Group Project Part 1

*Sezgin Yıldız, Alp Biberoglu, Gulnur Acar, Eren Demir*

*16 Nisan 2018*

Task 1 Q1

**library**(readr)

tr\_super\_league\_matches <- read\_csv("C:/Users/Sezgin/Desktop/tr\_super\_league\_matches.csv")

## Parsed with column specification:

## cols(

## matchid = col\_character(),

## season = col\_integer(),

## Home = col\_character(),

## Away = col\_character(),

## Match\_Date = col\_date(format = ""),

## Round = col\_integer(),

## Match\_Hour = col\_integer(),

## weekDay = col\_integer(),

## month = col\_integer(),

## AWA\_FLAG = col\_integer(),

## Home\_Score = col\_integer(),

## Away\_Score = col\_integer(),

## Match\_Result = col\_character()

## )

tr\_super\_league\_matches

## # A tibble: 2,322 x 13

## matchid season Home Away Match\_Date Round Match\_Hour weekDay month

## <chr> <int> <chr> <chr> <date> <int> <int> <int> <int>

## 1 04ouflDc 2010 karde~ manis~ 2010-08-15 1 20 1 8

## 2 0AUXYwsk 2010 eskis~ gencl~ 2010-08-14 1 21 7 8

## 3 6mVTZJRr 2010 gazia~ kasim~ 2010-08-14 1 19 7 8

## 4 6NaYWHB1 2010 sivas~ galat~ 2010-08-14 1 19 7 8

## 5 8UlmhSsA 2010 ankar~ trabz~ 2010-08-15 1 18 1 8

## 6 jgvhincG 2010 bucas~ besik~ 2010-08-14 1 21 7 8

## 7 MDnyeUci 2010 fener~ antal~ 2010-08-15 1 20 1 8

## 8 Uwjqg8S3 2010 basak~ kayse~ 2010-08-16 1 20 2 8

## 9 v1TyYcde 2010 bursa~ konya~ 2010-08-16 1 20 2 8

## 10 2ZASfL38 2010 konya~ eskis~ 2010-08-20 2 20 6 8

## # ... with 2,312 more rows, and 4 more variables: AWA\_FLAG <int>,

## # Home\_Score <int>, Away\_Score <int>, Match\_Result <chr>

**attach**(tr\_super\_league\_matches)

hist(Home\_Score, main= "Home Score Distribution",xlab= "Home Score",ylab= "Number of Games", col="blue")

hist(Away\_Score, main= "Away Score Distribution", xlab= "Away Score",ylab= "Number of Games", col="green")

hist(Home\_Score - Away\_Score, main= "Home Score - Away Score Distribution", xlab= "Home Score - Away Score", ylab= "Number of Games", col="red")

Task 1 Q2

What distribution do you think home and away goals are coming from? Calculate the expected number of games corresponding to each quantile (number of goals) by using sample means as distribution mean and plot these values on the histogram.

Answer: In accordance with the score difference distribution histogram, it can be seen that the distribution comes from standard normal distribution. The histogram also states that the score difference mostly 0, meaning that most games end as a draw, also could be indicating that the mean value is 0 . Games ending as a draw has the highest probability as a result of the number of games. The second quantile with the highest “probability” is +1 goal difference meaning that Home teams score 1 more goal than Away teams.

Task 2

**library**(readr)

tr\_super\_league\_odd\_details <- read\_csv("C:/Users/Sezgin/Desktop/tr\_super\_league\_odd\_details.csv")

