Stat 251 Course Project

Tyler Savage, Sofia Scribner

Introduction

Are the Major League Baseball players in the National League or the American League better at batting on average? There are many statistics that can be used to measure how good a baseball player or team is. For our project, we decided to look into the difference in batting average (AVG) between the two leagues in the MLB. For a long time, there was a difference in the rules between the two leagues. The American League had the "Designated Hitter" rule, where pitchers would have another player take their place in the batting order and bat for them. In the National League, pitchers would bat themselves. We are interested in looking into whether there was a significant difference between the leagues in batting performance due to this difference.

In order to do this we will need to approximate the mean and and standard deviation of the distributions of AVG for each league.

Method

Our data was obtained through kaggle, which obtained it's data from Sean Lahman's website: http://seanlahman.com . The baseball data itself is public information gathered from MLB games played.

For our likelihood, we decided to use a normal distribution, modeled as: $y_i \sim N(\mu, \sigma^2)$

This distribution is appropriate because batting average will typically be centered around the league average and symmetrical. Though there is a lower limit (you cannot bat less than 0), if we remove outliers and extremes it fits the distribution much better.

For our prior, we decided to use a normal distribution for the mean, modeled as: $\mu \sim N(0.25, 0.05^2)$

For the variance, we modeled an inverse gamma: $\sigma^2 \sim \text{Inv-Gamma}(2, 0.0025)$

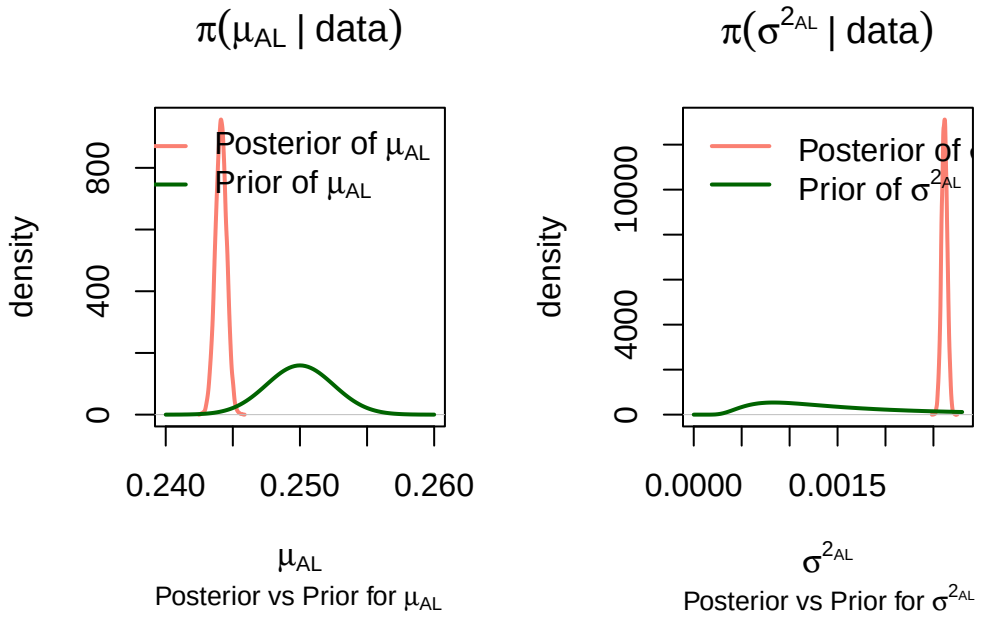
The normal distribution was chosen for the mean because it is conjugate to the normal likelihood, which means the posterior distribution for will also be normal and making calculations

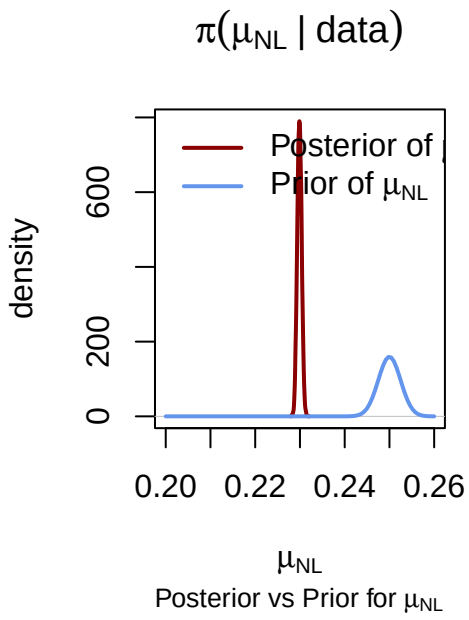
simpler. The inverse gamma was chosen for the variance because it is also conjugate to the normal likelihood, leading to an normal posterior for ^{2}. Additionally it is a flexible distribution for modeling variances.

