Conditional Probability

Divya Krishnan

Autumn 2015

If a baseball team scores X runs, what is the probability it will win the game?

This is the question we will explore in this lab (ddapted from Decision Science News, 2014). We will use a dataset of baseball game statistics from 2010-2013. Baseball is a played between two teams who take turns batting and fielding. A run is scored when a player advances around the bases and returns to home plate. More information about the dataset can be found at http://www.retrosheet.org/.

Data files can be found on Canvas in the lab folder. Download the files and load them into one data.frame in R as shown below. Comment this code to demonstrate you understand how it works.

```r
library(dplyr)
library(ggplot2)
colNames <- read.csv("cnames.txt", header=TRUE)
baseballData <- NULL
for (year in seq(2010,2013,by=1)){
  mypath <- paste('GL',year,'.txt',sep='')
  baseballData <- rbind(baseballData,read.csv(mypath,
  col.names=colNames$Name))
  baseballData <- tbl_df(baseballData)
}
```

Select the following relevant columns and create a new local data.frame to store the data you will use for your analysis.

- Date
- Home
- Visitor
- HomeLeague
- VisitorLeague
- HomeScore
- VisitorScore

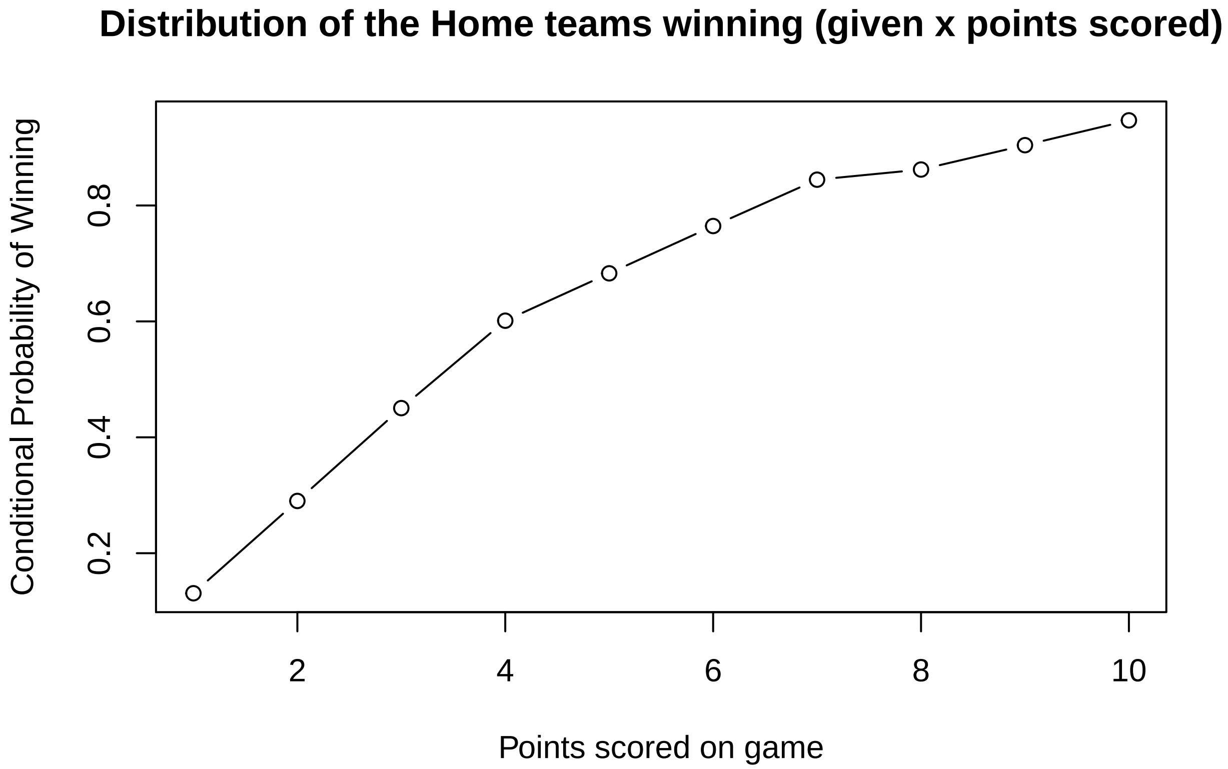
```r
bb<-tbl_df(data.frame(date=baseballData$Date,home=baseballData$Home,visitor=baseballData$Visitor,
       homeLeague=baseballData$HomeLeague,visitorLeague=baseballData$VisitorLeague,homeScore=baseballDa
       visitorScore=baseballData$VisitorScore))
```

Considering only games between two teams in the National League, compute the conditional probability of the team winning given X runs scored, for $X = 0, ..., 10$. Do this separately for Home and Visitor teams.

