Stat 251 Project: Steroid Era of Major League Baseball

By Jungeol Kim and Romnikh Ortega

Introduction

- Problem Statement

From the late ‘80s through the late 2000s, a number of players were believed to have used performance-enhancing drugs (PED) in Major League Baseball (MLB). As a result, offensive outputs such as Home Runs and Runs Batted In were believed to have increased, which also have increased the total runs (R, scores made by the team) and runs allowed (RA, scores made by the opponent team) during the “steroid era”. We are trying to find if there is difference between the average of total scores made (R + RA) for each MLB team between the steroid era (1994 - 2003) and the clean era (2004 - 2013), when MLB office implemented league-wide PED testing in 2004. If using PED was effective for offensive statistics in MLB, new regulations and PED testing from MLB office should reduce the total scores made.

- Parameters of Interest

We are interested in the posterior distribution of the difference in the means of the two groups: (total scores for all MLB teams during the steroid era) – (total scores for all MLB teams during the clean era).

Methods

- Random Variables

X = Total scores for each MLB team during the steroid era. The number of total scores for steroid era is 292 ($n_x = 292$).

Y = Total scores for each MLB team during the clean era. The number of total scores for clean era is 300 ($n_y = 300$).

- Likelihood

Steroid Era:

$$f(x_1,x_2,..,x_{nx}\,|\mu_x,\,\sigma_x^2) = \prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma_x}^{n_x/2}\; e^{\frac{\sum_{i=1}^{n_x}(x_i - u_x)^{\wedge}2}{2\sigma_x}}\right)$$

Clean Era:

$$f(y_1,y_2,..,y_{ny}\,|\mu_y,\,\sigma_y^2) = \prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma_y}^{n_y/2}\; e^{\frac{\sum_{i=1}^{n_y}(y_i - u_y)^{\wedge}2}{2\sigma_y}}\right)$$

- Histogram for total scores from steroid era and clean era

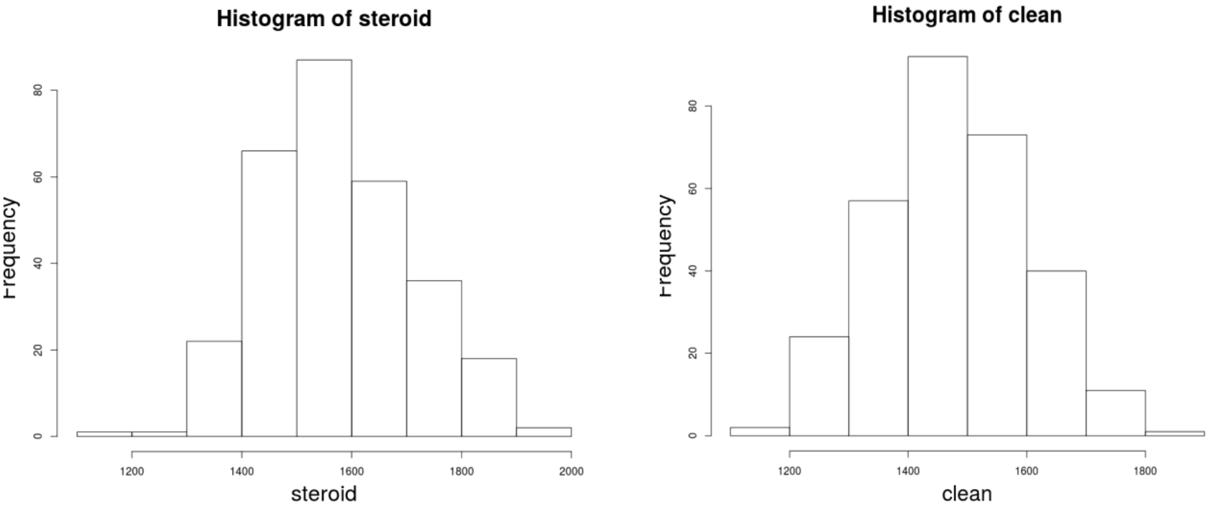


Figure 1 <Histograms for steroid data and clean data>

- Likelihood Justification

By plotting the histogram of the total scores from the steroid era and the clean era, we found out that both histograms looked approximately normal which led us to the conclusion that we should use the normal distribution for the likelihood.

- Prior Distribution

For λ(prior mean) we chose the normal distribution with the mean of 1500 and the variance of 27000

$\lambda \sim \mathbf{N(1500, 27000)}$.

For σ^2(prior variance), we chose the inverse gamma distribution with the rate parameter of 4 and the shape parameter of 81675.

$\boldsymbol{\sigma^2 \sim} \mathbf{IG(4, 81675)}$.

