#Range  
min(clean)

## [1] 1148

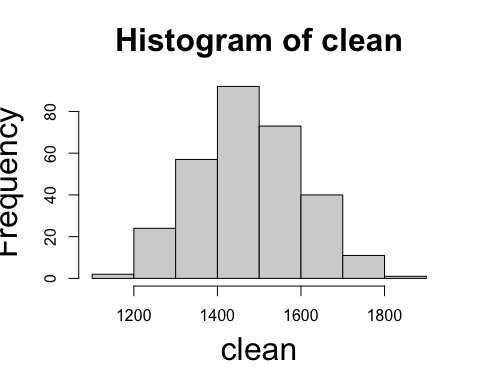
mean(clean)

## [1] 1473.427

max(clean)

## [1] 1868

# Histogram  
hist(cex.lab = 2, cex.main = 2, clean)



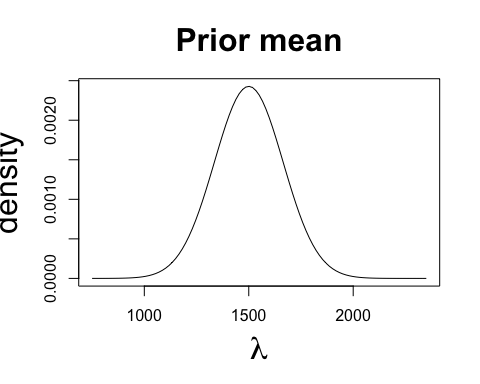
# Setting prior both steroid and clean  
n <- length(steroid)  
n2 <- length(clean)  
lambda <- 1500  
tau2 <- 27000  
gamma <- 4  
phi <- 81675  
  
phi / (gamma - 1)

## [1] 27225

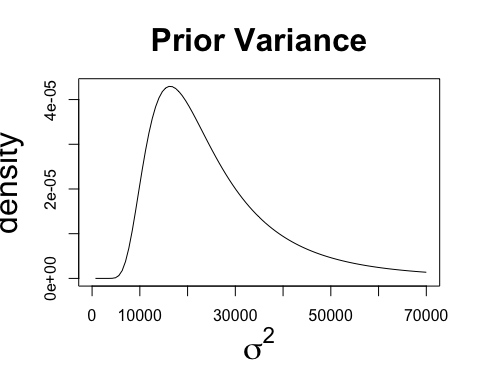
phi^2 / ((gamma-1)^2\*(gamma-2))

## [1] 370600312

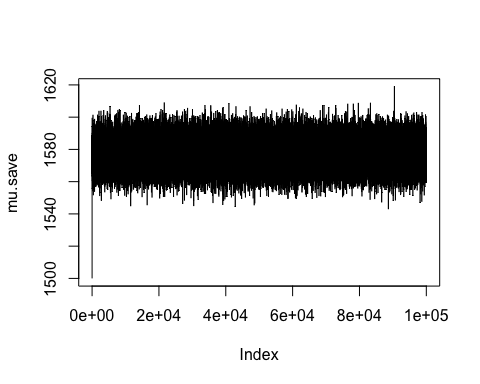
# Prior plots  
curve(cex.lab = 2, cex.main = 2, dnorm(x,lambda,sqrt(tau2)), xlim = c(750, 2350), xlab = expression(lambda), main = "Prior mean", ylab = "density")



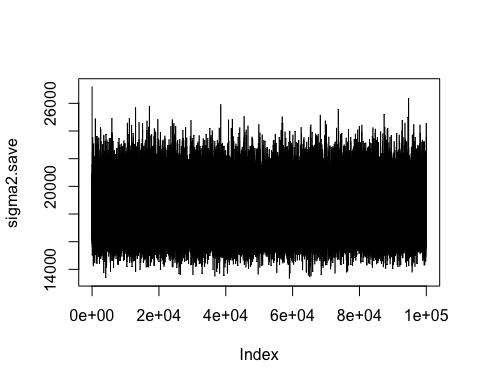
curve(cex.lab = 2, cex.main = 2, dinvgamma(x,gamma,phi), xlim = c(0, 70000), xlab = expression(sigma^2), main = "Prior Variance", ylab = "density")