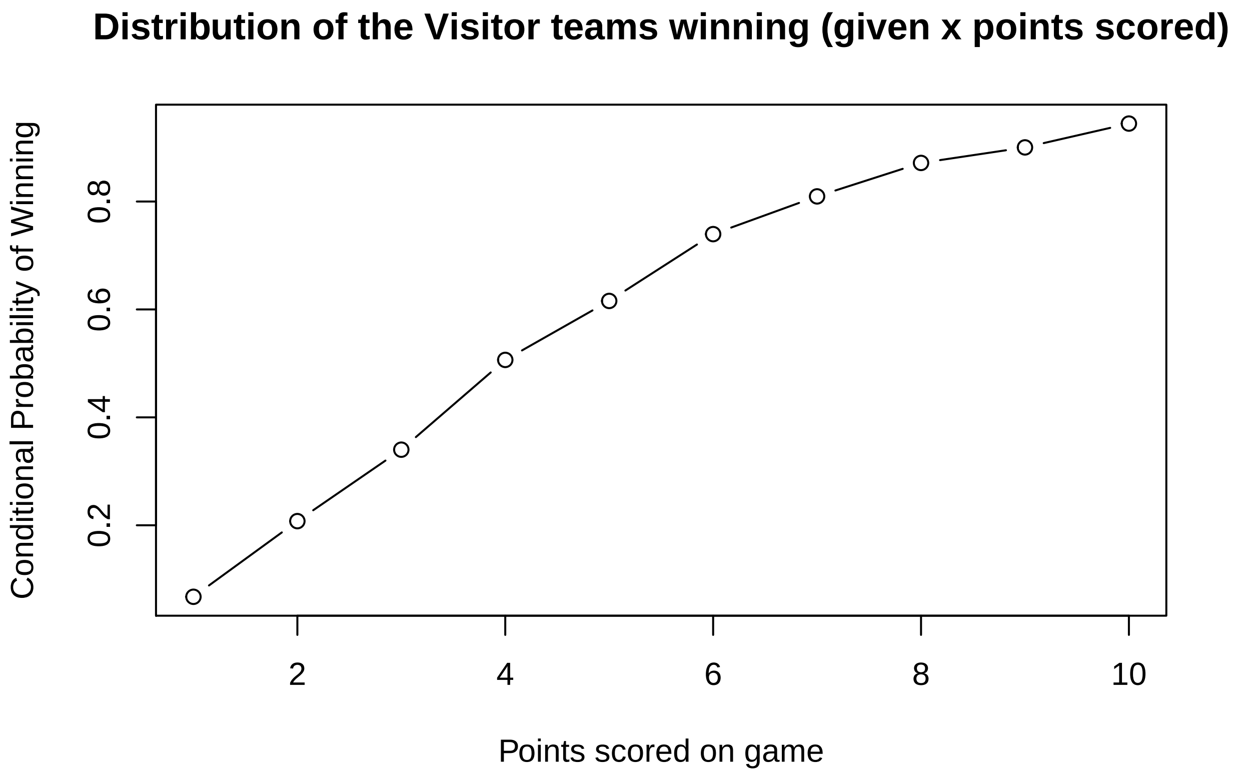
```r
# Conditional Probability of home teams
totalGames<-nrow(bb)
condHome<-vector()

for(i in 1:10)
{
  numScore<-nrow(subset(bb,bb$homeScore==i))
  pB<-numScore/totalGames
  numA<-nrow(subset(bb,bb$homeScore>bb$visitorScore & bb$homeScore==i))
  pBA<-numA/totalGames
  condHome[i]<-pBA/pB
}


# Conditional Probability of visitor teams
totalGames<-nrow(bb)
condVisitor<-vector()

for(i in 1:10)
{
  numScore<-nrow(subset(bb,bb$visitorScore==i))
  pB<-numScore/totalGames
  numA<-nrow(subset(bb,bb$visitorScore>bb$homeScore & bb$visitorScore==i))
  pBA<-numA/totalGames
  condVisitor[i]<-pBA/pB
}
```

- Design a visualization that shows your results.

