Code ▾

MATH1324 Assignment 2

Modelling the Distribution of Football Goals

Group Details

- Parikshit Sreedhar(s3643796)
- Arjun Balaji(s3629999)
- Ravi Madhurchand Pandey(s3638787)

Problem Statement

What are we trying to address?

Using Poisson distribution as a weapon, Can Analysts help bettors put their money on the total number of goals scored in a match?

- We will try to visualize and observe the distribution of total goals scored in:-
  - **every match in a season**
  - **every match in a combination of seasons(3 seasons & 5 seasons)**
- Try to **patternise** as well as examine whether a theoretical poisson distribution fits the data on the total number of goals scored in a match.

Load Packages

What packages were used to carry out this analysis?

Hide

```
install.packages("dplyr")
install.packages("calibrate")
library(calibrate)
library(dplyr)
```

Data

Where was the data collected from?

The data has been downloaded from http://www.football-data.co.uk/englandm.php (http://www.football-data.co.uk/englandm.php). The data contains the history of goals scored in English Premier League (EPL) for five consecutive seasons from 2010 to 2015.

An overivew of the data chosen for our analysis: **Date**, **HomeTeam**, **AwayTeam**, **FTHG (Full Time Home Team Goals)**, **FTAG (Full Time Away Team Goals)**

An extra attribute FTTG (Full Time Total Goals) has been calculated from FTHG and FTAG attributes.

Hide

```
##Setting the working directory
setwd("C:\\MS242 - Master of Analytics\\Intro to Statistics\\Assignment 2")
##Listing all the csv files within this directory
soccerdata<-list.files("C:\\MS242 - Master of Analytics\\Intro to Statistics\\A
ssignment 2", pattern=".csv", full.names=T)
##Reading Data
soccer_10_11<-read.csv(soccerdata[[1]])%>%select(Date,HomeTeam,AwayTeam,FTHG,FT
AG) #2010-2011
soccer_11_12<-read.csv(soccerdata[[2]])%>%select(Date,HomeTeam,AwayTeam,FTHG,FT
AG) #2011-2012
soccer_12_13<-read.csv(soccerdata[[3]])%>%select(Date,HomeTeam,AwayTeam,FTHG,FT
AG) #2012-2013
##Removing NA's
soccer_10_11<-soccer_10_11[!(is.na(soccer_10_11$FTHG))|!(is.na(soccer_10_11$FTA
G)),]
soccer_11_12<-soccer_11_12[!(is.na(soccer_11_12$FTHG))|!(is.na(soccer_11_12$FTA
G)),]
soccer_12_13<-soccer_12_13[!(is.na(soccer_12_13$FTHG))|!(is.na(soccer_12_13$FTA
G)),]
##Adding a new attribute - FTTG (Full Time Total Goals)
soccer_10_11$FTTG<-soccer_10_11$FTHG+soccer_10_11$FTAG
soccer_11_12$FTTG<-soccer_11_12$FTHG+soccer_11_12$FTAG
soccer_12_13$FTTG<-soccer_12_13$FTHG+soccer_12_13$FTAG
##Combining Data
soccer_s1_s2_s3<-rbind(soccer_10_11,soccer_11_12,soccer_12_13) #2010-2011,2011-
2012 & 2012-2013
datalist=list()#creating an empty list
for(i in 1:length(soccerdata)){
  datalist[[i]] <- read.csv(soccerdata[i]) %>% select(Date,HomeTeam,AwayTeam,FT
HG,FTAG)
}

soccer<-do.call("rbind",datalist) #Combining all seasons' data
##Removing NA's for the combined data (all 5 seasons)
soccer<-soccer[!(is.na(soccer$FTHG))|!(is.na(soccer$FTAG)),]
##Adding a new attribute - FTTG (Full Time Total Goals) for the combined data
(all 5 seasons)
soccer$FTTG<-soccer$FTHG+soccer$FTAG
```

