# PROJECT

*Asare Buahin*

*3/25/2018*

I. Introduction

My topic of interest for this project is the Player Efficiency Rating(PER) in professional basketball, and how it correlates to how good a player is. PER is a per minute rating that sums up a player's positive accomplishments and subtracts the player's negative accomplishments and ultimately depicts a player's performance. The PER is standardized so that every single year, the average NBA player in terms of performance has a PER of 15. I chose this as a topic of interest because I have always had an interest in advanced statistics used to measure basketball players' effectiveness. Readers of this report would be interested inmy results because much has been made of advanced statistics such as PER being used as a baseline to measure how good a basketball player is, as the calculation of PER does not properly take into account a player's defensive accomplishments. Readers would be interested in these results to see if a high PER actually translates to individual success (All NBA team selection).

To see the correlation between PER and player performance, I will find whether a player with a PER of 21 or over that plays more than 48 games in a season is selected for an ALL NBA teams selected at the end of the year. This study has not been done anywhere else. I chose 48 games as a minimum amount of games played because the smallest amount of games played by a player who was still selected for an ALL NBA team was 48 games. This happened in the 2006-07 season when Yao Ming played a total of 48 games and was selected to the ALL NBA 2nd team. This minimum was also chosen because eliminates the players who have a played a small amount of minutes and have an unusually high PER. For example, Naz Mitrou Long has played 1 minute for the Utah Jazz in the 2017-18 NBA season and has a PER of 135.4 A PER of 21 was chosen as a limit because a player with a PER of 21 or over is consiered a good player, considering an average player has a PER of 15 every year and the highest PER recorded over a season was 31.82(by Wilt Chamberlain).

The parameter of interest for this study is the probability that a player with a PER of 21 or over that plays more than 48 games in a season is selected for an ALL NBA team. The population of interest are individual seasons of NBA players with a PER of 21 or over that plays more than 48 games in a season. This population is taken from players who have played in the "Three Point era" which started in the 1979-00 season NBA season where the 3 point line was implemented into NBA basketball.My initial conjecture for the parameter of interest's is 2/3. This is based on my memory of the PERs of some players who have been selected for an ALL NBA Team in the past 3 years.

II: Data Colleciton Methods

To collect the data for my study, I use the Basketball Reference website, which has stats on all NBA players since the NBA-ABA merger(1976). My obersvational units for population of interest is NBA players with a PER of over 21 that played more than 48 games that I sample. Initially, the experiment was going to consider NBA players with a PER of over 22 who played more than 70 games.However, data from the NBA over its history shows that not too many players have a PER of 22 or over each season. For this reasons, reducing the minimum to 21 would be able to indicate that a player plays well while allowing for a sufficient amount of data to be collected for the experiment.While collecting data an unexpected trend that I noticed was that the further back in NBA history I went, there were less players with a PER of over 21 each season. For example, in the 2016-17 season, there were 28 players with a PER of over 21 and in 1982-83 season, there were only 13 players with a per of 21 or over. I feel like the reason for this trend is because a component of the calculation of PER is 3 pointers made and the pace a team plays at(the number of possesions a team has over a 48 minute span). With NBA basketball becoming faster over the years, along with players shooting more 3 pointers, PERs have increased as the way basketball is played has evolved.

The Basketball Reference website where the data was collected has a tool that allows you to get any desired data as an excel spreadsheet. So the site was used to collect every All NBA Team dating back to the 1979-1980

NBA season in an excel spreadsheet. The process of collecting data on PERs on players dating back to the 1979-1980 was a bit more complicated. To do this, I had to go to the database of advanced stats for every NBA player in each season dating back to the 1979-1980 season. On the site, you can organize the order of the players that show up on the database. So I chose to organize the data by having NBA players with the highest PER in that season at the top and the lowest PER at the bottom. When exporting the data to excel, each datasheet had the PERs of players for one season. For example one sheet had the PER for players in the 2016-17 season while the next sheet had the PER for players in the 2015-16 season. I deleted the data of players that did not meet my experiment conditions and combined the data I needed for the experiment from every sheet into a final data sheet.

