DATA ANALYSIS OF PREMIER LEAGUE PLAYERS

Anjali Chauhan, Nicolás Harrington, Chenming Ye,
Teila Raistrick, Madison Lacoursiere

1 Introduction

1.a Background

The Premier League is one of the most widely watched and wealthiest leagues in the world of soccer. Its clubs spend millions of dollars on players and development, and increasingly this development is focusing on data analysis. Several of the recent Premier League success stories, such as Liverpool and Southampton, have been attributed to data analysis departments in areas such as scouting. Analyzing the current makeup of the Premier League can help us understand the players that clubs are looking for, as well as comparing the league to others in Europe. Experience is a major factor when assessing a player, and experience within a certain league is especially valuable. One measure of this experience is appearances. This makes the target population of our study the currently registered players in the Premier League, which is currently 570 players with proper information available (one player has some missing data, including age). Each of these players is a possible sampling unit, with two samples being taken using simple random sampling (SRS) and stratified sampling (STR). The parameters of interest will be the mean number of appearances for the currently registered players and the proportion of players that have at least 100 Premier League appearances, which makes them fairly experienced.

1.b Objectives and Relevance

Our goal is to create a detailed analysis of players experience within the league by investigating the mean player appearances for actively registered players. We will do this using two separate sampling methods; Simple Random Sampling and Stratified Sampling. We will then compare our results and look at how they differ in statistical accuracy between the simple and stratified samples with the strata being based on a player's age.

Player performance statistics can be beneficial in many ways, for both those within the league and those watching from the sidelines. We can use the appearance data to find important statistics, such as the per-game performance statistics for each player. These statistics are very popular for player comparison and are used widely by both teams and fans within the league. Assuming players with more appearances perform better than those with fewer appearances, we are able to determine which players are prized amongst each team within the league. This can be beneficial to teams looking to improve their performance by acquiring more talented players.

Fans of the game may also use this data for examining a team's portfolio to determine their chances of winning the league that year. These performance statistics are crucial within any gambling environment, providing bookmakers and bettors key statistics to allow for educated bets and economic margins. Without them, there would be little data available for accurate projections, and the hobby would not exist. These statistics can also be helpful when creating merchandise for the teams, for example assigning the corresponding values to player-specific merchandise.

This data also has applications outside of the league we are sampling from. Analyzing the average number of performances a player has in the league, along with the proportion of players with at least 100 appearances, will allow us to begin comparisons between the major leagues of Europe. Since the top clubs in each league qualify for the UEFA Champions League, they will end up competing against one another, so analysts will want all the data available in order to help their team perform better in a given match. With rumors of a European Super League circulating currently, players who are well established in a league (with many appearances) could be sought after by teams looking for players who are proven to have the quality to play at the highest level.

2 Data Summaries and Sampling Method

2.a Data Summaries

The dataset includes the 571 currently registered Premier League players and their corresponding attributes and statistics, such as nationality, club, position, age, appearances, etc.

We must clean the data before we begin to sample. Since we will categorize the players' age into three sub-populations for stratified sampling, the players with no listed age should not be used in the following sampling. After cleaning the data, we found there are $N = 570$ players satisfying our requirements.

To explore the data, we will use both simple random sampling and stratified sampling, and compare these two sampling methods to examine gives a better result.

2.b Sampling Method

2.b.1 Simple Random Sampling (SRS)

Before deciding on the sample size, we set the significance level α to be 0.05 and the half-width δ of the $(1 - \alpha)$ confidence interval to be 0.1, as we do not want more error than that in our estimates. The sample variance S_s^2 is unknown, and we have no prior studies to refer to, so we can use the conservative case where $S_{guess}^2 = 0.25$.

$$n_0 = \frac{Z_{\alpha/2} S_{guess}^2}{\delta^2}$$

Since the population size N is already known and small, a finite population correction factor is used to minimize the sample size.

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

After rounding the number to an integer, the sample size n should be 82. Set the sample size $n = 82$, and pick the random sample from the population.

