Investigating NBA Draftee Priorities Over Time

Free Ducks: Ashir Raza, Vincent Huang, Mina Kim, Elizabeth Berenguer

2020-11-16

Section 1: Introduction and Data

Our dataset consists of various statistics for every player who has played in the NBA within the 1996-2020 seasons, ranging from their home country and college of origin, draft year and round, and various stats such as average points, rebounds, and assists per game.

This dataset was found on Kaggle (https://www.kaggle.com/justinas/nba-players-data) and was originally collected using the NBA Stats API by Justinas Cirtautus, who created this dataset by filling in missing rows of data manually using data from the Basketball Reference website. This data was collected in March 2020, during the midst of the 2019-2020 NBA season.

Each observation/case in the dataset represents a player and their corresponding qualities/ draft stats/game stats for a specific season. A player who played for 10 seasons within the analyzed timespan will have 10 observations. Some variables such as draft number remain constant, while others such as ppg change depending on the season. The variables in the dataset include different aspects pertaining to the player–including information about how/when they were drafted, what country/college they are from, physical characteristics (height, weight), and game stats (number of games played, rebounds).

Basketball is a constantly evolving game, and how NBA players played twenty years ago vs today is vastly different. From the rise of the three-point shooter to the fall of the big man, it is amazing how a game's rules can be so fluid over time. Our group sought to use this dataset to try and find more about how the highest flight of professional basketball has evolved over the years. Check the following article for more information: https://www.espn.com/nba/story/_/id/29113310/seven-ways-nba-changed-michael-jordan-bulls.

Research Question: How have the quality traits NBA teams desire in a draftee changed over time?

We will be looking at season to determine the time, the draft round and number to determine the relative importance of each player, and a way to combine a player's height, weight, and in-game statistical information in order to present an ideal position or skillset desired by teams for that draft year.

Our general hypothesis is that over the past twenty or so seasons, shorter and lighter players have been prioritized in drafts. Scoring, assisting, and true shooting averages will also likely have an upwards trend, while rebounding will likely trend downwards. Teams will have shifted from desiring a tall and heavy rebounder and shot-blocker towards a smaller and quicker shooter with more offensive capabilities.

Here are all the variables we used from the dataset and some brief descriptions of what they represent:

player_name: Name of the player

age: Age of the player

player_height: Height of the player (in centimeters)

player_weight: Weight of the player (in kilograms)

draft_year: The year the player was drafted

draft_round: The draft round the player was picked

draft_number: The number at which the player was picked in his draft round

reb: Average number of rebounds grabbed

ast: Average number of assists distributed

dreb_pct: Percentage of available defensive rebounds the player grabbed while he was on the floor

season: NBA season (number is the year the season started)

Section 2: Methodology and Data

We picked 1996 as the starting year for our analysis since that's the season during which the dataset begins collection. We then filtered out any recorded data from before 1996. We also capped our analysis before the 2019 season, as the 2019-2020 season was ongoing at the time of upload, so metrics such as games played and points per game would not be complete.

