Code ▾

# MATH1324 Assignment 2

### *Modelling the Distribution of Football Goals*

## Group Details

- Anshit Malik (s3631281)
- Namita Chhibba (s3631442)
- Mohammad (s3650497)

## Problem Statement

Comparing the performance of Club “Bayern Munchen” as a home team and as a visitor team using probability distribution for goals scored.

## Load Packages

Hide

```
#remove from comment if you need to install the desired packages
#install.packages("repmis")
#install.packages("dplyr")
library(repmis)
library(dplyr)
```

## Data

** Data has been imported from github using package “repmis”

Hide

```
source_data("https://github.com/jalapic/engsoccerdata/blob/master/data/germany.rda?raw=True")
```

```
Downloading data from: https://github.com/jalapic/engsoccerdata/blob/master/data/germany.rda?
raw=True

SHA-1 hash of the downloaded data file is:
2171a22b9405e2af28386fd7b3c8556f6113c5f8
```

```
[1] "germany"
```

Hide

```
main_data<-as.data.frame(germany)    #data with 16120 rows from 1963-2016
```

Variable Description

1. date - date of match/tie
2. Season - season (e.g. 2012 refers to 2012/13 season)
3. home - home team (note for games played at neutral venues this isn’t relevant)

4. visitor - visiting team (note for games played at neutral venues this isn’t relevant)
5. FT - final score. this is the final score even after extra time (i.e. not just after 90 minutes)
6. hgoal - number of goals scored by team when played as home team
7. vgoal - number of goals scored by team when played as visitor team

Two Dataframes are made by filtering imported data on home and visitor column for Bayern Munchen

Hide

```
Home_Team<-"Bayern Munchen"
Visitor_Team<-"Bayern Munchen"
Data_Club_as_Home<-main_data[main_data$home==Home_Team,]
Data_Club_as_Visitor<-main_data[main_data$visitor==Visitor_Team,]
```

# Distribution Fitting

In order to proceed further, first of all we need to investigate which distribution our data follows. As our problem statement is based on calculaing probability of scoring goals as home team and visitor team, therefore, we will consider Poisson distribution rather than any other distribution.

We worked on 3 cases to prove our consideration of Poisson distribution

Case 1. Assumptions for Poisson Distribution :

1. K is the number of times a goal occurs during a match and K can take values 0, 1, 2, .
2. The occurrence of scoring goals in one match does not affect the probability of scoring goals in the other match.
3. That is, events occur independently.
4. A team plays one match at a time. It cannot be participating in two matches at the same time.
5. The probability of scoring a goal is proportional to the duration of the match.

As our data meets the aforementioned assumptions, hence, K (number of goals in a match) is a Poisson random variable, and the distribution is a Poisson distribution.

Case 2. The mean of goals scored is almost same as variance of goals scored when Bayern Munchen played as Home Team and as Visitor Team, this also favours poisson distribution.

Hide

```
Main_data_Bayern_home_goal<-Data_Club_as_Home[,6]
home_mean<- round(Main_data_Bayern_home_goal %>% mean(),2)
home_var<-round(Main_data_Bayern_home_goal %>% var(),2)
home_sum<-round(Main_data_Bayern_home_goal %>% sum(),2)
Main_data_Bayern_visitor_goal<-Data_Club_as_Visitor[,7]
visitor_mean<-round(Main_data_Bayern_visitor_goal %>% mean(),2)
visitor_var<-round(Main_data_Bayern_visitor_goal %>% var(),2)
visitor_sum<-Main_data_Bayern_visitor_goal %>% sum()
stats_data_home<-rbind(home_sum,home_mean,home_var)
stats_data_visitor<-rbind(visitor_sum,visitor_mean,visitor_var)
statistic_name<-c("Sum","Mean","Variance")
stats_data<-cbind(statistic_name,stats_data_home,stats_data_visitor) %>% as.data.frame()
colnames(stats_data)<-c("Statistic","As Home","As Visitor")
rownames(stats_data)<-1:3
stats_data
```

| Statistic | As Home | As Visitor |
|---|---|---|
| <fctr> | <fctr> | <fctr> |

| | | | |
|---|---|---|---|
| 1 | Sum | 2313 | 1441 |
| 2 | Mean | 2.67 | 1.66 |
| 3 | Variance | 2.98 | 1.87 |