To gather the sample for the experiment, the systematic sampling method was used in excel. I first randomized the data by assigning a random number to each row of data(each NBA player season) by using the excel RAND() function. All of the data was selected and sorted in order of smallest to largest number assigned by the RAND() function. The sample to be collected has 30 data points, so a random number generator was used to choose a number between 0 and 59 and that number would be used as a starting point. A number between 0 and 49 was used because of the 629 total data points we have. Using systematic sampling, you divide the number of data points you have by the number of data points in your sample. Since 629 divided by 30 is not a whole number, I divided 600 by 30 to get a result of 20 and added the extra 29 data points not accounted for. By doing all of this, we have an interval at which a data point is selected at 20 and a starting point between player 0 and player 49.This method allows for all data points collected in the sample to be considered as opposed to just the first 600 data points. The random number generator gave me a starting point of 26. So the first data point for the sample was data point number 26. The next data point was 46 because of the interval of 20 we constructed. This was done until there were 30 data points for the sample and all the data was placed in one excel sheet.

III: Analysis of Results

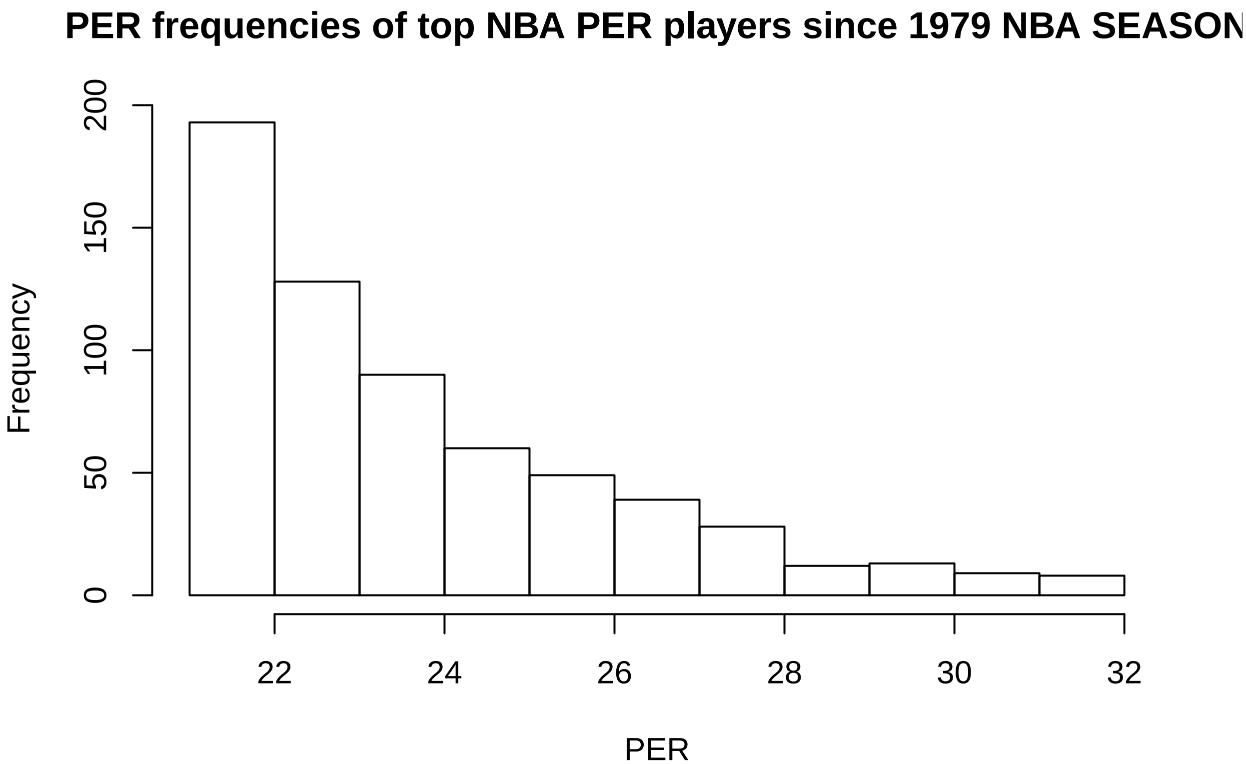
```r
library(readxl)
DATA_FOR_HISTOGRAM <- read_excel("~/Documents/Math 247 Fall 2017/PROJECT/DATA FOR HISTOGRAM.xlsx")
summary(DATA_FOR_HISTOGRAM)
```

```
##      Name                PER       
##  Length:629         Min.   :21.00  
##  Class :character   1st Qu.:21.80  
##  Mode  :character   Median :23.00  
##                     Mean   :23.73  
##                     3rd Qu.:25.10  
##                     Max.   :31.70  
```

```r
sd(DATA_FOR_HISTOGRAM$PER)
```

```
## [1] 2.419921
```

```r
hist(DATA_FOR_HISTOGRAM$PER, main="PER frequencies of top NBA PER players since 1979 NBA SEASON",
xlab="PER")
```




Above is a histogram of all the 692 data points collected for the experiment. We can see that the histogram has a unimodal shape that is skewed to the right, which makes sense as PERs in the 26 and above range belong to Hall of Fame level players. The data has a mean of 23.73, showing the high frequency of player seasons with PERs between 21 and 23. The population had a standard deviation of 2.42, showing the relatively large range of PERs in the population.