➢ Prior Plots

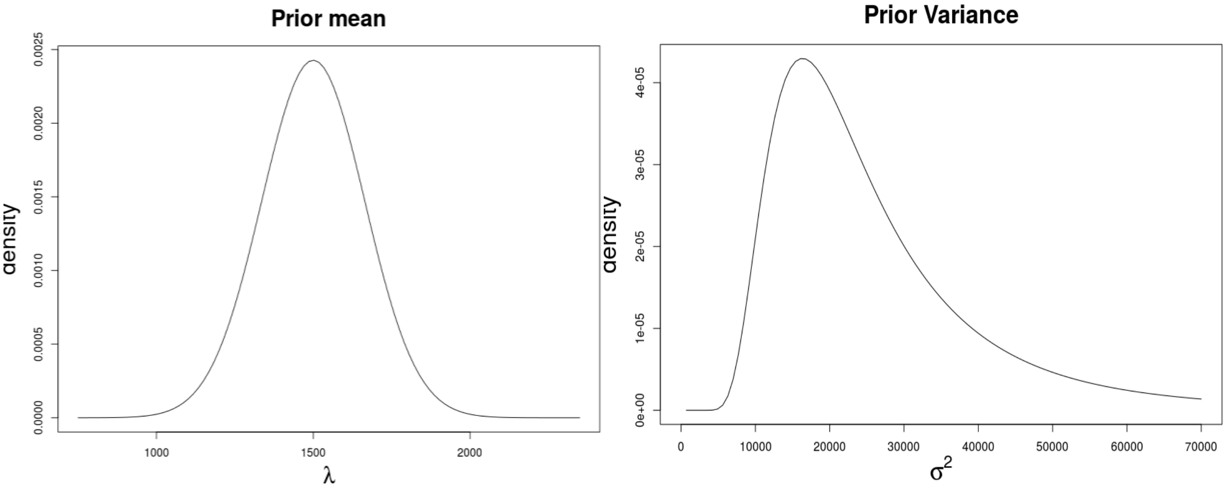


Figure 2 <Prior mean and Prior Variance>

➢ Prior Belief

We chose 1500 for prior mean because we expected the range of the total score to be between 1000 to 2000. By averaging these two numbers, we obtained the prior mean, λ. Also, we chose 27000 for our variance on λ because our estimated range from 1000 to 2000 gives us estimation of standard deviation around 165. Since squaring the estimated standard deviation gives us about 27000 of variance, we chose to set our prior mean with those parameters.

Also, using our prior belief that the range of the total scores is between 1000 and 2000 and by estimating the standard deviation of this distribution by taking the difference of the range and dividing by 6, we obtained the estimate for the variance as 165^2. By using the mean formula for inverse gamma, $\frac{\Phi}{\gamma - 1} = 165^2$, while keeping γ(rate parameter) as small as possible and Φ(shape parameter) as big as possible (to maximize the spread of the prior distribution), we chose the parameters $\gamma = 4$ and $\Phi = 81675$. This prior distribution is appropriate because the inverse Gamma distribution is the conjugate prior for our likelihood.

➢ Data collection

We have found our data online from “Kaggle.com”. This data contains various statistics for each MLB team from 1870s to 2016. Among these statistics, we have chosen two statistics, R and RA, and made a new column within the dataset for total scores made by summing the two statistics. Then, we grouped our data by years for steroid era and years for clean era.

During the analysis, we have found that 1994 and 1995 MLB seasons were shorter than others (having less games) due to strike and lockout. Each season had 114 games and 144 games per team only, while other years had 162 games per team. Therefore, we standardized total scores in the year 1994 and 1995 by diving totals scores by total games and multiply by 162. This way, we can approximate more accurate total score for the year 1994 and 1995 as if they had 162 games.

Also, we had eight fewer datapoints in the steroid era than the clean era because two teams, Arizona Diamondbacks and Tampa Bay Rays, were formed in 1998. Thus, the two teams were not in the league during 1994 ~ 1997 MLB seasons.

Total scores per team for each season during the steroid era range from 1130 to 1934 with mean of 1577.82, while total scores per team for each season during the clean era range from 1148 to 1868 with mean of 1473.427.

Results

- Full Conditional

Steroid era: μ_x(posterior mean), σ_x^2(posterior variance)

μ_x| data, $\sigma_x^2 \sim N(\lambda',(\tau^2)')$ where,

$\lambda' = \frac{(165^2)(\sum_{i=1}^{n_x} x_i) + 1500\sigma_x^2}{n_x 165^2 + \sigma_x^2}$ and $(\tau^2)' = \frac{165^2\sigma_x^2}{n_x 165^2 + \sigma_x^2}$ and $n_x = 292$ (8 fewer than n_y)

σ_x^2| data, $\mu \sim IG(\gamma',\Phi')$ where,

$\gamma' = 4 + \frac{n_x}{2}$ and $\Phi' = 81675 + \frac{\sum_{i=1}^{n_x}(x_i - \mu_x)}{2}$ and $n_y = 300$ (30 teams * 10 seasons)

Clean era: μ_y(posterior mean), σ_y^2(posterior variance)

μ_y| data, $\sigma_y^2 \sim N(\lambda',(\tau^2)')$ where,

$\lambda' = \frac{(165^2)(\sum_{i=1}^{n_y} y_i) + 1500\sigma_y^2}{n_y 165^2 + \sigma_y^2}$ and $(\tau^2)' = \frac{165^2\sigma_y^2}{n_y 165^2 + \sigma_y^2}$

σ_y^2| data, $\mu \sim IG(\gamma',\Phi')$ where,

$\gamma' = 4 + \frac{n_y}{2}$ and $\Phi' = 81675 + \frac{\sum_{i=1}^{n_y}(y_i - \mu_y)}{2}$

- Posterior Distribution

Note that the posterior distribution for the difference has no known closed form. Thus, we used Gibbs Sampling to estimate the joint posterior distribution (μ_x, σ_x^2) and (μ_y, σ_y^2). After getting the joint posterior distribution for both groups, we calculated $\mu_{x\text{ - }}\mu_y$ to obtain a posterior distribution for the difference in the means.