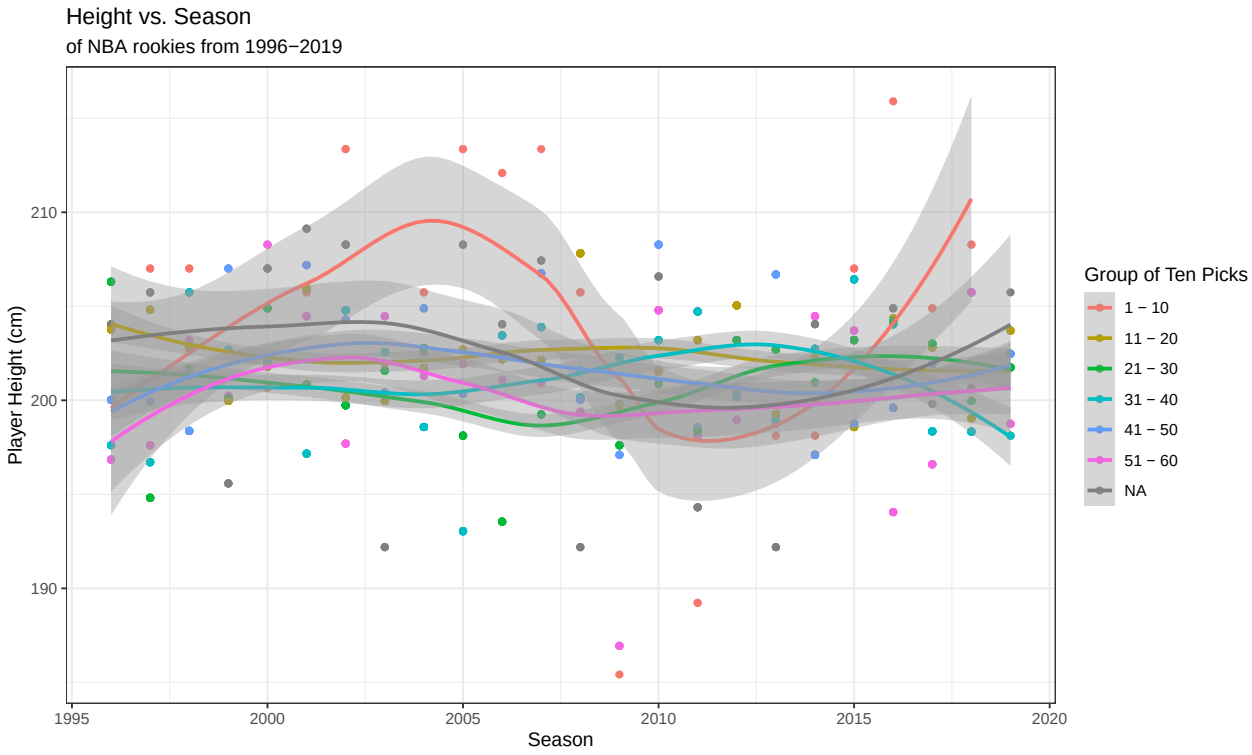
Note that this data set is biased towards the offensive side of the ball, as it lacks defensive statistics such as blocks or steals. Therefore, the contributions of players focused on the defensive side may be neglected.