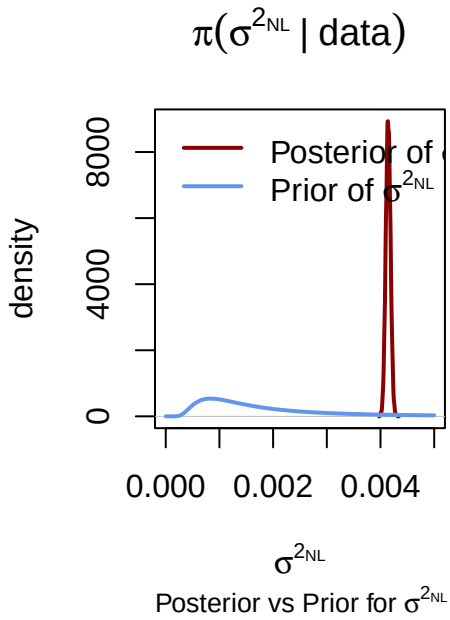
We chose our starting parameters by using the historical league average of 0.26 and for the variance we used a value that created an uninformative prior that was visually similar to the distribution of the actual data.

Table 1: Summary Statistics by League

lgID	Mean	SD	Min	Max	Range	Count
AL	0.2441129	0.0511376	0	0.5121951	0.5121951	15152
NL	0.2298922	0.0643792	0	0.5000000	0.5000000	17521







Results

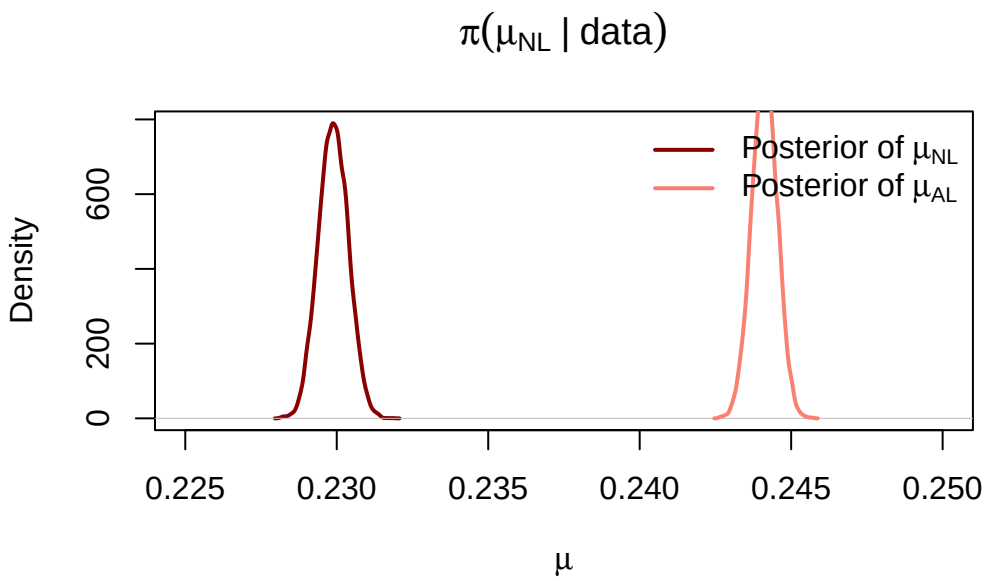
The full conditional distributions:

$\mu_{NL}|data, \sigma^2_{NL} \sim N(\frac{\sum_{i=1}^{n_{NL}} y_i}{n_{NL}}, \frac{\sigma^2_{NL}}{n_{NL}})$