- Posterior Plot and Credible Interval

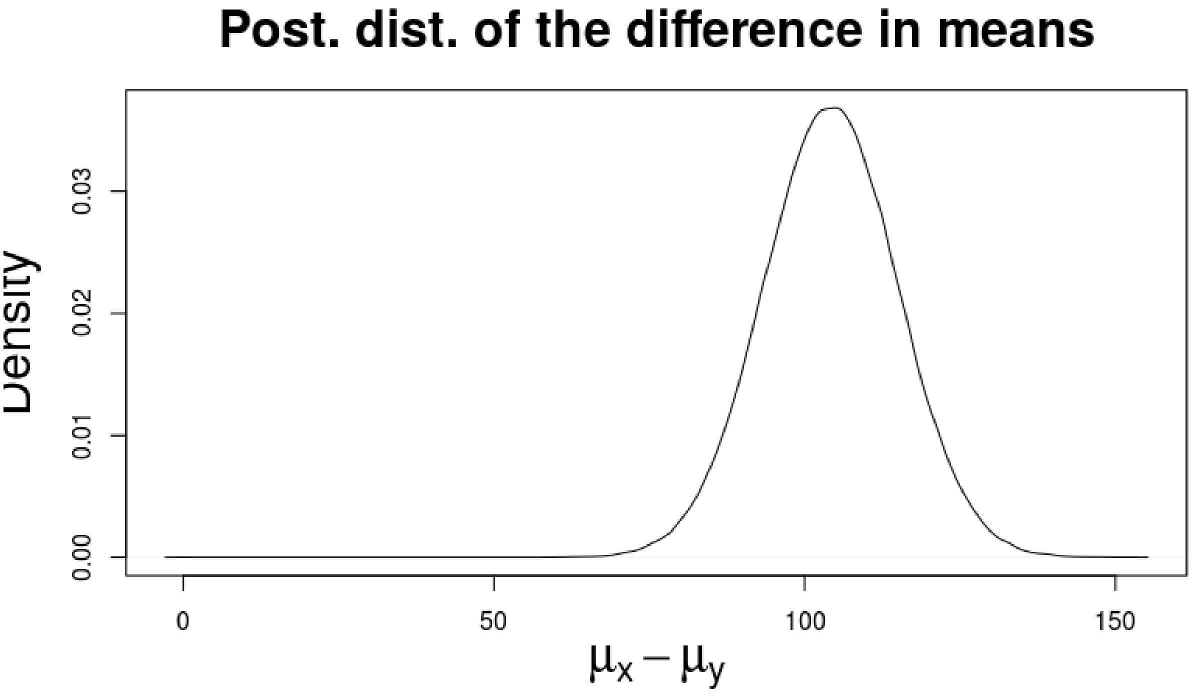


Figure 3 <Posterior Distribution of the difference in the means>

According to our analysis, the mean and the variance of the difference in the means between the two eras are 104.1607 and 115.6171. Also, from 95% posterior probability interval that we calculated, we can conclude that there is 95% probability that the total score per MLB team during the steroid era is higher than the total score per MLB team during the clean era between 82.9698 points and 125.2249 points.

Conclusions

From our analysis, we have found that the average total scores per team made during the steroid era is about 104.2 points higher than the average total scores per team made during the clean era. Thus, this answers our research question, and we can conclude that the total scores made per team during the steroid era is higher than the total scores made per team during the clean era.

When we first saw the data, distribution between two groups looked similar. Therefore, we set the same prior for both groups. Also, since there are pitchers who took performance enhancing drugs (PED) as well, we expected no difference between the two groups. However, after Gibbs sampling, we have found that mean and variance for the two groups are different from each other. This indicates that using PED was effective for batters to produce offensive outcomes in the league, and now they produce fewer offensive outputs due to new regulations and PED testing from MLB office.

For the future analysis, since we know that the average total scores each team made was about 104.2 points higher during the steroid era as opposed to the clean era, it will be interesting to do further analysis and determine if the winning average of a team with at least one player using steroid also increased during the steroid era.

Appendix (Data)

Accessible at:

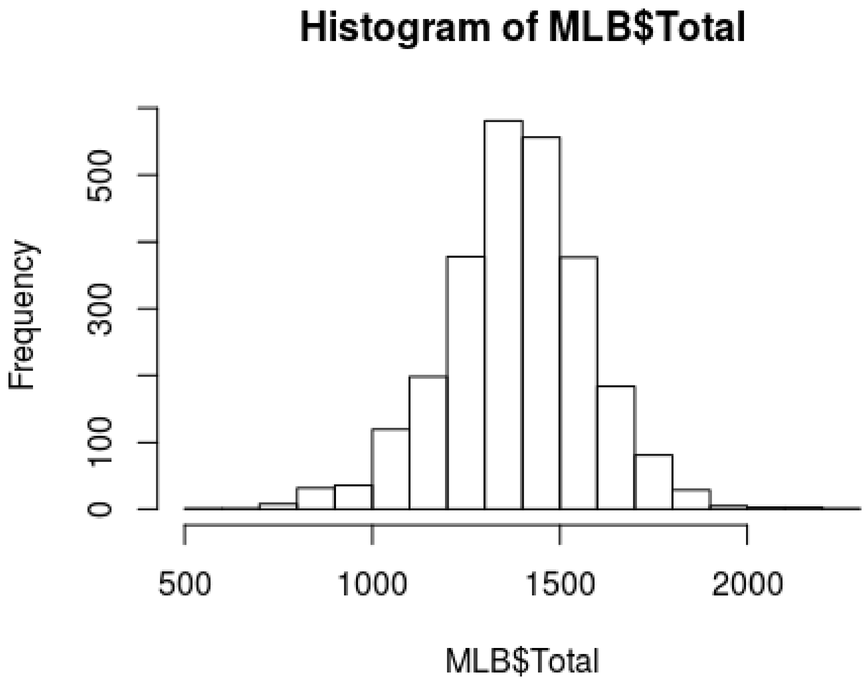
https://www.kaggle.com/timschutzyang/dataset1

Appendix (R code used)

Accessible at:

https://rstudio.cloud/project/645984

```r
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)
```



```r
# 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 14
4 144
## [18] 145 144 144 144 144 144 145 143 144 144 144
```

```
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), whi
ch(MLB$Year == 1996), which(MLB$Year == 1997),
                 which(MLB$Year == 1998), which(MLB$Year == 1999), whi
ch(MLB$Year == 2000), which(MLB$Year == 2001),
                 which(MLB$Year == 2002), which(MLB$Year == 2003))
steroid <- MLB[steroidyear, 5]
# Range
min(steroid)
```

```
## [1] 1130
```