3 rows

Case 3. We will compare Theoretical Distribution (using poisson function dpois,lambda=mean) and Empirical Distribution (by using probability formula, eg number of matches in which goals scored is 0/total number of matches….so on…till 10) as home team and visitor team separately.

——> EMPIRICAL VS THEORETICAL DISTRIBUTION (AS HOME)

Empirical & Theoretical Probability Function as Home

Hide

```
Home_Team_data<-Data_Club_as_Home
plot_df_practical_home<-data.frame(c(0:10))
for (i in 1:11){
  plot_df_practical_home[i,2]<-round((length(Home_Team_data[Home_Team_data$hgoal==i-1,6])/nro
w(Home_Team_data))*100,2)
}
plot_df_theoretical_home<-data.frame(c(0:10))
for (i in 1:11){
  plot_df_theoretical_home[i,2]<-round(dpois(i-1,home_mean)*100,2)
}
plot_df_theoretical_home_error<-data.frame(c(0:10))
for (i in 1:11){
  plot_df_theoretical_home_error[i,2]<-(plot_df_theoretical_home[i,2]-
plot_df_practical_home[i,2])^2
}
names(plot_df_theoretical_home)<-c("Goals","Theoretical Probability Scoring as Home")
names(plot_df_practical_home)<-c("Goals","Empirical Probability Scoring as Home")
names(plot_df_theoretical_home_error)<-c("Goals","Error Square")
df_plot_prac_theo_home<-cbind.data.frame(plot_df_practical_home,plot_df_theoretical_home$`The
oretical Probability Scoring as Home`,plot_df_theoretical_home_error$`Error Square`)
names(df_plot_prac_theo_home)<-c("Goals","Emp. Prob. Home","Theo. Prob. Home","Error Square")
df_plot_prac_theo_home
```

| Goals | Emp. Prob. Home | Theo. Prob. Home | Error Square |
|---:|---:|---:|---:|
| <int> | <dbl> | <dbl> | <dbl> |
| 0 | 7.61 | 6.93 | 0.4624 |
| 1 | 20.18 | 18.49 | 2.8561 |
| 2 | 22.49 | 24.68 | 4.7961 |
| 3 | 22.49 | 21.97 | 0.2704 |
| 4 | 12.57 | 14.66 | 4.3681 |
| 5 | 7.84 | 7.83 | 0.0001 |
| 6 | 4.38 | 3.48 | 0.8100 |
| 7 | 1.73 | 1.33 | 0.1600 |

| Goals<br>&lt;int&gt; | Emp. Prob. Home<br>&lt;dbl&gt; | Theo. Prob. Home<br>&lt;dbl&gt; | Error Square<br>&lt;dbl&gt; |
|---:|---:|---:|---:|
| 8 | 0.23 | 0.44 | 0.0441 |
| 9 | 0.35 | 0.13 | 0.0484 |

