Penalty Shootout Probabilities

Here, I calculate empirical frequencies of goals converted during penalty shootouts to estimate conditional probabilities of success for both scoring and winning the match.

```
# dependencies
library(dplyr)
library(tidyr) # to recast data
library(purrr) # for mapping to dataframes
library(forcats) # for manipulating factors
library(ggplot2) # for plotting
library(ggjoy) # for density plots
library(ggExtra) # for marginal histograms
library(scales) # for plotting opacity
```

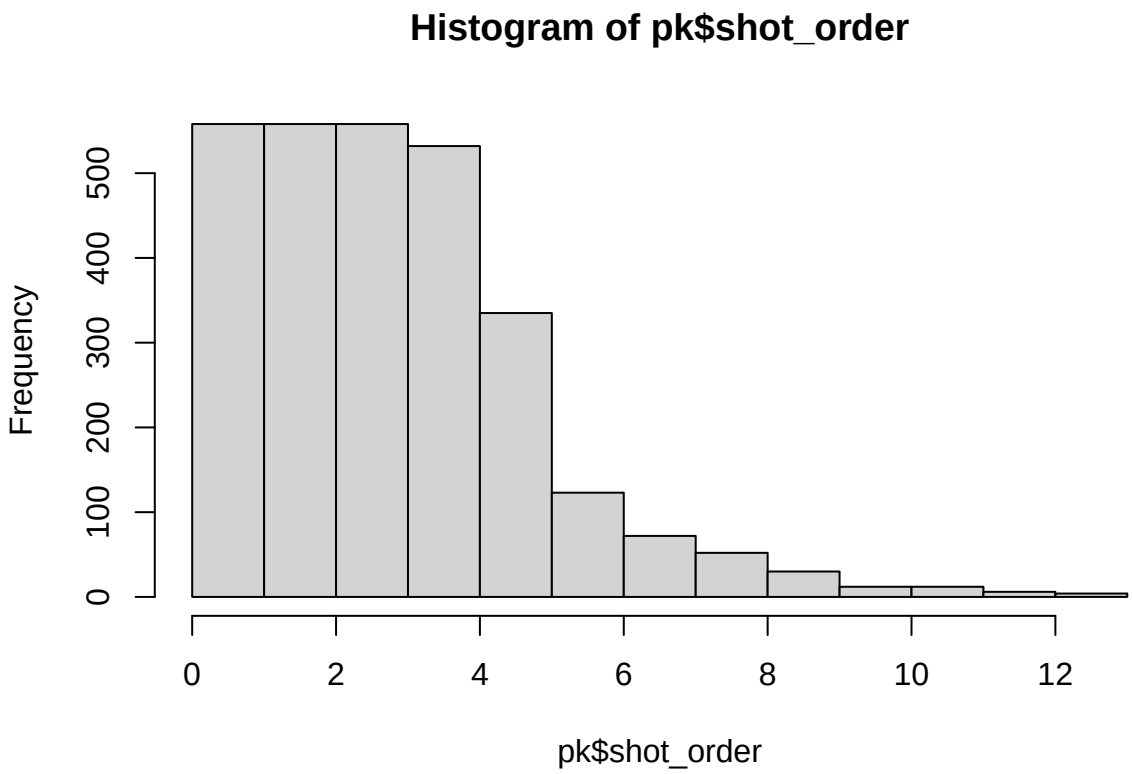
```
# read in data
pk<-read.csv("../00_data/02_processed/pks.csv",header=T)
pk$match<-as.factor(pk$match)
pk$take_first <- as.factor(pk$take_first)
```

First, let's look at how long shootouts actually tend to last. They can last as few as three rounds, but there's also no upper limit to how long they could go. What is the distribution of shootout duration in historical matches?

```
# check shot frequencies
table(pk$shot_order)
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13
## 558 558 558 532 335 123  72  52  30  12  12   6   4
```

```
hist(pk$shot_order,breaks=seq.int(0,(max(pk$shot_order))))
```



Not many shootouts last beyond the 9th round. We'll focus conditional probabilities on the first 9 rounds

```
pk<-pk[pk$shot_order<=9,]

# now with the desired # of rounds selected, set shot_order as a factor
pk$shot_order <- as.factor(pk$shot_order)

# and subset the data by team (take_order)
team_1<-pk[pk$take_first==1,]
team_2<-pk[pk$take_first==0,]
```