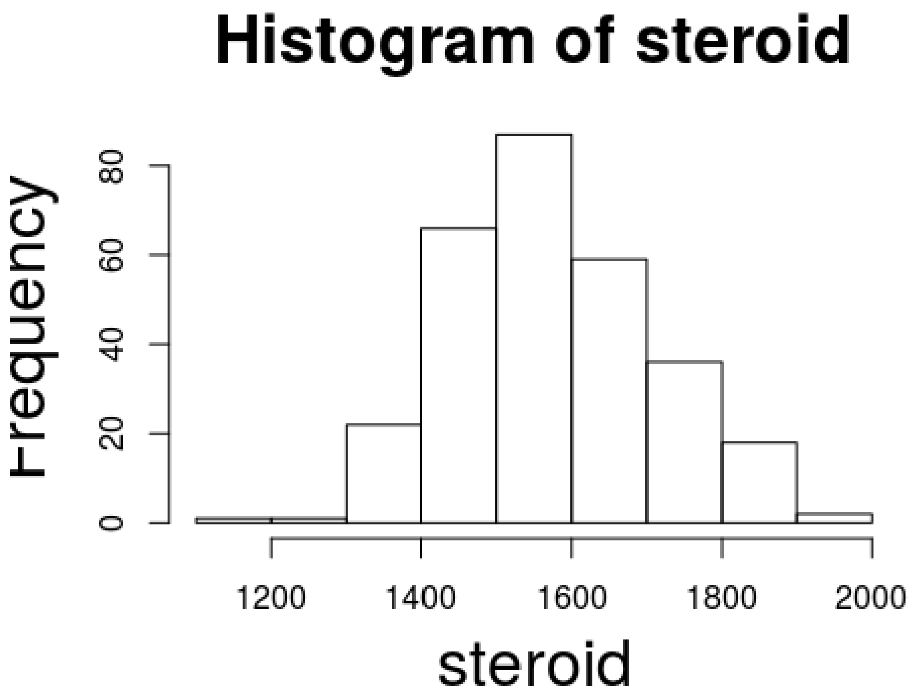
```
mean(steroid)
```

```
## [1] 1577.82
```

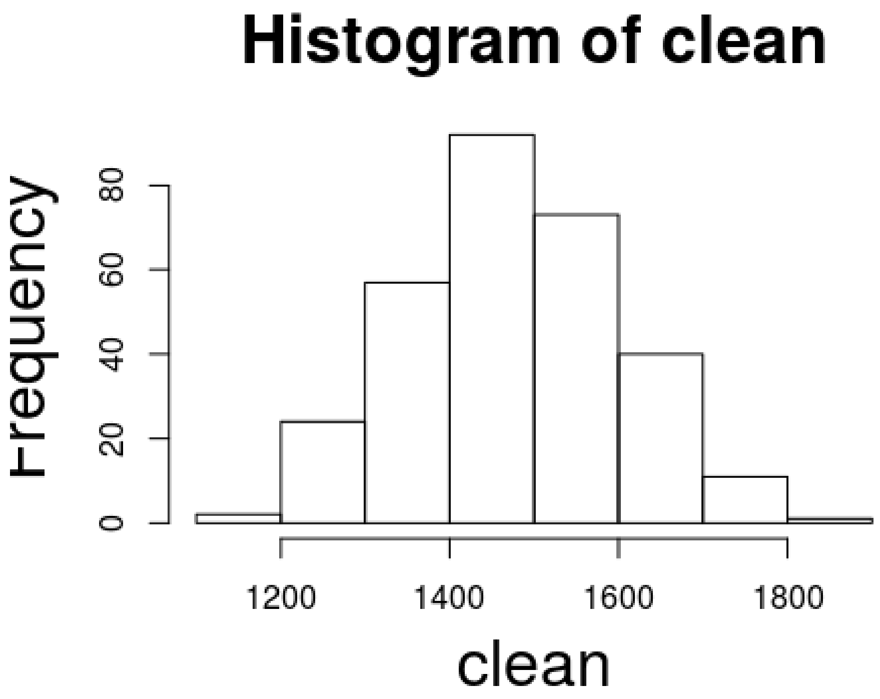
```
max(steroid)
```

```
## [1] 1934
```

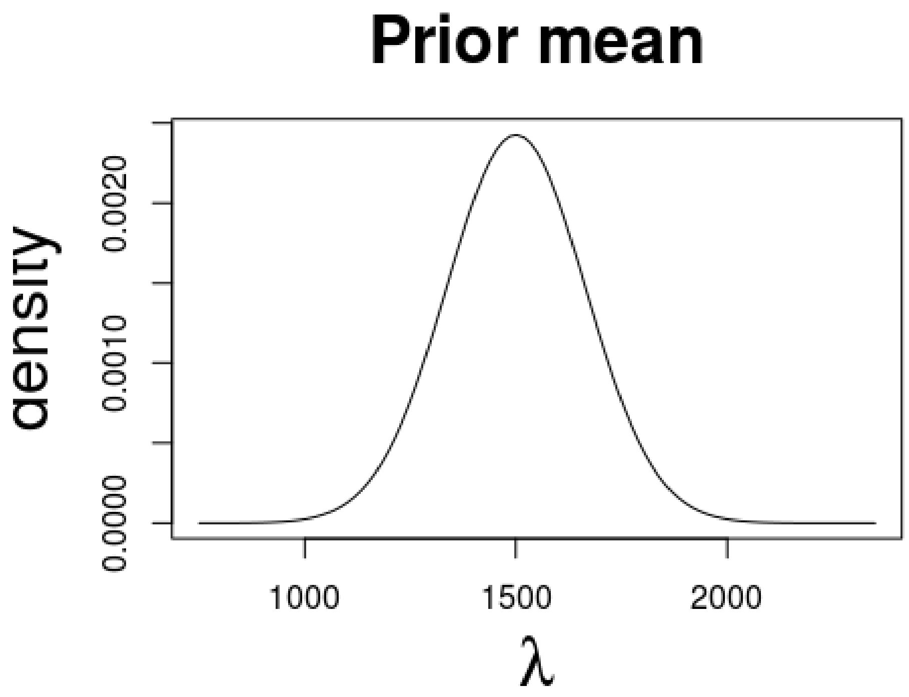
```
# Histogram
hist(cex.lab = 2, cex.main = 2, steroid)
```



