Scoring Runs in College Basketball

Introduction:

The research question I aim to answer is: Are college basketball teams in the power 6 conferences better at going on scoring runs when compared with the remaining teams in college basketball? The value in answering this question is that it could lead to a deeper dive into why one group or the other is better at going on scoring runs. It could be a talent gap in players, or it could have to do with coaching, play style, opponent strength, etc. It could also be really valuable to do another study on scoring runs allowed, and then compare that with this research to see if a team is really streaky, good at defending against runs, or very susceptible to runs.

Methods:

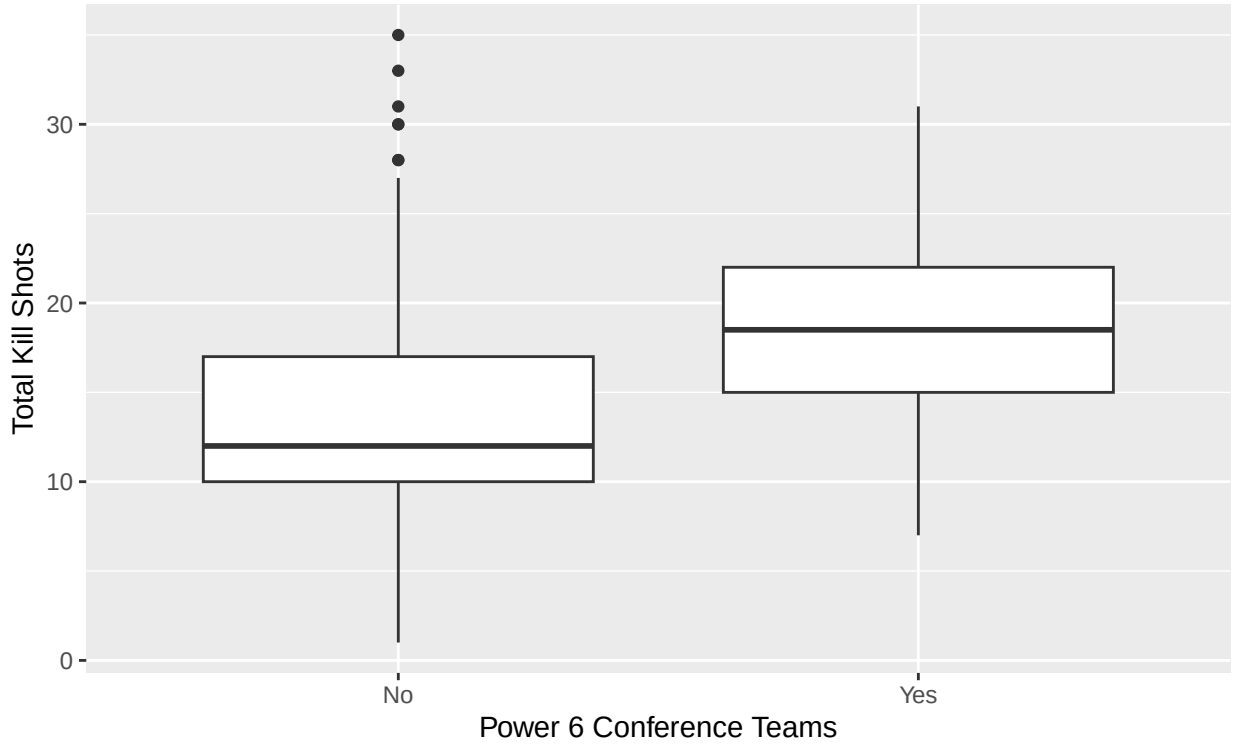
I'm using data from EvanMiya.com (it's a college basketball advanced statistics website based on Bayesian statistics) for this study. The statistic that I will specifically be using is total kill shots. As defined by the website, total kill shots is the total number of double digit scoring runs produced in a season. I will be using the data from the most recent season (2022-23), and it includes all division 1 basketball teams. I am using a gamma-poisson data model to analyze the data. I chose this because the data is a count, which fits many of the assumptions made for the Poisson distribution. It seems reasonable to assume that if one team is on a scoring run, it will be independent of another team's chance in a different game to go on a scoring run. Also, it would be impossible for a team to go on two scoring runs at the same time, meaning that only one occurrence can happen in an interval of time.