## Parsed with column specification:

## cols(

## matchid = col\_character(),

## Home = col\_character(),

## Away = col\_character(),

## Bookmaker = col\_character(),

## HomeOdd = col\_double(),

## AwayOdd = col\_double(),

## TieOdd = col\_double()

## )

tr\_super\_league\_odd\_details

## # A tibble: 10,072 x 7

## matchid Home Away Bookmaker HomeOdd AwayOdd TieOdd

## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>

## 1 0029LnfI konyaspor sivassp~ Betsson 2.82 2.56 3.28

## 2 0029LnfI konyaspor sivassp~ Pinnacle 2.73 2.76 3.42

## 3 0029LnfI konyaspor sivassp~ bet365 2.75 2.56 3.35

## 4 0029LnfI konyaspor sivassp~ bwin 2.72 2.45 3.32

## 5 00ApPQui trabzonspor galatas~ Betfair 2.52 2.75 3.20

## 6 00ApPQui trabzonspor galatas~ Betsson 2.17 2.90 3.20

## 7 00ApPQui trabzonspor galatas~ Pinnacle 2.67 2.88 3.42

## 8 00ApPQui trabzonspor galatas~ bet365 2.25 3.00 3.25

## 9 00ApPQui trabzonspor galatas~ bwin 2.33 2.98 3.20

## 10 00IbEw7m akhisar-genclik-spor gaziant~ Betsson 2.04 3.75 3.32

## # ... with 10,062 more rows

**attach**(tr\_super\_league\_odd\_details)

## The following objects are masked from tr\_super\_league\_matches:

##

## Away, Home, matchid

**library**(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## <U+221A> ggplot2 2.2.1 <U+221A> purrr 0.2.4

## <U+221A> tibble 1.4.2 <U+221A> dplyr 0.7.4

## <U+221A> tidyr 0.8.0 <U+221A> stringr 1.2.0

## <U+221A> ggplot2 2.2.1 <U+221A> forcats 0.2.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --

## x dplyr::filter() masks stats::filter()

## x dplyr::lag() masks stats::lag()

**library**(ggplot2)

*#Calculation of P(home win), P(tie) and P(away win) with formula P(x)=1/odd*

probabilities <- tr\_super\_league\_odd\_details %>% group\_by(matchid, Bookmaker) %>% summarise(Phomewin= 1/HomeOdd, Pawaywin = 1/AwayOdd, Ptie= 1/TieOdd) %>% ungroup()

probabilities

## # A tibble: 10,072 x 5

## matchid Bookmaker Phomewin Pawaywin Ptie

## <chr> <chr> <dbl> <dbl> <dbl>

## 1 0029LnfI bet365 0.364 0.390 0.299

## 2 0029LnfI Betsson 0.354 0.391 0.305

## 3 0029LnfI bwin 0.367 0.408 0.301

## 4 0029LnfI Pinnacle 0.366 0.362 0.292

## 5 00ApPQui bet365 0.444 0.333 0.308

## 6 00ApPQui Betfair 0.396 0.364 0.312

## 7 00ApPQui Betsson 0.460 0.345 0.312

## 8 00ApPQui bwin 0.430 0.336 0.312

## 9 00ApPQui Pinnacle 0.375 0.347 0.292

## 10 00IbEw7m bet365 0.460 0.299 0.308

## # ... with 10,062 more rows

*#Calculation of the probabilities via normalization approach*

probabilities\_norm <- tr\_super\_league\_odd\_details %>% group\_by(matchid, Bookmaker) %>% summarise(Pnorhomewin= (1/HomeOdd)\*(1/((1/HomeOdd)+(1/AwayOdd)+(1/TieOdd))), Pnorawaywin= (1/AwayOdd)\*(1/((1/HomeOdd)+(1/AwayOdd)+(1/TieOdd))), Pnortie= (1/TieOdd)\*(1/((1/HomeOdd)+(1/AwayOdd)+(1/TieOdd)))) %>% ungroup()

probabilities\_norm

## # A tibble: 10,072 x 5

## matchid Bookmaker Pnorhomewin Pnorawaywin Pnortie

## <chr> <chr> <dbl> <dbl> <dbl>

## 1 0029LnfI bet365 0.346 0.371 0.284

## 2 0029LnfI Betsson 0.337 0.372 0.291

## 3 0029LnfI bwin 0.341 0.379 0.280

## 4 0029LnfI Pinnacle 0.359 0.354 0.287

## 5 00ApPQui bet365 0.409 0.307 0.283

## 6 00ApPQui Betfair 0.369 0.339 0.291

## 7 00ApPQui Betsson 0.412 0.309 0.280

## 8 00ApPQui bwin 0.399 0.312 0.290

## 9 00ApPQui Pinnacle 0.370 0.342 0.288

## 10 00IbEw7m bet365 0.431 0.280 0.289

## # ... with 10,062 more rows

*#Plotting P(home win) - P(away win) vs. P(tie) utilizing both probability calculation approach*

calculation1 <- probabilities %>% mutate(diff1= Phomewin - Pawaywin)

calculation2 <- probabilities\_norm %>% mutate(diff2= Pnorhomewin - Pnorawaywin)

ggplot(data= calculation1, aes(x=diff1, y=Ptie)) + geom\_point(stat= "identity") + labs(x= "", y= "", title= "P(home win) - P(away win) vs. P(tie) with First Prabability Calculation Approach") + theme\_bw() + theme(legend.position = "none", axis.text.x = element\_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12))

ggplot(data= calculation2, aes(x=diff2, y=Pnortie)) + geom\_point(stat= "identity") + labs(x= "", y= "", title= "P(home win) - P(away win) vs. P(tie) with Second Prabability Calculation Approach") + theme\_bw() + theme(legend.position = "none", axis.text.x = element\_text(angle = 45, vjust = 0.5, hjust = 0.5, size = 12))