On the next page is the histogram for the sample made up of 30 data points randomly chosen with systematic sampling. The histogram has a bimodal shape, with 2 peaks at 21 and 25. This bimodal shape in comparison to the unimodal skewed right shape of the first histogram shows the randomness of the sample collected.The sample has a mean of 24.26, which is 0.5 PER points higher than the mean of the entire population. The sample has a standard deviation of 2.46 which is just a 0.24 higher than the standard deviation of the entire population collected. The closeness of the means and standard deviations shows that the sample collected is representative of the data.

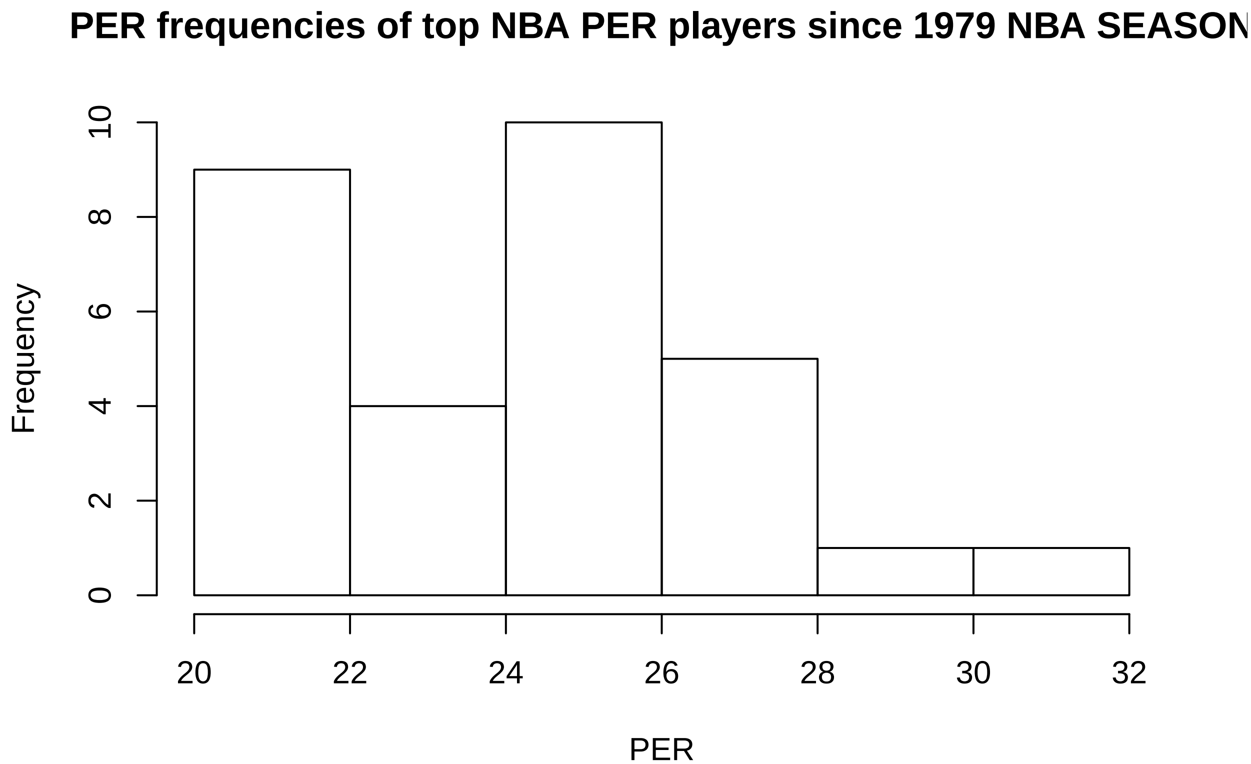
```r
library(readxl)
SAMPLE_FOR_HISTOGRAM <- read_excel("~/Documents/Math 247 Fall 2017/PROJECT/SAMPLE FOR HISTOGRAM.xlsx")
summary(SAMPLE_FOR_HISTOGRAM)
```

```
##      Name                PER
##  Length:30          Min.   :21.00
##  Class :character   1st Qu.:21.93
##  Mode  :character   Median :24.10
##                     Mean   :24.26
##                     3rd Qu.:25.60
##                     Max.   :30.60
```

```r
sd(SAMPLE_FOR_HISTOGRAM$PER)
```

```
## [1] 2.458551
```

```r
hist(SAMPLE_FOR_HISTOGRAM$PER, main="PER frequencies of top NBA PER players since 1979 NBA SEASON",
xlab="PER")
```




