Statistical analysis of Liverpool's 2024 25 Season

Emmanuel Bungei - 3524891

2025-10-10

Introduction

This project analyses Liverpool FC's **2024/25 Premier League** season using data captured from all the 38 matches both home and away. The dataset consist of match date, opponent, venue, result, goals scored (GS), goals conceded (GC), possession (Poss) and attendance. I captured the dataset from https://fbref.com/en/squads/822bd0ba/2024-2025/Liverpool-Stats#all_matchlogs.

Over the years I have developed an interest in the football sport and decided to combine it with analytics for this portfolio. The 2024-2025 Premier League season was Liverpool's winning campaign which provided me with a good dataset for applying the course's statistical methods. The photo below shows LFC team as they lifted the trophy of the season.

Figure 1: Liverpool team lifting the trophy

```
lfc <- read.table("LiverpoolFC_2024_25.txt", header = TRUE, sep = ",")
lfc
```

```
##          Date  Time        Opponent Venue Result GS GC Poss Attendance
## 1  17/08/2024 12:30    Ipswich Town  Away      W  2  0   62      30014
## 2  25/08/2024 16:30       Brentford  Home      W  2  0   62      60017
## 3  01/09/2024 16:00  Manchester Utd  Away      W  3  0   47      73738
## 4  14/09/2024 15:00 Nottham Forest  Home      L  0  1   68      60344
## 5  21/09/2024 15:00     Bournemouth  Home      W  3  0   58      60347
## 6  28/09/2024 17:30          Wolves  Away      W  2  1   55      31413
## 7  05/10/2024 12:30 Crystal Palace  Away      W  1  0   68      25185
## 8  20/10/2024 16:30         Chelsea  Home      W  2  1   43      60277
## 9  27/10/2024 16:30         Arsenal  Away      D  2  2   55      60383
## 10 02/11/2024 15:00        Brighton  Home      W  2  1   49      60331
## 11 09/11/2024 20:00     Aston Villa  Home      W  2  0   62      60292
## 12 24/11/2024 14:00     Southampton  Away      W  3  2   62      31278
## 13 01/12/2024 16:00 Manchester City  Home      W  2  0   44      60248
## 14 04/12/2024 19:30   Newcastle Utd  Away      D  3  3   58      52237
## 15 14/12/2024 15:00          Fulham  Home      D  2  2   61      60333
## 16 22/12/2024 16:30       Tottenham  Away      W  6  3   48      61439
## 17 26/12/2024 20:00  Leicester City  Home      W  3  1   68      60300
## 18 29/12/2024 17:15        West Ham  Away      W  5  0   54      62476
## 19 05/01/2025 16:30  Manchester Utd  Home      D  2  2   53      60275
## 20 14/01/2025 20:00  Nottham Forest  Away      D  1  1   70      30249
## 21 18/01/2025 15:00       Brentford  Away      W  2  0   60      17215
## 22 25/01/2025 15:00    Ipswich Town  Home      W  4  1   70      60420
## 23 01/02/2025 15:00     Bournemouth  Away      W  2  0   51      11239
## 24 12/02/2025 19:30         Everton  Away      D  2  2   63      39280
## 25 16/02/2025 14:00          Wolves  Home      W  2  1   50      60248
## 26 19/02/2025 19:30     Aston Villa  Away      D  2  2   48      41910
## 27 23/02/2025 16:30 Manchester City  Away      W  2  0   34      52803
## 28 26/02/2025 20:15   Newcastle Utd  Home      W  2  0   61      60374
## 29 08/03/2025 15:00     Southampton  Home      W  3  1   71      60399
## 30 02/04/2025 20:00         Everton  Home      W  1  0   73      60457
## 31 06/04/2025 14:00          Fulham  Away      L  2  3   63      27770
## 32 13/04/2025 14:00        West Ham  Home      W  2  1   55      60376
## 33 20/04/2025 16:30  Leicester City  Away      W  1  0   58      30402
## 34 27/04/2025 16:30       Tottenham  Home      W  5  1   61      60415
## 35 04/05/2025 16:30         Chelsea  Away      L  1  3   64      39829
## 36 11/05/2025 16:30         Arsenal  Home      D  2  2   45      60324
## 37 19/05/2025 20:00        Brighton  Away      L  2  3   51      31611
## 38 25/05/2025 16:00  Crystal Palace  Home      D  1  1   69      60382
```

