Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***

Name: Alexander Hayek_ahh767
Lab day and time: Wednesday 10am

Introduction

My favorite sport to watch is basketball. The most exciting part of the game for me is when rookies compete with veterans of the game. I want to research the rookies and their likelihood of having a good season based on their draft pick number. This is because I do fantasy basketball with my friends. This is where you draft players from different teams to have your 'Fantasy Team'. The better the player does (higher points scored per game, assist, rebounds, etc) the more points you get. Your goal is to beat everyone in your fantasy league to win. My stakeholders in this research are people who bet on individual players in software's such as Fantasy Basketball, professional teams with lottery picks, and basketball rookies themselves. My population of interest is all NBA rookies that were drafted into the NBA from the past 6 years.

Research Questions

RQ1 – Does NBA draft pick number influence rookie rankings for rookies' first season in the NBA?

RQ2 – Does an all-star NBA player starting over a rookie influence his chances of higher ranking in rookie season

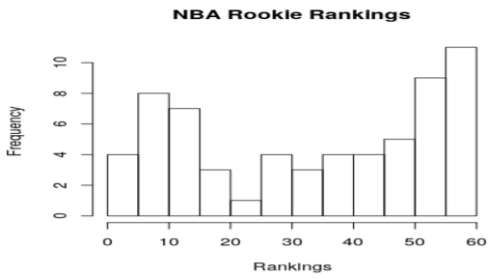
Data Collection Summary

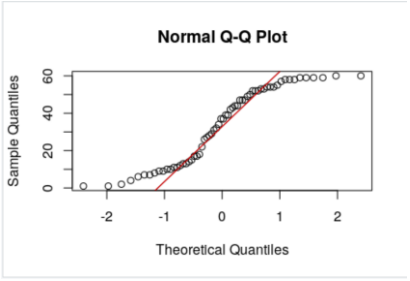
For my individual sampling unit, I have the NBA rookie, their draft pick, whether or not there was an all-star in their position and what they were ranked at the end of their rookie season. For my sampling unit, it does not represent my population. This is due to the fact that rookies can include individuals who come up from the G-league (minor league NBA). People who don't get selected from the NBA draft or who go undrafted can join a summer league for an NBA team then hopefully from there get a contract to play pro. Their first year in the NBA will count as their rookie season but will not come right after they are drafted. My final sample size is 63. I needed to remove Joel Embiid given that he was injured his first season so he could not play. He was a high draft pick but ranked very low because he had no game statistics his first season so this would be considered an outlier.

Descriptive Analysis of Response Variable *(include graph(s) of distribution here)*

Biostatistics SDS 328M – Preliminary Analysis Report

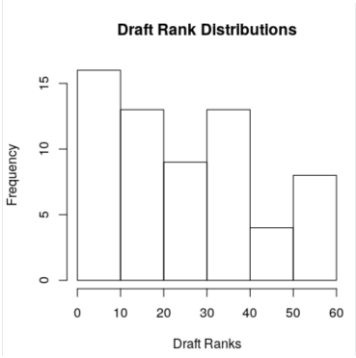
Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***

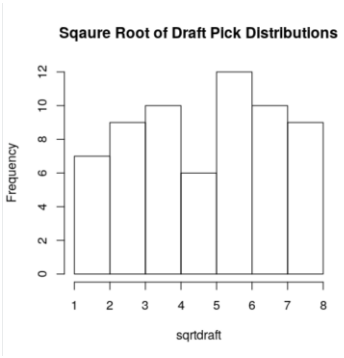


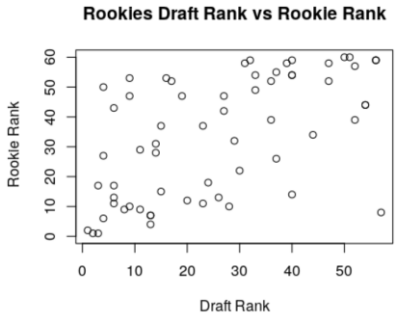


The shape appears to be non-symmetrically bimodal; this can be most from the random sampling. The qqplot indicates that my graph is skewed on the tails which is seen in the histogram. With a higher sample size, I would expect to see a more normal distribution. Joel Embiid is currently a front runner for MVP this season. He was drafted at a relatively high draft pick but because he was injured his rookie seasons, he did not have any game statistics. Henceforth, I removed him from the data given he caused skew to the data. Other than him, all the other samples seemed somewhat reasonable. For my response variable, I tried all the transformations possible (ie log, 1/x, sqrt(x), ln(x)) and I still could not get a normal distribution of my data. Given the data is not normally distributed, the median is 37 and the IQR is 40 respectively.