The parameter of interest in this case is the rate parameter or λ. This represents the expected number of kill shots per team per season, based on conference. Understanding more about this rate parameter will allow for an easier comparison of power 6 and non-power six conference. It will help in making inferences about the two groups, such as determining the likelihood that one group is better at scoring runs than the other. The prior I will use for the power 6 conferences is a Gamma(95,5). I chose this because of experience watching college basketball and the skill level of power 6 conference teams. This prior leads to an expected value of 19 kill shots per season per team, or about two kill shots in every three games. The prior I will use for the non-power 6 conferences is a Gamma(70,5). I chose this because of experience watching college basketball and the lower skill level of most non-power 6 conference teams. This prior leads to an expected value of 14 kill shots per season per team, or about one kill shot in every other game.

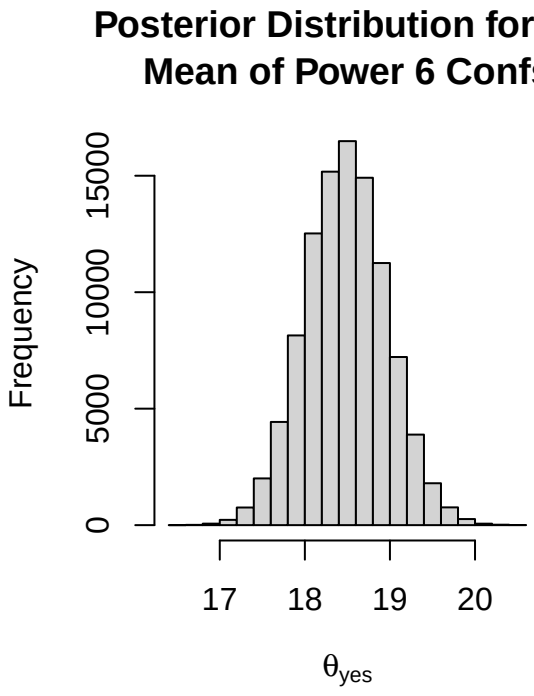
Results:

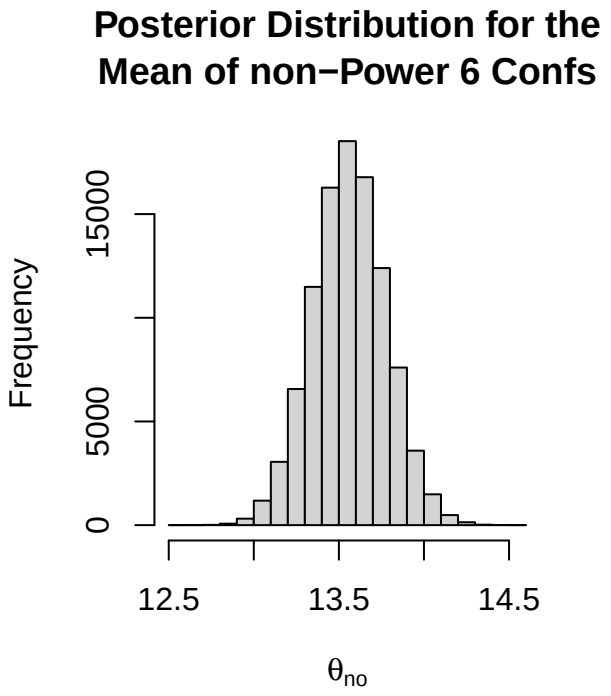
As a brief summary of the kill shot data for the power 6 conference teams: the mean is 18.4473684, the variance is 29.9838596, and the number of teams is 76.