As a first general question: What is the probability of scoring each given shot in a penalty shootout? Let's answer by looking at the empirical frequencies of goal conversions per shot

```
team_1_freqs <- tapply(team_1$goal,team_1$shot_order,mean)
names(team_1_freqs) <- paste0("team_1_", 1:9)

team_2_freqs <- tapply(team_2$goal,team_2$shot_order,mean)
names(team_2_freqs) <- paste0("team_2_", 1:9)

# calculate the standard error for each frequency
get_se <- function(x) (sd(x)/sqrt(length(x)))^2
team_1_se <- tapply(team_1$goal, team_1$shot_order, get_se)
team_2_se <- tapply(team_2$goal, team_2$shot_order, get_se)


# check shot frequencies for each team
print(team_1_freqs)
```

```
##  team_1_1  team_1_2  team_1_3  team_1_4  team_1_5  team_1_6  team_1_7  team_1_8
## 0.7455197 0.7849462 0.7562724 0.7374101 0.7115385 0.6451613 0.7500000 0.6538462
```

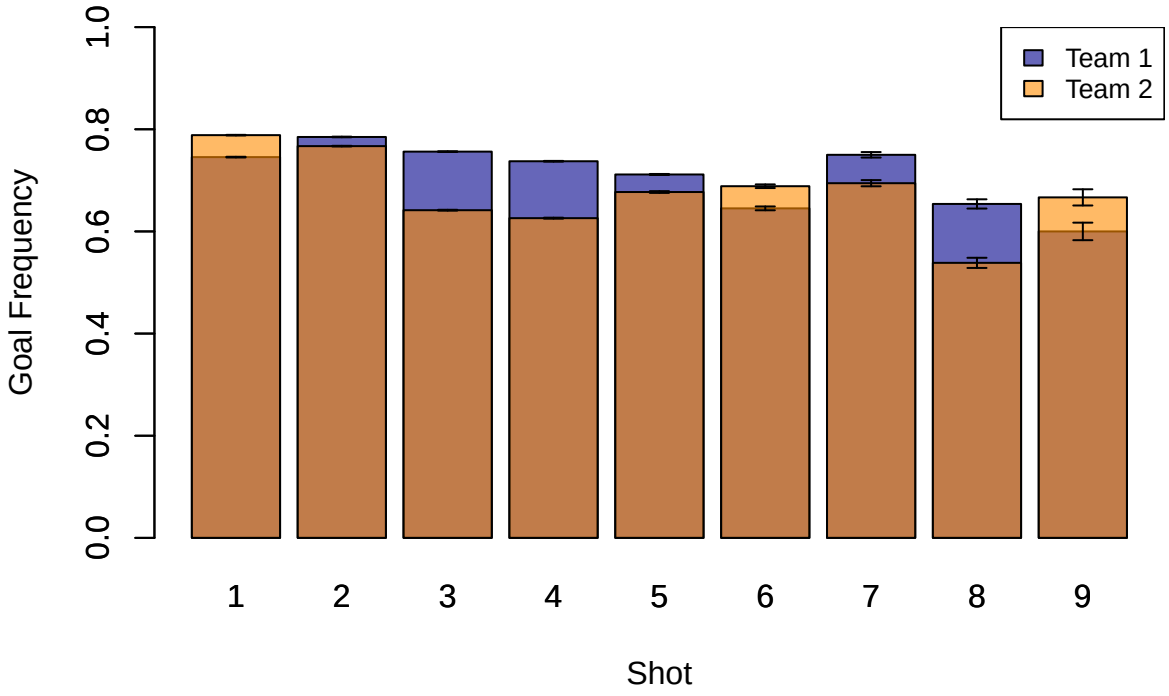
```
##  team_1_9
## 0.6000000
```

```
print(team_2_freqs)
```

```
##  team_2_1  team_2_2  team_2_3  team_2_4  team_2_5  team_2_6  team_2_7  team_2_8
## 0.7885305 0.7670251 0.6415771 0.6259843 0.6771654 0.6885246 0.6944444 0.5384615
##  team_2_9
## 0.6666667
```