2.b.2 Stratified Sampling (STR)

How we choose to stratify the data is very important, as using different stratification conditions will affect the final result. We conduct a separate SRS in each domain, stratifying the population P by age into three sub-populations. We set the sub-population with ages greater than or equal to 30 as "Oldest", denoted as P_{Oldest}, the sub-population with ages less than or equal to 23 as "Young", denoted as P_{Young}, and the remaining population as a sub-population "Prime", denoted P_{Prime}. We have the following relationships:

$$P_{Oldest} \cup P_{Prime} \cup P_{Young} = P$$

$$N_{Oldest} + N_{Prime} + N_{Young} = N$$

With different sub-population sizes but no assumed difference in variance, we use proportional allocation in order to determine strata size. The sub-population weights are calculated as:

$$w_{Oldest} = \frac{N_{Oldest}}{N}$$

$$w_{Prime} = \frac{N_{Prime}}{N}$$
$$w_{Young} = \frac{N_{Young}}{N}$$

By multiplying the sample size n by the corresponding weight for each stratum, we get the necessary sample size for each stratum. After rounding the result to an integer, the sample size of strata should be:

$$n_{Oldest} = \frac{N_{Oldest}}{N} \cdot n \approx 17$$
$$n_{Prime} = \frac{N_{Prime}}{N} \cdot n \approx 37$$
$$n_{Young} = \frac{N_{Young}}{N} \cdot n \approx 28$$

3 Data Analysis

3.a Assumptions

The statistical analysis done is based on the following assumptions:

1. Normality of the Data (and of estimators)
2. Proportional Allocation (under equal assumed variance)
3. Finite Population Correction Applies (knowledge of population size)

Remark: No cost allocation in this scenario.

3.b Formulas

3.b.1 SRS Estimates:

To analyze the SRS, we will use a vanilla estimate of the mean and its corresponding standard error. These directly estimate the population mean and standard deviation, and their formulas are listed below.

1. $\hat{p} = \sum_i^N \frac{x_i}{n}$ with x_i being an indicator variable for a condition. In this case, the condition is having at least 100 appearances.
2. $SE(\hat{p}) = \sqrt{(1 - \frac{n}{N}) \cdot \frac{\hat{p}(1-\hat{p})}{n}}$
3. $\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i$
4. $SE(\bar{y}) = \sqrt{(1 - \frac{n}{N}) \cdot \frac{s_S^2}{n}}$

3.b.2 SRS CIs:

We are calculating the 95% Confidence Intervals for proportion and mean of the SRS sample:

1. $\bar{y} \pm z^*_{0.975} SE(\bar{y}) \implies \bar{y} \pm z^*_{0.975} \sqrt{(1 - \frac{n}{N}) \cdot \frac{s_S^2}{n}}$
2. $\hat{p} \pm z^*_{0.975} SE(\hat{p}) \implies \hat{p} \pm z^*_{0.975} \sqrt{(1 - \frac{n}{N}) \cdot \frac{\hat{p}(1-\hat{p})}{n}}$

   where the critical normal value at the .05 significance level is $z^*_{0.975} \approx 1.96$

3.b.3 STR Estimates:

For the Stratified Sample, we can analyze its results using the sample strata means and standard deviations. This will result in a more precise estimate for the population mean, as each sample strata will estimate only the within-strata variance. This analysis, seen in formula 4, makes the standard error of our estimate the weighted sum of the sample strata, which will generally lead to a smaller estimate variance than the SRS.