1. Student's *t*-test

For the T-test, I will investigate whether playing at home resulted to scoring more goals compared to playing away. The test conducted is a one sided test since a team tends to win more games at home due to the 'Home advantage' and support from the big number of fans. My null hypothesis is that the average goals scored at home is less than or equal to the average goals scored away.

```
home_goals <- lfc$GS[lfc$Venue == "Home"]
away_goals <- lfc$GS[lfc$Venue == "Away"]

home_goals
```

```
##  [1] 2 0 3 2 2 2 2 2 3 2 4 2 2 3 1 2 5 2 1
```

```
away_goals
```

```
##  [1] 2 3 2 1 2 3 3 6 5 1 2 2 2 2 2 2 1 1 2
```

```
mean(home_goals)
```

```
## [1] 2.210526
```

```r
mean(away_goals)
```

```
## [1] 2.315789
```

```r
t.test(home_goals, away_goals,alternative = "greater")
```

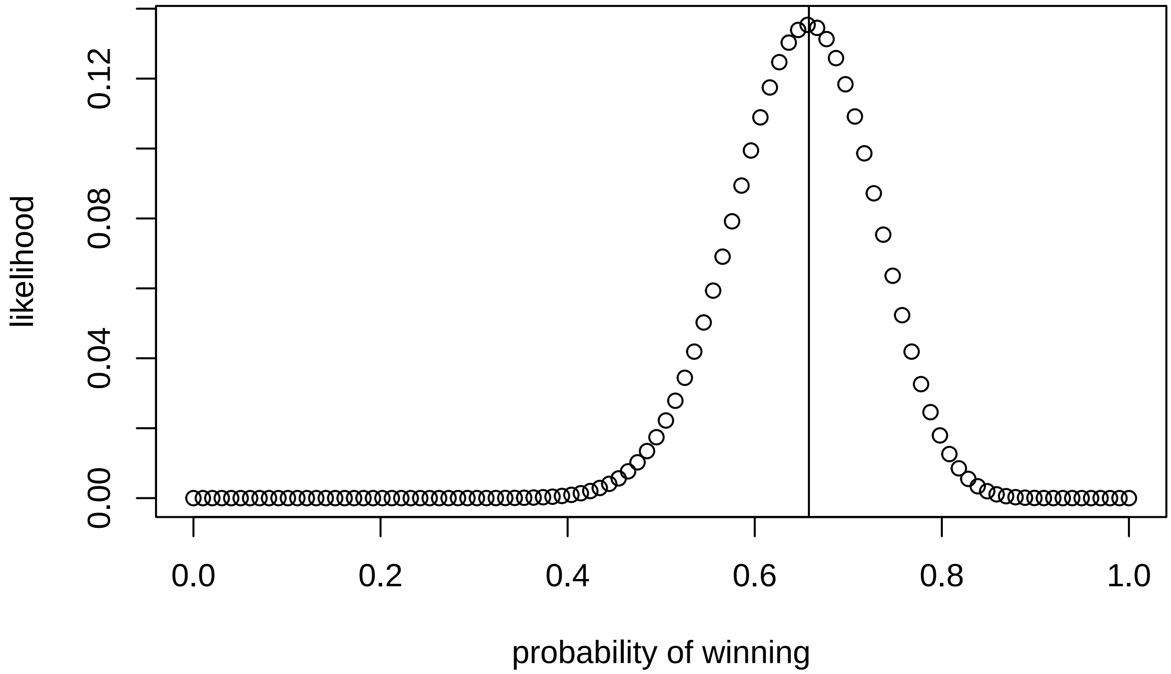
```
##
##  Welch Two Sample t-test
##
## data:  home_goals and away_goals
## t = -0.27189, df = 34.936, p-value = 0.6063
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.7594294        Inf
## sample estimates:
## mean of x mean of y
##  2.210526  2.315789
```

The p-value = 0.6063 means that there is a 60.63% chance of observing the data, which is way higher than 5% making the result not statistically significant. It is evident that playing at home does not result to scoring more goals and any observed difference is due to random chance. We therefore fail to reject the null hypothesis since we do not have enough evidence and may require more data.

