GROUP 3.5USTU

PROJECT PART2

R PROJECT PART2

TASK 1

In this part of the Project, we first filtered the matches regarding the seasons they were played. Starting from the 1th of July 2010 till the 1th of July 2019, we filtered the data into 9 seasons. Since there were different football club names for the same club, we filtered the names as well. We removed NA cells that cause misleading the results.

We take the difference between HomeGoal and AwayGoal which are dependent to each other so we made our calculations based on paired observation and since the distribution of sample is not known we would use T distribution and take (number of matches-1) as degrees of freedom however, since the degrees of freedom is so high, with Central Limit Theorem we used normal distribution. We calculated the mean of the difference of number of home and away goals for 9 seasons. Then we calculated the standard deviation of the difference of number of home and away goals by using sample standard deviation and dividing it by the squared number of sample size (number of matches) for 9 seasons.

Our Null Hypothesis is

$$\mathbf{H_0: \mu = 0.}$$

It means that the mean of the distribution that we obtained by taking the difference of the HomeGoal and AwayGoal is zero.

And our Alternative Hypothesis is

$$\mathbf{H_1: \mu > 0.}$$

It means that the mean of the distribution that we obtained by taking the difference of the HomeGoal and AwayGoal is more than zero.

We calculated the p value by using "pnorm" command in R and since this command calculates cdf, we subtract the value from 1 since we analyze the part larger than the part which is calculated.

By using P-value approach, for all of the seasons calculated probabilities were smaller than the significant value (α=0.1), we rejected all the null hypotheses.