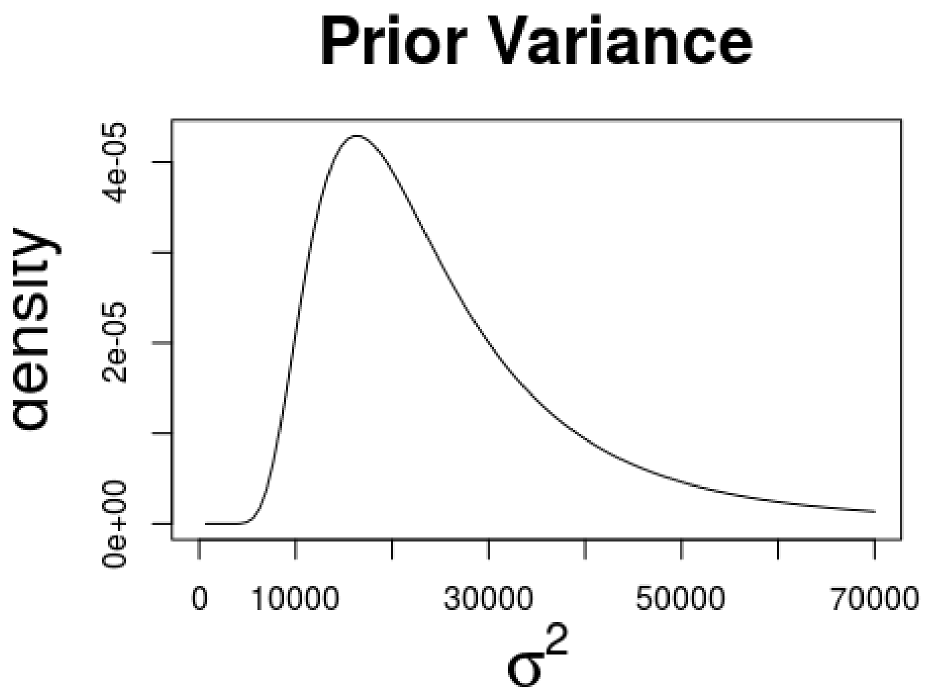
```
# Clean data
cleanyear <- c(which(MLB$Year == 2004), which(MLB$Year == 2005), which
(MLB$Year == 2006), which(MLB$Year == 2007),
               which(MLB$Year == 2008), which(MLB$Year == 2009), which
(MLB$Year == 2010), which(MLB$Year == 2011),
               which(MLB$Year == 2012), which(MLB$Year == 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)
```



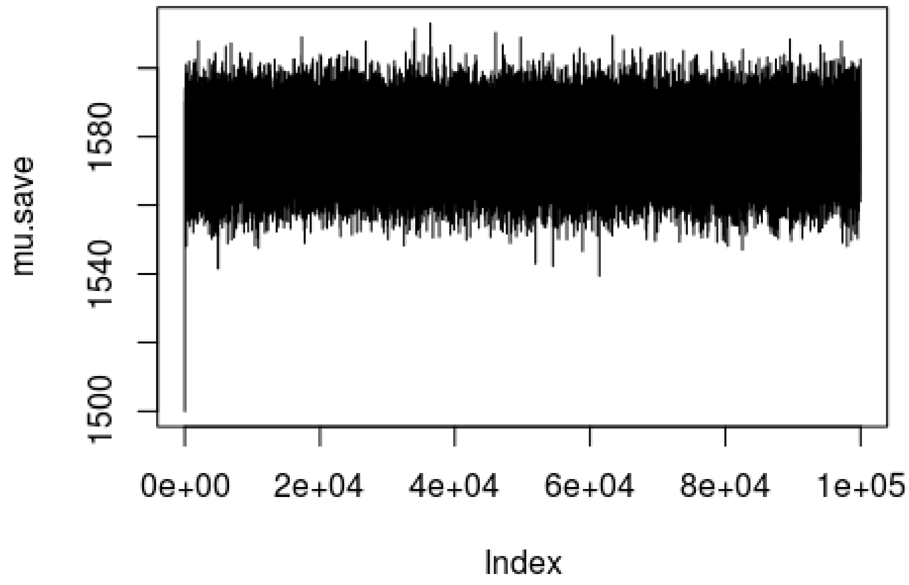
```
# Setting prior both steroid and clean
n <- length(steroid)
n2 <- length(clean)
lambda <- 1500
tau2 <- 27000
gamma <- 4
phi <- 81675

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 = "d
ensity")
```