2. Likelihood function

Here my goal is to get the estimate probability of Liverpool winning a match in the season. They played 38 matches and won 25 of them giving a probability of 0.6579. Below is a distribution graph that assesses the probability of different values across p.

```r
p <- seq(from=0,to=1, len=100)
plot(p,dbinom(25,38,p), xlab="probability of winning", ylab="likelihood")
abline(v=25/38)
```

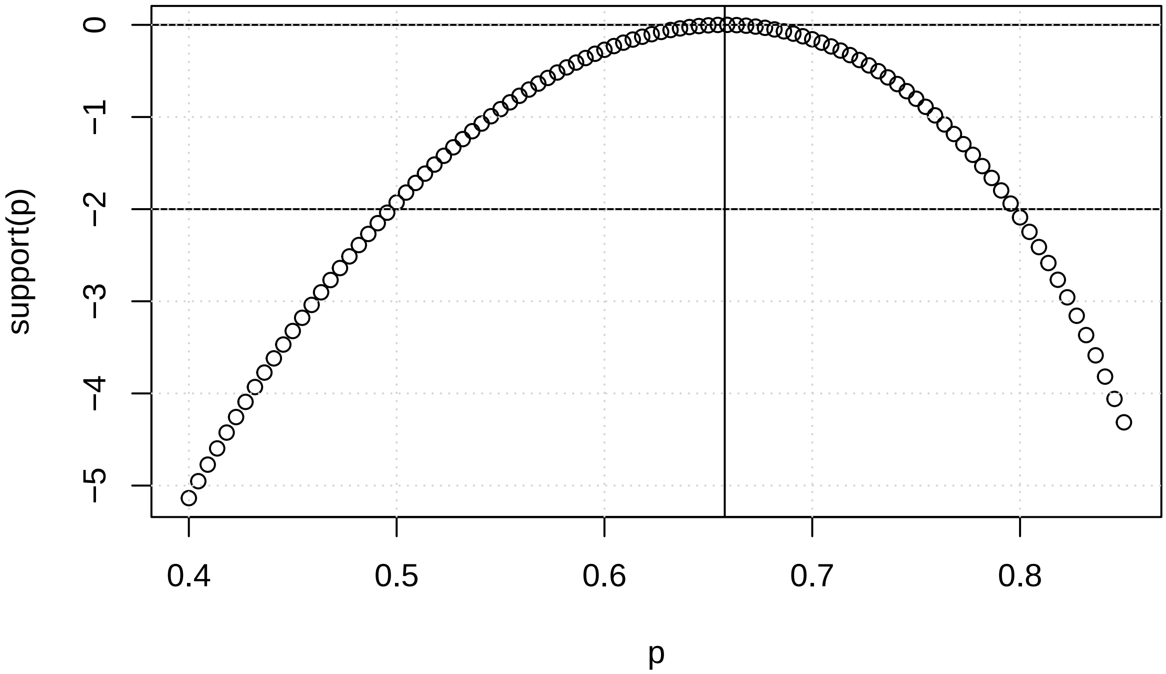


The likelihood function peaks at approximately 0.65 where the graph intersects with the perpendicular abline marking the maximum likelihood estimate(MLE).

The support function

From the support function, I get to determine the credible interval acceptable for a range of p values.

```r
p<-seq(from=0.4,to=0.85,len=100)
support<-function(p){dbinom(25,38,p,log=TRUE)-dbinom(25,38,25/38,log=TRUE)}
plot(p,support(p))
abline(h=-2)
abline(h=0)
abline(v=25/38)
grid()
```



```
f<-function(p){support(p)+2}
uniroot(f,c(0.4,0.6))
```

```
## $root
## [1] 0.4970105
##
## $f.root
## [1] 1.54721e-05
##
## $iter
## [1] 5
##
## $init.it
## [1] NA
##
## $estim.prec
## [1] 6.276826e-05
```

```
uniroot(f,c(0.75,0.85))
```

```
## $root
## [1] 0.7973163
##
## $f.root
## [1] 3.021924e-05
```

```
##
## $iter
## [1] 5
##
## $init.it
## [1] NA
##
## $estim.prec
## [1] 6.103516e-05
```

The line at -2 cuts off the graph at the lower and upper limits where credible p values lie. From the function, our lower limit is 0.497 while our upper limit is 0.797. Therefore a credible interval for the probability of winning a match is between (0.497, 0.797)