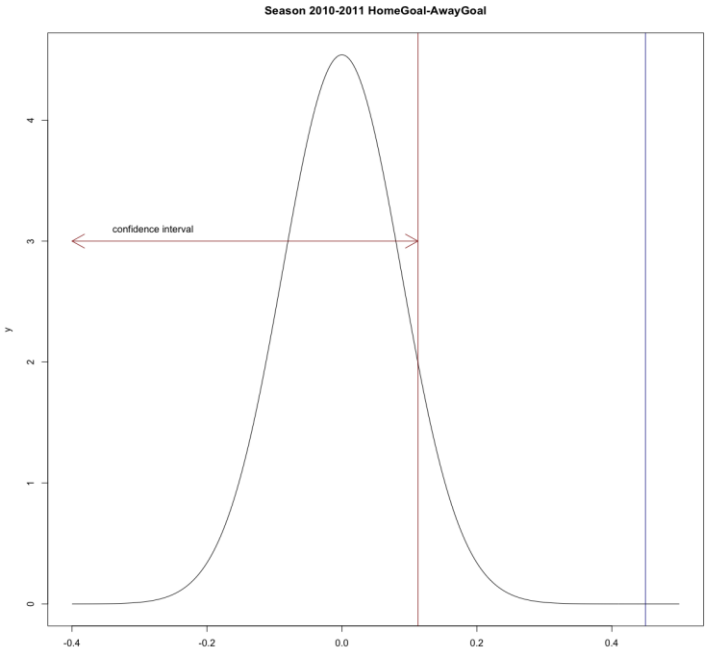

Figure 1: Graph of home and away goal difference

Confidence interval for season 2010-2011

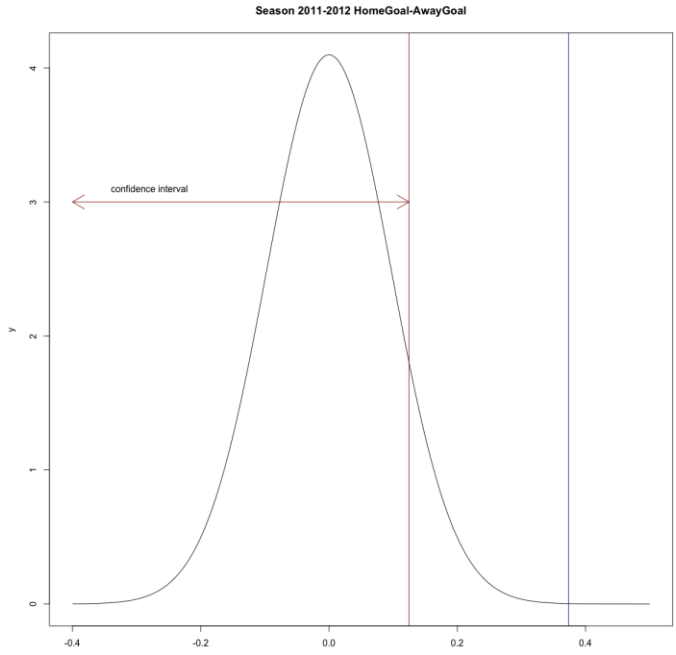

Figure 2: Graph of home and away goal difference

Confidence interval for season 2011-2012

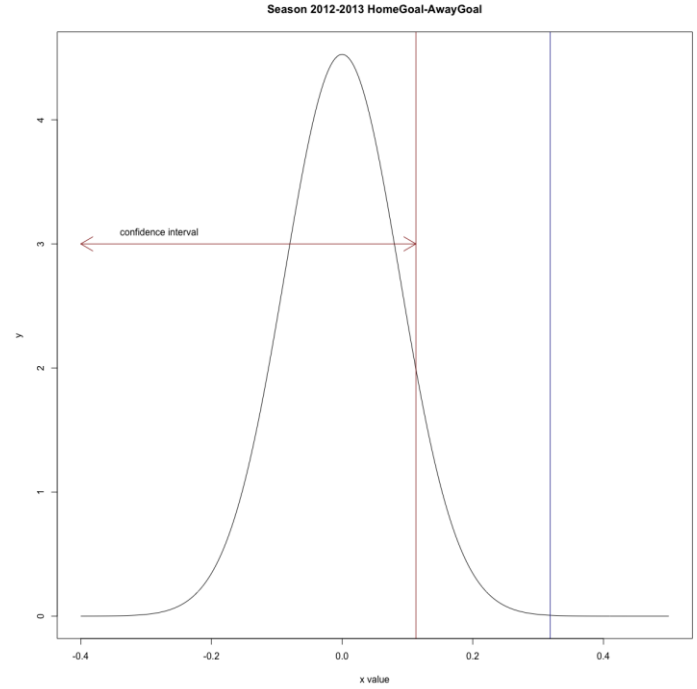

Figure 3: Graph of home and away goal difference

Confidence interval for season 2012-2013

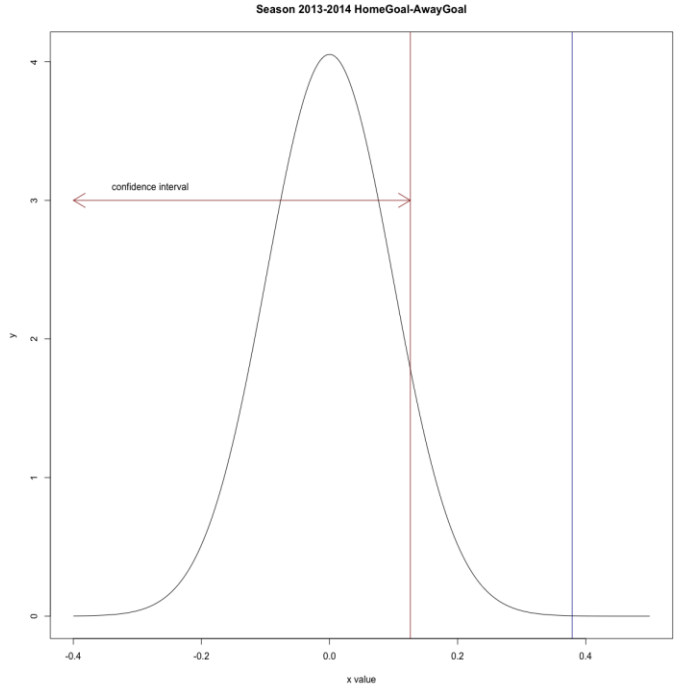

Figure 4 : Graph of home and away goal difference

Confidence interval for season 2013-2014

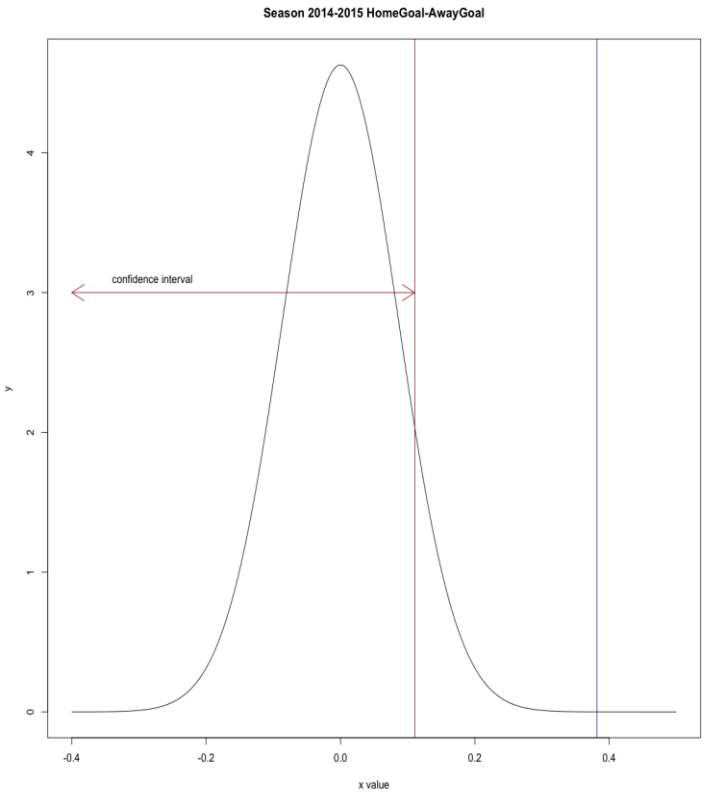

Figure 5 : Graph of home and away goal difference

Confidence interval for season 2014-2015

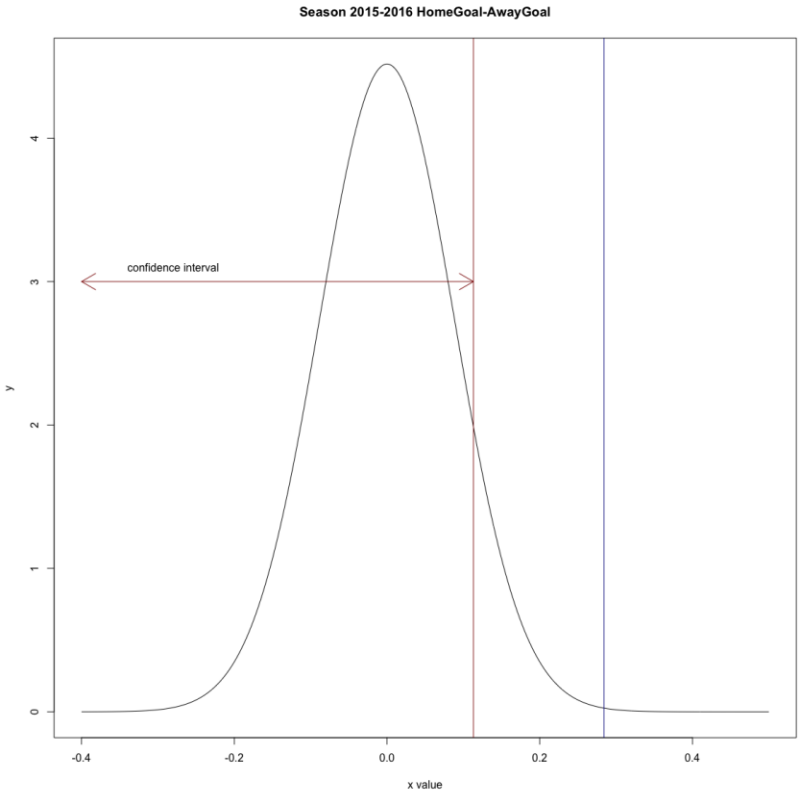

Figure 6 : Graph of home and away goal difference

Confidence interval for season 2015-2016

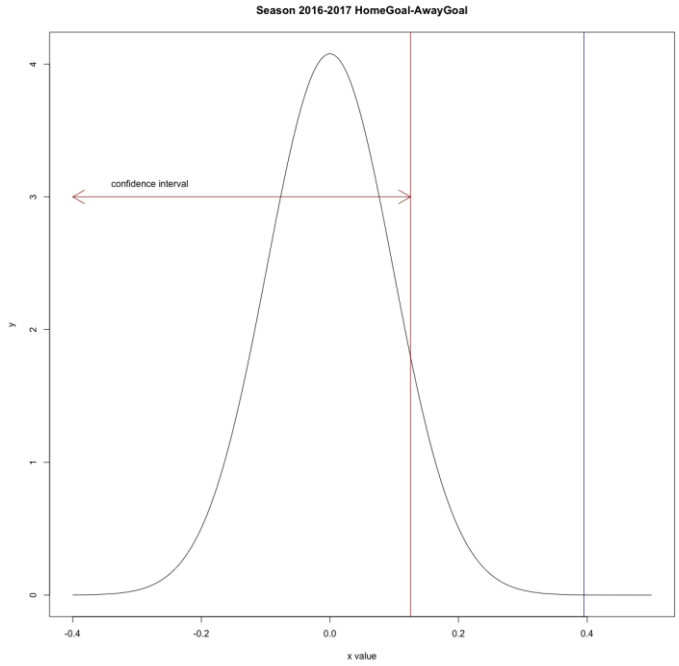

Figure 7 : Graph of home and away goal difference

Confidence interval for season 2016-2017

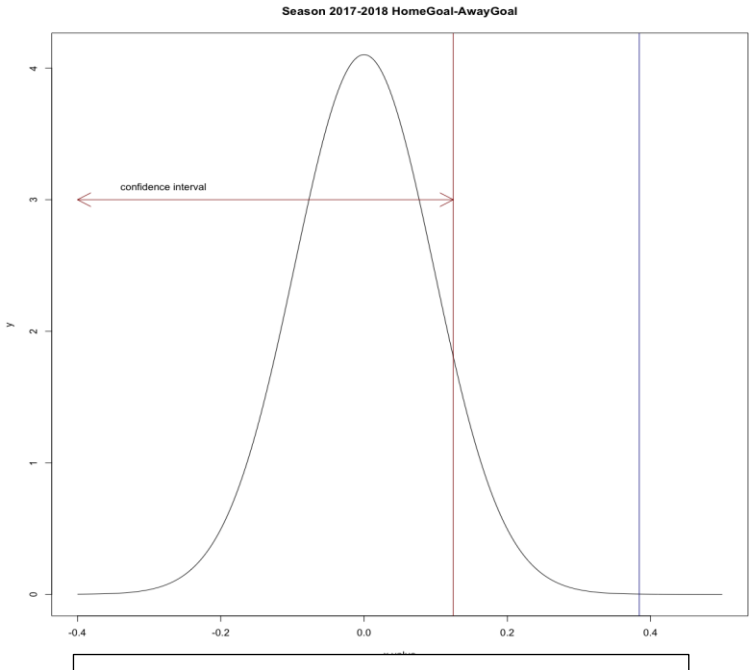

Figure 8: Graph of home and away goal difference

Confidence interval for season 2017-2018

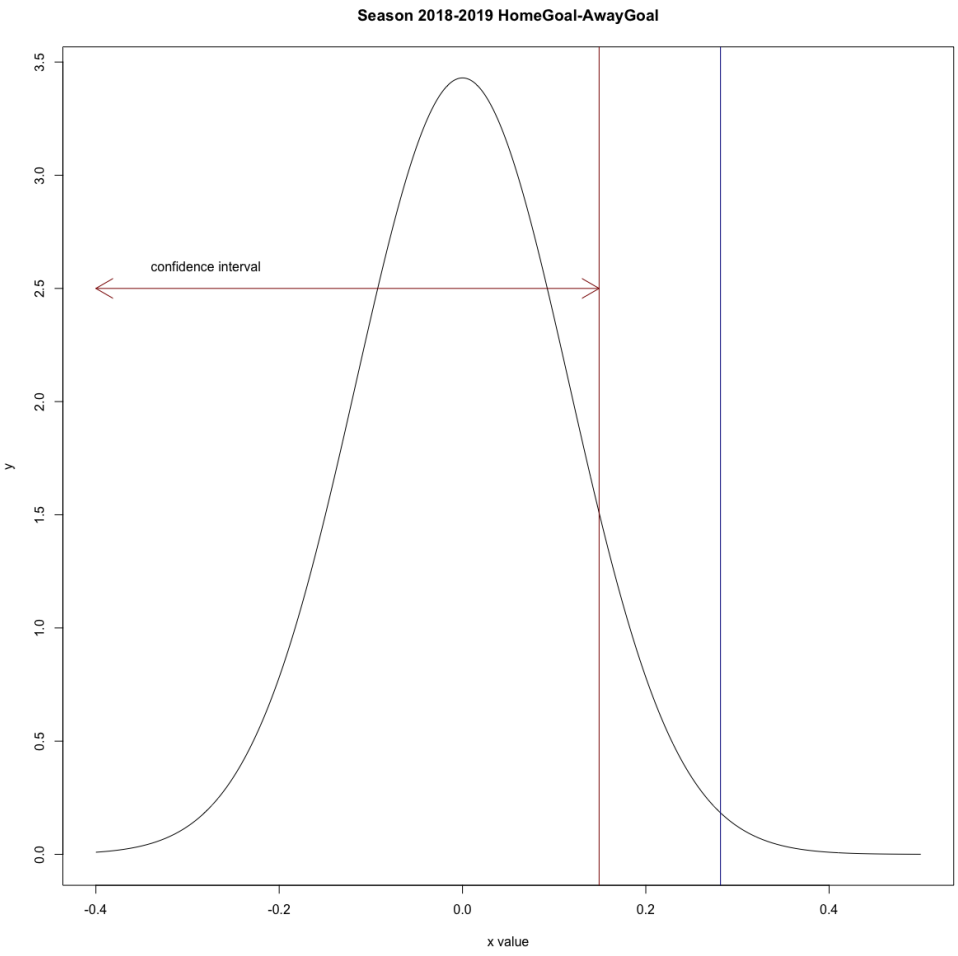

Figure 9 : Graph of home and away goal difference

Confidence interval for season 2018-2019

TASK2

In this part, 5 bookmakers are chosen, namely 888sport, Bet365, 10Bet, BetVictor and Betsafe. First we filtered the 2010 and 2016 years' data for each bookmaker that is chosen and we filtered the odds, however we worked with the latest_odds data.

Then we calculated the relevant probabilities with those odds and after we normalized the calculated probabilities for season 2010. We created a difference column for home goal probabilities and away goal probabilities since they are paired.

We did the same calculations for season 2016. To find the distribution of the differences of probabilities we took the mean as zero (H_0: ($\mu_1 - \mu_2 = 0$)). Then we calculated the standard deviation of the probability differences and then divide the result by the square root of sample size (number of matches).

Since we did not know the distribution of the sample we used T distribution and by using "pt" command we found a probability value, however; since "pt" command calculate cdf, we subtract the value form 1 and found the p value.

$d_{diff} = \mu_1 - \mu_2$

Our null hypothesis is

H_0: $d_{diff} = 0$.

It means that the differences of the mean of probabilities of the home goals and away goals is equal to zero.

And our Alternative Hypothesis is

H_1: $d_{diff} \neq 0$.

It means that the differences of the mean of probabilities of the home goals and away goals are not equal to zero.

We reject the null hypothesis since the p value we found is less than $\alpha/2 = 0.05/2$. We used $\alpha/2$ since we analyzed two sided interval. We rejected all the null hypotheses by using the P-value

approach, calculated probabilities in 2010 and 2016 for all bookmakers were smaller than the significant value ($\alpha/2 = 0.05/2$) .

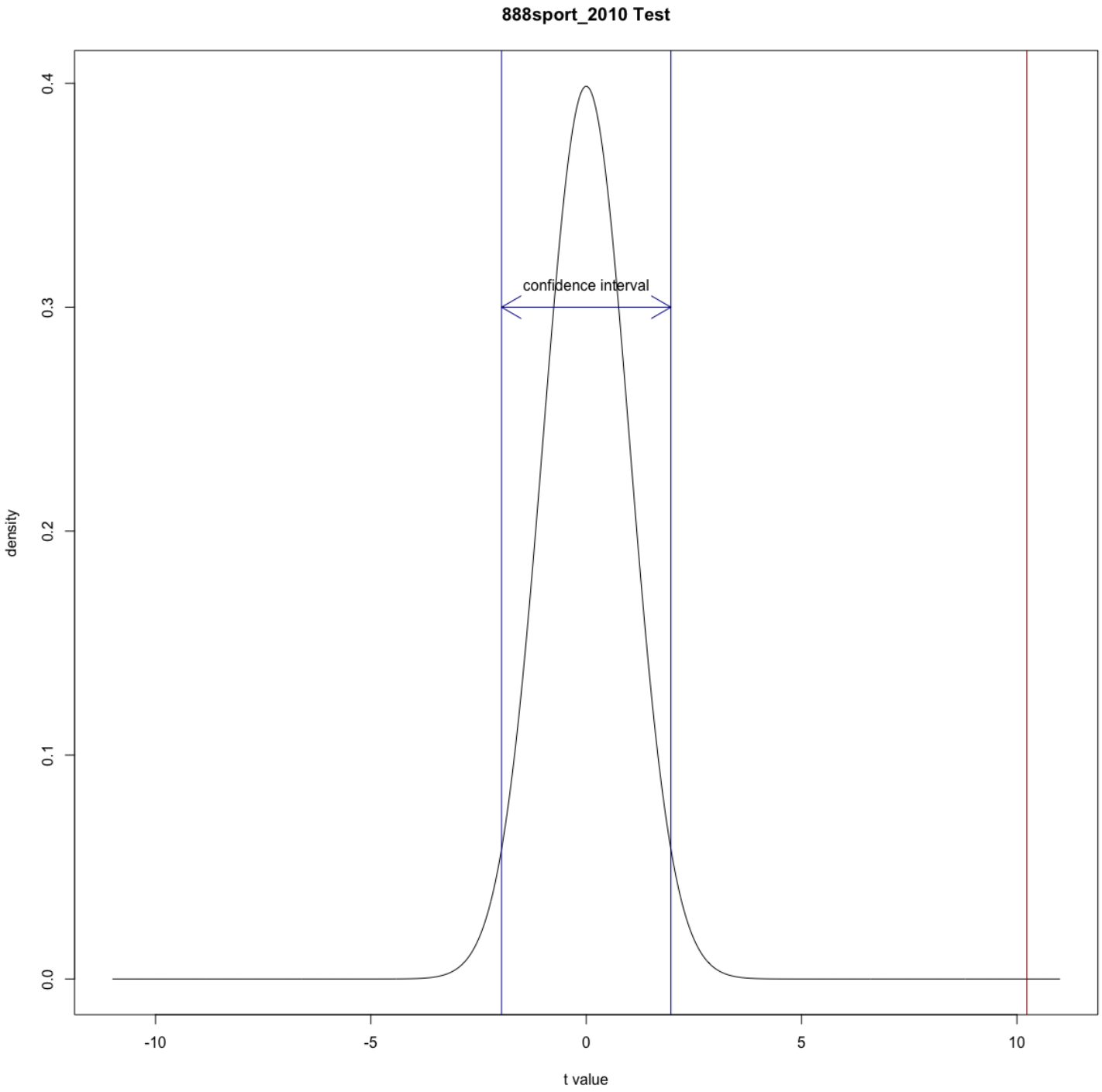

Figure 10 : Graph of confidence interval and calculated value of 888Sport for 2010 season

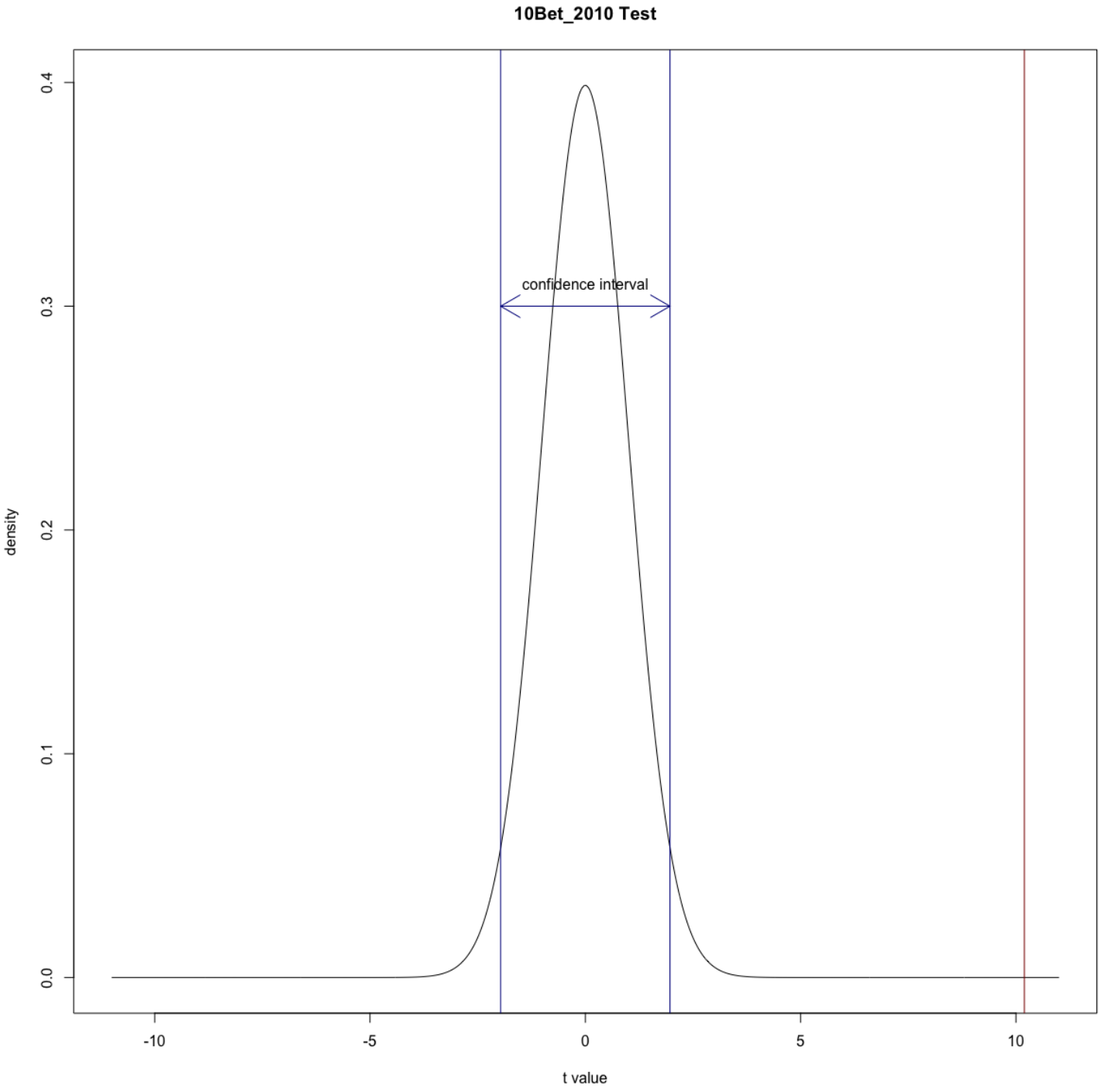

Figure 11 : Graph of confidence interval and calculated value of 10Bet for 2010 season

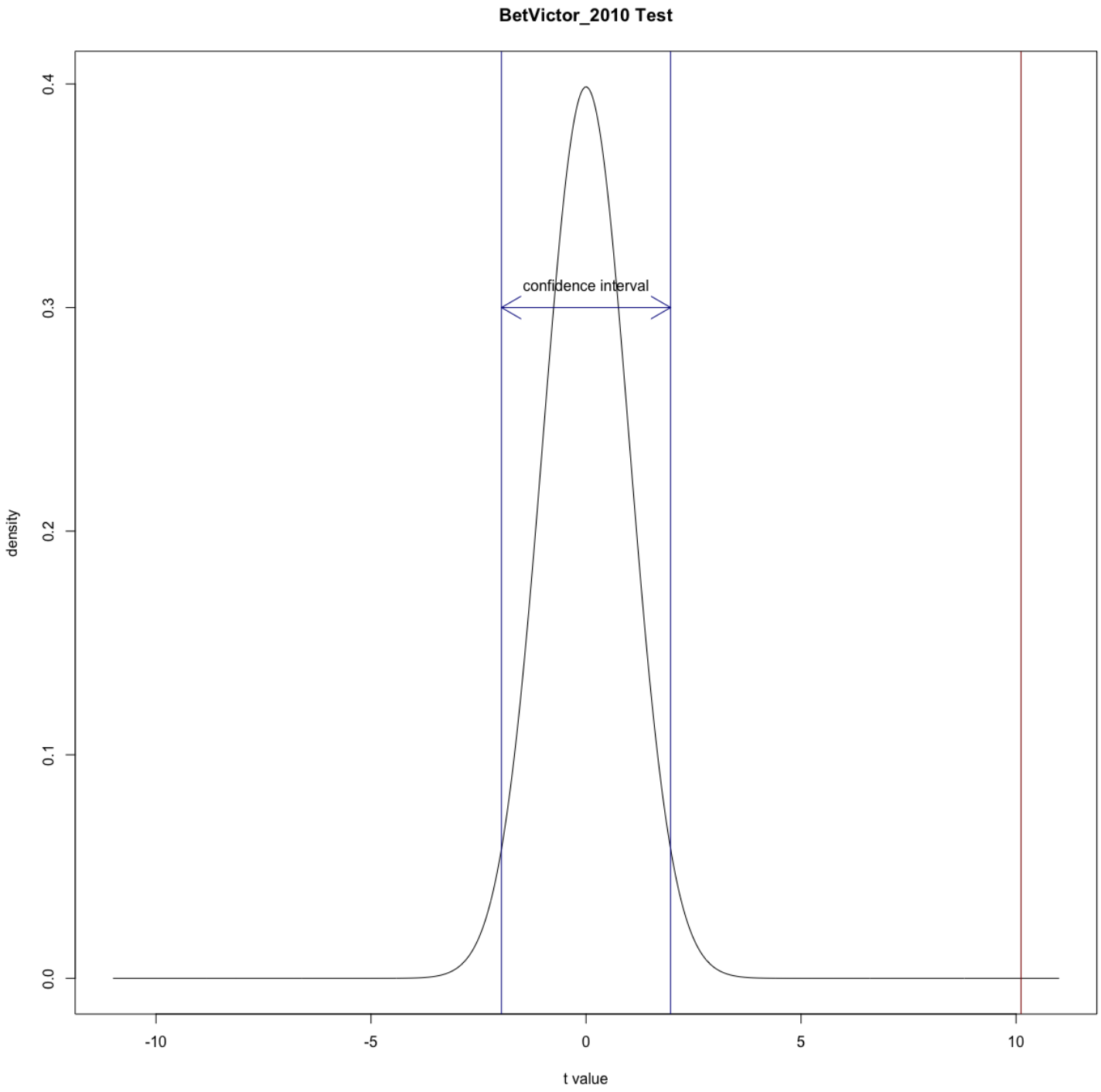

Figure 12 : Graph of confidence interval and calculated value of BetVictor for 2010 season