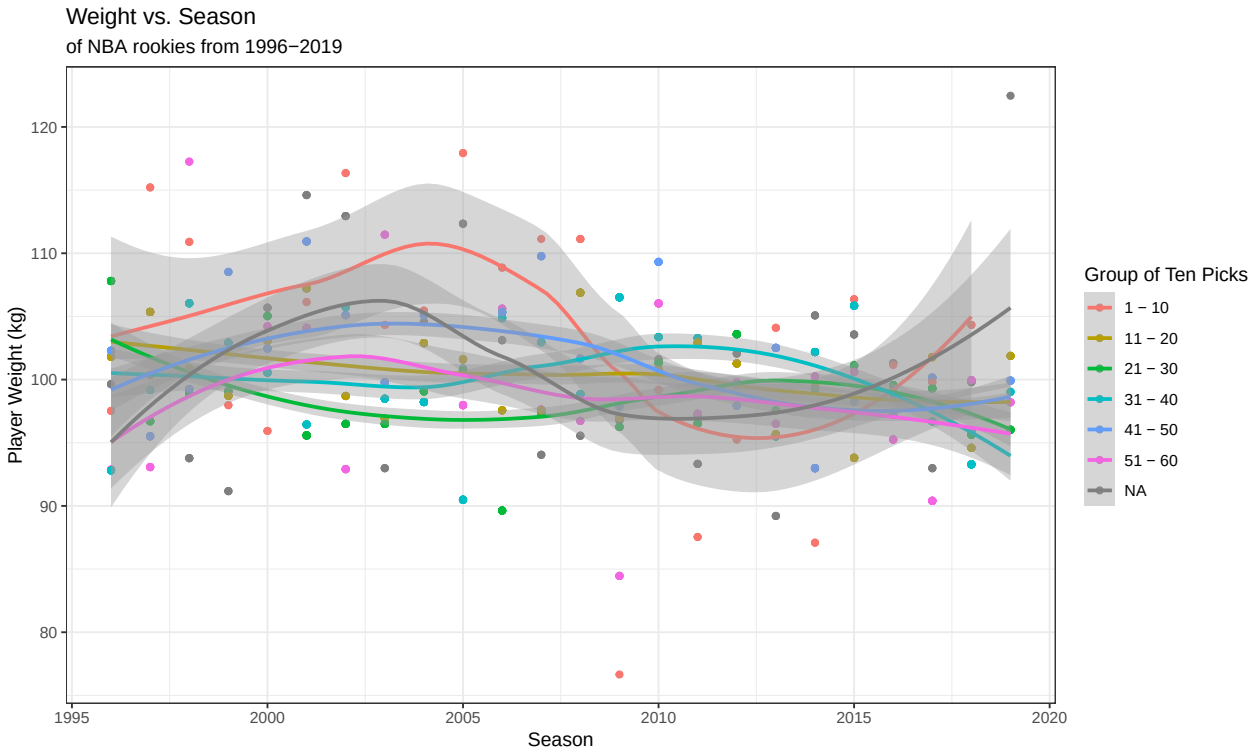
To begin, we create a new dataset called 'draft_data', which contains only newly drafted players. With the removal of all returning players, we are able to examine how drafting has changed from 1981 to 2018 and study what qualities NBA teams value over time. We essentially only took in the rookie seasons of players in the dataset by filtering in only the observations that had a player's smallest age, which would correspond to their rookie season.

Next, we grouped all of the draft picks by multiples of ten (ie: picks 1-10, 11-20, etc. . .) in order to organize the data more efficiently. These groups will be referred to as 'draft groups' for the remainder of this report.

We then grouped by both season and draft group, and found the mean of both height and weight–the two most obvious physical characteristics of an NBA player–for each season and draft group. We graphed these over time, with the draft groups differentiated by color, in order to determine if NBA teams' value of height and weight have changed at all over time within the past couple of years.

We used scatterplots with smoothing lines to find how height and weight change over time, which we deemed to be the most appropriate graph because height and weight are both quantitative variables. It is important to note that although the season (our measurement of time) is technically stored as categorical information in the dataset, we can treat it as quantitative because R will order it numerically regardless. This is because we mutated the season variable to include just the first four characters, which is just the first year of the NBA season as NBA seasons run between two years.





Note that "NA" refers to players who were undrafted, but still picked up by an NBA team. Based on the two visualizations, it seems that both height and weight similarly changed over time. Both appeared to increase for the first few seasons, decrease for the next few seasons, and then increase once again for the latest few seasons from 1996 to 2019. This pattern was especially noticeable for picks 1-10, which are the most valuable picks. There appears to be no overall trend that occurs over the past twenty or so seasons for both weight and height.

Weight and height traditionally correlate with different positions in basketball, with the guards, or backcourt, being smaller and lighter and the forwards and centers, or frontcourt, being heavier and taller. To move forward, we will attempt to test how the value of both the frontcourt and the backcourt has changed over time, using statistical testing to determine how the selection and prioritization of guards or forwards has changed over time.

We will be using simulations of null distributions, t-tests using the Central Limit Theorem, and linear modeling to find p-values and confidence intervals to look for more insights into our data set. Null simulations will be used so we can test differences in proportions and means between various stats to find different preferences in draftees. CLT t-tests will be used to check if different time periods had different preferences in draftees. A quick linear modeling test will be used to determine a good statistic that has a linear relationship with height or weight.

Section 3: Tests and Results

Backcourt players or guards typically move the ball around frequently in order to deliver it to forwards or centers, who then score with the ball. They typically have many assists and fewer defensive rebounds, since they are usually located at the back of the court during an offensive play.