1. $\hat{p}_{str} = \sum_{h=1}^{H} (\frac{N_h}{N}) \cdot \hat{p}_{S_h}$
2. $SE(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N}) \cdot SE^2(\hat{p}_{S_h})}$
3. $\bar{y}_{str} = \sum_{h=1}^{H} (\frac{N_h}{N}) \cdot \bar{y}_{S_h}$
4. $SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N}) \cdot SE^2(\bar{y}_{S_h})}$; where $SE^2(\bar{y}_{S_h}) = (1 - \frac{n_h}{N_h}) \cdot \frac{s^2_{S_h}}{n_h}$

3.b.4 STR CIs:

We are calculating the 95% Confidence Intervals for proportion and mean of the Stratified sample:

1. $\bar{y} \pm z^*_{0.975} SE(\bar{y}_{str}) \implies \bar{y} \pm z^*_{0.975} \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N}) \cdot (1 - \frac{n_h}{N_h}) \cdot \frac{s^2_{S_h}}{n_h}}$
2. $\hat{p} \pm z^*_{0.975} SE(\hat{p}_{str}) \implies \hat{p} \pm z^*_{0.975} \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N}) \cdot SE^2(\hat{p}_{S_h})}$

   where the critical normal value at the .05 significance level is $z^*_{0.975} \approx 1.96$

3.c Estimation

Estimate	SRS	STR
Proportion	0.26829	0.29412
SE($\hat{p}$)	0.04527	0.03624
Mean	62.34146	74.77977
SE($\bar{y}$)	7.73289	5.93186

Table 1: Estimation Results (rounded off to 5 decimal places)

3.d Confidence Intervals

CI	SRS	STR
Proportion	[0.17956, 0.35703]	[0.22310, 0.36514]
Mean	[47.18527,77.49765]	[63.15353, 86.40600]

Table 2: Confidence Intervals (rounded off to 5 decimal places)

3.d.1 Interpretation:

The confidence intervals are all constructed using a 5% significance level, which makes them 95% confidence intervals. In 19 out of the 20 repeated studies, a confidence interval constructed in the same manner would contain the true parameter value it is constructed for.

Proportion intervals:
For the SRS and STR, we are 95% confident the true proportion lies between the values of the table. The SRS interval has a smaller lower bound than the STR interval, and both have similar upper bounds, with the STR's being slightly higher. The SRS's proportion interval of 0.17956 to 0.35703 and the STR's of 0.22310 to 0.36514 both indicate the majority of players do not have over 100 appearances. Since there is 95% confidence this proportion is in these ranges, we can conclude the movement of players between other leagues and the Premier League makes it so most do not play in the league for long, or they are bit-part players for a team over a longer period of time.

Mean Intervals:
The confidence intervals are [47.18527,77.49765] for the SRS and [63.15353, 86.40600] both agree with the proportional findings. Since we have 95% confidence that the mean number of appearances fall in these ranges, active players appear to have more than one season of appearances on average (38 games per season).

3.e Comparison of different methods used

The two methods used for sampling are SRS and Stratified. Based on our computations and results from Table 1 and Table 2, we observe that the Stratified method is a better-suited method for the data because:

- The Standard Error of Proportion SE($\hat{p}_{str}$) and Means $SE(\bar{y}_{str})$ for Stratified Sampling is 0.03624 and 5.93186 respectively which is smaller than the Standard Error of Proportion SE($\hat{p}$) and Means $SE(\bar{y})$ for SRS, 0.04527 and 7.73289 respectively.
- The Confidence Intervals of Proportion and Means for Stratified Sampling are [0.22310, 0.36514] and [63.15353, 86.40600] respectively, which are narrower than the Confidence Intervals of Proportion and Means for SRS, [0.17956, 0.35703] and [47.18527, 77.49765] respectively.

Therefore, the data obtained supports the claim that Stratified Sampling is better than SRS.

4 Conclusion and Discussion

As stated in our objectives, the results of our study provide crucial data for the Premier League players and teams, fans of the sport, as well as outside leagues looking to recruit new players. Importantly, our data contains strong ties to a player's ability, assuming that increased appearances imply improved abilities. For both of our binary estimators, the proportion of players with more than 100 appearances was approximately 0.28. The relatively small size of this proportion implies that it could actually be applied to our review of player ability. If, for example, this proportion were closer to 1, it would have very few applications since almost every player would be considered a "good" player under this framework.