3. Fisher's exact test

This test examines whether the probability of winning differs between home and away matches. The test is a one-sided test since we expect the team to perform better at home because of the home advantage and large number of fan attendance. The null hypothesis therefore is that the probability of winning at home is the same as winning away.

```r
M = matrix(c(14,11,5,8),2,2)
dimnames(M) <- list(Venue=c("Home","Away"),Result=c("Win","Not Win"))
M
```

```
##       Result
## Venue  Win Not Win
##   Home  14       5
##   Away  11       8
```

```r
dhyper(14,25,13,19)
```

```
## [1] 0.1623039
```

```r
sum(dhyper(14:25,25,13,19))
```

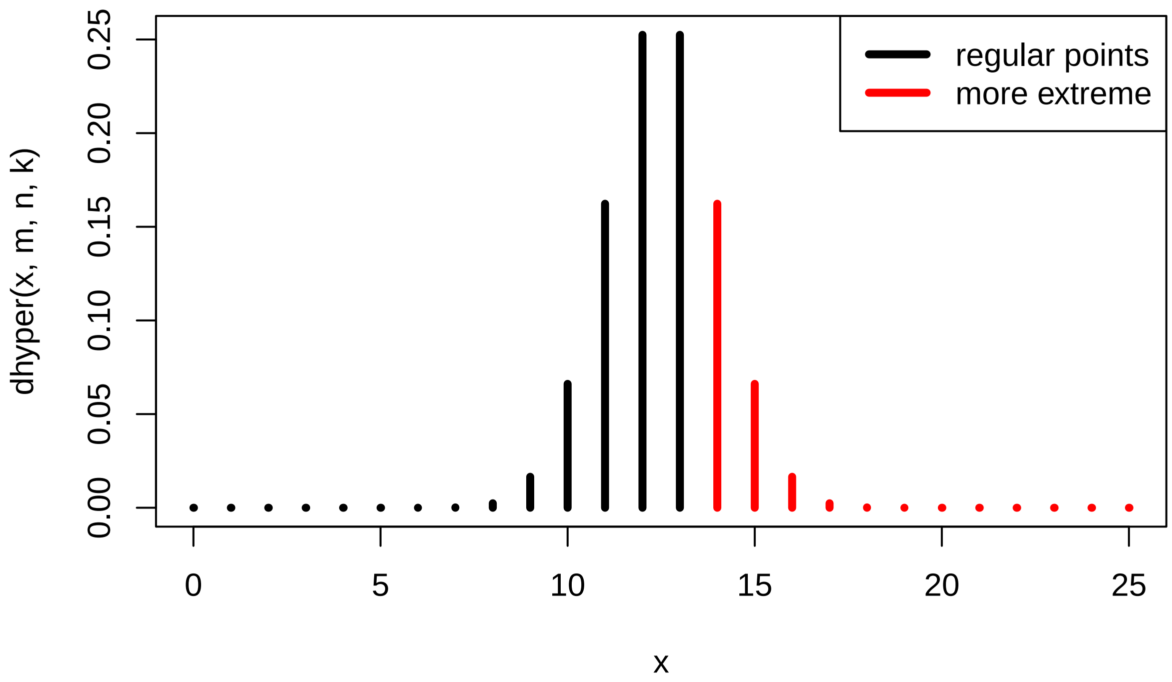
```
## [1] 0.2475273
```

```r
m <- 25
n <- 13
k <- 19

x <- 0:25

plot(x, dhyper(x, m, n, k), type = "h", lwd = 4, col = c(rep("black",14),rep("red",12)))

legend("topright", lwd = 4, col = c("black", "red"),
       legend = c("regular points", "more extreme"))
```



```r
fisher.test(M,alternative="greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  M
## p-value = 0.2475
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.5290407       Inf
## sample estimates:
## odds ratio
##   1.998097
```

The plot highlights the probability distribution of home wins and the red bars marking the values more extreme than the observed home wins.

p-value = 0.2475. We fail to reject the null hypothesis since there is no strong evidence that Liverpool is more likely to win playing at home compared to away.