Thus, we have created a new variable called 'guards', which classifies if a rookie player is a guard or not. Our cutoff for this classification is 5 assists per game and a 20% defensive rebounding percentage. If a player has an average of 5 or more assists per game and a defensive rebounding percentage of 20% or less, then they are considered a guard. Otherwise, they are not a guard and are a forward or center.

We found these estimated benchmarks by comparing to the stats of a few famous guards such as Chris Paul and Stephen Curry, who on average had about 6 or 7 assists and defensive rebounding percentages that were lower than 20%. We felt that 5 assists per game and 20% defensive rebounding percentage were good estimates of standard values for a rookie guard.

Null Simulation for Overall Guard Preference (Difference in Proportions)

For this null simulation, our question is this: Are more guards picked in the first draft round? Has the preference for guards shifted over time?

Null Hypothesis: The proportion of guards picked in the first round is equal to the proportion of guards picked in the second round.

Alternate Hypothesis: The proportion of guards picked in the first round is greater than the proportion of guards picked in the second round.

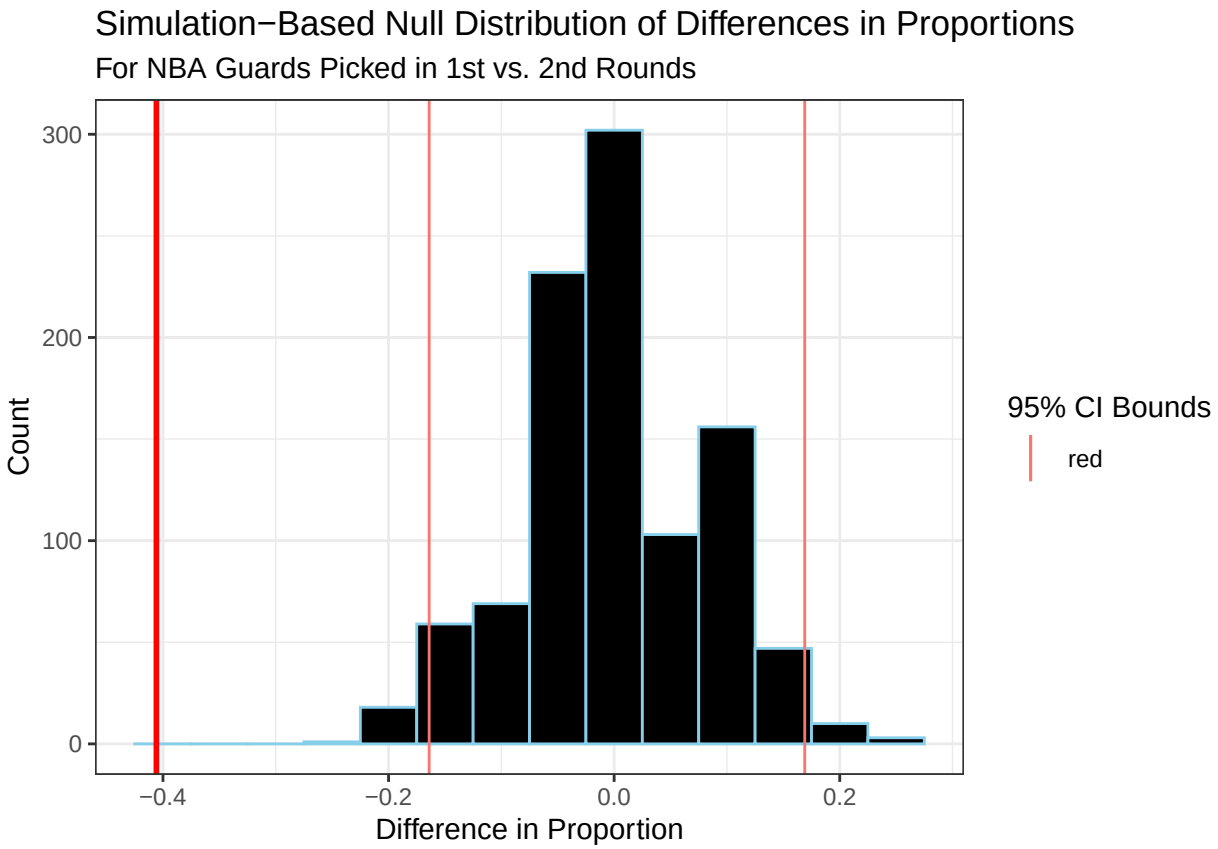
$H_0 : p_1 = p_2$

$H_a : p_1 > p_2$

To do this, we will use a simulation of a null distribution of difference in proportions. We will use 1000 samples and visualize them on a graph, along with a 95% confidence interval.

```