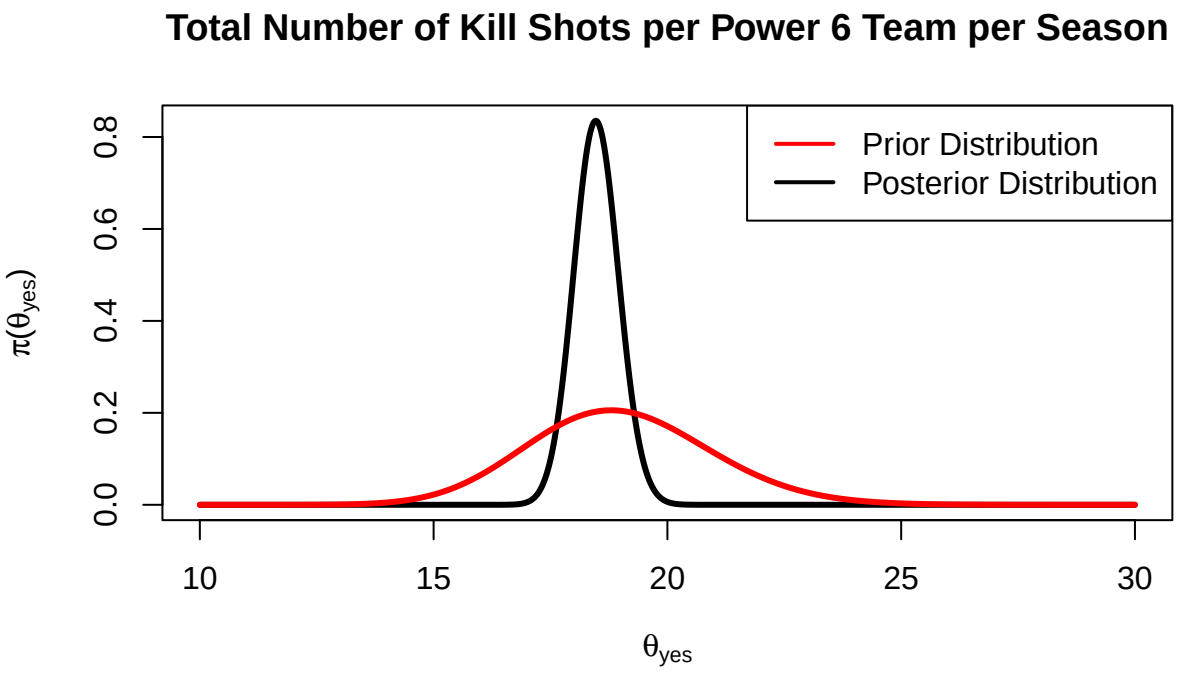
As a brief summary of the kill shot data for the non-power 6 conference teams: the mean is 13.554007, the variance is 32.8213737, and the number of teams is 287.