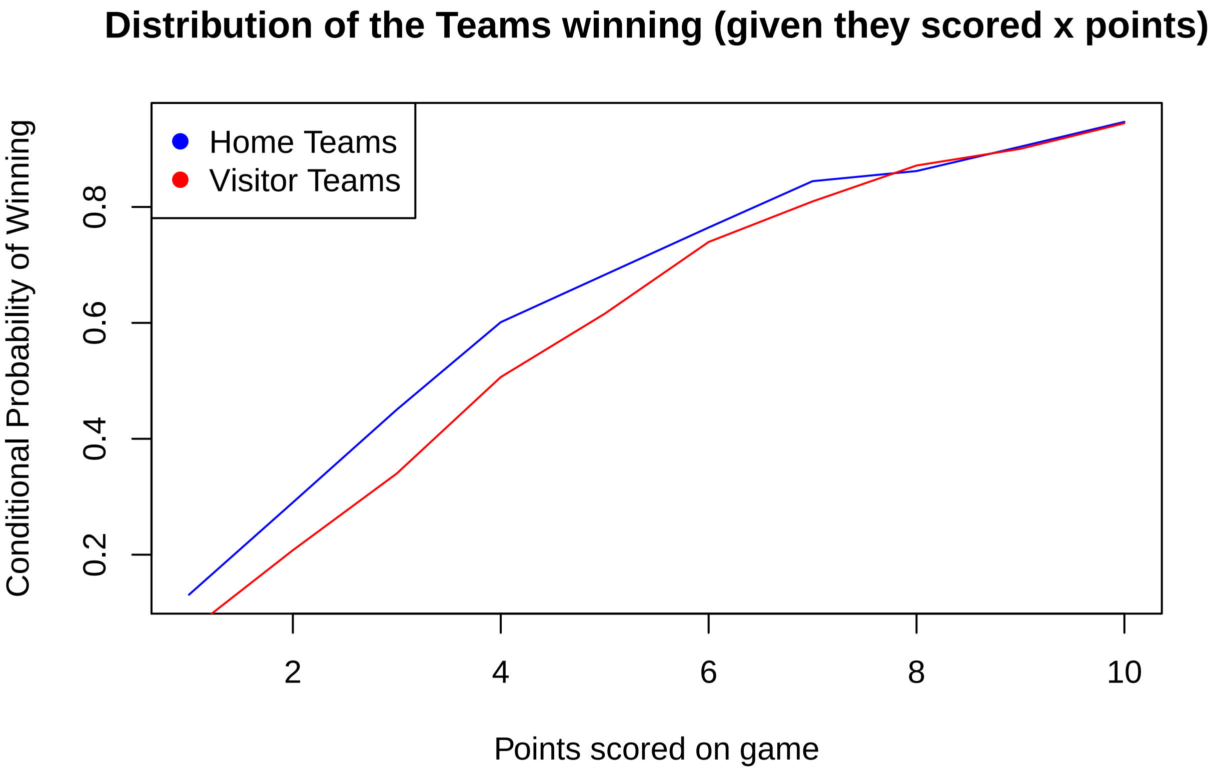
```
par(mfrow=c(1,1))
plot(condHome,ylab="Conditional Probability of Winning",
     xlab="Points scored on game",main="Distribution of the Home teams winning (given x points scored)"
```



```
plot(condVisitor,ylab="Conditional Probability of Winning",
     xlab="Points scored on game",main="Distribution of the Visitor teams winning (given x points score
```



```
# Plotting both distributions together
plot(condHome,type="n",ylab="Conditional Probability of Winning",
     xlab="Points scored on game",main="Distribution of the Teams winning (given they scored x points)"
lines(condHome,col="blue",xlab="")
lines(condVisitor,col="red",pch=4)
legend("topleft",legend=c("Home Teams","Visitor Teams"),col=c("blue","red"),pch=c(19,19),cex=1)
```



- Discuss what you find.

I find that the distribution of home team winning given they scored 0 to 10 points is increasing linearly along with the conditional probabililty of them winning. A similar distribution can be observed for the visitor team winning given they scored 0 to 10 points.

Plotting both the distributions together, we observe that given both home and visitor teams scored the same number of points, the home teams have a higher probability of winning. This is similar to what we may expect. The home teams are more likely to win than visitor teams, provided they score the same number of points. This goes with the belief that home teams score better than visitor teams, due to the atmosphere being more conducive to the home teams.