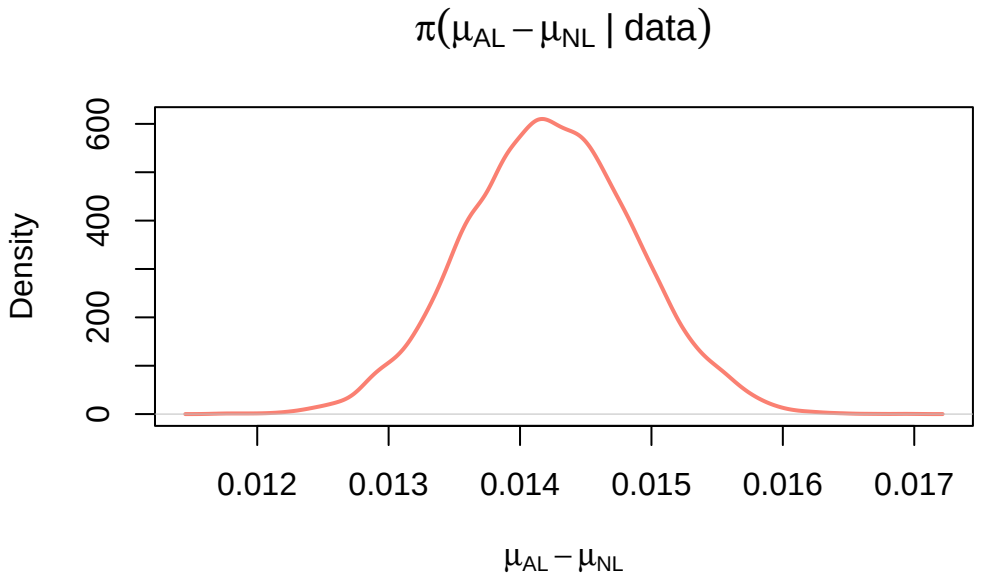
$\mu_{AL}|data, \sigma^2_{AL} \sim N(\frac{\sum_{i=1}^{n_{AL}} y_i}{n_{AL}}, \frac{\sigma^2_{AL}}{n_{AL}})$

$\sigma^2_{NL}|\mu_{NL}, data \sim \text{Inv-Gamma}(\alpha + \frac{n_{NL}}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n_{NL}}(y_i - \mu_{NL})^2)$

$\sigma^2_{AL}|\mu_{AL}, data \sim \text{Inv-Gamma}(\alpha + \frac{n_{AL}}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n_{AL}}(y_i - \mu_{AL})^2)$



Comparing posterior of μ between leagues



Posterior for $\mu_{AL} - \mu_{NL}$

Table 2: 95% Credible Intervals

Parameter	Lower	Upper
mu[NL]	0.2290	0.2308
mu[AL]	0.2433	0.2449
sigma^2[NL]	0.0041	0.0042
sigma^2[AL]	0.0026	0.0027
Difference in mu[AL]-mu[NL]	0.0130	0.0155

There is a 95% probability that the true mean for the National League is between 0.2290 and 0.2309. There is a 95% probability that the true mean for the American League is between 0.2433 and 0.2449. There is a 95% probability that the true standard deviation for the National League is between 0.0041 and 0.0042. There is a 95% probability that the true standard deviation for the American League is between 0.0026 and 0.0027. There is a 95% probability that the true difference in means for the leagues is between 0.0130 and 0.0155.

Table 3: Posterior Means and Variances

Parameter	Mean	Variance
NL	0.2299	2.410e-07
AL	0.2441	1.739e-07

Conclusion

After our analysis, the data has shown us how the mean is different between each league. Using the charts and probability intervals, we can see that there is no overlap between the graphs and intervals, showing statistical significance in their difference.

Given that we started with the same prior for both leagues, it is interesting to note how the data for each league changed the distribution enough to create no overlap and a significant amount of concentration. We believe this is due to the amount of data, which informs our analysis and weakens the power of the prior.

Further exploration could be done using additional features like hits, runs, and homeruns to get a better idea of batter performance, allowing us to enhance our knowledge with more detail. Additionally, we could perform the same analysis on pitcher performance to see if the leagues have a difference in that regard. One limit to our analysis is that if there was in fact a difference between the leagues in pitcher performance, that would impact batter performance as well, which means the difference in batter performance could be the result of a difference in pitcher skill rather than batter skill.