## # A tibble: 1 x 2
##    lower upper
##    <dbl> <dbl>
## 1 -0.164 0.169
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1     0
```



Our p-value was about 0, and so we reject the null hypothesis at the a = 0.05 level. This decision is further supported by our 95% confidence interval, [-0.164, 0.169], and our null distribution graph, which shows that our observed difference is outside the confidence interval. There is sufficient evidence to suggest that a larger proportion of guards were picked in the first round compared to the second round, indicating that in general, skilled guards are likely more valued than forwards and centers within the past seasons starting from 1996.

Null Simulation for Guard or Frontcourt Preference by Year (Difference in Means)

For this null simulation, our question is this: Has the preference for guards or frontcourt players shifted over time?

We now test whether guards were consistently valued over the past 20 seasons, or if the players NBA teams valued were different in certain time periods. We can do this by comparing the mean of the season for both guards and frontcourt players to see which is higher or lower, if either. We will simulate a null distribution of differences in means, using 2500 samples.

Null Hypothesis: The mean year of guards who were drafted is not different from that of non-guards (frontcourt players) who were drafted.

Alternative Hypothesis: The mean year of guards who were drafted is different from that of non-guards (frontcourt players) who were drafted.

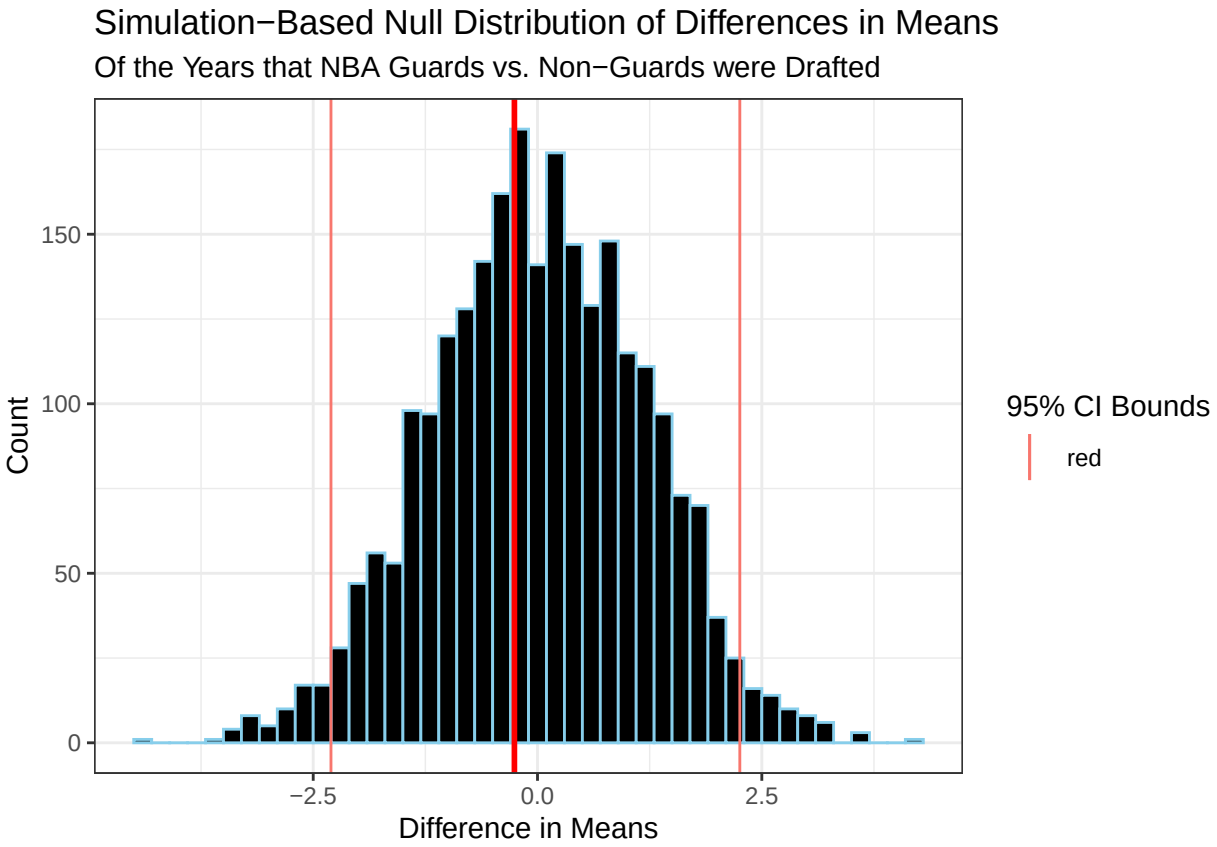
$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 -2.30  2.26
```