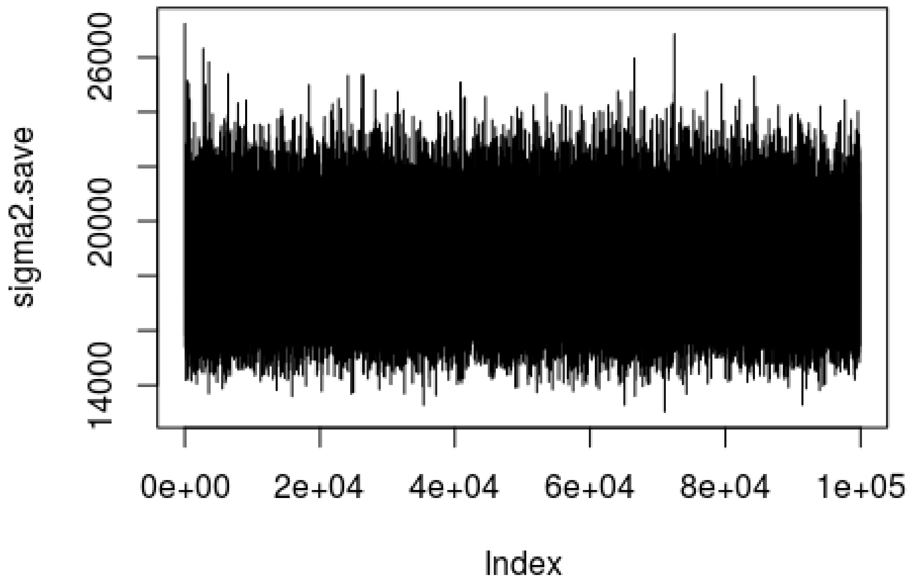
```r
curve(cex.lab = 2, cex.main = 2, dinvgamma(x,gamma,phi), xlim = c(0, 70000), xlab = expression(sigma^2), main = "Prior Variance", ylab = "density")
```



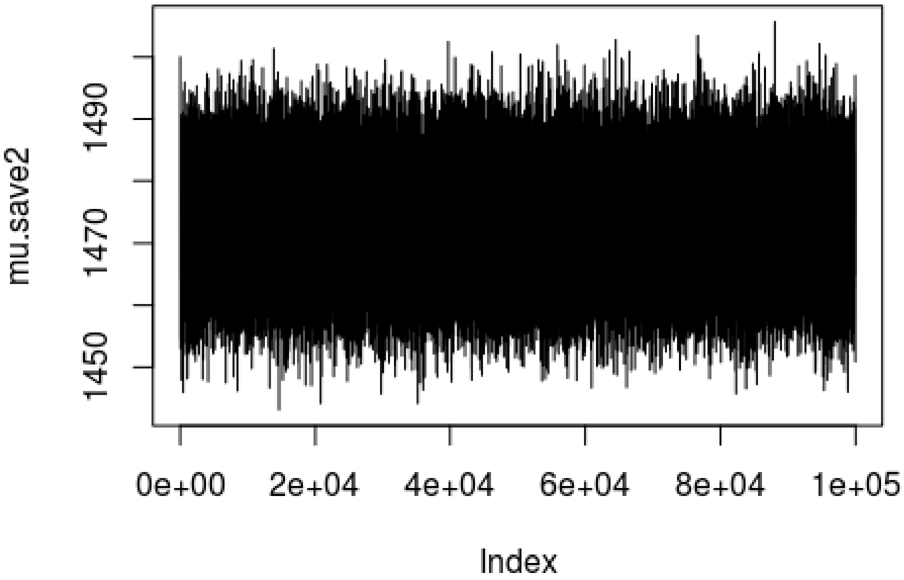
```r
# Gibbs sampling method
mu <- lambda
mu2 <- lambda
sigma2 <- phi / (gamma - 1)
sigma2.2 <- phi / (gamma - 1)

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 + si
gma2)
  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")
```