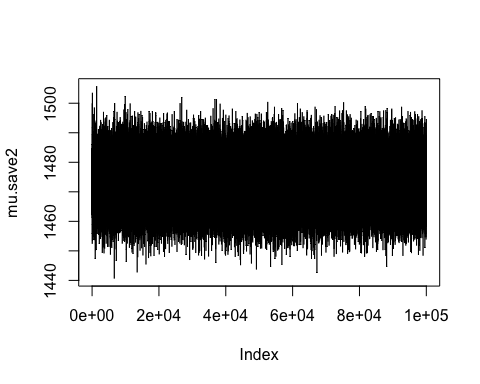
# Gibbs sampling method  
mu <- lambda  
mu2 <- lambda  
sigma2 <- phi / (gamma - 1)  
sigma2.2 <- phi / (gamma - 1)  
  
iters <- 100000  
  
mu.save <- rep(0, iters)  
mu.save2 <- rep(0, iters)  
  
mu.save[1] <- mu  
mu.save2[1] <- mu2  
  
sigma2.save <- rep(0, iters)  
sigma2.save[1] <- sigma2  
  
sigma2.save2 <- rep(0, iters)  
sigma2.save2[1] <- sigma2.2  
  
for (i in 2:iters){  
 lambda.p <- (tau2 \* sum(steroid) + sigma2 \* lambda) / (tau2 \* n + sigma2)  
 tau2.p <- sigma2 \* tau2 / (tau2 \* n + sigma2)  
   
 lambda.p2 <- (tau2 \* sum(clean) + sigma2.2 \* lambda) / (tau2 \* n2 + sigma2.2)  
 tau2.p2 <- sigma2.2 \* tau2 / (tau2 \* n2 + sigma2.2)  
   
 mu <- rnorm(1, lambda.p, sqrt(tau2.p))  
 mu.save[i] <- mu  
   
 mu2 <- rnorm(1, lambda.p2, sqrt(tau2.p2))  
 mu.save2[i] <- mu2  
   
 gamma.p <- gamma + n/2  
 phi.p <- phi + sum((steroid - mu)^2) / 2  
   
 gamma.p2 <- gamma + n2/2  
 phi.p2 <- phi + sum((clean - mu2)^2) / 2  
   
 sigma2 <- rinvgamma(1,gamma.p, phi.p)  
 sigma2.save[i] <- sigma2  
   
 sigma2.2 <- rinvgamma(1, gamma.p2, phi.p2)  
 sigma2.save2[i] <- sigma2.2  
}  
  
# Trace plot (Steroid)  
plot(mu.save, type = "l")



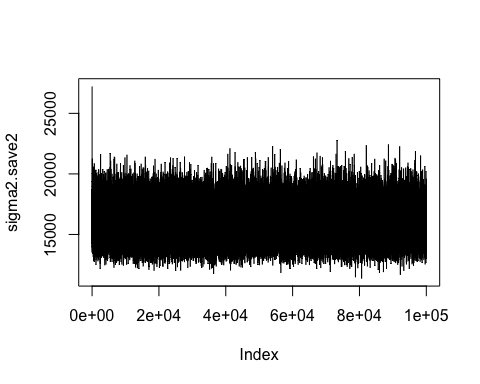
plot(sigma2.save, type = "l")



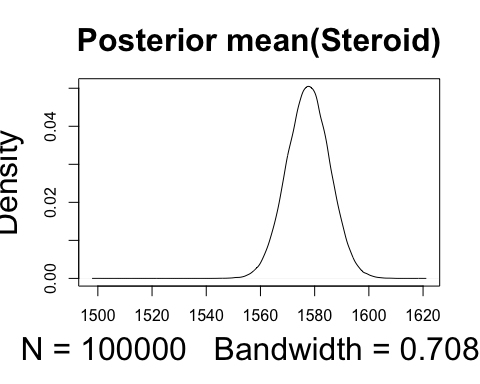
# Trace plot (Clean)  
plot(mu.save2, type = "l")



plot(sigma2.save2, type = "l")



# Posterior Distributions for each group  
plot(cex.lab = 2, cex.main = 2, density(mu.save), main = "Posterior mean(Steroid)")