1-10 of 11 rows Previous **1** 2 Next

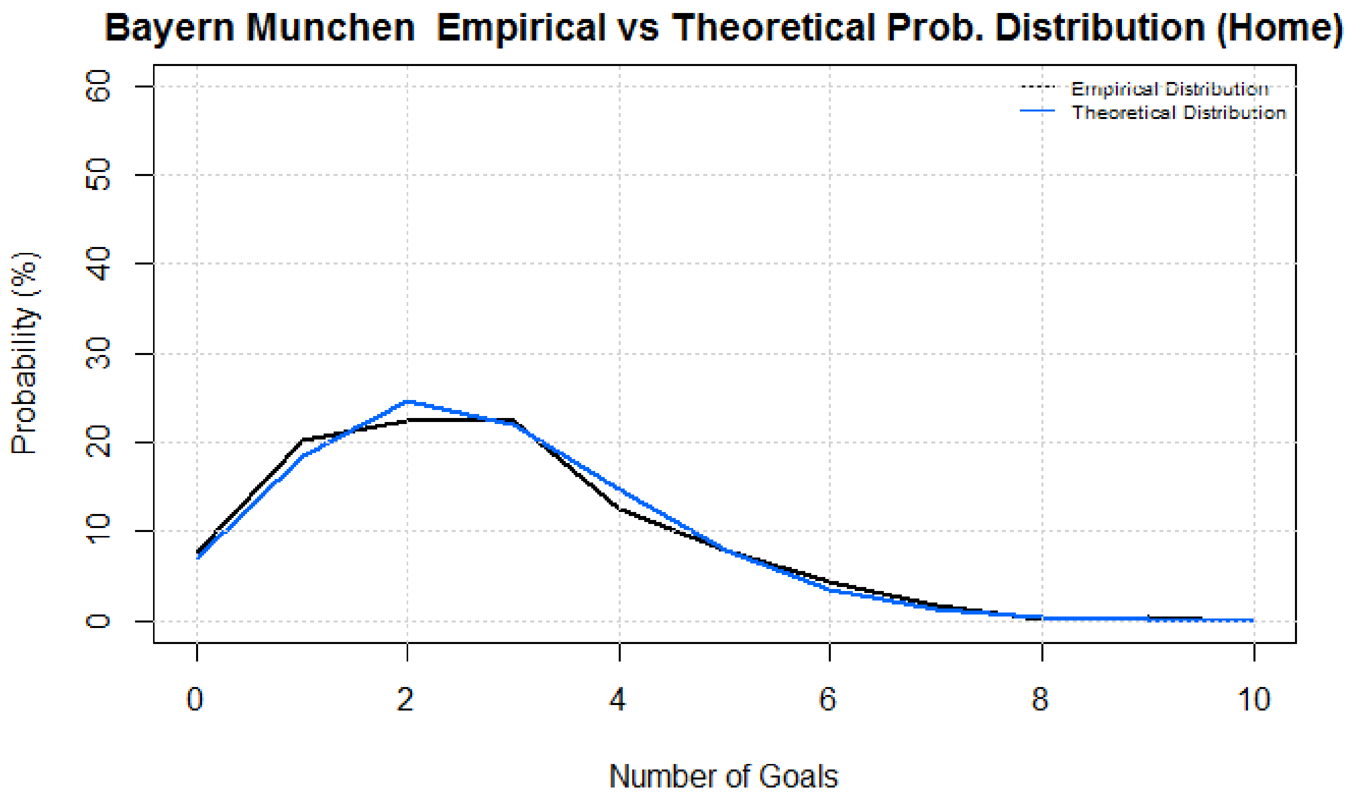
Comparison of Empirical & Theoretical Distribution As Home on single plot

Hide

```
plot(plot_df_practical_home$Goals,plot_df_practical_home[,2],ylim=c(0,60),ylab="Probability (%)",xlab="Number of Goals",type="l",col="black",lwd=2,main=paste(Home_Team," Empirical vs Theoretical Prob. Distribution (Home)"))
lines(plot_df_theoretical_home$Goals,plot_df_theoretical_home[,2],ylim=c(0,60),ylab="Probability (%)",xlab="Number of Goals",col="#0066ff",lwd=2)
```

Hide

```
legend('topright', c("Empirical Distribution","Theoretical Distribution") ,
   lty=1, col=c( "black","#0066ff"), bty='n', cex=.6)
grid()
```



——> EMPIRICAL VS THEORETICAL DISTRIBUTION (AS VISITOR)