By estimating our parameters with two distinct sampling methods, simple random sampling and stratified sampling, we were able to compare the benefits of either method in a new environment.

Both sampling methods provided similar estimates; however, as expected, stratified sampling resulted in a smaller standard error, making it the preferred sampling method. In particular, the stratified sample's standard error is 23.3% smaller than the SRS's standard error for the mean number of appearances estimate and 19.9% smaller for the proportion estimate. As discussed in our lectures, stratified sampling limits standard error by reducing the sampling variability present and limiting estimation to within strata variance, thus providing more precise results.

Our 95% confidence interval for the mean number of player appearances in the Premier League was in line with our expectations. For reference, the confidence interval using the stratified sample was [63, 86], rounded to the nearest whole number. To place this into context, there are 38 Premier League games each season, and most players would not play in every game, thus demonstrating our results are relatively feasible. As such, the results of our study are generalizable to all of the players currently registered in the Premier League.

4.a Limitations

Upon reviewing our data, there are some limitations to be noted along with recommendations for future studies. The primary limitation we faced was the availability of data, and as a result of this, the identification of appropriate auxiliary variables. Although we were pleased with the results of our stratified estimates in stratifying by age, this variable is likely not the most representative indicator for estimating the mean number of appearances for players registered in the Premier League. For example, two players could be 30 years old, with one player having joined the league at 17, while the other having just joined this year, which could lead to high within strata variance and thus undermine the efficacy of stratified estimation.

Therefore, it would be our recommendation for future researchers to gather further information relating to the number of years each player has been in the league and use those results as their stratifying variable. This would help improve the estimates because a player's time in the league is more closely related to their number of appearances in the league. Furthermore, a stronger positive correlation between variables would improve the performance of a ratio estimate over a vanilla estimate. In our study, a ratio estimate would not have significantly improved our results since the correlation between age and number of appearances was only around 0.6. The R code for finding the correlation coefficient can be found in the appendix below.

5 Appendix

5.a Data

Our population data was sourced from Kaggle, showing the currently registered Premier League player statistics. The data was harvested on 6-11-20 but continues to be updated weekly throughout the season to reflect current player performance. Players may be missing data when not applicable to their positions, though as we are not testing using any role-reliant parameters this did not affect our results. However, we did experience players with incomplete data (when relevant information is omitted) and accounted for this by removing them from our sample. Similar data can be found pertaining to other leagues in Europe that can be used to cross-validate our results.
Here is a link to our SRS and STR sampled data in two tables.

5.b R Code

Sampling Code:

```
PremierLeague <- read.csv("PremierLeague.csv", header = T)

#Cleaning data using min and max age range
PremierLeague <- PremierLeague[which(PremierLeague$Age >= 17 & PremierLeague$Age
    <= 38),]
set.seed(136)
n = ((.1/qnorm(.975))^-2)*.25
N = nrow(PremierLeague)
#Simple random sample of Registered players
SRS.Index <- sample.int(N, n, replace = F)
SRS.samp <- PremierLeague[SRS.Index, ]

#Correlation
cor(PremierLeague$Age, PremierLeague$Appearances)

#Stratified Sample using proportional allocation
#Age Groups
Oldest <- PremierLeague[which(PremierLeague$Age >= 30),]; N.Old <- nrow(Oldest)
Prime <- PremierLeague[which(PremierLeague$Age <= 29 & PremierLeague$Age >= 24),];
     N.Prime <- nrow(Prime)
Young <- PremierLeague[which(PremierLeague$Age <= 23),]; N.Young <- nrow(Young)
weight <- c((N.Old/N), (N.Prime/N), (N.Young/N))

#Proportional allocation
n.h <- round(weight*n, 0)
N.h <- c(N.Old, N.Prime, N.Young)

#Sampling
Old.samp.index <- sample.int(N.Old, n.h[1], replace = F); Old.samp <- Oldest[Old.
    samp.index,]
Prime.samp.index <- sample.int(N.Prime, n.h[2], replace = F); Prime.samp <- Prime[
    Prime.samp.index,]
Young.samp.index <- sample.int(N.Young, n.h[3], replace = F); Young.samp <- Young[
    Young.samp.index,]
```