```
## # A tibble: 1 x 1
```

```
##   p_val
##   <dbl>
## 1 0.585
```



Our p-value was 0.5848, which is much greater than a = 0.05, and so we fail to reject our null hypothesis. There is not sufficient evidence to suggest that the mean year of guards who were drafted is different from that of non-guards (frontcourt players) who were drafted. In the context of our research question, there is not sufficient evidence to suggest that NBA teams' preferences have shifted towards picking guards during the first draft round over the past few seasons starting from 1996.

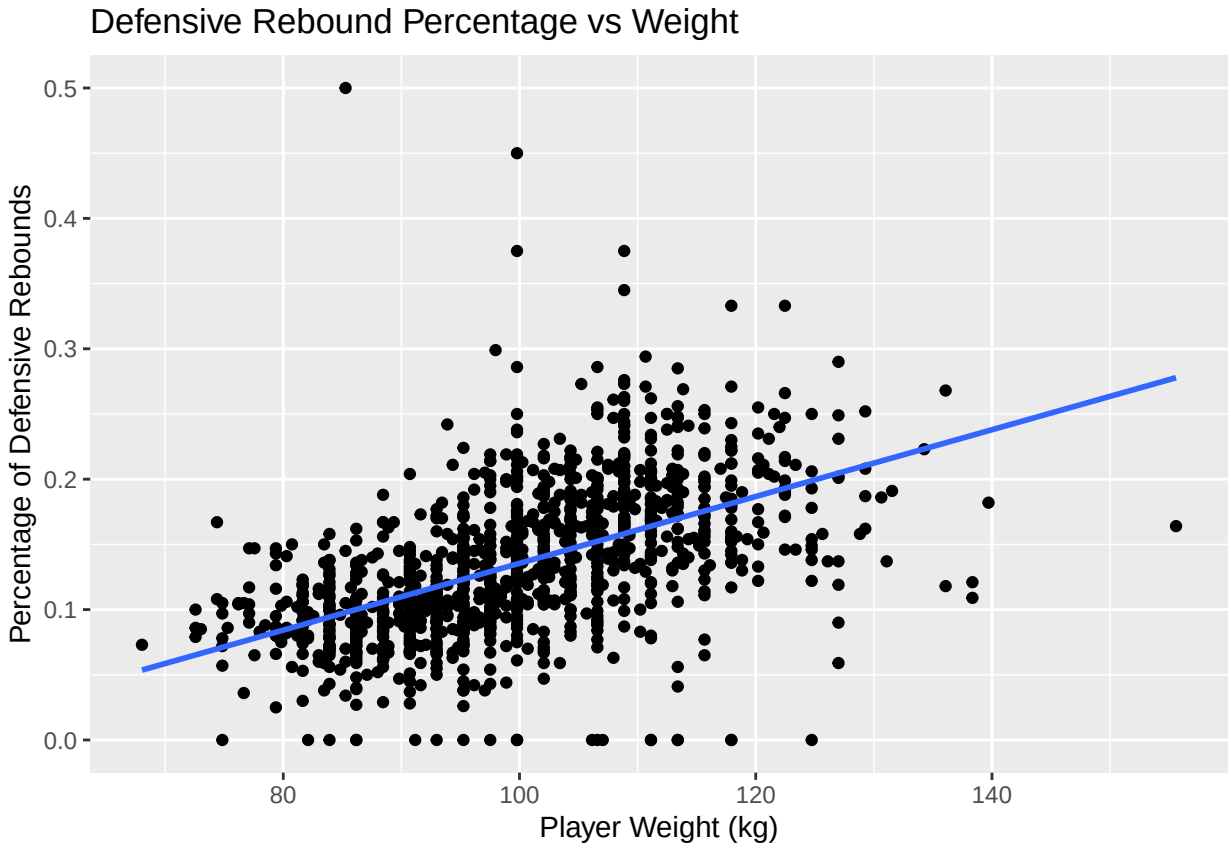
Because our findings have so far not been conclusive, we will now attempt to analyze whether frontcourt players were more valued in certain seasons based on our visualizations in Section 1.

Linear Regression for Defensive Rebound Percentage vs Weight

For this linear regression, our question is this: Does weight have a linear relationship with defensive rebounding percentage?

From our height and weight graphs, we observed that there was a stronger preference for players in the first draft group who were taller and weighed more in the early 2000s (from about 2000 to 2006). Taller and heavier players usually receive more rebounds. However, we will test this using linear regression so that we do not rely upon an assumption. We are also using defensive rebound percentage rather than rebounds since the percentage is a better indicator of rebounding accuracy and efficiency than just a total number of rebounds.

We fit the logistic regression model with rebound percentage as the response variable with the predictor being the player's weight. We chose to prioritize weight over height as heavier players tend to be more powerful and are able to use their mass to force opposing players away so they can grab more rebounds.



Null Hypothesis: The slope for weight is equal to 0.

Alternate Hypothesis: The slope for weight is not equal to 0.

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1207	0.0119	-10.1537	0
player_weight	0.0026	0.0001	21.7061	0

Using a two-sided t-test with 1198-2 = 1196 degrees freedom and a test statistic of 21.7061 for weight, we obtained a p-value that rounds to about 0. This is smaller than our alpha value of 0.05, which means that we can reject the null hypothesis. In order words, we have sufficient evidence to say that there is a slope for weight and there is a relationship between player weight and rebound percentage.

CLT Two-Sided T-Test for Defensive Rebounding Percentage differences from 2000-2006

For this t-test, our question is this: Are mean defensive rebounding percentages from 2000-2006 different than those for the entire dataset?

Now we know that defensive rebound percentage is an accurate indicator of how heavy a player is. We then used a t-test on our data to see if there was a higher defensive rebound percentage for players picked in the first round during the 2000-2006 seasons. If this is true, then that would mean the rebound percentage would be higher in those 6 seasons (and would likely be a consequence of valuing taller and heavier players during that time period).