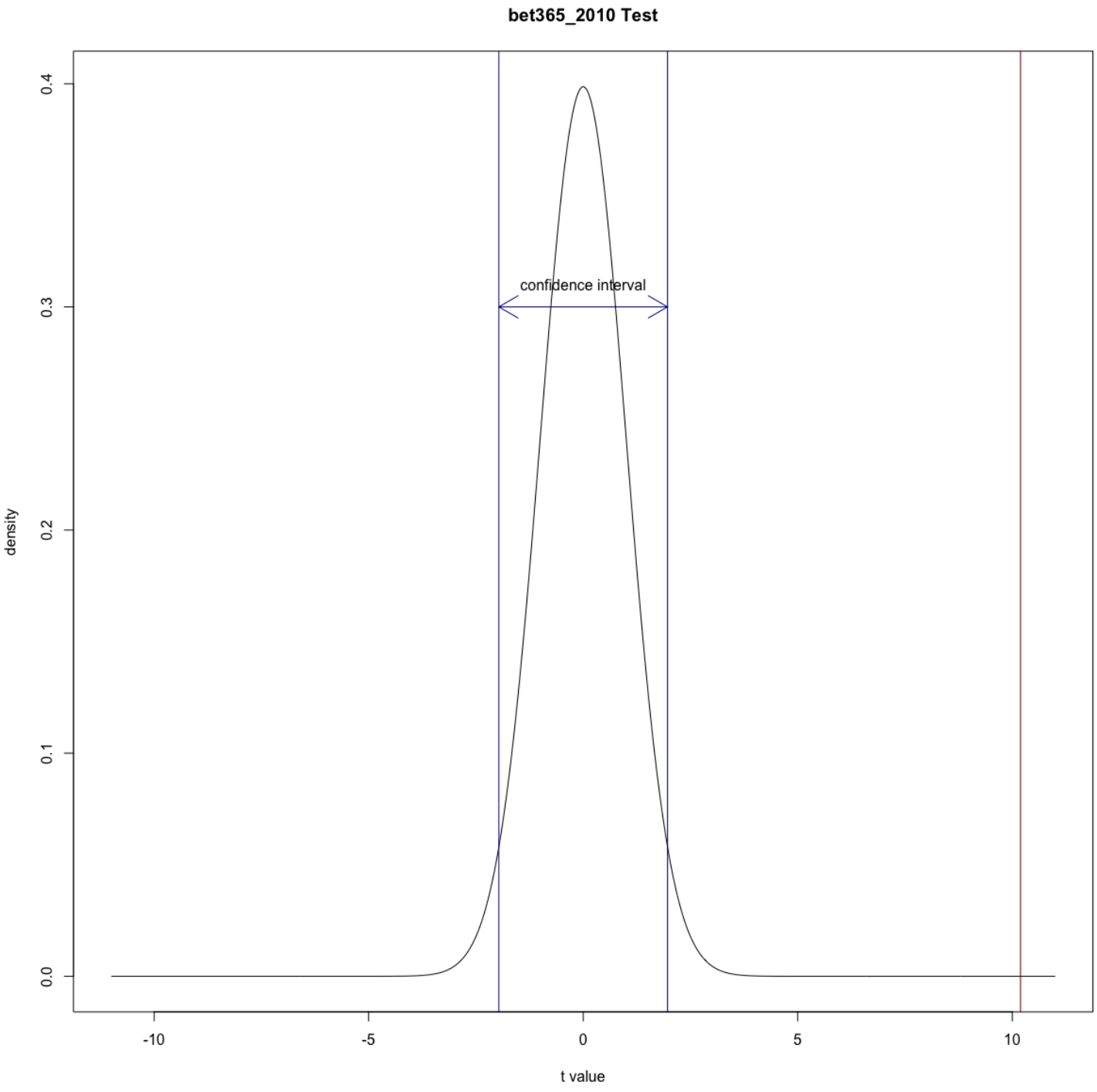

Figure 13 : Graph of confidence interval and calculated value of bet365 for 2010 season

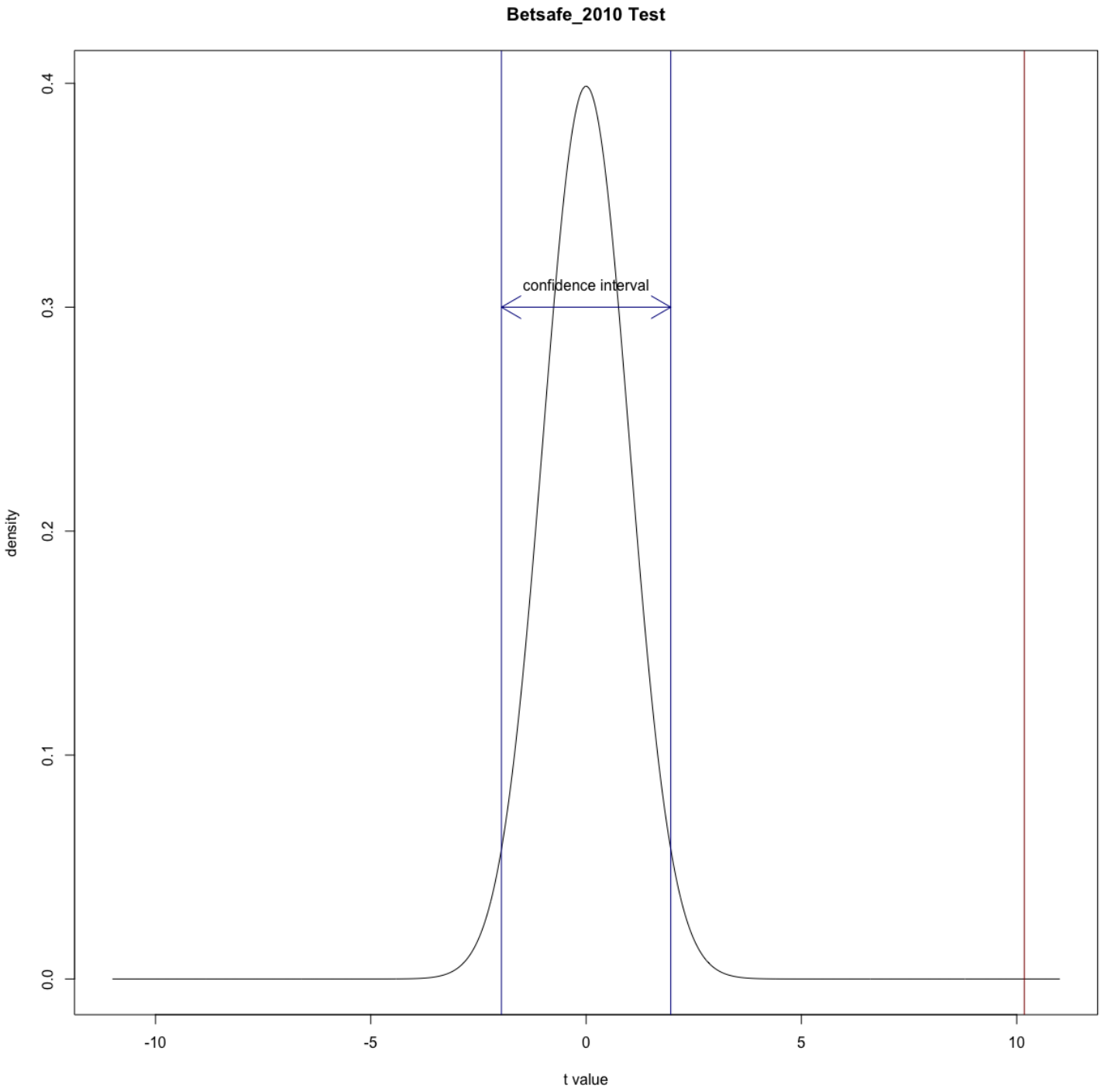

Figure 14 : Graph of confidence interval and calculated value of Betsafe for 2010 season

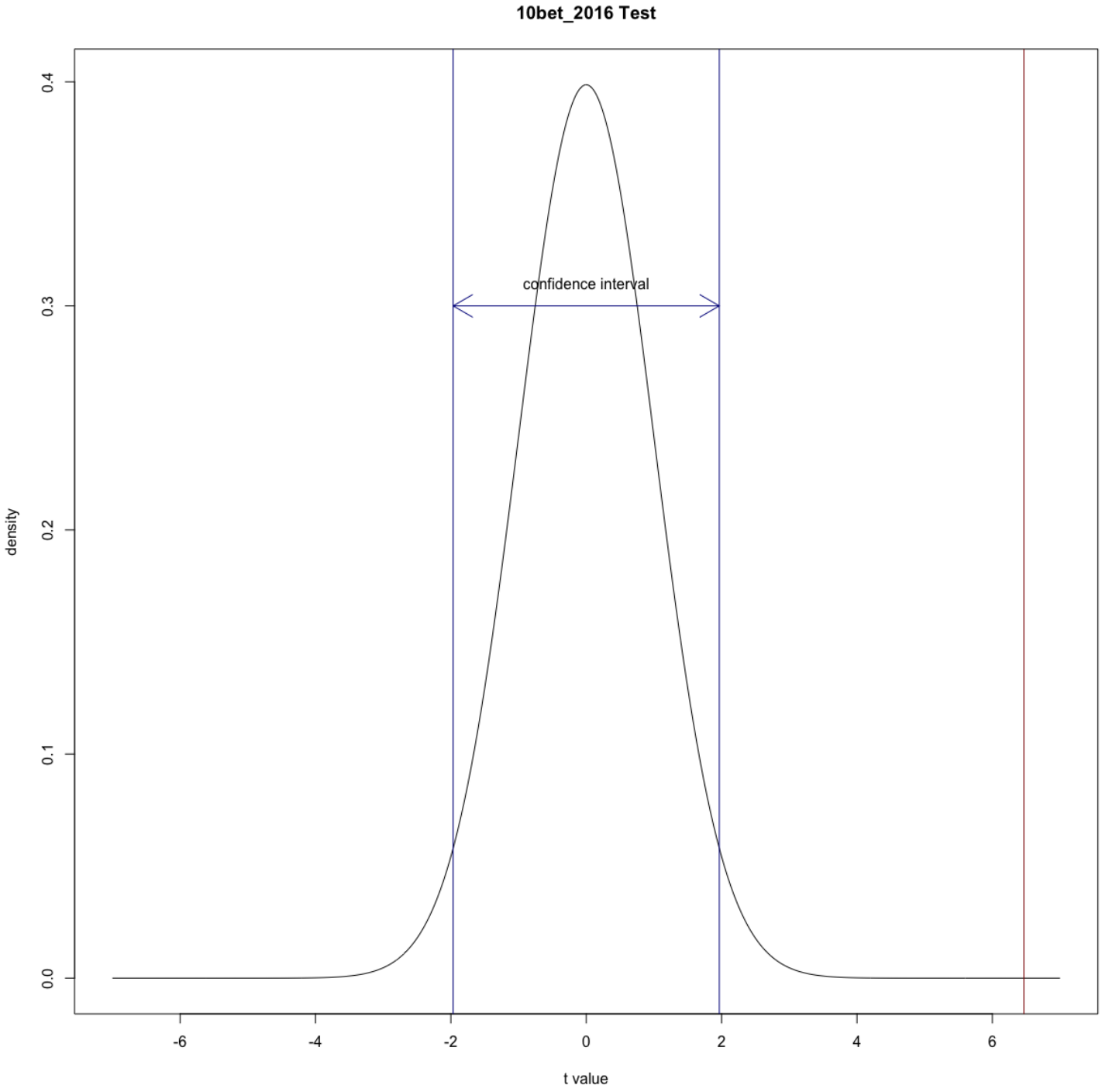

Figure 15 : Graph of confidence interval and calculated value of 10Bet for 2016 season

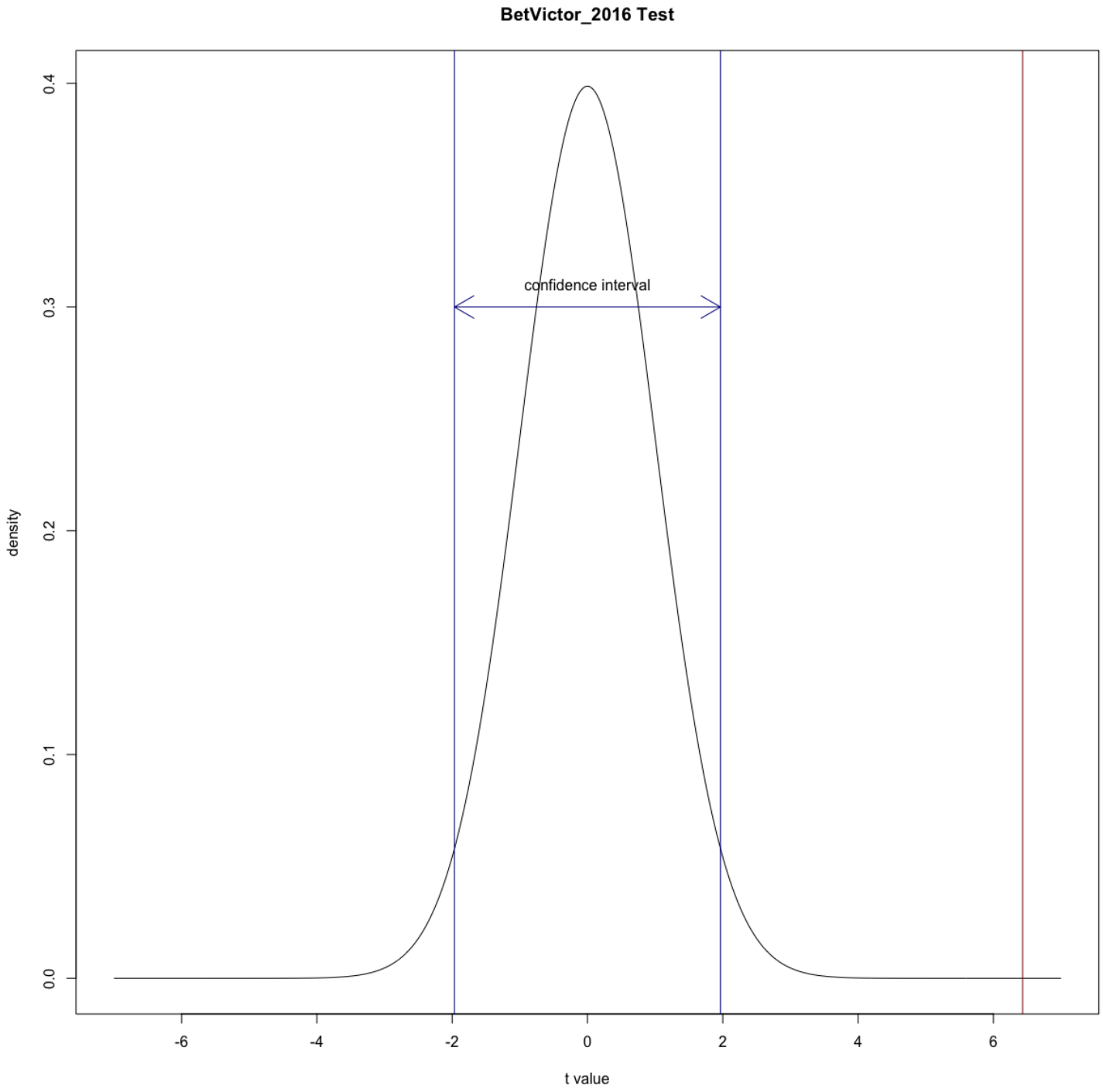

Figure 16 : Graph of confidence interval and calculated value of BetVictor for 2016 season