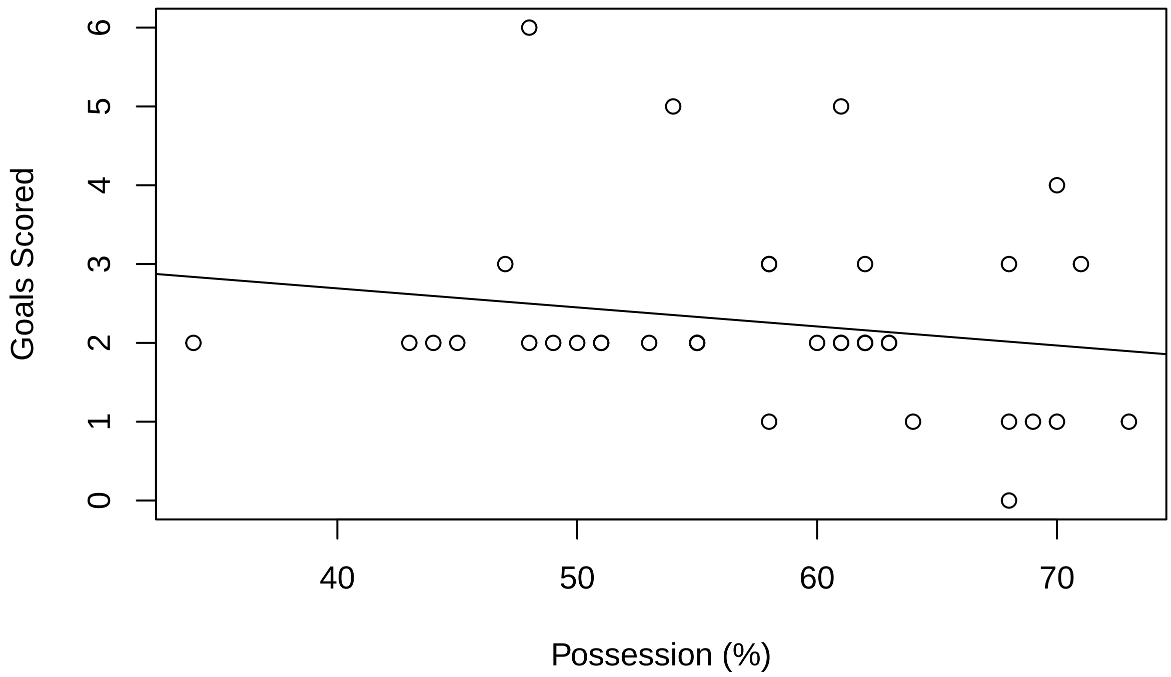
4. Linear Regression

For linear regression, I investigate whether the ball possession has a significant effect on the number of goals scored.

```
model <- lm(GS ~ Poss, data = lfc)
model
```

```
##
## Call:
## lm(formula = GS ~ Poss, data = lfc)
##
## Coefficients:
## (Intercept)         Poss
##     3.65753     -0.02415
```

```
plot(lfc$Poss, lfc$GS, xlab = "Possession (%)", ylab = "Goals Scored")
abline(model)
```



```
summary(model)
```

```
##
## Call:
## lm(formula = GS ~ Poss, data = lfc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0153 -0.5889 -0.3293  0.3241  3.5017
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.65753    1.23568   2.960  0.00541 **
## Poss        -0.02415    0.02115  -1.142  0.26097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.173 on 36 degrees of freedom
## Multiple R-squared:  0.03496,    Adjusted R-squared:  0.008157
## F-statistic: 1.304 on 1 and 36 DF,  p-value: 0.261
```

From the plot, there is a negative correlation between the possession and goals scored denoted by the regression line. The higher possession does not necessarily lead to scoring more goals. The regression line is slightly off from the points because our y data is the number of goals which are integers and not continuous making it harder to plot. The negative correlation may have resulted from several factors including counter-attacking. Despite having lower possession against stronger teams, Liverpool may have relied on quick counter attacks scoring more goals with less possession. Also scoring early and holding on to the ball for the rest of the match increases their possession and while maintaining a lower number of goals.

Conclusion

I have learnt to apply the 4 statistical techniques to analyse the real football data and critically evaluate the results. In my evaluation, I was surprised to find out that possession did not have a positive effect on goals scored. Similarly, on the Fisher's exact test there was no significant difference between home and away performance despite having the home advantage on their home games. The outcome challenged normal beliefs on football games through the statistical results generated. Even though Liverpool had a strong win rate, the statistical evidence showed that the possession and venue are not strong predictors of a successful game.