Null Hypothesis: The defensive rebounding percentage mean during the 2000-2006 seasons is equal to the

percentage of rebounds during all other years.

Alternative Hypothesis: The defensive rebounding percentage mean during the 2000-2006 seasons is different than the percentage of rebounds during all other years.

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

statistic	t_df	p_value	alternative
-0.239	30.898	0.813	two.sided

Using a two-sided t-test with a test statistic of -0.239 and 31 degrees of freedom, we found a p-value of 0.813. Because our p_value is larger than our alpha value of 0.05, we fail to reject the null hypothesis. We have insufficient evidence to say that the mean defensive rebounding percentage during the 2000-2006 seasons is different than the mean defensive rebounding percentage during all other years. We also have insufficient evidence to say that the mean defensive rebounding percentage during the 2000-2006 seasons is equal to the mean defensive rebounding percentage during all other years.

Section 4: Discussion and Conclusion

We learned that heavier NBA players were not more preferred in the 2000-2006 seasons, despite the implications of our height and weight graphs over seasons, categorized by draft group. We reached this conclusion through our CLT test results, whose p-value, 0.813, was much larger than our alpha value of 0.05.

We also learned that the proportion of guards picked in the first round was greater than that of guards picked in the second round. Because our p-value for the first null simulation test used was 0, which was lower than our significance level of 0.05, we concluded that guards are generally prioritized by NBA teams during the draft process over the recent years.

However, we also learned that there is no relation between positional preference (guard vs frontcourt) and time. Because our p-value for the second null simulation test was 0.585, which was much larger than our significance level of 0.05, we realized that there is no discernible relation between choosing players of different positions by NBA teams based off of the time. Since we failed to reject the null, we cannot determine if there is truly no relation between positional preference and season either.

In conclusion, to answer our research question of how have the quality traits NBA teams desire in a draftee changed over time, our response is that although guards were consistently more desired over the past couple of seasons, there were no noticeable relations in positional differences or weight differences among players drafted based off of different seasons, implying that the quality traits NBA teams desire in draftees have stayed relatively consistent from the 1996-2018 seasons. However, more testing is needed before we can make definite conclusions on the lack of drafting patterns since most of our statistical tests resulted in failure to reject the null hypothesis.