Empirical & Theoretical Probability Function as Visitor

Hide

```
Visitor_Team_data<-Data_Club_as_Visitor
plot_prac_theo_visitor<-data.frame(c(0:10))
for (i in 1:11){
  plot_prac_theo_visitor[i,2]<-round((length(Visitor_Team_data[Visitor_Team_data$hgoal==i-
1,6])/nrow(Visitor_Team_data))*100,2)
}
for (i in 1:11){
  plot_prac_theo_visitor[i,3]<-round(dpois(i-1,visitor_mean)*100,2)
}
for (i in 1:11){
  plot_prac_theo_visitor[i,4]<-(plot_prac_theo_visitor[i,3]-plot_prac_theo_visitor[i,2])^2
}
names(plot_prac_theo_visitor)<-c("Goals","Emp. Prob. Visitor","Theo. Prob. Visitor","Error Sq
uare")
plot_prac_theo_visitor
```

| Goals<br>&lt;int&gt; | Emp. Prob. Visitor<br>&lt;dbl&gt; | Theo. Prob. Visitor<br>&lt;dbl&gt; | Error Square<br>&lt;dbl&gt; |
|---:|---:|---:|---:|
| 0 | 29.99 | 19.01 | 120.5604 |
| 1 | 34.95 | 31.56 | 11.4921 |
| 2 | 18.92 | 26.20 | 52.9984 |
| 3 | 10.03 | 14.50 | 19.9809 |
| 4 | 3.69 | 6.02 | 5.4289 |
| 5 | 1.27 | 2.00 | 0.5329 |
| 6 | 0.81 | 0.55 | 0.0676 |
| 7 | 0.35 | 0.13 | 0.0484 |
| 8 | 0.00 | 0.03 | 0.0009 |
| 9 | 0.00 | 0.01 | 0.0001 |

1-10 of 11 rows Previous **1** 2 Next

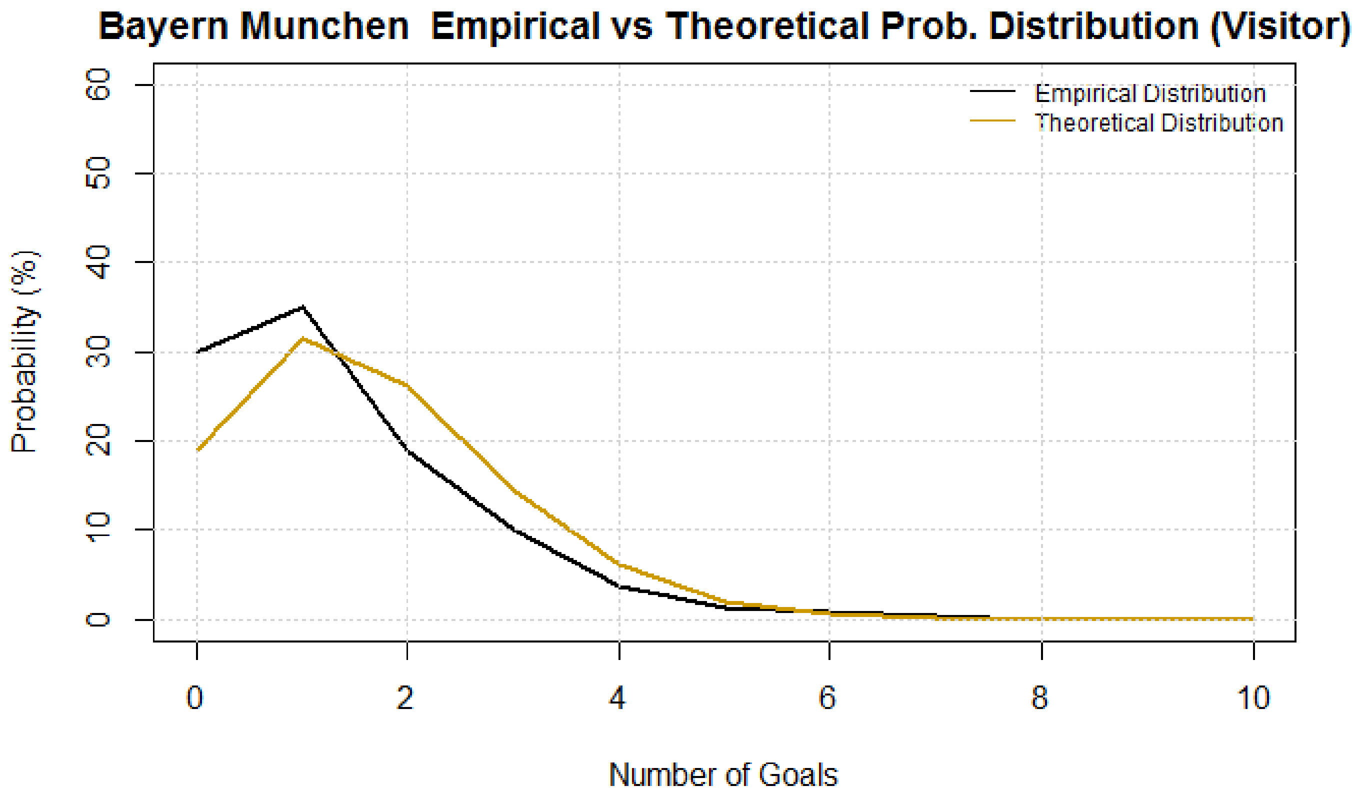
Comparison of Empirical & Theoretical Distribution As Visitor on single plot