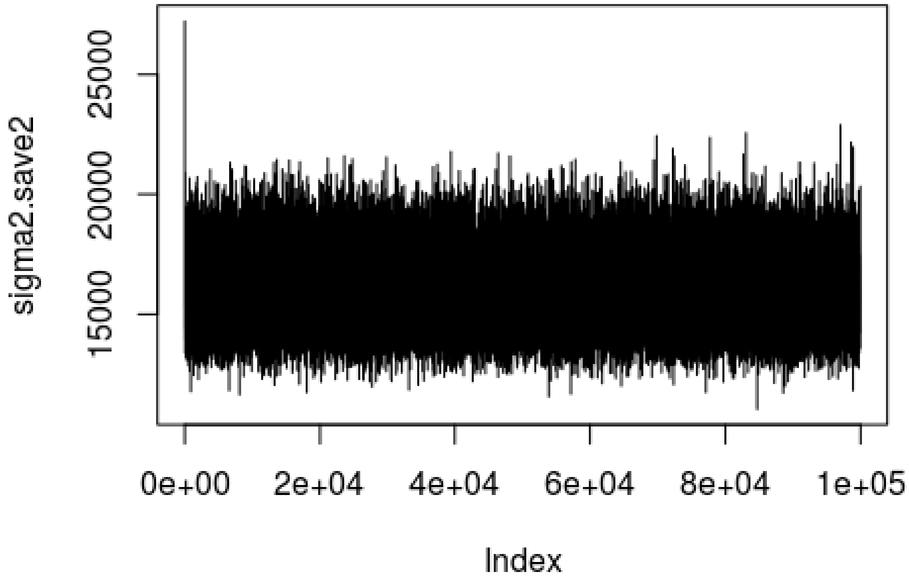
```r
plot(sigma2.save, type = "l")
```



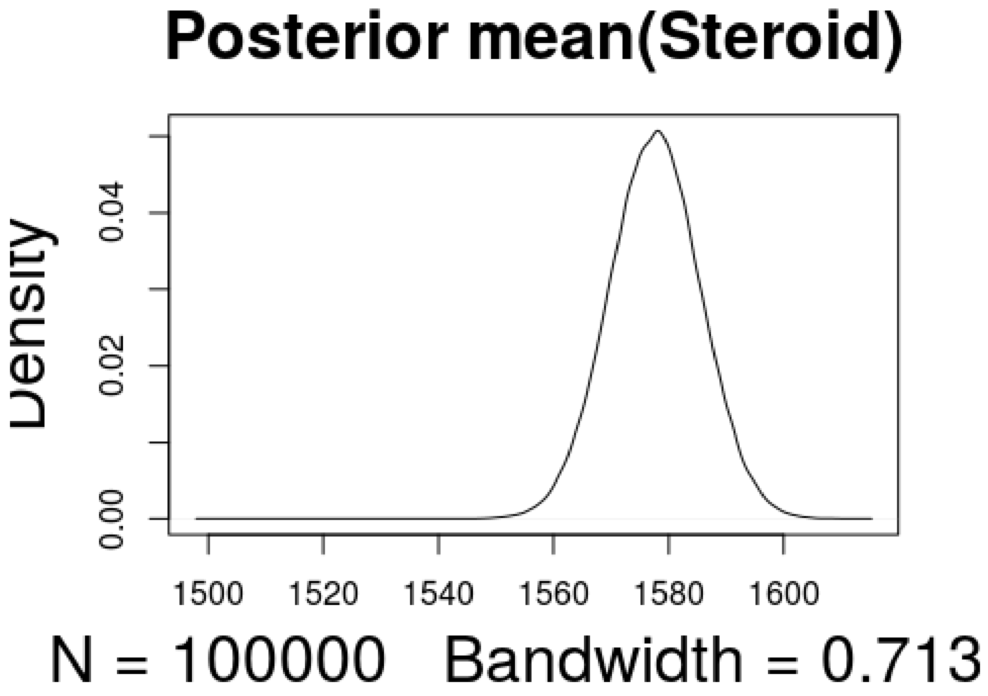
```r
# Trace plot (Clean)
plot(mu.save2, type = "l")
```



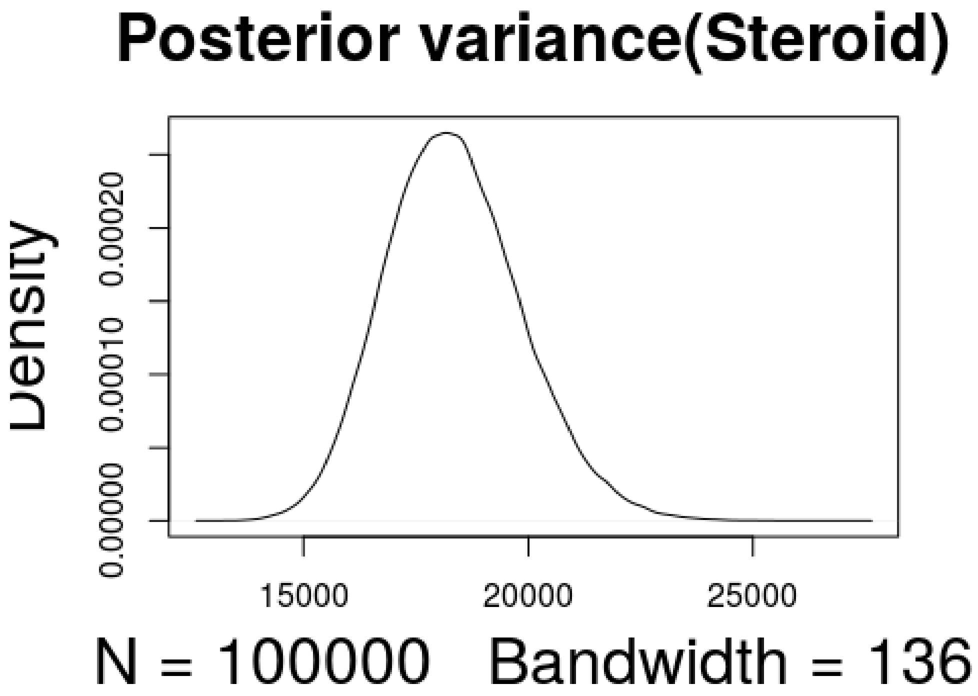
```r
plot(sigma2.save2, type = "l")
```



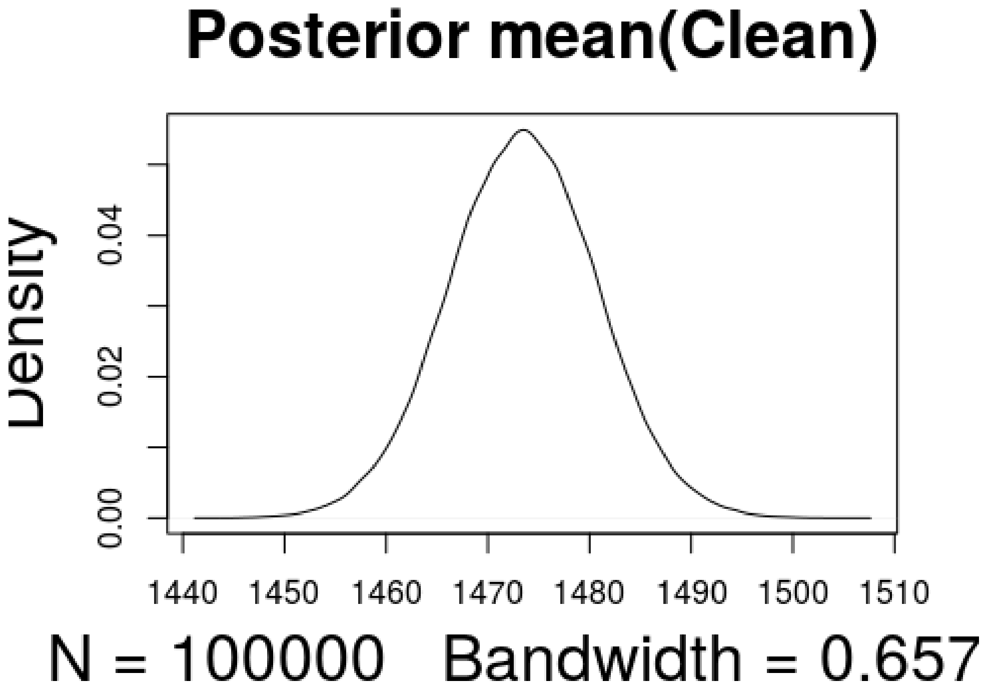
```
# Posterior Distributions for each group
plot(cex.lab = 2, cex.main = 2, density(mu.save), main = "Posterior me
an(Steroid)")
```