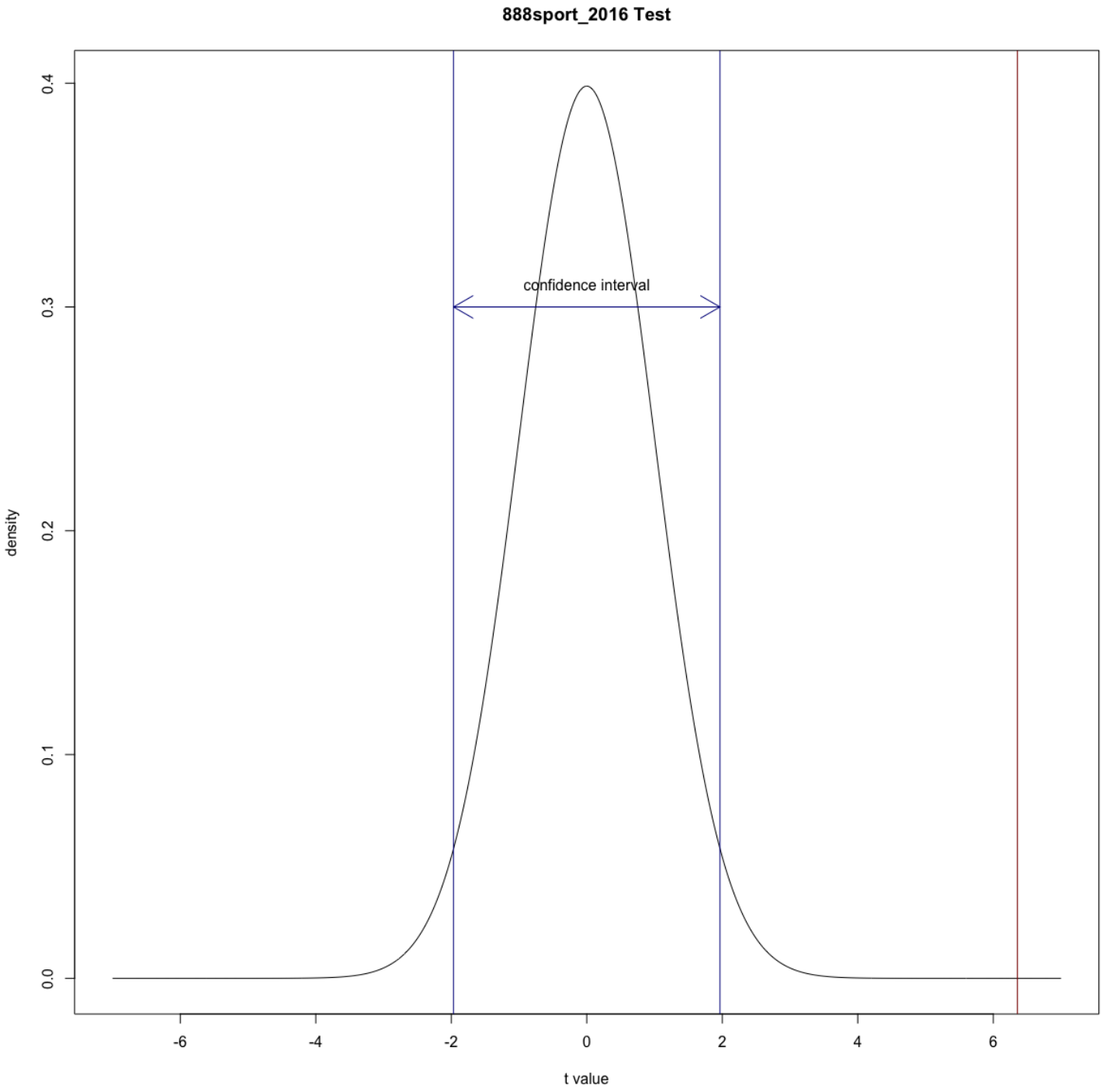

Figure 17 : Graph of confidence interval and calculated value of 88sport for 2016 season

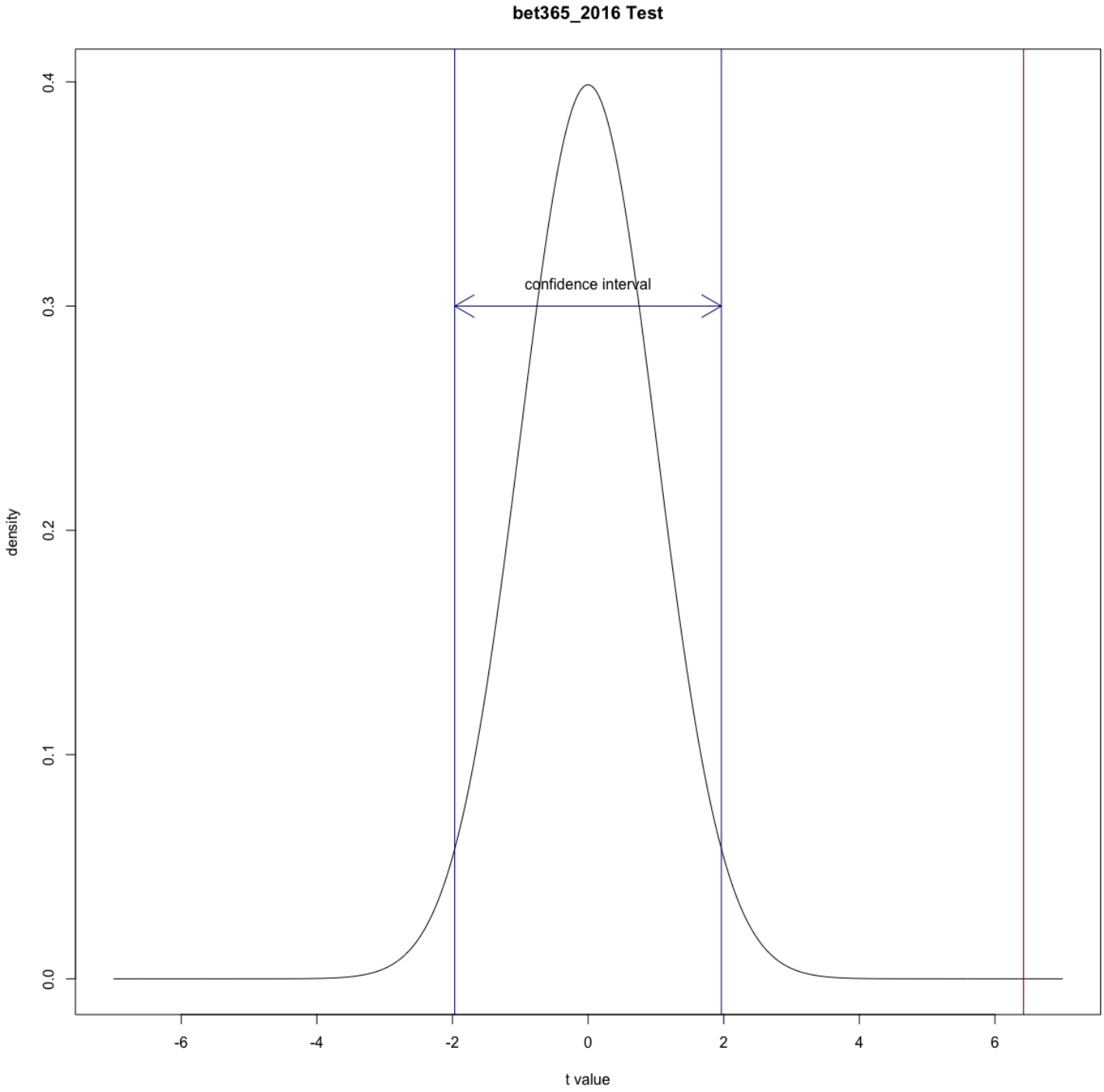

Figure 18 : Graph of confidence interval and calculated value of bet365 for 2016 season

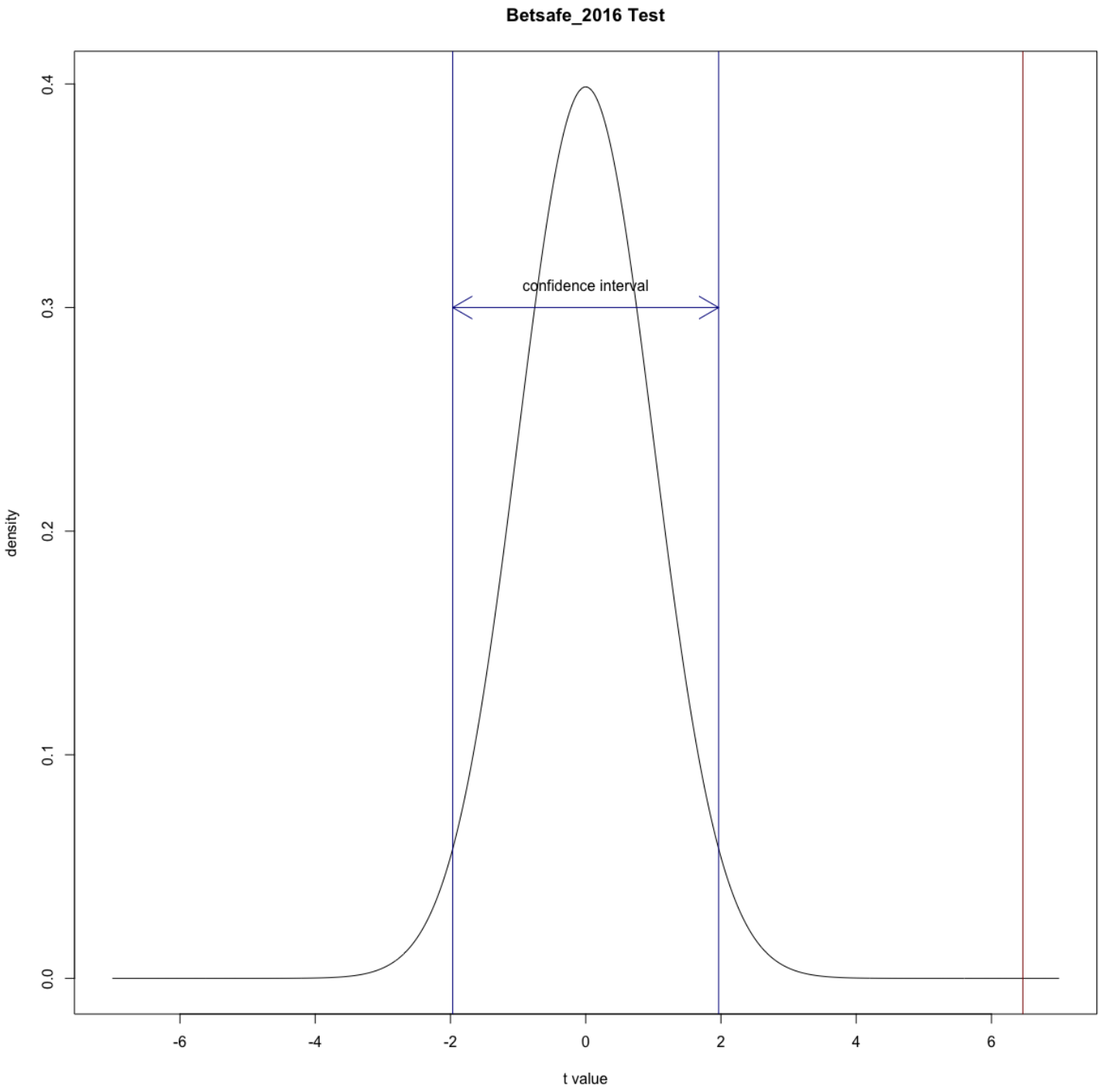

Figure 19 : Graph of confidence interval and calculated value of Betsafe for 2016 season

TASK3

In this part, using the column of AwayGoal we obtained in the task 1, we calculated the standard deviation of the away goals in season 2010 and we did the sample operations for season 2015.

Our null hypothesis is

H_0: $\sigma_1^2 / \sigma_2^2 = 1$

Our alternative hypothesis is

H_1: $\sigma_1^2 / \sigma_2^2 \neq 1$

We used f distribution to calculate probabilities since we can manipulate σ_1^2 / σ_2^2 by multiplying with S_2^2 / S_1^2.

$P(\sigma_1^2 / \sigma_2^2 < 1) = P((\sigma_1^2 S_2^2 / \sigma_2^2 S_1^2) < S_2^2 / S_1^2)$

$P((\sigma_1^2 S_2^2 / \sigma_2^2 S_1^2) < S_2^2 / S_1^2) = P(\sigma_2^2 S_1^2 / \sigma_1^2 S_2^2 > S_1^2 / S_2^2)$

Degrees of freedom:

V1= (nrow(season_2010)-1)

V2= (nrow(season_2015)-1)

Then we calculate the probability for one side by using “pf” command. We found p value=0.029.

For α=0.01, α/2=0.005. We fail to reject null hypothesis since p value (0.029) line is in the confidence interval.

For α=0.1, α/2=0.05. We reject null hypothesis since p value is outside of the confidence interval.

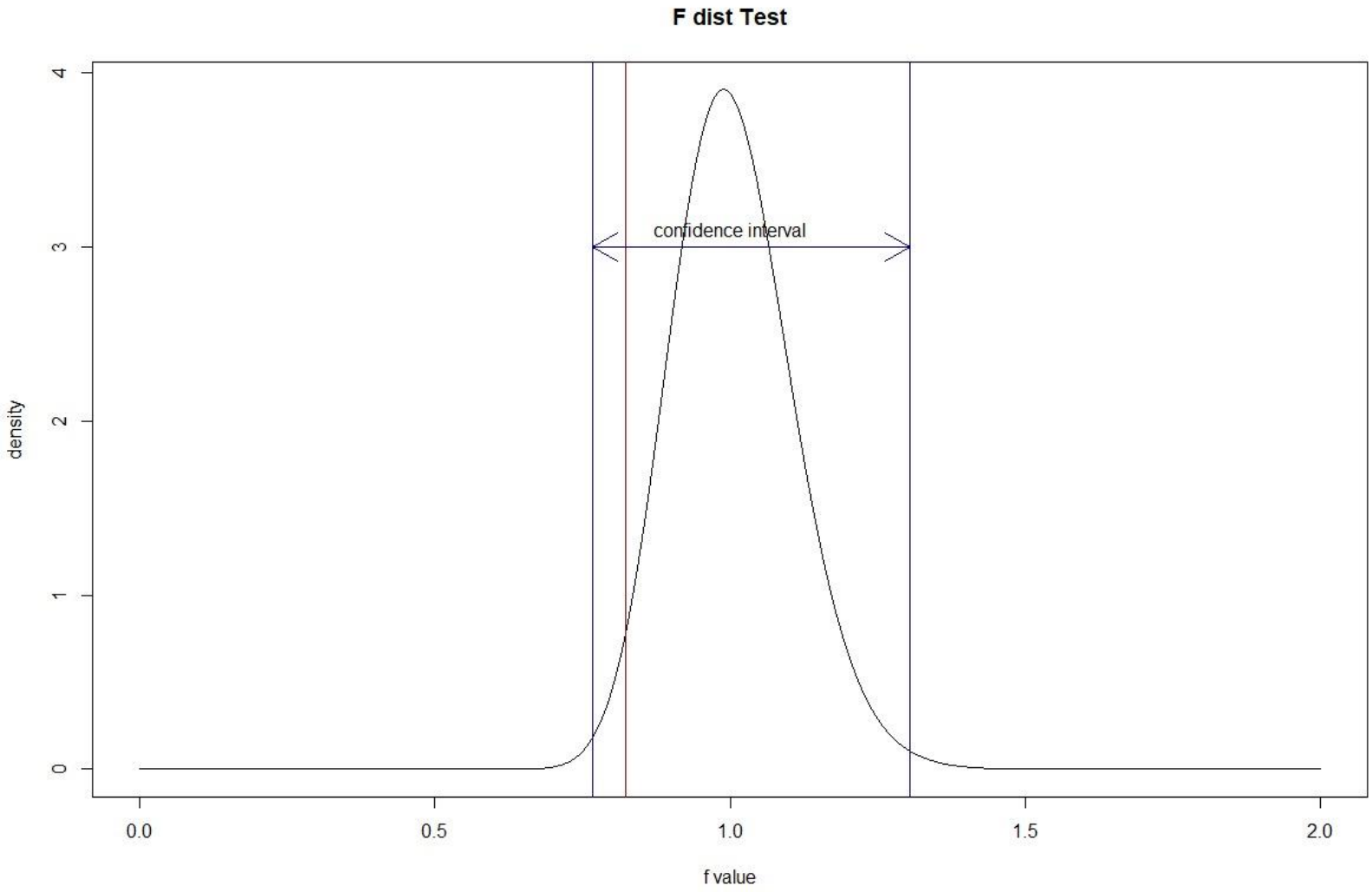

Figure 20: Confidence interval for fail to reject and p value line

For α=0.01 We fail to reject null hypothesis.