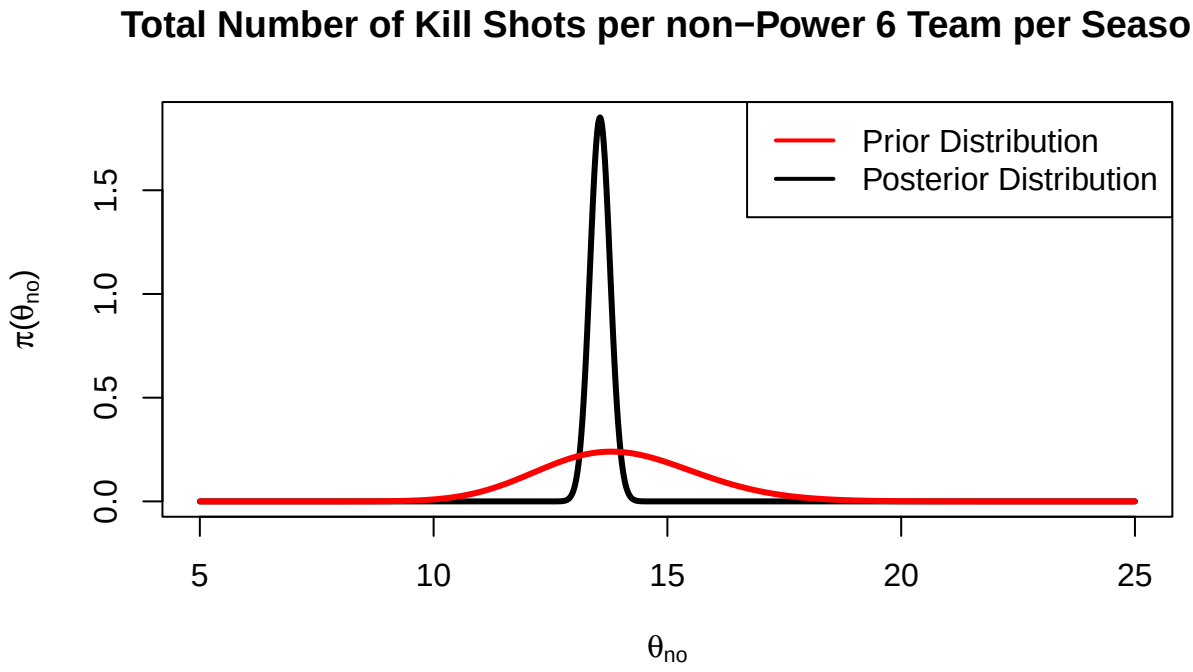


These box plots show a good amount of overlap between the two groups, while also showing that the power 6 teams tend to have higher numbers of kill shots. It is also interesting to note that the range and variability for the non-power 6 teams are higher than the range and variance for the power 6 teams.







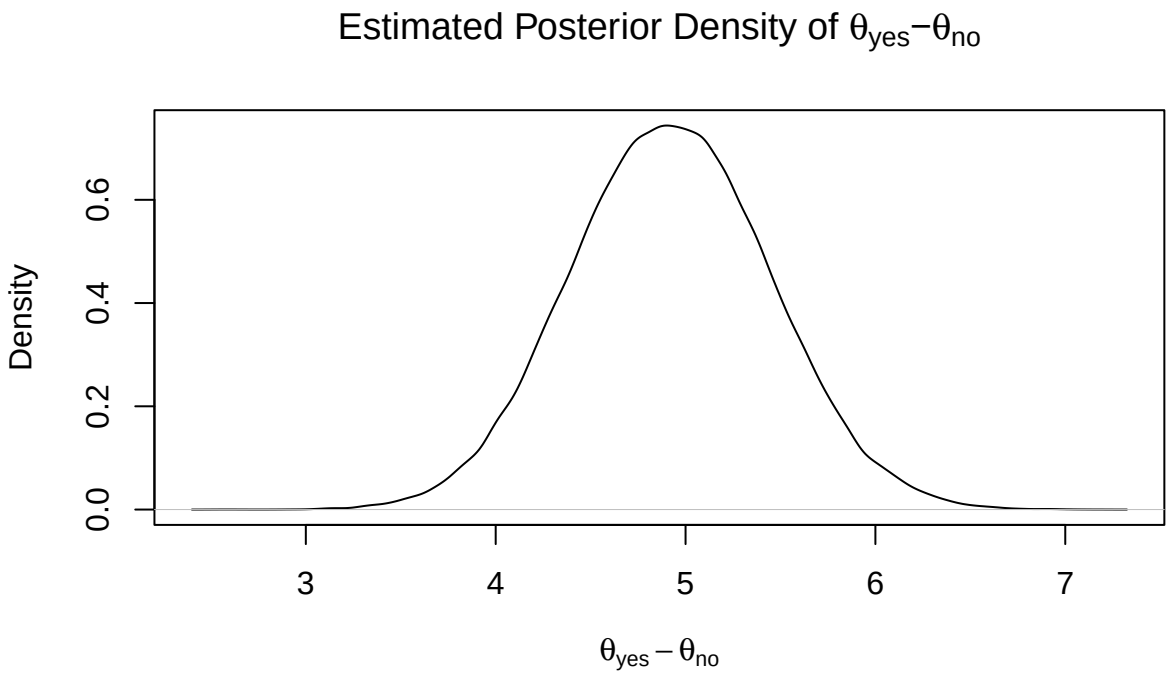


The posterior probability that the average total number of kill shots per power 6 team per season is between 17.6 and 19.4 is 95%.

The posterior probability that the average total number of kill shots per non-power 6 team per season is between 13.1 and 14 is 95%.

I ran a difference of means test on the rate of kill shots for power 6 teams versus the rate of kill shots for non-power 6 teams and found the probability to be equal to one. This means that the rate of kill shots for power 6 conference teams is higher than the rate for non-power 6 conference teams, or in other words, college basketball teams in the power 6 conferences are better at going on scoring runs when compared with the remaining teams in division 1.

The posterior probability that the average total number of kill shots per power 6 team per season is between 3.91 and 5.95 kill shots higher than the average total number of kill shots per non-power 6 team per season is 95%.