Hide

```
#plot.new()
plot(plot_prac_theo_visitor$Goals,plot_prac_theo_visitor[,2],ylim=c(0,60),ylab="Probability
 (%)",xlab="Number of Goals",type="l",col="black",lwd=2,main=paste(Home_Team," Empirical vs T
heoretical Prob. Distribution (Visitor)"))
lines(plot_prac_theo_visitor$Goals,plot_prac_theo_visitor[,3],ylim=c(0,60),ylab="Probability
 (%)",xlab="Number of Goals",col="#cc9900",lwd=2)
```

Hide

```
legend('topright', c("Empirical Distribution","Theoretical Distribution") ,
   lty=1, col=c("black","#cc9900"), bty='n', cex=.75)
grid()
```



Error Analysis

Hide

```
Home_team_prob_Error<-round(df_plot_prac_theo_home[,4]%>%mean()%>%sqrt(),2)
Visitor_Team_prob_error<-round(plot_prac_theo_visitor[,4]%>%mean()%>%sqrt(),2)
Error_df<-cbind.data.frame(Home_team_prob_Error,Visitor_Team_prob_error)
names(Error_df)<-c("As Home Team","As Visitor Team")
rownames(Error_df)<-"Error %"
Error_df
```

| | As Home Team | As Visitor Team |
|---|---|---|
| | <dbl> | <dbl> |
| Error % | 1.12 | 4.38 |

1 row

As all 3 cases nearly favours Poisson Distribution, hence we will use Poisson Distribution for the interpretation of our problem statement.