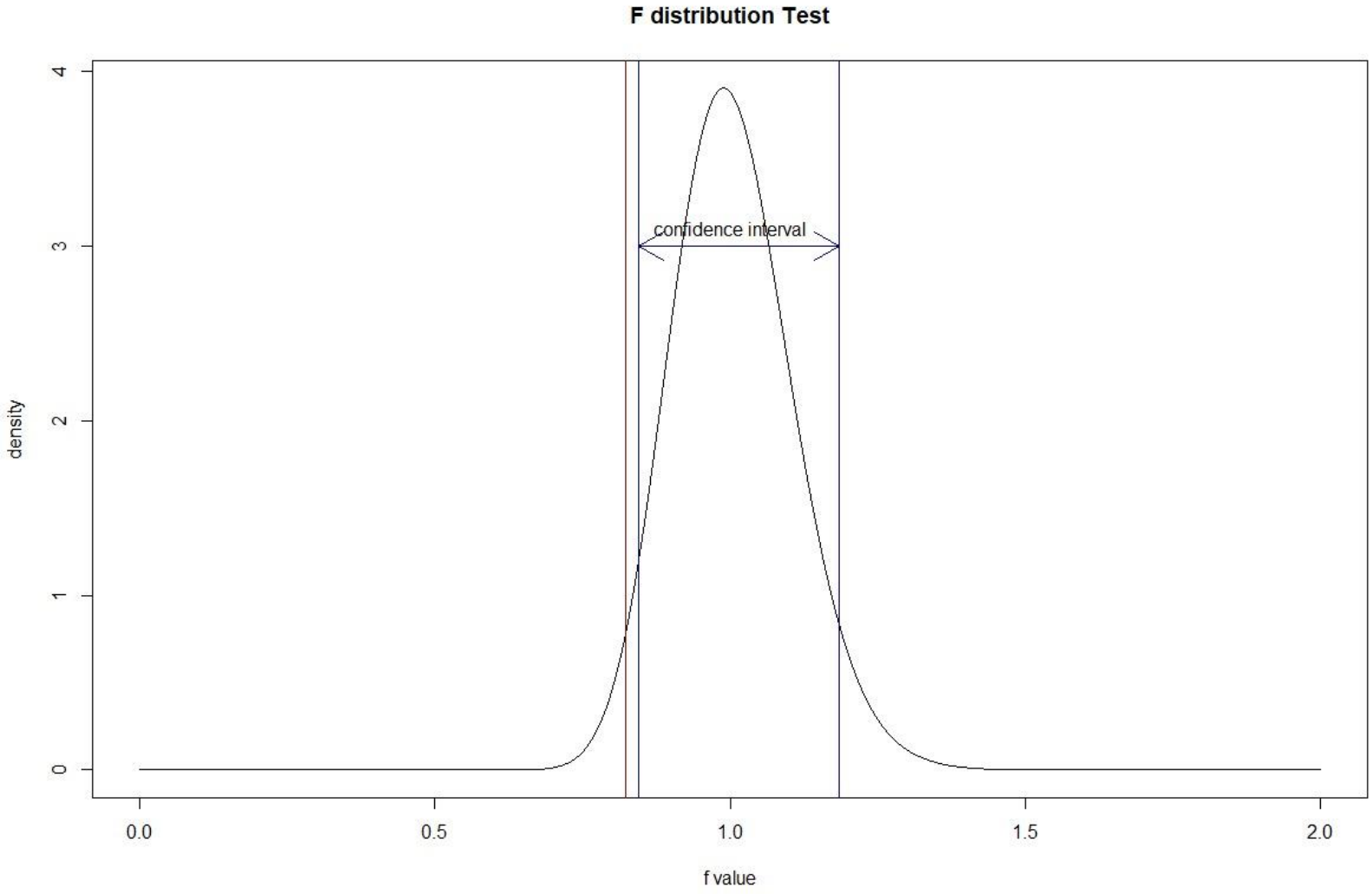

Figure 21: Confidence interval for fail to reject and p value line

a=0.1 We reject null hypothesis.

TASK4

In this part, we used the data from our previous work. In previous part, we calculated the means of HomeGoal and AwayGoal, in this part we took these values as the lambdas of poisson distributions.

Next, we calculated the expected values with our theoretical poisson distribution. Then, in order to apply Goodness of Fit Test we take the difference between HomeGoal and HomeGoal_pois and then take the square of it, divide by the expected value and summed them. We performed these steps in for loops. We considered this summation as calculated value.

Our null hypothesis and alternative hypothesis for HomeGoal:

$H_{0:}$ The distribution of HomeGoal resembles to Poisson distribution

$H_{1:}$ The distribution of HomeGoal does not resemble to Poisson distribution

We used the Critical Value Approach.

With $\alpha=0.05$ we calculated the critical value as 15.50731 for degrees of freedom 8 and the calculated value as 23.14755. Since the calculated value is larger than the critical value, we rejected the H_0

Our null hypothesis and alternative hypothesis for AwayGoal:

$H_{0:}$ The distribution of AwayGoal resembles to Poisson distribution

$H_{1:}$ The distribution of AwayGoal does not resemble to Poisson distribution

With $\alpha=0.05$ we calculated the critical value as 14.06714 for degrees of freedom 7 and the p value as 37.80014. Since the calculated value is larger than the critical value, we rejected the $H_{0.}$

.