For the experiment, there were 629 total data points collected. Meaning, from the 1979-80 to 2016-17 NBA seasons, there were 629 different player seasons in which the player had a PER of 21 or over and had played at least 48 games in that season. From the sample collected, 19/30 player seasons were a success, having a PER of 21 or over that made the All NBA Team in the respective season. The population of interest for the experiment is seasons of NBA player who had a PER of 21 or over who played over 48 in the 3 point era(1979-2017). The parameter for this experiment is the probability that a player who had a season with a PER of 21 or over and played at least 48 games makes an All NBA team.

Below are the null and alternative hypotheses for the experiment

$H_o : \pi{=}2/3$ $H_a : \pi{<}2/3$

The null hypothesis for this study states that an NBA player with a PER of 21 or over who played at least 48 games in the season will be more than likely to make an all NBA team. The alternate hypothesis for this study states that a lower number than expected of NBA player with a PER of 21 or over who played at least 48 games in a season will make an All NBA Team. In this setting, a type 1 error would be that a player with a PER of over 21 and has played at least 48 games who is more likely to make an All NBA Team does not make an All NBA Team. Meaning, a player who has a higher PER does not make an All NBA team. A type 2 error would be that a player that a player with a PER of 21 or over that is less likely to make the All NBA team does not make an All NBA team.

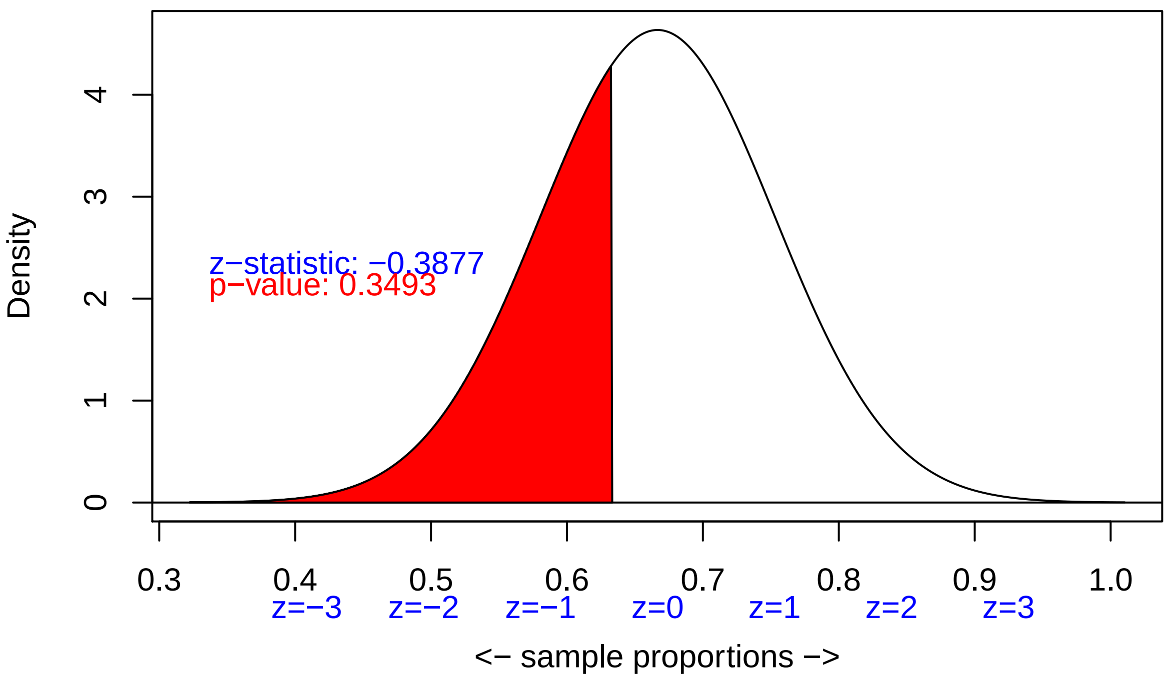
The sample collected is representative of the population because the PERs collected from the population can reasonably be considered a representative sample of the population because the sample had 19 successes out of 30. This success rate from the sample is very close to the null hypothesis stated of 2/3 for the whole population. Along with this, the standard deviation and mean of both the sample and the population are very close in value.