Estimating mean appearances:

```
SRS.mean <- mean(SRS.samp$Appearances)
SRS.se <- sqrt((1 - n/N) * (var(SRS.samp$Appearances)/n))
ybar.h <- c(mean(Old.samp$Appearances), mean(Prime.samp$Appearances), mean(Young.
    samp$Appearances))
```

```r
var.h <- c(var(Old.samp$Appearances), var(Prime.samp$Appearances), var(Young.samp$Appearances))
se.h <- sqrt((1 - n.h / N.h) * var.h / n.h)
rbind(ybar.h, se.h)
ybar.str <- sum(weight * ybar.h)
se.str <- sqrt(sum(weight^2 * se.h^2))
c(ybar.str, se.str); c(SRS.mean, SRS.se)
```

Estimating proportion of players with at least 100 appearances:

```r
#Parameter 2: Proportion of players with at least  100 appearances
#SRS Sample
Experienced.samp <- SRS.samp[which(SRS.samp$Appearances >= 100),]
prop.exp <- nrow(Experienced.samp)/nrow(SRS.samp)
se.prop.exp <- sqrt((1 - n/N)*((prop.exp)*(1-prop.exp))/n)

#Stratified Sample

#Old
Old.samp.exp <- Old.samp[which(Old.samp$Appearances >= 100),]
Old.prop.exp <- nrow(Old.samp.exp)/nrow(Old.samp)
Old.var.prop.exp <- (Old.prop.exp)*(1-Old.prop.exp)

#Prime
Prime.samp.exp <- Prime.samp[which(Prime.samp$Appearances >= 100),]
Prime.prop.exp <- nrow(Prime.samp.exp)/nrow(Prime.samp)
Prime.var.prop.exp <- (Prime.prop.exp)*(1-Prime.prop.exp)

#Young
Young.samp.exp <- Young.samp[which(Young.samp$Appearances >= 100),]
Young.prop.exp <- nrow(Young.samp.exp)/nrow(Young.samp)
Young.var.prop.exp <- (Young.prop.exp)*(1-Young.prop.exp)

#Together
prop.h <- c(Old.prop.exp, Prime.prop.exp, Young.prop.exp)
var.prop.h <- c(Old.var.prop.exp, Prime.var.prop.exp, Young.var.prop.exp)
se.prop.h <- sqrt((1 - n.h / N.h) * var.prop.h / n.h)
rbind(prop.h,se.prop.h)
prop.str.hat <- sum(weight * prop.h)
se.prop.str <- sqrt(sum(weight^2 * se.prop.h^2))
print(c(prop.exp, se.prop.exp))
print(c(prop.str.hat, se.prop.str))
```

Confidence Intervals:

```r
#Confidence Intervals
#SRS mean appearances
lb1 <- SRS.mean - qnorm(.975)*SRS.se; ub1 <- SRS.mean + qnorm(.975)*SRS.se
CI.SRS <- c(lb1, ub1); CI.SRS

#Stratified mean appearances
lb2 <- ybar.str - qnorm(.975)*se.str; ub2 <- ybar.str + qnorm(.975)*se.str
CI.STR <- c(lb2, ub2); CI.STR

#SRS proportion experienced
lb3 <- prop.exp - qnorm(.975)*se.prop.exp; ub3 <- prop.exp + qnorm(.975)*se.prop.exp
CI.SRS.prop <- c(lb3, ub3); CI.SRS.prop

#STR proportion experienced
lb4 <- prop.str.hat - qnorm(.975)*se.prop.str; ub4 <- prop.str.hat + qnorm(.975)*se.prop.str
CI.STR.prop <- c(lb4, ub4); CI.STR.prop
```