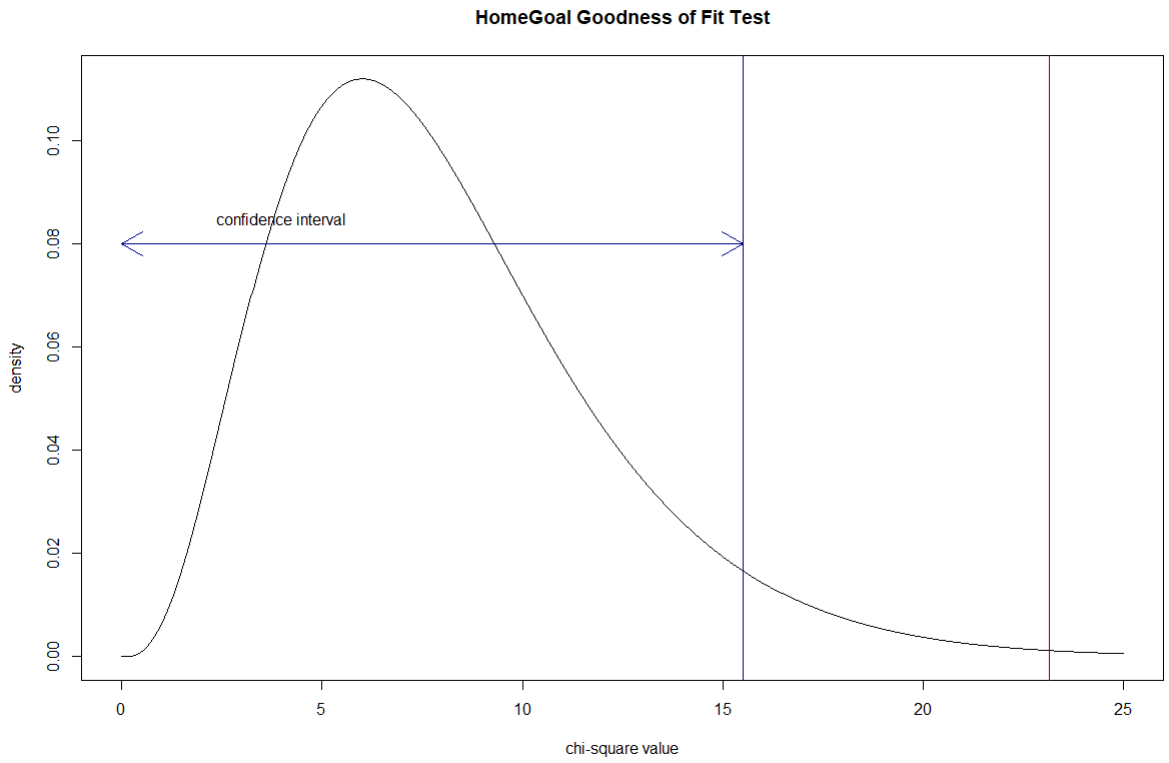

Figure 22: Confidence interval and calculated value line for HomeGoal

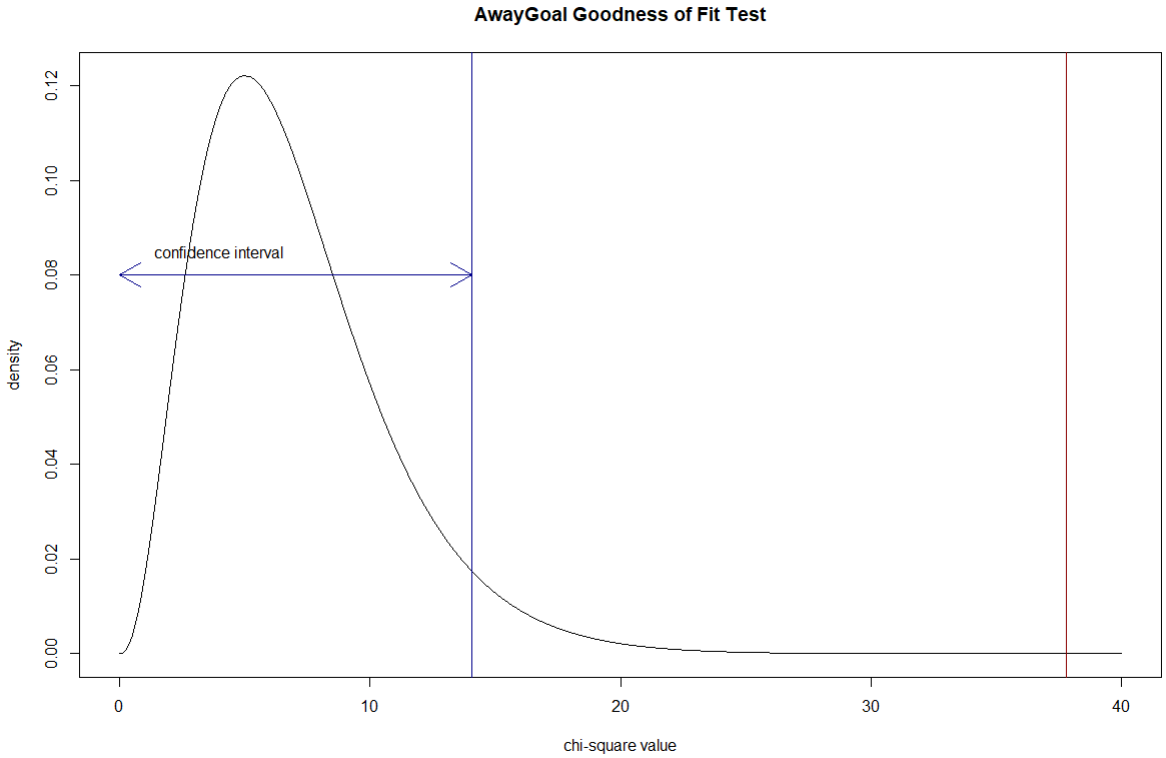

Figure 23: Confidence interval and calculated value line for AwayGoal

R codes for Plotting Figures i. Code for Figure 1:

```
library(gridExtra)
library(data.table)
library(lubridate)
setwd("/Users/enesozeren/Downloads/")
matches=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_matches.rds')
odds=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_odd_details.rds')

matches[,c('HomeGoal','AwayGoal'):=tstrsplit(score,':')]
matches[,HomeGoal:=as.numeric(HomeGoal)]
matches[,AwayGoal:=as.numeric(AwayGoal)]
matches[,leagueId:=NULL]
matches[,Difference:=HomeGoal-AwayGoal]
#Tarih
matches[,Date:=as_datetime(date,tz='Turkey')]
matches[,date:=NULL]
matches[,type:=NULL]

#geç
matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
matches[home=="crystal-palace"]$home="crystal palace"
matches[home=="stoke"]$home="stock city"
sort(unique(matches$home))
matches[away=="manchester-united"]$away="manchester united"
```

```
matches[away=="manchester-utd"]$away="manchester united"
matches[away=="manchester utd"]$away="manchester united"
matches[away=="manchester-city"]$away="manchester city"
matches[away=="west-ham"]$away="west ham"
matches[away=="crystal-palace"]$away="crystal palace"
matches[away=="stoke"]$away="stock city"
sort(unique(matches$away))


#remove NA
temp=matches[,list(Difference, Date, matchId)]
temp<-temp[complete.cases(temp),]

#Sezonlara göre Maçlar
season_2010_2011=temp[Date>'2010-07-01'& Date<'2011-07-01']
season_2011_2012=temp[Date>'2011-07-01'& Date<'2012-07-01']
season_2012_2013=temp[Date>'2012-07-01'& Date<'2013-07-01']
season_2013_2014=temp[Date>'2013-07-01'& Date<'2014-07-01']
season_2014_2015=temp[Date>'2014-07-01'& Date<'2015-07-01']
season_2015_2016=temp[Date>'2015-07-01'& Date<'2016-07-01']
season_2016_2017=temp[Date>'2016-07-01'& Date<'2017-07-01']
season_2017_2018=temp[Date>'2017-07-01'& Date<'2018-07-01']
season_2018_2019=temp[Date>'2018-07-01'& Date<'2019-07-01']


#calculating the sd and mean of seasons
sd_of_season_2010_2011=sd(season_2010_2011$Difference)/sqrt(nrow(season_2010_2011))
mean_of_season_2010_2011=mean(season_2010_2011$Difference)
```

```
sd_of_season_2011_2012=sd(season_2011_2012$Difference)/sqrt(nrow(season_2011_2012))
mean_of_season_2011_2012=mean(season_2011_2012$Difference)

sd_of_season_2012_2013=sd(season_2012_2013$Difference)/sqrt(nrow(season_2012_2013))
mean_of_season_2012_2013=mean(season_2012_2013$Difference)

sd_of_season_2013_2014=sd(season_2013_2014$Difference)/sqrt(nrow(season_2013_2014))
mean_of_season_2013_2014=mean(season_2013_2014$Difference)

sd_of_season_2014_2015=sd(season_2014_2015$Difference)/sqrt(nrow(season_2014_2015))
mean_of_season_2014_2015=mean(season_2014_2015$Difference)

sd_of_season_2015_2016=sd(season_2015_2016$Difference)/sqrt(nrow(season_2015_2016))
mean_of_season_2015_2016=mean(season_2015_2016$Difference)

sd_of_season_2016_2017=sd(season_2016_2017$Difference)/sqrt(nrow(season_2016_2017))
mean_of_season_2016_2017=mean(season_2016_2017$Difference)

sd_of_season_2017_2018=sd(season_2017_2018$Difference)/sqrt(nrow(season_2017_2018))
mean_of_season_2017_2018=mean(season_2017_2018$Difference)

sd_of_season_2018_2019=sd(season_2018_2019$Difference)/sqrt(nrow(season_2018_2019))
mean_of_season_2018_2019=mean(season_2018_2019$Difference)

#graphs of normal distributions of seasons
pval_2010_2011=1-pnorm(mean_of_season_2010_2011,mean = 0,sd =
sd_of_season_2010_2011)
```

```
pval_2011_2012=1-pnorm(mean_of_season_2011_2012,mean = 0,sd = sd_of_season_2011_2012)
pval_2012_2013=1-pnorm(mean_of_season_2012_2013,mean = 0,sd = sd_of_season_2012_2013)
pval_2013_2014=1-pnorm(mean_of_season_2013_2014,mean = 0,sd = sd_of_season_2013_2014)
pval_2014_2015=1-pnorm(mean_of_season_2014_2015,mean = 0,sd = sd_of_season_2014_2015)
pval_2015_2016=1-pnorm(mean_of_season_2015_2016,mean = 0,sd = sd_of_season_2015_2016)
pval_2016_2017=1-pnorm(mean_of_season_2016_2017,mean = 0,sd = sd_of_season_2016_2017)
pval_2017_2018=1-pnorm(mean_of_season_2017_2018,mean = 0,sd = sd_of_season_2017_2018)
pval_2018_2019=1-pnorm(mean_of_season_2018_2019,mean = 0,sd = sd_of_season_2018_2019)


#standard normal distribution data
x <- seq(-0.4, 0.5, length=1000)
y <- dnorm(x, mean=0, sd=sd_of_season_2010_2011)
alpha_2010_2011 <- qnorm(0.9,mean=0, sd=sd_of_season_2010_2011)
#plot a standard normal distribution
plot(x, y, type="l",lty=1, main="Season 2010-2011 HomeGoal-AwayGoal",xlab="x value")

#plot a vertical line
abline(v= mean_of_season_2010_2011, col='dark blue')
abline(v= alpha_2010_2011, col='dark red')
#make the arrow
arrows(x0=-0.4, y0=3, x1=alpha_2010_2011, y1=3, code=3, col='dark red')
#plot the text
```

```
text(x=-0.28, y=3.1,labels='confidence interval', col='black')



x <- seq(-0.4, 0.5, length=1000)
y <- dnorm(x, mean=0, sd=sd_of_season_2011_2012)
alpha_2011_2012 <- qnorm(0.9,mean=0, sd=sd_of_season_2011_2012)
plot(x, y, type="l",lty=1,  main="Season 2011-2012 HomeGoal-AwayGoal", xlab="x value")

abline(v= mean_of_season_2011_2012, col='dark blue')
abline(v= alpha_2011_2012, col='dark red')

arrows(x0=-0.4, y0=3, x1=alpha_2011_2012, y1=3, code=3, col='dark red')

text(x=-0.28, y=3.1,labels='confidence interval', col='black')



x <- seq(-0.4, 0.5, length=1000)
y <- dnorm(x, mean=0, sd=sd_of_season_2012_2013)
alpha_2012_2013 <- qnorm(0.9,mean=0, sd=sd_of_season_2012_2013)

plot(x, y, type="l",lty=1,  main="Season 2012-2013 HomeGoal-AwayGoal", xlab="x value")
abline(v= mean_of_season_2012_2013, col='dark blue')
abline(v= alpha_2012_2013, col='dark red')
arrows(x0=-0.4, y0=3, x1=alpha_2012_2013, y1=3, code=3, col='dark red')
```

```
text(x=-0.28, y=3.1,labels='confidence interval', col='black')



x <- seq(-0.4, 0.5, length=1000)
y <- dnorm(x, mean=0, sd=sd_of_season_2013_2014)
alpha_2013_2014 <- qnorm(0.9,mean=0, sd=sd_of_season_2013_2014)
plot(x, y, type="l",lty=1,  main="Season 2013-2014 HomeGoal-AwayGoal", xlab="x value")

abline(v= mean_of_season_2013_2014, col='dark blue')
abline(v= alpha_2013_2014, col='dark red')
arrows(x0=-0.4, y0=3, x1=alpha_2013_2014, y1=3, code=3, col='dark red')
text(x=-0.28, y=3.1,labels='confidence interval', co…
[14:49, 1.5.2019] Enes  Özeren: TASK 2
[14:49, 1.5.2019] Enes  Özeren: library(ggplot2)
library(gridExtra)
library(data.table)
library(lubridate)
setwd("/Users/enesozeren/Downloads/")
matches=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_matches.rds')
odds=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_odd_details.rds')



matches[,c('HomeGoal','AwayGoal'):=tstrsplit(score,':')]
matches[,HomeGoal:=as.numeric(HomeGoal)]
matches[,AwayGoal:=as.numeric(AwayGoal)]
matches[,leagueId:=NULL]
matches[,Difference:=HomeGoal-AwayGoal]
```

```
#Tarih
matches[,Date:=as_datetime(date,tz='Turkey')]
matches[,date:=NULL]
matches[,type:=NULL]

#geç
matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
matches[home=="crystal-palace"]$home="crystal palace"
matches[home=="stoke"]$home="stoke city"
sort(unique(matches$home))
matches[away=="manchester-united"]$away="manchester united"
matches[away=="manchester-utd"]$away="manchester united"
matches[away=="manchester utd"]$away="manchester united"
matches[away=="manchester-city"]$away="manchester city"
matches[away=="west-ham"]$away="west ham"
matches[away=="crystal-palace"]$away="crystal palace"
matches[away=="stoke"]$away="stoke city"
sort(unique(matches$away))


[14:49, 1.5.2019] Enes  Özeren: TASK 2
[14:49, 1.5.2019] Enes  Özeren: library(ggplot2)
library(gridExtra)
library(data.table)
```

```
library(lubridate)
setwd("/Users/enesozeren/Downloads/")
matches=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_matches.rds')
odds=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_odd_details.rds')


matches[,c('HomeGoal','AwayGoal'):=tstrsplit(score,':')]
matches[,HomeGoal:=as.numeric(HomeGoal)]
matches[,AwayGoal:=as.numeric(AwayGoal)]
matches[,leagueId:=NULL]
matches[,Difference:=HomeGoal-AwayGoal]
#Tarih
matches[,Date:=as_datetime(date,tz='Turkey')]
matches[,date:=NULL]
matches[,type:=NULL]

#geç
matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
matches[home=="crystal-palace"]$home="crystal palace"
matches[home=="stoke"]$home="stoke city"
sort(unique(matches$home))
matches[away=="manchester-united"]$away="manchester united"
matches[away=="manchester-utd"]$away="manchester united"
matches[away=="manchester utd"]$away="manchester united"
```

```
matches[away=="manchester-city"]$away="manchester city"
matches[away=="west-ham"]$away="west ham"
matches[away=="crystal-palace"]$away="crystal palace"
matches[away=="stoke"]$away="stoke city"
sort(unique(matches$away))


#TASK2
temp=matches[,list(matchId,Date)]

#bet365
odds_bet365=odds[betType=="1x2" & bookmaker == "bet365"]
latest_odds_bet365=odds_bet365[,list(final_odd=odd[.N]),by=list(matchId,oddtype)]
latest_odds_bet365=dcast(latest_odds_bet365,matchId~oddtype,value.var = "final_odd")

latest_odds_bet365[,prob_home:=1/odd1]
latest_odds_bet365[,prob_away:=1/odd2]
latest_odds_bet365[,prob_draw:=1/oddX]
latest_odds_bet365[,total_prob:=prob_home+prob_away+prob_draw]

latest_odds_bet365[,prob_home_normalized:=prob_home/total_prob]
latest_odds_bet365[,prob_away_normalized:=prob_away/total_prob]
latest_odds_bet365[,prob_draw_normalized:=prob_draw/total_prob]

latest_odds_bet365=merge(temp,latest_odds_bet365,by="matchId")

odds_2010_bet365=latest_odds_bet365[Date>'2010-07-01'& Date<'2011-07-01']
odds_2010_bet365=odds_2010_bet365[,difference:=prob_home_normalized-
prob_away_normalized]
```

```
odds_2016_bet365=latest_odds_bet365[Date>'2016-07-01'& Date<'2017-07-01']

odds_2016_bet365=odds_2016_bet365[,difference:=prob_home_normalized-
prob_away_normalized]


mean_2010_diff_bet365 = mean(odds_2010_bet365$difference)

sd_2010_diff_bet365 = sd(odds_2010_bet365$difference)


prob_2010_bet365 = 1- pt(
(mean_2010_diff_bet365/(sd_2010_diff_bet365/sqrt(nrow(odds_2010_bet365)))),((nrow(odds_2
010_bet365))-1))


mean_2016_diff_bet365 = mean(odds_2016_bet365$difference)

sd_2016_diff_bet365 = sd(odds_2016_bet365$difference)


prob_2016_bet365 = 1-pt(
(mean_2016_diff_bet365/(sd_2016_diff_bet365/sqrt(nrow(odds_2016_bet365)))),((nrow(odds_2
016_bet365))-1))


##bet365_2010 graph

x <- seq(-11, 11, length=1000)

y <- dt(x, df=(nrow(odds_2010_bet365)-1))

c1 <- qt(0.975,df=nrow(odds_2010_bet365)-1)

c2 <- qt(0.025,df=nrow(odds_2010_bet365)-1)

pvalue_bet365_2010<-
(mean_2010_diff_bet365/(sd_2010_diff_bet365/sqrt(nrow(odds_2010_bet365))))


plot(x, y, type="l",lty=1,main="bet365_2010 Test", ylab="density",xlab="t value")

abline(v= c1, col='dark blue')

abline(v= c2, col='dark blue')
```

```
abline(v= pvalue_bet365_2010, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')


##bet365_2016 graph
x <- seq(-7, 7, length=1000)
y <- dt(x, df=(nrow(odds_2016_bet365)-1))
c1 <- qt(0.975,df=nrow(odds_2016_bet365)-1)
c2 <- qt(0.025,df=nrow(odds_2016_bet365)-1)
pvalue_bet365_2016<-
(mean_2016_diff_bet365/(sd_2016_diff_bet365/sqrt(nrow(odds_2016_bet365))))

plot(x, y, type="l",lty=1,main="bet365_2016 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_bet365_2016, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')



#10Bet
odds_10Bet=odds[betType=="1x2" & bookmaker == "10Bet"]
latest_odds_10Bet=odds_10Bet[,list(final_odd=odd[.N]),by=list(matchId,oddtype)]
```

```
latest_odds_10Bet=dcast(latest_odds_10Bet,matchId~oddtype,value.var = "final_odd")

latest_odds_10Bet[,prob_home:=1/odd1]
latest_odds_10Bet[,prob_away:=1/odd2]
latest_odds_10Bet[,prob_draw:=1/oddX]
latest_odds_10Bet[,total_prob:=prob_home+prob_away+prob_draw]

latest_odds_10Bet[,prob_home_normalized:=prob_home/total_prob]
latest_odds_10Bet[,prob_away_normalized:=prob_away/total_prob]
latest_odds_10Bet[,prob_draw_normalized:=prob_draw/total_prob]

latest_odds_10Bet=merge(temp,latest_odds_10Bet,by="matchId")

odds_2010_10Bet=latest_odds_10Bet[Date>'2010-07-01'& Date<'2011-07-01']
odds_2010_10Bet=odds_2010_10Bet[,difference:=prob_home_normalized-
prob_away_normalized]

odds_2016_10Bet=latest_odds_10Bet[Date>'2016-07-01'& Date<'2017-07-01']
odds_2016_10Bet=odds_2016_10Bet[,difference:=prob_home_normalized-
prob_away_normalized]

mean_2010_diff_10Bet = mean(odds_2010_10Bet$difference)
sd_2010_diff_10Bet = sd(odds_2010_10Bet$difference)

prob_2010_10Bet = 1- pt(
(mean_2010_diff_10Bet/(sd_2010_diff_10Bet/sqrt(nrow(odds_2010_10Bet)))),((nrow(odds_201
0_10Bet))-1))

mean_2016_diff_10Bet = mean(odds_2016_10Bet$difference)
sd_2016_diff_10Bet = sd(odds_2016_10Bet$difference)
```

```
prob_2016_10Bet = 1-pt( (mean_2016_diff_10Bet/(sd_2016_diff_10Bet/sqrt(nrow(odds_2016_10Bet)))),((nrow(odds_2016_10Bet))-1))


##10Bet_2010 graph

x <- seq(-11, 11, length=1000)

y <- dt(x, df=(nrow(odds_2010_10Bet)-1))

c1 <- qt(0.975,df=nrow(odds_2010_10Bet)-1)

c2 <- qt(0.025,df=nrow(odds_2010_10Bet)-1)

pvalue_10Bet_2010<- (mean_2010_diff_10Bet/(sd_2010_diff_10Bet/sqrt(nrow(odds_2010_10Bet))))


plot(x, y, type="l",lty=1,main="10Bet_2010 Test", ylab="density",xlab="t value")

abline(v= c1, col='dark blue')

abline(v= c2, col='dark blue')

abline(v= pvalue_10Bet_2010, col='dark red')

arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')

text(x=0, y=0.31,labels='confidence interval', col='black')


##10bet_2016 graph

x <- seq(-7, 7, length=1000)

y <- dt(x, df=(nrow(odds_2016_10Bet)-1))

c1 <- qt(0.975,df=nrow(odds_2016_10Bet)-1)

c2 <- qt(0.025,df=nrow(odds_2016_10Bet)-1)

pvalue_10Bet_2016<- (mean_2016_diff_10Bet/(sd_2016_diff_10Bet/sqrt(nrow(odds_2016_10Bet))))
```

```
plot(x, y, type="l",lty=1,main="10bet_2016 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_10Bet_2016, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')
```

```
#BetVictor
odds_BetVictor=odds[betType=="1x2" & bookmaker == "BetVictor"]
latest_odds_BetVictor=odds_BetVictor[,list(final_odd=odd[.N]),by=list(matchId,oddtype)]
latest_odds_BetVictor=dcast(latest_odds_BetVictor,matchId~oddtype,value.var = "final_odd")

latest_odds_BetVictor[,prob_home:=1/odd1]
latest_odds_BetVictor[,prob_away:=1/odd2]
latest_odds_BetVictor[,prob_draw:=1/oddX]
latest_odds_BetVictor[,total_prob:=prob_home+prob_away+prob_draw]

latest_odds_BetVictor[,prob_home_normalized:=prob_home/total_prob]
latest_odds_BetVictor[,prob_away_normalized:=prob_away/total_prob]
latest_odds_BetVictor[,prob_draw_normalized:=prob_draw/total_prob]

latest_odds_BetVictor=merge(temp,latest_odds_BetVictor,by="matchId")
```

```
odds_2010_BetVictor=latest_odds_BetVictor[Date>'2010-07-01'& Date<'2011-07-01']

odds_2010_BetVictor=odds_2010_BetVictor[,difference:=prob_home_normalized-
prob_away_normalized]


odds_2016_BetVictor=latest_odds_BetVictor[Date>'2016-07-01'& Date<'2017-07-01']

odds_2016_BetVictor=odds_2016_BetVictor[,difference:=prob_home_normalized-
prob_away_normalized]


mean_2010_diff_BetVictor = mean(odds_2010_BetVictor$difference)

sd_2010_diff_BetVictor = sd(odds_2010_BetVictor$difference)


prob_2010_BetVictor = 1- pt(
(mean_2010_diff_BetVictor/(sd_2010_diff_BetVictor/sqrt(nrow(odds_2010_BetVictor)))),((nro
w(odds_2010_BetVictor))-1))


mean_2016_diff_BetVictor = mean(odds_2016_BetVictor$difference)

sd_2016_diff_BetVictor = sd(odds_2016_BetVictor$difference)


prob_2016_BetVictor = 1-pt(
(mean_2016_diff_BetVictor/(sd_2016_diff_BetVictor/sqrt(nrow(odds_2016_BetVictor)))),((nro
w(odds_2016_BetVictor))-1))




##BetVictor_2010 graph

x <- seq(-11, 11, length=1000)

y <- dt(x, df=(nrow(odds_2010_BetVictor)-1))

c1 <- qt(0.975,df=nrow(odds_2010_BetVictor)-1)

c2 <- qt(0.025,df=nrow(odds_2010_BetVictor)-1)
```

```
pvalue_BetVictor_2010<-
(mean_2010_diff_BetVictor/(sd_2010_diff_BetVictor/sqrt(nrow(odds_2010_BetVictor))))

plot(x, y, type="l",lty=1,main="BetVictor_2010 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_BetVictor_2010, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')


##BetVictor_2016 graph
x <- seq(-7, 7, length=1000)
y <- dt(x, df=(nrow(odds_2016_BetVictor)-1))
c1 <- qt(0.975,df=nrow(odds_2016_BetVictor)-1)
c2 <- qt(0.025,df=nrow(odds_2016_BetVictor)-1)
pvalue_BetVictor_2016<-
(mean_2016_diff_BetVictor/(sd_2016_diff_BetVictor/sqrt(nrow(odds_2016_BetVictor))))

plot(x, y, type="l",lty=1,main="BetVictor_2016 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_BetVictor_2016, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')
```