Considering the nature of this study, it would not be reasonable to model this experiment as a binomial distribution. The observations, which are the NBA player seasons, is kept constant throughout the experiment. Along with this each NBA player season is independent of each other, as players do not play exactly the same way over multiple seasons due to factors such as injury, improvement, changing of teams etc. Each NBA Player during the season makes an All NBA team or does not, meaning there are only two outcomes. For each outcome the probability that a player with a PER of 21 or over does not remain constant across observations, because from the 1979-80 season to the 1988-89 seaon there were only 2 All NBA teams per season whereas past that time period there were 3 All NBA teams per season. Since this condition is not met, it would not be reasonable to model this experiment as a binomial distribution.

Below is a one proportion-z-test for the sample. We are assuming the null hypothesis is true. Since $n\pi$=20 which is greater than 10 and n(1-$\pi$) is equal to ten, the sampling distribution of P-hat is normally distributed. The p-value calculated in this one proportion-z-test is 0.35 and the z statistics was calculated to be -0.3877. In terms of the experiment, the p-value of 0.35 is the probabilty that a player with a PER of 21 or over who player at least 48 games in the season that makes the all star team is observed 19 or less times in 30 observations. Since the p-value is greater than a significance level of 0.05, there is enough evidence to conlcude that an NBA player with a PER of 21 or over who has played at least 48 games is more than likely to make an ALL NBA team.

```
load("~/Documents/Math 247 Fall 2017/ISCAM.RData")
n = 30
x = 19
p.hat = x/n
pi = 2/3
iscamonepropztest(x, n, pi, "less")
```

```
##
## One Proportion z test
##
## Data: observed successes = 19, sample size = 30, sample proportion = 0.6333
##
## Null hypothesis       : pi = 0.666666666666667
## Alternative hypothesis: pi < 0.666666666666667
## z-statistic: -0.3877
## p-value: 0.3493
```




Considering a significance level of 0.05, we can find a confidence interval of all the plausible values of our parameter for the experiment. Below are calculations for this confidence interval

```
k = 19
n = 30
phat = k/n
Critvalue = 1.96
SEpihat= ((phat*(1-phat))/n)^0.5
lowerbound = phat-SEpihat*Critvalue
upperbound = phat+SEpihat*Critvalue
lowerbound
```

```
## [1] 0.4608896
```

```
upperbound
```

```
## [1] 0.805777
```

From these calculations, we can say with 95% confidence that the probability an NBA player with a PER of 21 or over that has played at least 48 games for the season considered makes an All NBA team is between 0.461 and 0.806.

IV: Conclusion

The study conducted shows that the null hypothesis stated is true and the probability a player with a PER of 21 or over who has played at least 48 games in the season being considered makes an All NBA team is 2/3. The study shows that we are 95% confident that this probability lies between 0.461 and 0.806. From this study, I have learnt more about how the PER works. For example, I have noticed that forwards tend to have a higher PER because of their versatility. Along with this, I noticed how much higher the top players in the NBA's PERs now than they were before due to the more advanced shooters and higher pace of the modern NBA era. Considering the study showed that a player with a PER of 21 or over is more than likely to make the All NBA team, the PER is a very useful advanced statistic for casual followers of the NBA who do not watch players as often to determine how good a player is. The sample constructed is represntative of the population data because the sample proportion was almost equal to the null hypothesis. If I were to do anything differently, I would consider players before the three point era in addition to the players considered for this experiment. This would allow me to compare the two and three point era's and make conclusions as to how the use and computation of advanced statistics such as PER has changed over time.

Works Cited

Basketball-Reference. "Basketball Statistics and History." Basketball-Reference.com, www.basketball-reference.com/.

Basketball-Reference. "Calculating PER." Basketball-Reference.com, www.basketball-reference.com/about/per.html.

Basketball-Reference. "Glossary." Basketball-Reference.com, www.basketball-reference.com/about/glossary.html.