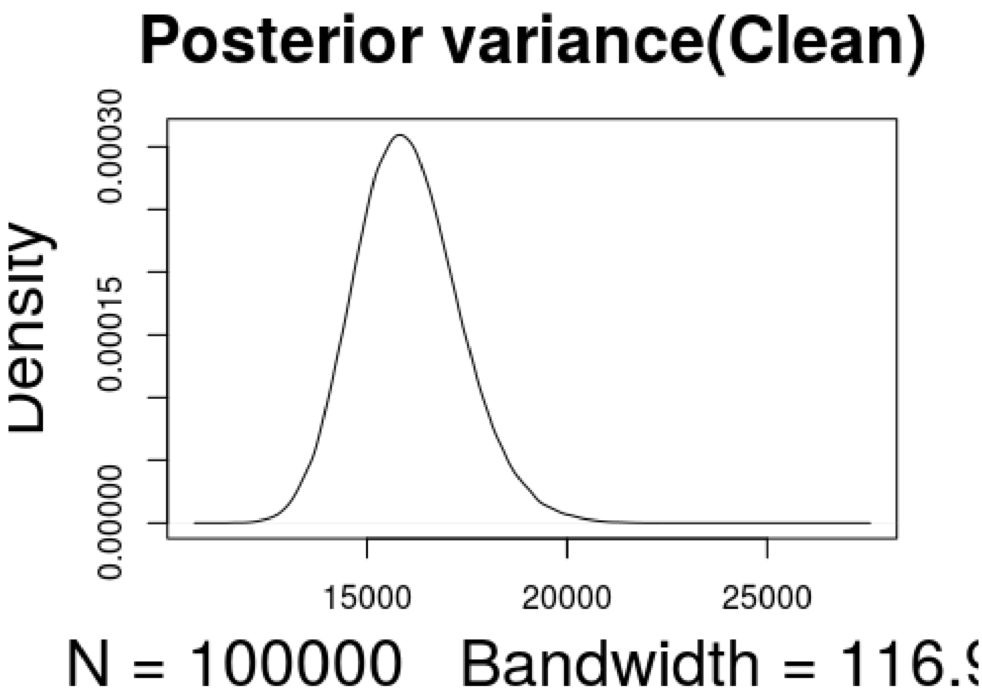
```
quantile(mu.save, probs = c(0.025, 0.975))

##     2.5%    97.5% 
## 1562.092 1593.151

mean(mu.save)

## [1] 1577.632

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



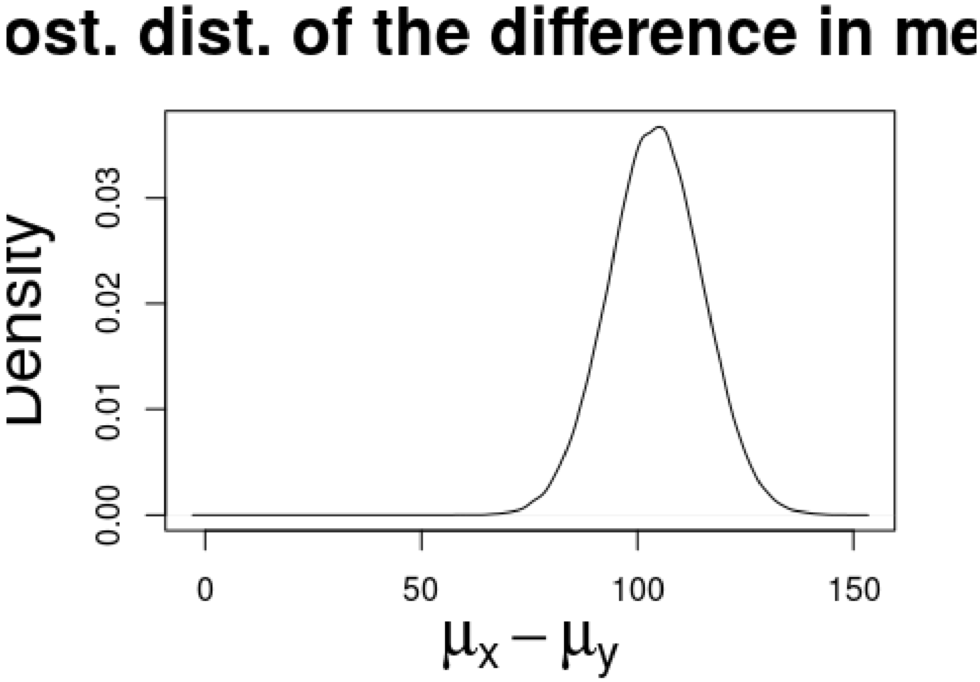
```
quantile(sigma2.save, probs = c(0.025, 0.975))

##     2.5%    97.5% 
## 15626.93 21545.20

mean(sigma2.save)

## [1] 18347.79

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% 
## 1459.063 1487.778

mean(mu.save2)

## [1] 1473.461

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



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

##     2.5%      50%    97.5%
## 13662.39 15951.86 18782.34

mean(sigma2.save2)

## [1] 16020.52

# 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]), dens
ity(diff), main = "Post. dist. of the difference in means")
```



ost. dist. of the difference in me

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

##      2.5%       50%     97.5%
##  82.98207 104.17774 125.28865

mean(diff)

## [1] 104.1709

var(diff)

## [1] 116.6596
```