Distribution Fitting

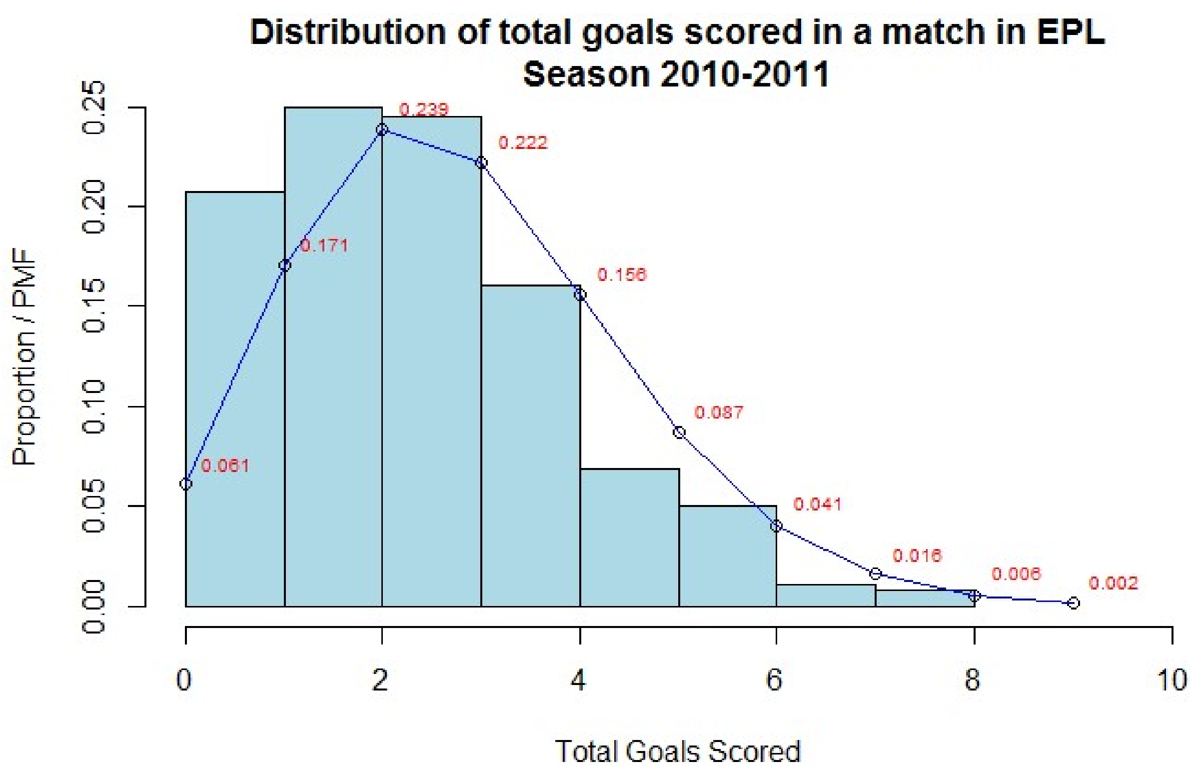
How are we going to patternise?

Hide

```
##This is for the 2010-2011 season
mu1<-mean(soccer_10_11$FTTG,na.rm=T)
##Generating a sequence of values for the random variable X1 and plotting Full
time total goals
##for season 1 and overlaying the theoretical distribution (with mean being tha
t
## of average total goals scored in season 1)
X1<-seq(ifelse(sign(round(mu1-sqrt(mu1)*4))==-1,0,round(mu1-sqrt(mu1)*4)),round
(mu1+sqrt(mu1)*4))
PMF1<-dpois(X1,mu1)
PMF1<-as.list(PMF1)
dplist1=list()
for(i in 1:length(PMF1)){dplist1[[i]]<-round(PMF1[[i]],3)}
hist(soccer_10_11$FTTG,main="Distribution of total goals scored in a match in E
PL\nSeason 2010-2011",xlab="Total Goals Scored",ylab="Proportion / PMF",freq =
FALSE, col="lightblue", xlim = range(0:10, na.rm = FALSE))
points(X1,dpois(X1,mu1),type="p",col="black")
```

Hide

```
lines(X1,dpois(X1,mu1),type="l",col="blue")
for(i in 1:length(PMF1)){text(X1[i],dplist1[[i]]+0.0100, dplist1[[i]], cex = .
6, pos = 4, col="red")}
```