#Betsafe

```
odds_Betsafe=odds[betType=="1x2" & bookmaker == "Betsafe"]
latest_odds_Betsafe=odds_Betsafe[,list(final_odd=odd[.N]),by=list(matchId,oddtype)]
latest_odds_Betsafe=dcast(latest_odds_Betsafe,matchId~oddtype,value.var = "final_odd")

latest_odds_Betsafe[,prob_home:=1/odd1]
latest_odds_Betsafe[,prob_away:=1/odd2]
latest_odds_Betsafe[,prob_draw:=1/oddX]
latest_odds_Betsafe[,total_prob:=prob_home+prob_away+prob_draw]

latest_odds_Betsafe[,prob_home_normalized:=prob_home/total_prob]
latest_odds_Betsafe[,prob_away_normalized:=prob_away/total_prob]
latest_odds_Betsafe[,prob_draw_normalized:=prob_draw/total_prob]

latest_odds_Betsafe=merge(temp,latest_odds_Betsafe,by="matchId")

odds_2010_Betsafe=latest_odds_Betsafe[Date>'2010-07-01'& Date<'2011-07-01']
odds_2010_Betsafe=odds_2010_Betsafe[,difference:=prob_home_normalized-
prob_away_normalized]

odds_2016_Betsafe=latest_odds_Betsafe[Date>'2016-07-01'& Date<'2017-07-01']
odds_2016_Betsafe=odds_2016_Betsafe[,difference:=prob_home_normalized-
prob_away_normalized]

mean_2010_diff_Betsafe = mean(odds_2010_Betsafe$difference)
sd_2010_diff_Betsafe = sd(odds_2010_Betsafe$difference)
```

```
prob_2010_Betsafe = 1- pt( (mean_2010_diff_Betsafe/(sd_2010_diff_Betsafe/sqrt(nrow(odds_2010_Betsafe)))),((nrow(odds_2010_Betsafe))-1))

mean_2016_diff_Betsafe = mean(odds_2016_Betsafe$difference)

sd_2016_diff_Betsafe = sd(odds_2016_Betsafe$difference)

prob_2016_Betsafe = 1-pt( (mean_2016_diff_Betsafe/(sd_2016_diff_Betsafe/sqrt(nrow(odds_2016_Betsafe)))),((nrow(odds_2016_Betsafe))-1))

##Betsafe_2010 graph

x <- seq(-11, 11, length=1000)

y <- dt(x, df=(nrow(odds_2010_Betsafe)-1))

c1 <- qt(0.975,df=nrow(odds_2010_Betsafe)-1)

c2 <- qt(0.025,df=nrow(odds_2010_Betsafe)-1)

pvalue_Betsafe_2010<- (mean_2010_diff_Betsafe/(sd_2010_diff_Betsafe/sqrt(nrow(odds_2010_Betsafe))))

plot(x, y, type="l",lty=1,main="Betsafe_2010 Test", ylab="density",xlab="t value")

abline(v= c1, col='dark blue')

abline(v= c2, col='dark blue')

abline(v= pvalue_Betsafe_2010, col='dark red')

arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')

text(x=0, y=0.31,labels='confidence interval', col='black')

##Betsafe_2016 graph
```

```
x <- seq(-7, 7, length=1000)
y <- dt(x, df=(nrow(odds_2016_Betsafe)-1))
c1 <- qt(0.975,df=nrow(odds_2016_Betsafe)-1)
c2 <- qt(0.025,df=nrow(odds_2016_Betsafe)-1)
pvalue_Betsafe_2016<-
(mean_2016_diff_Betsafe/(sd_2016_diff_Betsafe/sqrt(nrow(odds_2016_Betsafe))))

plot(x, y, type="l",lty=1,main="Betsafe_2016 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_Betsafe_2016, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')
```

```
#888sport

odds_888sport=odds[betType=="1x2" & bookmaker == "888sport"]
latest_odds_888sport=odds_888sport[,list(final_odd=odd[.N]),by=list(matchId,oddtype)]
latest_odds_888sport=dcast(latest_odds_888sport,matchId~oddtype,value.var = "final_odd")

latest_odds_888sport[,prob_home:=1/odd1]
latest_odds_888sport[,prob_away:=1/odd2]
latest_odds_888sport[,prob_draw:=1/oddX]
```

```
latest_odds_888sport[,total_prob:=prob_home+prob_away+prob_draw]

latest_odds_888sport[,prob_home_normalized:=prob_home/total_prob]
latest_odds_888sport[,prob_away_normalized:=prob_away/total_prob]
latest_odds_888sport[,prob_draw_normalized:=prob_draw/total_prob]

latest_odds_888sport=merge(temp,latest_odds_888sport,by="matchId")

odds_2010_888sport=latest_odds_888sport[Date>'2010-07-01'& Date<'2011-07-01']
odds_2010_888sport=odds_2010_888sport[,difference:=prob_home_normalized-
prob_away_normalized]

odds_2016_888sport=latest_odds_888sport[Date>'2016-07-01'& Date<'2017-07-01']
odds_2016_888sport=odds_2016_888sport[,difference:=prob_home_normalized-
prob_away_normalized]

mean_2010_diff_888sport = mean(odds_2010_888sport$difference)
sd_2010_diff_888sport = sd(odds_2010_888sport$difference)

prob_2010_888sport = 1- pt(
(mean_2010_diff_888sport/(sd_2010_diff_888sport/sqrt(nrow(odds_2010_888sport)))),((nrow(o
dds_2010_888sport))-1))

mean_2016_diff_888sport = mean(odds_2016_888sport$difference)
sd_2016_diff_888sport = sd(odds_2016_888sport$difference)

prob_2016_888sport = 1-pt(
(mean_2016_diff_888sport/(sd_2016_diff_888sport/sqrt(nrow(odds_2016_888sport)))),((nrow(o
dds_2016_888sport))-1))
```

```
##888sport_2010 graph
x <- seq(-11, 11, length=1000)
y <- dt(x, df=(nrow(odds_2010_888sport)-1))
c1 <- qt(0.975,df=nrow(odds_2010_888sport)-1)
c2 <- qt(0.025,df=nrow(odds_2010_888sport)-1)
pvalue_888sport_2010<-
(mean_2010_diff_888sport/(sd_2010_diff_888sport/sqrt(nrow(odds_2010_888sport))))

plot(x, y, type="l",lty=1,main="888sport_2010 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_888sport_2010, col='dark red')
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')


##888sport_2016 graph
x <- seq(-7, 7, length=1000)
y <- dt(x, df=(nrow(odds_2016_888sport)-1))
c1 <- qt(0.975,df=nrow(odds_2016_888sport)-1)
c2 <- qt(0.025,df=nrow(odds_2016_888sport)-1)
pvalue_888sport_2016<-
(mean_2016_diff_888sport/(sd_2016_diff_888sport/sqrt(nrow(odds_2016_888sport))))

plot(x, y, type="l",lty=1,main="888sport_2016 Test", ylab="density",xlab="t value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= pvalue_888sport_2016, col='dark red')
```

```
arrows(x0=c1, y0=0.3, x1=c2, y1=0.3, code=3, col='dark blue')
text(x=0, y=0.31,labels='confidence interval', col='black')

[14:49, 1.5.2019] Enes  Özeren: TASK 3
[14:49, 1.5.2019] Enes  Özeren: library(ggplot2)
library(gridExtra)
library(data.table)
library(lubridate)
setwd("/Users/enesozeren/Downloads/")
matches=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_matches.rds')
odds=readRDS('df9b1196-e3cf-4cc7-9159-f236fe738215_odd_details.rds')

matches[,c('HomeGoal','AwayGoal'):=tstrsplit(score,':')]
matches[,HomeGoal:=as.numeric(HomeGoal)]
matches[,AwayGoal:=as.numeric(AwayGoal)]
matches[,leagueId:=NULL]
matches[,Difference:=HomeGoal-AwayGoal]
#Tarih
matches[,Date:=as_datetime(date,tz='Turkey')]
matches[,date:=NULL]
matches[,type:=NULL]

#geç
matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
```

```
matches[home=="crystal-palace"]$home="crystal palace"
matches[home=="stoke"]$home="stoke city"
sort(unique(matches$home))
matches[away=="manchester-united"]$away="manchester united"
matches[away=="manchester-utd"]$away="manchester united"
matches[away=="manchester utd"]$away="manchester united"
matches[away=="manchester-city"]$away="manchester city"
matches[away=="west-ham"]$away="west ham"
matches[away=="crystal-palace"]$away="crystal palace"
matches[away=="stoke"]$away="stoke city"
sort(unique(matches$away))


#TASK3
season_2010 = matches[Date>'2010-07-01'& Date<'2011-07-01']
sd_2010 = sd(season_2010$AwayGoal)


season_2016 = matches[Date>'2016-07-01'& Date<'2017-07-01']
sd_2016 = sd(season_2016$AwayGoal)


#calculating the F probability
prob_variance_ratio = pf( (sd_2010*2/sd_2016*2) , nrow(season_2010)-1, nrow(season_2016)-1)


#calculating the F probability
```

```
prob_variance_ratio = pf( (sd_2010^2/sd_2016^2) , nrow(season_2010)-1, nrow(season_2016)-1)

#for alpha=0.1
x <- seq(0, 2, length=1000)
y <- df(x,  nrow(season_2010)-1, nrow(season_2016)-1)
c1 <- qf(0.95,nrow(season_2010)-1, nrow(season_2016)-1)
c2 <- qf(0.05,nrow(season_2010)-1, nrow(season_2016)-1)
c3 <- (sd_2010^2/sd_2016^2)

plot(x, y, type="l",lty=1,main="F distribution Test", ylab="density",xlab="f value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= c3, col='dark red')
arrows(x0=c1, y0=3, x1=c2, y1=3, code=3, col='dark blue')
text(x=1, y=3.1,labels='confidence interval', col='black')



#for alpha=0.01
x <- seq(0, 2, length=1000)
y <- df(x,  nrow(season_2010)-1, nrow(season_2016)-1)
c1 <- qf(0.995,nrow(season_2010)-1, nrow(season_2016)-1)
c2 <- qf(0.005,nrow(season_2010)-1, nrow(season_2016)-1)
c3 <- (sd_2010^2/sd_2016^2)

plot(x, y, type="l",lty=1,main="F distribution Test", ylab="density",xlab="f value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
```

```
abline(v= c3, col='dark red')
arrows(x0=c1, y0=3, x1=c2, y1=3, code=3, col='dark blue')
text(x=1, y=3.1,labels='confidence interval', col='black')



[14:50, 1.5.2019] Enes  Özeren: TASK 4
[14:50, 1.5.2019] Enes  Özeren: library(ggplot2)
library(gridExtra)
library(data.table)
library(lubridate)
setwd("/Users/enesozeren/Downloads/")

matches=readRDS("df9b1196-e3cf-4cc7-9159-f236fe738215_matches.rds")
matches[, c("HomeGoal", "AwayGoal"):=tstrsplit(score, ":")]

matches$leagueId=NULL
matches
matches$type=NULL
matches
sort(unique(matches$home))

matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
matches[home=="crystal-palace"]$home="crystal palace"
matches[home=="stoke"]$home="stoke city"
```

```
sort(unique(matches$home))
matches[away=="manchester-united"]$away="manchester united"
matches[away=="manchester-utd"]$away="manchester united"
matches[away=="manchester utd"]$away="manchester united"
matches[away=="manchester-city"]$away="manchester city"
matches[away=="west-ham"]$away="west ham"
matches[away=="crystal-palace"]$away="crystal palace"
matches[away=="stoke"]$away="stoke city"
sort(unique(matches$away))

matches[,c("HomeGoal", "AwayGoal"):=tstrsplit(score, ":")]
matches

matches[,HomeGoal:=as.numeric(HomeGoal)]
matches[,AwayGoal:=as.numeric(AwayGoal)]
matches[,MatchResult:=ifelse(HomeGoal>AwayGoal,"home",
ifelse(HomeGoal==AwayGoal,"draw", "away"))]
matches

require(lubridate)
matches[,timestamp:=as_datetime(date)]
matches



#Summarize by home goal
summary_by_homegoal=matches[,list(count=.N),by=list(matchId,HomeGoal)]

factor(summary_by_homegoal$HomeGoal)
table_for_homegoal=table(summary_by_homegoal$HomeGoal)
```

```
table_for_homegoal


hist(summary_by_homegoal$HomeGoal,main = "HomeGoal Table", xlab = "Home Goals", ylab = "Number of Games", las =1, breaks = 30,col='light blue')

mean_homegoal=mean(matches$HomeGoal,na.rm = T)

mean_homegoal

par(new=TRUE)

plot(dpois(x=0:8,lambda=mean_homegoal), xlab = "Home Goals",ylab="Number of Games",axes=F,col='dark red',pch=19)



HomeGoal_pois=c(dpois(0,mean_homegoal)*sum(table_for_homegoal),
                dpois(1,mean_homegoal)*sum(table_for_homegoal),
                dpois(2,mean_homegoal)*sum(table_for_homegoal),
                dpois(3,mean_homegoal)*sum(table_for_homegoal),
                dpois(4,mean_homegoal)*sum(table_for_homegoal),
                dpois(5,mean_homegoal)*sum(table_for_homegoal),
                dpois(6,mean_homegoal)*sum(table_for_homegoal),
                dpois(7,mean_homegoal)*sum(table_for_homegoal),
                dpois(8,mean_homegoal)*sum(table_for_homegoal))


real_vs_poison_homegoal=data.table(Real_HomeGoal=table_for_homegoal,Poison_HomeGoal= HomeGoal_pois)


ggplot(real_vs_poison_homegoal, aes(real_vs_poison_homegoal$Real_HomeGoal.V1, cumsum(real_vs_poison_homegoal$Real_HomeGoal.N))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Real Home Goals")+
  xlab("Number of Home Goals")+
  ylab("Cumulative HomeGoals")+
```

```
  ylim(500, 3500)



ggplot(real_vs_poison_homegoal, aes(real_vs_poison_homegoal$Real_HomeGoal.V1,
cumsum(real_vs_poison_homegoal$Poison_HomeGoal))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Poisson Home Goals")+
  xlab("Number of Home Goals")+
  ylab("Cumulative HomeGoals")+
  ylim(500, 3500)



#Summerize by awaygoals
summary_by_awaygoal=matches[,list(count=.N),by=list(matchId,AwayGoal)]
factor(summary_by_awaygoal$AwayGoal)
table_for_awaygoal=table(summary_by_awaygoal$AwayGoal)
table_for_awaygoal


hist(summary_by_awaygoal$AwayGoal,main = "AwayGoal Table", xlab = "Away Goals", ylab =
"Number of Games",las=1, breaks = 30,col='light blue')


mean_awaygoal=mean(matches$AwayGoal,na.rm = T)
par(new=TRUE)
plot(dpois(x=0:7,lambda=mean_awaygoal), xlab = "Away Goals",ylab="Number of
Games",axes=F,col='dark red',pch=19)


AwayGoal_pois=c(dpois(0,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(1,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(2,mean_awaygoal)*sum(table_for_awaygoal),
```

```
        dpois(3,mean_awaygoal)*sum(table_for_awaygoal),
        dpois(4,mean_awaygoal)*sum(table_for_awaygoal),
        dpois(5,mean_awaygoal)*sum(table_for_awaygoal),
        dpois(6,mean_awaygoal)*sum(table_for_awaygoal),
        dpois(7,mean_awaygoal)*sum(table_for_awaygoal))


real_vs_poison_awaygoal=data.table(Real_AwayGoal=table_for_awaygoal,Poison_AwayGoal=
AwayGoal_pois)


ggplot(real_vs_poison_awaygoal, aes(real_vs_poison_awaygoal$Real_AwayGoal.V1,
cumsum(real_vs_poison_awaygoal$Real_AwayGoal.N))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Real Away Goals")+
  xlab("Number of Away Goals")+
  ylab("Cumulative Away Goals")+
  ylim(500, 3500)


ggplot(real_vs_poison_awaygoal, aes(real_vs_poison_awaygoal$Real_AwayGoal.V1,
cumsum(real_vs_poison_awaygoal$Poison_AwayGoal))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Poisson Away Goals")+
  xlab("Number of Away Goals")+
  ylab("Cumulative Away Goals")+
  ylim(500, 3500)



#(HomeGoals-AwayGoals)
```

```
summary_by_homegoal_and_awaygoal=matches[,list(count=.N),by=list(matchId,HomeGoal,AwayGoal)]

summary_by_homegoal_and_awaygoal

homegoal_minus_awaygoal=summary_by_homegoal_and_awaygoal[,list(count=.N),by=list(matchId,HomeGoal-AwayGoal)]

homegoal_minus_awaygoal[,c("HomeGoal-AwayGoal"):=HomeGoal]

homegoal_minus_awaygoal$HomeGoal=NULL

homegoal_minus_awaygoal$count=NULL

homegoal_minus_awaygoal

hist(homegoal_minus_awaygoal$`HomeGoal-AwayGoal`,main = "HomeGoal - AwayGoal Table", xlab = "Home Goals - Away Goals", ylab = "Number of Games", las =1, breaks = 60,col='light blue')



table_for_homegoal

sum(table_for_homegoal)



qchisq(0.95,df=8)



sum_away<-sum(((table_for_awaygoal[1:8]-AwayGoal_pois[1:8])^2)/(AwayGoal_pois[1:8]))



x <- seq(0, 40, length=1000)

y <- dchisq(x, df=7)


c <- qchisq(0.95,df=7)

plot(x, y, type="l",lty=1,main="Away Test", ylab="chi-square",xlab="x value")
```

```
abline(v= c, col='dark blue')
abline(v= sum_away, col='dark red')

arrows(x0=0, y0=0.08, x1=c, y1=0.08, code=3, col='dark blue')

text(x=4, y=0.085,labels='confidence interval', col='black')



sum_home<-sum(((table_for_homegoal[1:9]-HomeGoal_pois[1:9])^2)/(HomeGoal_pois[1:9]))

x <- seq(0, 25, length=1000)
y <- dchisq(x, df=8)

c_home <- qchisq(0.95,df=8)
plot(x, y, type="l",lty=1,main="Home Test", ylab="chi-square",xlab="x value")

abline(v= c_home, col='dark blue')
abline(v= sum_home, col='dark red')

arrows(x0=0, y0=0.08, x1=c_home, y1=0.08, code=3, col='dark blue')
arrows(x0=0, y0=0.06, x1=sum_home, y1=0.06, code=3, col='dark red')

text(x=4, y=0.085,labels='confidence interval', col='black')
text(x=2, y=0.065,labels= 'p value', col='black')
```