<u>Investigation of Explanatory Variable 1</u> *(include univariate and bivariate graphs here)*





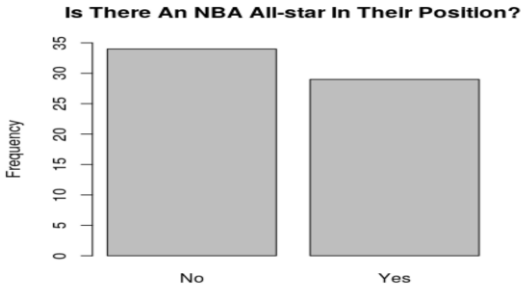


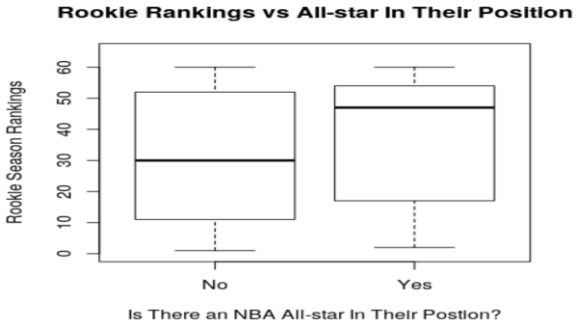
For the primary explanatory variable, the initial distribution of the data appears right skewed. I then did a square root transformation for the graph to appear more symmetrically bimodal. I also tried log(x) and 1/x transformations; however, they were not nearly as symmetric. In reference to the scatterplot, there appears to be a positive correlation between draft pick rank and rookie rank which is expected. The higher the rank a rookie is drafted, the higher their rookie season rank should be. Given the data is not normally distributed, the median is 24 and the IQR is 29.5 respectively.

<u>Investigation of Explanatory Variable 2</u> *(include univariate and bivariate graphs here)*

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***





The second explanatory variable is between rookie rankings and whether or not there was an all-star in their position. With a sample size of 63, there were 34 rookies with no all-star in their position and 29 yeses with an all-star in their position. Looking at the boxplot, there is a clear indication that for those rookies that did not have an all-star in their position, also had a higher rank their rookie season. This indicates that an all-star can affect the game statistics one a rookie has their first season which henceforth, affects their ranking. This was also expected. This represents a relationship between an all-star in a rookie's position and their rookie ranking.

R Code *(organized by variable without output or extraneous syntax)*

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit.***

Univariate Code:

```
hist(nba$Rookie.Rank, main='NBA Rookie Rankings', xlab='Rankings', breaks=15)
qqnorm(nba$Rookie.Rank)
qqline(nba$Rookie.Rank, col='red')
median(nba$Rookie.Rank)
IQR(nba$Rookie.Rank)

hist(nba$Draft.Rank, main='Draft Rank Distributions', xlab='Draft Ranks')
sqrtdraft <- sqrt(nba$Draft.Rank)
hist(sqrtdraft, main='Sqaure Root of Draft Pick Distributions')
median(nba$Draft.Rank)
IQR(nba$Draft.Rank)

plot(nba$all.star.starter, main='Is There An NBA All-star In Their Position?',
    names=c('No', 'Yes'), ylab='Frequency', ylim=c(0,35))
summary(nba$all.star.starter)
```

Bivariate Code:

```
plot(nba$Draft.Rank, nba$Rookie.Rank, main='Rookies Draft Rank vs Rookie Rank',
    xlab='Draft Rank', ylab=' Rookie Rank', ylim=c(0,60))

boxplot(nba$Rookie.Rank~nba$all.star.starter, main='Rookie Rankings vs All-star In Their
Position',
       ylab='Rookie Season Rankings', xlab='Is There an NBA All-star In Their Postion?',
       ylim=c(0,65), names=c('No', 'Yes'))
```

Python Code for Random Sampling:

```
def main():

        draft_14 = []
        draft_15 = []
        new_draft_15 =[]
        draft_16 = []
        new_draft_16 = []
        draft_17 = []
        new_draft_17 = []
        draft_18 = []
        new_draft_18 =  []
        draft_19 = []
        new_draft_19 = []

        num_ranks = 360
```

```
num_needed = 100

ranks = random.sample(range(1, num_ranks), num_needed)
ranks.sort()

print()
print()

for i in ranks:
        if i <= 60:
                draft_14.append(i)
        elif i <= 120:
                draft_15.append(i)
        elif i <= 180:
                draft_16.append(i)
        elif i <= 240:
                draft_17.append(i)
        elif i <= 300:
                draft_18.append(i)
        elif i <= 360:
                draft_19.append(i)

lis = []


for i in range(len(draft_15)):
        lis.append(draft_15[i])
        for j in lis:
                x = j - 60
                new_draft_15.append(x)
        lis.remove(draft_15[i])

for i in range(len(draft_16)):
        lis.append(draft_16[i])
        for j in lis:
                x = j - 120
                new_draft_16.append(x)
        lis.remove(draft_16[i])

for i in range(len(draft_17)):
        lis.append(draft_17[i])
        for j in lis:
                x = j - 180
                new_draft_17.append(x)
        lis.remove(draft_17[i])

for i in range(len(draft_18)):
```

```
                lis.append(draft_18[i])
                #draft_15.remove(draft_15[i])
                for j in lis:
                        x = j - 240
                        new_draft_18.append(x)
                lis.remove(draft_18[i])

        for i in range(len(draft_19)):
                lis.append(draft_19[i])
                for j in lis:
                        x = j - 300
                        new_draft_19.append(x)
                lis.remove(draft_19[i])


        print('This is the random ranks for 2014 draft: ', draft_14)

        print('This is the random ranks for 2015 draft: ', new_draft_15)

        print('This is the random ranks for 2016 draft: ', new_draft_16)

        print('This is the random ranks for 2017 draft: ', new_draft_17)

        print('This is the random ranks for 2018 draft: ', new_draft_18)

        print('This is the random ranks for 2019 draft: ', new_draft_19)

        print()

        sumrookies = len(draft_14) + len(new_draft_15) + len(new_draft_16) +
len(new_draft_17) + len(new_draft_18) + len(new_draft_19)

        print('total rookies: ', sumrookies)

main()
```