Appendix

```
# Initializing chunk

set.seed(54321)

bat <- read.csv("Batting.csv", header=T)

summary(bat)


cleaned_bat <- na.omit(bat)

cleaned_bat$lgID <- as.factor(cleaned_bat$lgID)

cleaned_bat <- subset(cleaned_bat, AB >= 30)
cleaned_bat$AVG <- cleaned_bat$H/cleaned_bat$AB

hist(cleaned_bat$AVG)


# chunk for AL Gibbs

# Filter the data for players in the American League
al_data <- subset(cleaned_bat, lgID == "AL")

# Create a vector of batting averages for the AL league
al_avg <- al_data$AVG

al_n <- length(al_avg)

# Set base parameters
al_lambda <- 0.25
al_tau2 <- 0.05^2

al_gamma <- 2
al_phi <- 0.0025

al_mu <- .2
al_sigma2 <-.003

iters <- 10000
```

```r
al_mu_save <- rep(0, iters)
al_mu_save[1] <- al_mu
al_sigma2_save <- rep(0, iters)
al_sigma2_save[1] <- al_sigma2

for(t in 2:iters){

al_lambda_p <- (al_tau2*sum(al_avg) +
                  al_sigma2*al_lambda)/(al_tau2*al_n + al_sigma2)
  al_tau2_p <- al_sigma2*al_tau2/(al_tau2*al_n + al_sigma2)

  al_mu <- rnorm(1, al_lambda_p, sqrt(al_tau2_p))

  al_mu_save[t] <- al_mu



  al_gamma_p <- al_gamma + al_n/2
  al_phi_p <- al_phi + sum((al_avg - al_mu)^2 )/2

  al_sigma2 <- rinvgamma(1, al_gamma_p, al_phi_p)

  al_sigma2_save[t] <- al_sigma2

}

# Generate plots
burn <- 100
al_mu_use <- al_mu_save[-(1:burn)]
al_sigma2_use <- al_sigma2_save[-(1:burn)]


# chunk for NL Gibbs

# Filter the data for players in the National League
nl_data <- subset(cleaned_bat, lgID == "NL")

# Create a vector of batting averages for the NL league
nl_avg <- nl_data$AVG

nl_n <- length(nl_avg)
```

```r
# Set base parameters
nl_lambda <- 0.25
nl_tau2 <- 0.05^2

nl_gamma <- 2
nl_phi <- 0.0025

nl_mu <- .2
nl_sigma2 <-.003

iters <- 10000

nl_mu_save <- rep(0, iters)
nl_mu_save[1] <- nl_mu
nl_sigma2_save <- rep(0, iters)
nl_sigma2_save[1] <- nl_sigma2

for(t in 2:iters){

nl_lambda_p <- (nl_tau2*sum(nl_avg) +
                  nl_sigma2*nl_lambda)/(nl_tau2*nl_n + nl_sigma2)
  nl_tau2_p <- nl_sigma2*nl_tau2/(nl_tau2*nl_n + nl_sigma2)

  nl_mu <- rnorm(1, nl_lambda_p, sqrt(nl_tau2_p))

  nl_mu_save[t] <- nl_mu



  nl_gamma_p <- nl_gamma + nl_n/2
  nl_phi_p <- nl_phi + sum((nl_avg - nl_mu)^2 )/2

  nl_sigma2 <- rinvgamma(1, nl_gamma_p, nl_phi_p)

  nl_sigma2_save[t] <- nl_sigma2

}

# Generate plots
burn <- 100
nl_mu_use <- nl_mu_save[-(1:burn)]
nl_sigma2_use <- nl_sigma2_save[-(1:burn)]
```

```r
# Plots for AL prior and posterior

par(mfrow=c(1,2))
plot(density(al_mu_use), xlab=expression(mu[AL]), ylab="density",
     main=expression(pi(mu[AL]~"|"~data)), xlim = c(0.24, 0.26),
     col = 'salmon', lwd = 2)
curve(dnorm(x, .25, .05^2), add =T, col = 'darkgreen', lwd = 2)
legend("topright",
       legend = c(expression(paste("Posterior of ", mu[AL])),
                  expression(paste("Prior of ", mu[AL]))),
       col = c("salmon", "darkgreen"),
       lwd = 2,
       bty = "n")
text(expression(paste("Posterior vs Prior for ", mu[AL])),
      side = 1, # Place the caption on the bottom (side 1)
      line = 4, # Adjust the vertical position as needed
      adj = 0,  # Left align the caption
      cex = 0.8) # Adjust the font size as needed


plot(density(al_sigma2_use), xlab=expression(sigma^2[AL]),
     ylab="density",
     main=expression(pi(sigma^2[AL]~"|"~data)), xlim = c(0, .0028),
     col = 'salmon', lwd = 2)
curve(dinvgamma(x, 2, .0025), add =T, col = 'darkgreen', lwd = 2)
legend("topleft",
       legend = c(expression(paste("Posterior of ", sigma^2[AL])),
                  expression(paste("Prior of ", sigma^2[AL]))),
       col = c("salmon", "darkgreen"),
       lwd = 2,
       bty = "n")
text(expression(paste("Posterior vs Prior for ", sigma^2[AL])),
      side = 1, # Place the caption on the bottom (side 1)
      line = 4, # Adjust the vertical position as needed
      adj = 0,  # Left align the caption