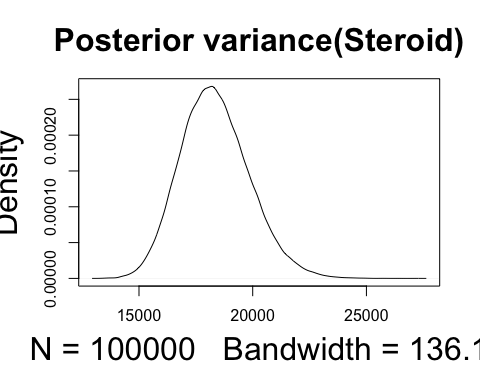
quantile(mu.save, probs = c(0.025, 0.975))

## 2.5% 97.5%   
## 1562.207 1592.965

mean(mu.save)

## [1] 1577.634

plot(cex.lab = 2, cex.main = 2, density(sigma2.save), main = "Posterior variance(Steroid)")



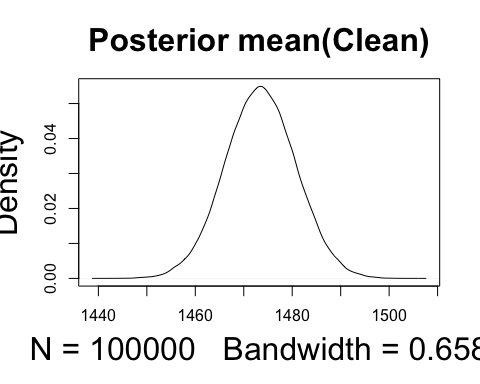
quantile(sigma2.save, probs = c(0.025, 0.975))

## 2.5% 97.5%   
## 15627.36 21563.41

mean(sigma2.save)

## [1] 18345.95

plot(cex.lab = 2, cex.main = 2, density(mu.save2), main = "Posterior mean(Clean)")



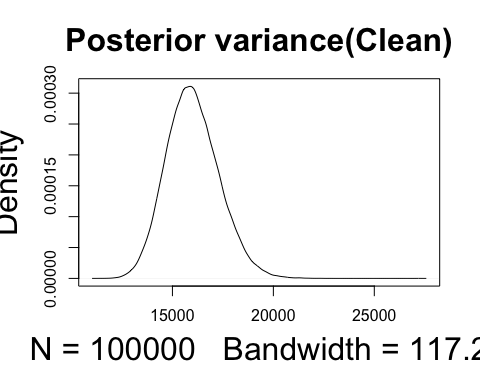
quantile(mu.save2, probs = c(0.025, 0.975))

## 2.5% 97.5%   
## 1459.084 1487.836

mean(mu.save2)

## [1] 1473.464

plot(cex.lab = 2, cex.main = 2, density(sigma2.save2), main = "Posterior variance(Clean)")



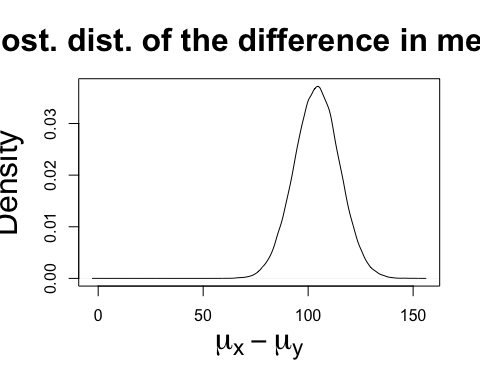
quantile(sigma2.save2, probs = c(0.025, .5, 0.975))

## 2.5% 50% 97.5%   
## 13640.94 15940.82 18741.88

mean(sigma2.save2)

## [1] 16009.76

# Postrior Distribution for the difference in the means  
diff <- mu.save - mu.save2  
plot(cex.lab = 2, cex.main = 2, xlab = expression(mu[x] - mu[y]), density(diff), main = "Post. dist. of the difference in means")



quantile(diff, probs = c(0.025, .5, 0.975))

## 2.5% 50% 97.5%   
## 83.0495 104.2001 125.2992

mean(diff)

## [1] 104.1703

var(diff)

## [1] 115.997