x.Code for Figures 10 and 11:

```
matches[,c('HomeGoal','AwayGoal'):=tstrsplit(score,':')]
matches[,HomeGoal:=as.numeric(HomeGoal)]
matches[,AwayGoal:=as.numeric(AwayGoal)]
matches[,leagueId:=NULL]
matches[,Difference:=HomeGoal-AwayGoal]
#Date
matches[,Date:=as_datetime(date,tz='Turkey')]
matches[,date:=NULL]
matches[,type:=NULL]

#Filter
matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
matches[home=="crystal-palace"]$home="crystal palace"
matches[home=="stoke"]$home="stoke city"
sort(unique(matches$home))
matches[away=="manchester-united"]$away="manchester united"
matches[away=="manchester-utd"]$away="manchester united"
matches[away=="manchester utd"]$away="manchester united"
matches[away=="manchester-city"]$away="manchester city"
matches[away=="west-ham"]$away="west ham"
matches[away=="crystal-palace"]$away="crystal palace"
matches[away=="stoke"]$away="stoke city"
sort(unique(matches$away))
```

```
season_2010 = matches[Date>'2010-07-01'& Date<'2011-07-01']
sd_2010 = sd(season_2010$AwayGoal)

season_2016 = matches[Date>'2016-07-01'& Date<'2017-07-01']
sd_2016 = sd(season_2016$AwayGoal)

#calculating the F probability
prob_variance_ratio = pf( (sd_2010*2/sd_2016*2) , nrow(season_2010)-1, nrow(season_2016)-1)
#for alpha=0.1
x <- seq(0, 2, length=1000)
y <- df(x,  nrow(season_2010)-1, nrow(season_2016)-1)
c1 <- qf(0.95,nrow(season_2010)-1, nrow(season_2016)-1)
c2 <- qf(0.05,nrow(season_2010)-1, nrow(season_2016)-1)
c3 <- (sd_2010^2/sd_2016^2)

plot(x, y, type="l",lty=1,main="F distribution Test", ylab="density",xlab="f value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= c3, col='dark red')
arrows(x0=c1, y0=3, x1=c2, y1=3, code=3, col='dark blue')
text(x=1, y=3.1,labels='confidence interval', col='black')

  #for alpha=0.01
x <- seq(0, 2, length=1000)
y <- df(x,  nrow(season_2010)-1, nrow(season_2016)-1)
c1 <- qf(0.995,nrow(season_2010)-1, nrow(season_2016)-1)
c2 <- qf(0.005,nrow(season_2010)-1, nrow(season_2016)-1)
c3 <- (sd_2010^2/sd_2016^2)
```

```
plot(x, y, type="l",lty=1,main="F distribution Test", ylab="density",xlab="f value")
abline(v= c1, col='dark blue')
abline(v= c2, col='dark blue')
abline(v= c3, col='dark red')
arrows(x0=c1, y0=3, x1=c2, y1=3, code=3, col='dark blue')
text(x=1, y=3.1,labels='confidence interval', col='black')
```

xi.Codes for Figures 12 and 13:

```
setwd("C:/Users/ASUS/Documents")
matches=readRDS("df9b1196-e3cf-4cc7-9159-f236fe738215_matches.rds")
library(ggplot2)
library(data.table)
matches[, c("HomeGoal", "AwayGoal"):=tstrsplit(score, ":")]
matches$leagueId=NULL
matches
matches$type=NULL
matches
sort(unique(matches$home))
matches[home=="manchester-united"]$home="manchester united"
matches[home=="manchester-utd"]$home="manchester united"
matches[home=="manchester utd"]$home="manchester united"
matches[home=="manchester-city"]$home="manchester city"
matches[home=="west-ham"]$home="west ham"
matches[home=="crystal-palace"]$home="crystal palace"
sort(unique(matches$home))
matches[,c("HomeGoal", "AwayGoal"):=tstrsplit(score, ":")]
matches[,HomeGoal:=as.numeric(HomeGoal)]
```

```
matches[,AwayGoal:=as.numeric(AwayGoal)]

matches[,MatchResult:=ifelse(HomeGoal>AwayGoal,"home",
ifelse(HomeGoal==AwayGoal,"draw", "away"))]

require(lubridate)

matches[,timestamp:=as_datetime(date)]

matches


#Summarize by home goal

summary_by_homegoal=matches[,list(count=.N),by=list(matchId,HomeGoal)]


factor(summary_by_homegoal$HomeGoal)

table_for_homegoal=table(summary_by_homegoal$HomeGoal)

table_for_homegoal


hist(summary_by_homegoal$HomeGoal,main = "HomeGoal Table", xlab = "Home Goals", ylab
= "Number of Games", las =1, breaks = 30,col='light blue')

mean_homegoal=mean(matches$HomeGoal,na.rm = T)

mean_homegoal

par(new=TRUE)

plot(dpois(x=0:8,lambda=mean_homegoal), xlab = "Home Goals",ylab="Number of
Games",axes=F,col='dark red',pch=19)


HomeGoal_pois=c(dpois(0,mean_homegoal)*sum(table_for_homegoal),

        dpois(1,mean_homegoal)*sum(table_for_homegoal),

        dpois(2,mean_homegoal)*sum(table_for_homegoal),

        dpois(3,mean_homegoal)*sum(table_for_homegoal),

        dpois(4,mean_homegoal)*sum(table_for_homegoal),

        dpois(5,mean_homegoal)*sum(table_for_homegoal),

        dpois(6,mean_homegoal)*sum(table_for_homegoal),
```

```
        dpois(7,mean_homegoal)*sum(table_for_homegoal),
        dpois(8,mean_homegoal)*sum(table_for_homegoal))

real_vs_poison_homegoal=data.table(Real_HomeGoal=table_for_homegoal,Poison_HomeGoal=
HomeGoal_pois)

ggplot(real_vs_poison_homegoal, aes(real_vs_poison_homegoal$Real_HomeGoal.V1,
cumsum(real_vs_poison_homegoal$Real_HomeGoal.N))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Real Home Goals")+
  xlab("Number of Home Goals")+
  ylab("Cumulative HomeGoals")+
  ylim(500, 3500)

ggplot(real_vs_poison_homegoal, aes(real_vs_poison_homegoal$Real_HomeGoal.V1,
cumsum(real_vs_poison_homegoal$Poison_HomeGoal))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Poisson Home Goals")+
  xlab("Number of Home Goals")+
  ylab("Cumulative HomeGoals")+
  ylim(500, 3500)



#Summerize by awaygoals
summary_by_awaygoal=matches[,list(count=.N),by=list(matchId,AwayGoal)]
factor(summary_by_awaygoal$AwayGoal)
table_for_awaygoal=table(summary_by_awaygoal$AwayGoal)
table_for_awaygoal
```

```
hist(summary_by_awaygoal$AwayGoal,main = "AwayGoal Table", xlab = "Away Goals", ylab = "Number of Games",las=1, breaks = 30,col='light blue')


mean_awaygoal=mean(matches$AwayGoal,na.rm = T)

par(new=TRUE)

plot(dpois(x=0:7,lambda=mean_awaygoal), xlab = "Away Goals",ylab="Number of Games",axes=F,col='dark red',pch=19)


AwayGoal_pois=c(dpois(0,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(1,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(2,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(3,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(4,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(5,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(6,mean_awaygoal)*sum(table_for_awaygoal),
                dpois(7,mean_awaygoal)*sum(table_for_awaygoal))



real_vs_poison_awaygoal=data.table(Real_AwayGoal=table_for_awaygoal,Poison_AwayGoal=AwayGoal_pois)


ggplot(real_vs_poison_awaygoal, aes(real_vs_poison_awaygoal$Real_AwayGoal.V1, cumsum(real_vs_poison_awaygoal$Real_AwayGoal.N))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Real Away Goals")+
  xlab("Number of Away Goals")+
  ylab("Cumulative Away Goals")+
  ylim(500, 3500)
```

```
ggplot(real_vs_poison_awaygoal, aes(real_vs_poison_awaygoal$Real_AwayGoal.V1, cumsum(real_vs_poison_awaygoal$Poison_AwayGoal))) +
  geom_step(aes(group=1))+
  ggtitle("CDF of Poisson Away Goals")+
  xlab("Number of Away Goals")+
  ylab("Cumulative Away Goals")+
  ylim(500, 3500)


#(HomeGoals-AwayGoals)
summary_by_homegoal_and_awaygoal=matches[,list(count=.N),by=list(matchId,HomeGoal,AwayGoal)]
summary_by_homegoal_and_awaygoal
homegoal_minus_awaygoal=summary_by_homegoal_and_awaygoal[,list(count=.N),by=list(matchId,HomeGoal-AwayGoal)]
homegoal_minus_awaygoal[,c("HomeGoal-AwayGoal"):=HomeGoal]
homegoal_minus_awaygoal$HomeGoal=NULL
homegoal_minus_awaygoal$count=NULL
homegoal_minus_awaygoal
hist(homegoal_minus_awaygoal$`HomeGoal-AwayGoal`,main = "HomeGoal - AwayGoal Table", xlab = "Home Goals - Away Goals", ylab = "Number of Games", las =1, breaks = 60,col='light blue')


table_for_homegoal
sum(table_for_homegoal)


qchisq(0.95,df=8)
sum_away<-sum(((table_for_awaygoal[1:8]-AwayGoal_pois[1:8])^2)/(AwayGoal_pois[1:8]))


x <- seq(0, 40, length=1000)
y <- dchisq(x, df=7)
```

```
c <- qchisq(0.95,df=7)
plot(x, y, type="l",lty=1,main="Away Test", ylab="chi-square",xlab="x value")

abline(v= c, col='dark blue')
abline(v= sum_away, col='dark red')

arrows(x0=0, y0=0.08, x1=c, y1=0.08, code=3, col='dark blue')
text(x=4, y=0.085,labels='confidence interval', col='black')
sum_home<-sum(((table_for_homegoal[1:9]-HomeGoal_pois[1:9])^2)/(HomeGoal_pois[1:9]))
x <- seq(0, 25, length=1000)
y <- dchisq(x, df=8)

c_home <- qchisq(0.95,df=8)
plot(x, y, type="l",lty=1,main="Home Test", ylab="chi-square",xlab="x value")

abline(v= c_home, col='dark blue')
abline(v= sum_home, col='dark red')

arrows(x0=0, y0=0.08, x1=c_home, y1=0.08, code=3, col='dark blue')
arrows(x0=0, y0=0.06, x1=sum_home, y1=0.06, code=3, col='dark red')

text(x=4, y=0.085,labels='confidence interval', col='black')
text(x=2, y=0.065,labels= 'p value', col='black'
```