      cex = 0.8) # Adjust the font size as needed


# Plots for NL prior and posterior

par(mfrow=c(1,2))
```

```r
plot(density(nl_mu_use), xlab=expression(mu[NL]), ylab="density",
     main=expression(pi(mu[NL]~"|"~data)), xlim = c(0.2, 0.26),
     col = 'darkred', lwd = 2)
curve(dnorm(x, .25, .05^2), add =T, col = 'cornflowerblue', lwd = 2)
legend("topleft",
       legend = c(expression(paste("Posterior of ", mu[NL])),
                  expression(paste("Prior of ", mu[NL]))),
       col = c("darkred", "cornflowerblue"),
       lwd = 2,
       bty = "n")
text(expression(paste("Posterior vs Prior for ", mu[NL])),
      side = 1, # Place the caption on the bottom (side 1)
      line = 4, # Adjust the vertical position as needed
      adj = 0,  # Left align the caption
      cex = 0.8) # Adjust the font size as needed


plot(density(nl_sigma2_use), xlab=expression(sigma^2[NL]), ylab="density",
     main=expression(pi(sigma^2[NL]~"|"~data)), xlim = c(0, .005),
     col = 'darkred', lwd = 2)
curve(dinvgamma(x, 2, .0025), add =T, col = 'cornflowerblue', lwd = 2)
legend("topleft",
       legend = c(expression(paste("Posterior of ", sigma^2[NL])),
                  expression(paste("Prior of ", sigma^2[NL]))),
       col = c("darkred", "cornflowerblue"),
       lwd = 2,
       bty = "n")
mtext(expression(paste("Posterior vs Prior for ", sigma^2[NL])),
      side = 1, # Place the caption on the bottom (side 1)
      line = 4, # Adjust the vertical position as needed
      adj = 0,  # Left align the caption
      cex = 0.8) # Adjust the font size as needed

# Plots for difference between AL and NL posteriors

plot(density(nl_mu_use),
     xlab=expression(mu[NL]),
     ylab="Density",
     main=expression(pi(mu[NL]~"|"~data)),
     xlim = c(0.225, 0.25),
     col = 'darkred',
     lwd = 2)
```

```r
lines(density(al_mu_use),
      col = 'salmon',
      lwd = 2)

legend("topright",
       legend = c(expression(paste("Posterior of ", mu[NL])),
                  expression(paste("Posterior of ", mu[AL]))),
       col = c("darkred", "salmon"),
       lwd = 2,
       bty = "n")
mtext(expression(paste("Comparing posterior of ", mu, " between leagues")),
      side = 1, # Place the caption on the bottom (side 1)
      line = 4, # Adjust the vertical position as needed
      adj = 0,  # Left align the caption
      cex = 0.8) # Adjust the font size as needed

mu_dif<- al_mu_use-nl_mu_use

plot(density(mu_dif),
     xlab=expression(mu[AL]-mu[NL]),
     ylab="Density",
     main=expression(pi(mu[AL]-mu[NL]~"|"~data)),
     col = 'salmon',
     lwd = 2)
mtext(expression(paste("Posterior for ", mu[AL]-mu[NL])),
      side = 1, # Place the caption on the bottom (side 1)
      line = 4, # Adjust the vertical position as needed
      adj = 0,  # Left align the caption
      cex = 0.8) # Adjust the font size as needed


# Table of intervals

mu_dif_conf<- quantile(mu_dif, c(.0275, .975))
al_mu_conf<- quantile(al_mu_use, c(.0275, .975))
nl_mu_conf<- quantile(nl_mu_use, c(.0275, .975))
al_var_conf<- quantile(al_sigma2_use, c(.0275, .975))
nl_var_conf<- quantile(nl_sigma2_use, c(.0275, .975))

ci_table <- data.frame(
  Parameter = c("mu[NL]", "mu[AL]", "sigma^2[NL]", "sigma^2[AL]",
                "Difference in mu[AL]-mu[NL]"),
```

```
  Lower = c(nl_mu_conf[1], al_mu_conf[1], nl_var_conf[1],
            al_var_conf[1], mu_dif_conf[1]),
  Upper = c(nl_mu_conf[2], al_mu_conf[2], nl_var_conf[2],
            al_var_conf[2], mu_dif_conf[2])
)


kable(ci_table, caption = "95% Credible Intervals", digits = 3)


# Posterior means and variance table

post_table <- data.frame(
    Parameter = c("NL", "AL"),
  Mean = c(mean(nl_mu_use), mean(al_mu_use)),
  Variance = c(var(nl_mu_use), var(al_mu_use))
)

kable(post_table, caption = "Posterior Means and Variances", digits = c(10, 4))
```