Visualize these frequencies as a bar graph

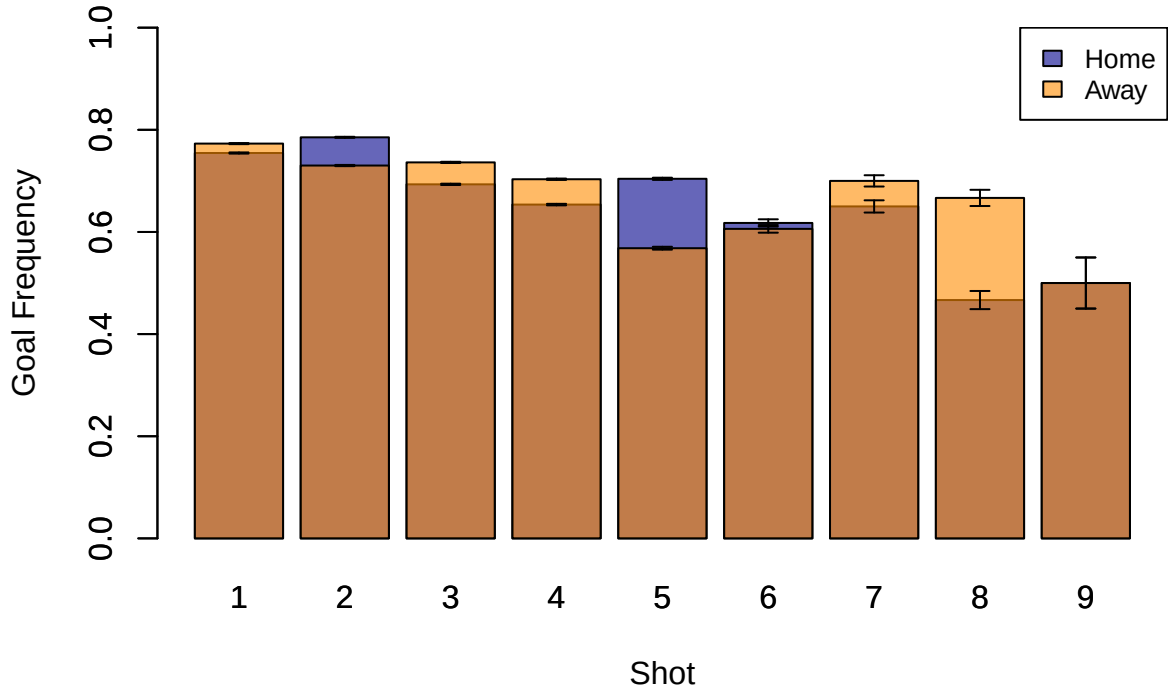
```
# use a function to add arrows on the chart
error.bar <- function(x, y, upper, lower=upper, length=0.1,...){
  arrows(x,y+upper, x, y-lower, angle=90, code=3, length=0.5*length, ...)
}

# plot frequencies
bp<-barplot(team_1_freqs ~ seq(1,9), ylim = c(0,1), col = alpha('darkblue', 0.6),
        xlab='Shot', ylab='Goal Frequency')
legend('topright', legend=c('Team 1', 'Team 2'), fill=c(alpha('darkblue', 0.6), alpha('darkorange', 0.6)
barplot(team_2_freqs ~ seq(1,9), col = alpha('darkorange', 0.6), add=T)

# add error bars
error.bar(bp,team_1_freqs, tapply(team_1$goal, team_1$shot_order, get_se))
error.bar(bp,team_2_freqs, tapply(team_2$goal, team_2$shot_order, get_se))
```



Visually, goal frequencies appear generally similar between both teams, with the team that takes first showing a slightly higher goal turnover rate in most rounds. Though goal frequencies are similar between teams, low standard errors around these estimates mean these differences could still be meaningful enough to provide an edge.

For another general question: How much does home field advantage matter? How do shot frequencies compare for the home vs away team?

```r
# subset data by home vs away team
home_team <- pk[pk$neutral_stadium == 0 & pk$attacker_home == 1,]
away_team <- pk[pk$neutral_stadium == 0 & pk$attacker_home == 0,]

# shot frequencies for either team
home_freqs <- tapply(home_team$goal,home_team$shot_order,mean)
names(home_freqs) <- paste0("home_", 1:9)

away_freqs <- tapply(away_team$goal,away_team$shot_order,mean)
names(away_freqs) <- paste0("away_", 1:9)

# calculate the standard error for each frequency
home_team_se <- tapply(home_team$goal, home_team$shot_order, get_se)
away_team_se <- tapply(away_team$goal, away_team$shot_order, get_se)

# plot frequencies
bp<-barplot(home_freqs ~ seq(1,9), ylim = c(0,1), col = alpha('darkblue', 0.6),
        xlab='Shot', ylab='Goal Frequency')
legend('topright', legend=c('Home', 'Away'), fill=c(alpha('darkblue', 0.6), alpha('darkorange', 0.6)),
barplot(away_freqs ~ seq(1,9), col = alpha('darkorange', 0.6), add=T)

# add error bars
error.bar(bp,home_freqs, tapply(home_team$goal, home_team$shot_order, get_se))
error.bar(bp,away_freqs, tapply(away_team$goal, away_team$shot_order, get_se))
```



These results look noisier. This probably suggests there's no shot-by-shot advantage to the home team during a shootout.