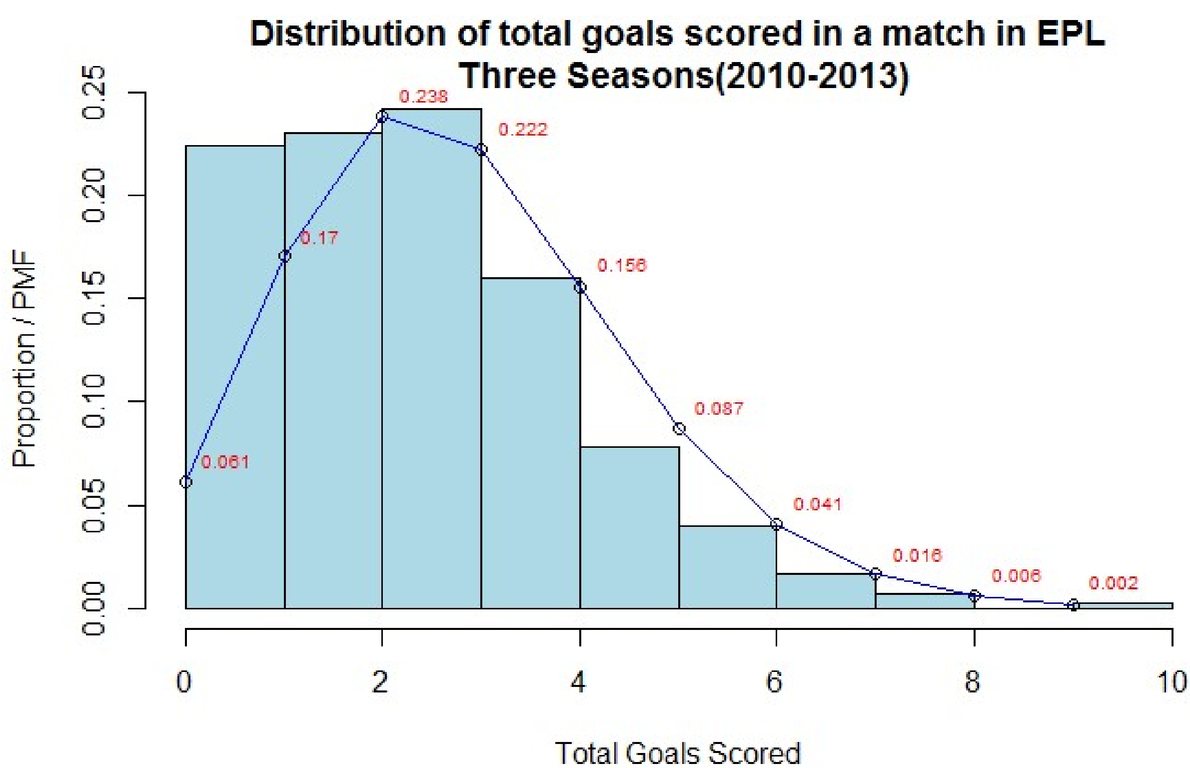


Hide

```
##Season 1 , Season 2 and Season 3 combined
mu123<-mean(soccer_s1_s2_s3$FTTG,na.rm=TRUE)
X123<-seq(ifelse(sign(round(mu123-sqrt(mu123)*4))==-1,0,round(mu123-sqrt(mu123)
*4)),round(mu123+sqrt(mu123)*4))
PMF123<-dpois(X123,mu123)
PMF123<-as.list(PMF123)
dplist123=list()
for(i in 1:length(PMF123)){dplist123[[i]]<-round(PMF123[[i]],3)}
hist(soccer_s1_s2_s3$FTTG,main="Distribution of total goals scored in a match i
n EPL\n Three Seasons(2010-2013)",xlab="Total Goals Scored",ylab="Proportion /
PMF",freq = FALSE, col="lightblue")
points(X123,dpois(X123,mu123),type="p",col="black")
```

Hide

```
lines(X123,dpois(X123,mu123),type="l",col="blue")
for(i in 1:length(PMF123)){text(X123[i],dplist123[[i]]+0.0100, dplist123[[i]],
cex = .6, pos=4, col = "red")}
```