# Interpretation

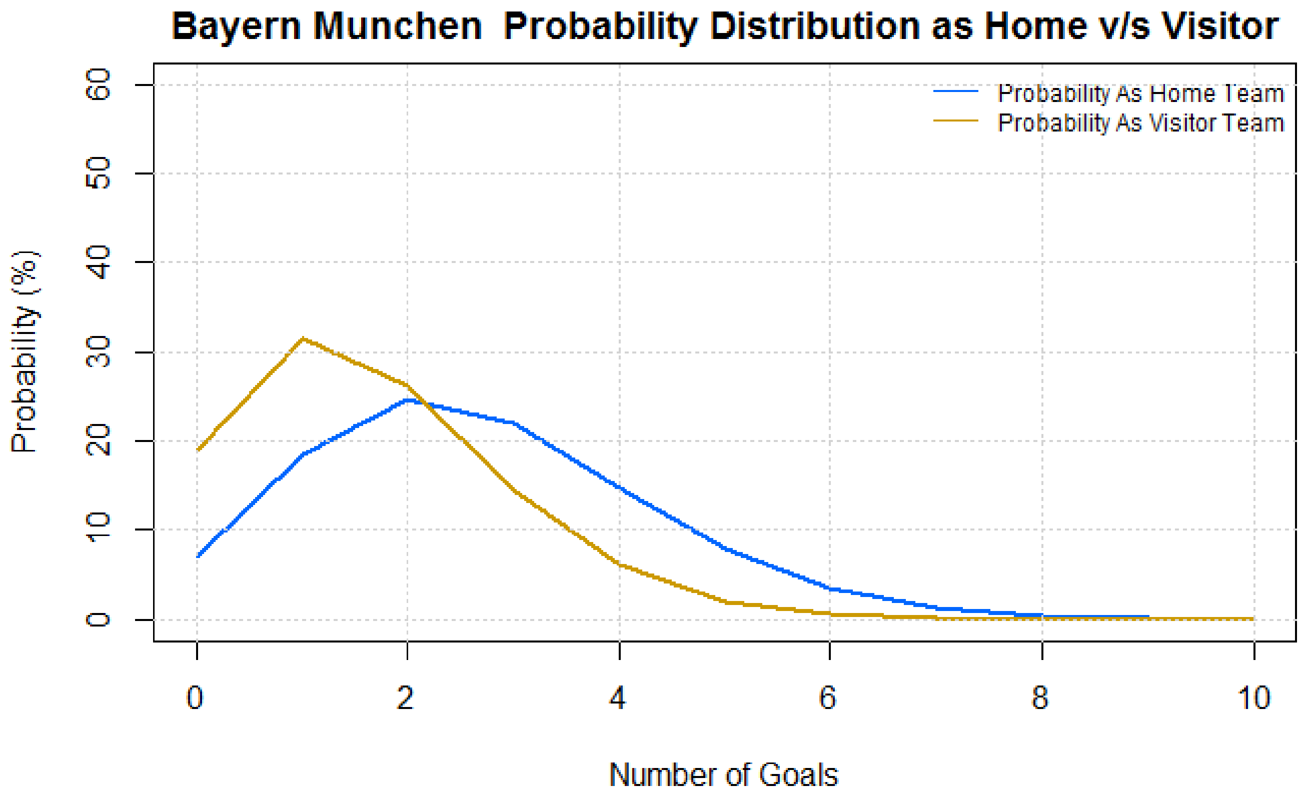
Empirical Data Distribution (As Home vs As Visitor)

Hide

```
Home_Team<-"Bayern Munchen"
plot_df_practical<-data.frame(c(0:10))
for (i in 1:11){
  plot_df_practical[i,2]<-round(dpois(i-1,home_mean)*100,2)
}
for (i in 1:11){
  plot_df_practical[i,3]<-round(dpois(i-1,visitor_mean)*100,2)
}
names(plot_df_practical)<-c("Goals","Probability of Home Team Scoring","Probability of Visito
r Team Scoring")
#plot.new()
plot(plot_df_practical$Goals,plot_df_practical$`Probability of Home Team
Scoring`,ylim=c(0,60),ylab="Probability (%)",xlab="Number of Goals",type="l",col="#0066ff",lw
d=2,main=paste(Home_Team," Probability Distribution as Home v/s Visitor"))
lines(plot_df_practical$Goals,plot_df_practical$`Probability of Visitor Team
Scoring`,ylim=c(0,60),ylab="Probability (%)",xlab="Number of Goals",col="#cc9900",lwd=2)
```

Hide

```
legend('topright', c("Probability As Home Team","Probability As Visitor Team") ,
   lty=1, col=c("#0066ff","#cc9900"), bty='n', cex=.75)
grid()
```



# Insights

Majorly 2 insights can be derived from the above plot,

1. Bayern Munchen has more probability of scoring goals under 2 being a visitor team rather being as home team.
2. As the number of goals crosses the 2 mark, Bayern Munchen has greater probability of scoring goals being a home team.