Discussion:

In summary, college basketball teams in a power 6 conference more often go on double digit scoring runs when compared with the teams outside of power 6 conferences. Teams in the power 6 conferences tend to be the better teams in college basketball, so I'm not too surprised to see that they are also better at going on scoring runs, on average. However, the shortcomings of this analysis are that the data only reflects numbers from this season and there isn't much understanding beyond player talent as to why the power 6 conferences go on more scoring runs. A very useful follow-up analysis to this one could be focused on the individual differences between teams that go on a lot of scoring runs and teams that go on few scoring runs. Understanding the intangibles and other statistics that explain more about scoring runs could help the top teams continue to perfect what they already do well and help the lower teams adapt new strategies that will lead to more scoring runs.

Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library(readxl)
library(tidyverse)

# Reading in the data
kill <- read_excel("Kill Shot Data.xlsx")
# The data means
mu.yes <- mean(kill$TKS[kill$P6=="Yes"])
mu.no <- mean(kill$TKS[kill$P6=="No"])

# The data n's
n.yes <- length(kill$TKS[kill$P6=="Yes"])
n.no <- length(kill$TKS[kill$P6=="No"])

# The data variances
var.yes <- var(kill$TKS[kill$P6=="Yes"])
var.no <- var(kill$TKS[kill$P6=="No"])
# Side-by-side box plots
ggplot(data = kill, mapping = aes(x = P6, y = TKS)) +
  geom_boxplot() +
  labs(y = "Total Kill Shots",
       x = "Power 6 Conference Teams")
# Prior for power 6
a.yes <- 95
b.yes <- 5

# Prior for non power 6
a.no <- 70
b.no <- 5

# Posterior for power 6
astar.yes <- a.yes + sum(kill$TKS[kill$P6=="Yes"])
bstar.yes <- b.yes + length(kill$TKS[kill$P6=="Yes"])

# Posterior for non power 6
astar.no <- a.no + sum(kill$TKS[kill$P6=="No"])
bstar.no <- b.no + length(kill$TKS[kill$P6=="No"])

# Posterior distributions for theta before and theta after
theta.yes <- rgamma(100000, shape=astar.yes, rate=bstar.yes)
theta.no <- rgamma(100000, shape=astar.no, rate=bstar.no)

par(mfrow=c(1,2))
# Plotting the posterior distribution for power 6
hist(theta.yes, xlab=expression(theta[yes]),
     main="Posterior Distribution for the\nMean of Power 6 Confs")
# Plotting the posterior distribution for non power 6
hist(theta.no, xlab=expression(theta[no]),
     main="Posterior Distribution for the\nMean of non-Power 6 Confs")
thetas <- seq(10, 30, length=1001)
plot(thetas, dgamma(thetas, astar.yes, bstar.yes), col="black",
```

```
   lwd=3, main="Total Number of Kill Shots per Power 6 Team per Season",
   xlab=expression(theta[yes]), type="l",
   ylab=expression(paste(pi, "(", theta[yes], ")", sep="")))
lines(thetas, dgamma(thetas, a.yes, b.yes), col="red",
      lwd=3)
legend("topright", c("Prior Distribution", "Posterior Distribution"), lwd=2, col=c("red", "black"))
thetas <- seq(5, 25, length=1001)
plot(thetas, dgamma(thetas, astar.no, bstar.no), col="black",
   lwd=3, main="Total Number of Kill Shots per non-Power 6 Team per Season",
   xlab=expression(theta[no]), type="l",
   ylab=expression(paste(pi, "(", theta[no], ")", sep="")))
lines(thetas, dgamma(thetas, a.no, b.no), col="red",
      lwd=3)
legend("topright", c("Prior Distribution", "Posterior Distribution"), lwd=2, col=c("red", "black"))
# Credible intervals for both groups
cred.yes <- quantile(theta.yes, c(0.025, 0.975))
cred.no <- quantile(theta.no, c(0.025, 0.975))
# Probability that the (population) rate of kill shots for power 6 teams
# is higher than the rate for non-power 6 teams
prob <- mean(theta.yes > theta.no)
# Credible interval for the difference
cred <- quantile(theta.yes - theta.no, c(0.025, 0.975))
# Graphic comparing the means
diff <- theta.yes - theta.no
plot(density(diff), xlab = expression(theta[yes] - theta[no]),
     main = expression(paste("Estimated Posterior Density of ",
                             theta[yes], "-", theta[no]), sep=""))
```