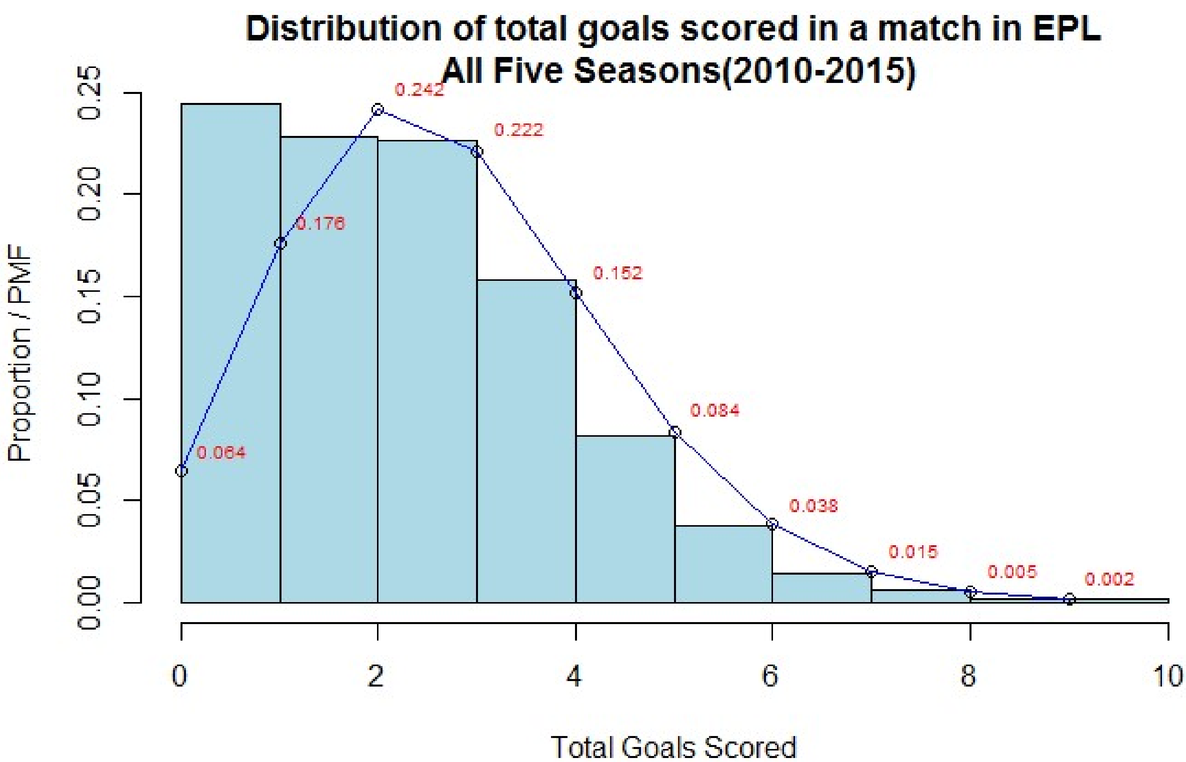


Hide

```
##For all the 5 seasons
mu<-mean(soccer$FTTG,na.rm=T)
X<-seq(ifelse(sign(round(mu-sqrt(mu)*4))==-1,0,round(mu-sqrt(mu)*4)),round(mu+s
qrt(mu)*4))
PMF<-dpois(X,mu)
PMF<-as.list(PMF)
dplist=list()
for(i in 1:length(PMF)){dplist[[i]]<-round(PMF[[i]],3)}
hist(soccer$FTTG,main="Distribution of total goals scored in a match in EPL\n A
ll Five Seasons(2010-2015)",xlab="Total Goals Scored",ylab="Proportion / PMF",f
req = FALSE, col="lightblue")
points(X,dpois(X,mu),type="p",col="black")
```

Hide

```
lines(X,dpois(X,mu),type="l",col="blue")
for(i in 1:length(PMF)){text(X[i],dplist[[i]]+0.0100, dplist[[i]], cex = .6,pos
=4,col="red")}
```



Hide

```
##Calculating Average Total Goals
Scenario1<-round(mean(soccer_10_11$FTTG),3) #Considering 2010-2011 Season
Scenario2<-round(mean(soccer_s1_s2_s3$FTTG),3)#Considering 2010-2011,2011-2012
& 2012-2013 Season
Scenario3<-round(mean(soccer$FTTG),3) #Considering all 5 Seasons
```

Table 1: Average Total_Goals scored under each of the scenarios

SCENARIO	AVG_TOTAL_GOALS
Scenario1	2.797
Scenario2	2.800
Scenario3	2.747

From the above 3 scenarios, though the number of data points vary, a near perfect fit of the theoretical poisson distribution to the empirical data is achieved in all three. This is because the average total number of goals scored is approximately the same across all 3 scenarios.

Interpretation

What message can bettors drive home from this analysis?

The theoretical poisson distribution fits(reasonably well) the distribution of total number of goals scored in an EPL match and bettors can use this blueprint to clearly work out the chance of scoring X(an arbitary discrete value) no.of total goals scored in a match and use this as a basis to put their money (with a fair degree of certainty) on the total number of goals that would be scored in a match in the coming EPL seasons.Also ,bettors can comfortably dock their baseline prediction on the total number of goals scored in a match to around 3 goals.

Limitations Though bettors can wield the Poisson weapon to have a less hazier view on the total number of goals scored in a match in a coming EPL season , they will not be able to obtain visibility on the result(win/loss) of the match itself.

Suggestions Assuming that there is a lot of money at stake , bettors may want to expand the horizon of the data under consideration and leverage its multiple features/attributes(Half-Time Goals,Posession,No.of Substitutions due to injuries etc..) to get a near accurate prediction on the total number of goals scored in an EPL match as well as the result of the same.

Reference: Anon 2017, “England Football Results Betting Odds | Premiership Results & Betting Odds”, Football-data.co.uk, viewed 23 April, 2017, http://www.football-data.co.uk/englandm.php (http://www.football-data.co.uk/englandm.php).