Our data visualization methods in Section 1 were not as reliable or precise as we would have hoped. Although we were able to map out draft picks over twenty years groups by draft groups, the margins of error were fairly large as seen by the shaded regions around our smoothing lines. We might address this issue by creating visualizations for several seasons at a time, rather than trying to represent the entire dataset with one graph.

Our null distribution simulations are probably the most accurate and reliable part of our statistical analysis since the sample sizes we used were fairly large. To improve in this aspect, we might increase the number of samples per simulation.

However, a potential error with our first null simulation is how we determined who was a guard and who was not. We did this by estimating values based off of a few famous guards such as Chris Paul and Stephen Curry. However, looking at just a few data points is not a reliable method of finding benchmarks like this, as they could have easily been outliers. We could have been more accurate by using a larger sample size to determine the standard assists per game and defensive rebounding percentage for a guard.

Our linear modeling between weight and defensive rebound percentage appears to be of moderate accuracy, as there is a fairly wide spread of the data points. However, the linear model does seem to line up with the data somewhat well. Our p-value for our t-test was also extremely small, which showed that defensive rebound percentage is affected by weight.

Our CLT t-test may be somewhat inaccurate since we filtered for only the top ten picks for each season. We did this because we wanted to emphasize the priority of draft picks, but this may have also lowered the sample size of data and thus rendered our data inaccurate.

If we were to redo this analysis, we would try to filter our data more. Even after we filtered out every non-rookie player in our large dataset, we still had about 1200 observations, which is large enough to increase the probability of errors and outliers. We would also try to facet our data and then graph it to facilitate ease of analysis and interpretation.

In the future, we want to analyze age and how the age of NBA draftees has–or has not–changed over time. The debate between going pro as soon as possible or staying four years at a Division 1 college is a major point of contention, and it would be an interesting question to explore.

We also analyzed how weight has a linear relationship with defensive rebound percentage and performed a CLT double-sided t-test on whether weight changed depending on whether the draftee was drafted within the 2000-2006 seasons, and resulted in no conclusions. Thus, another action would be to check if height is linear with defensive rebounding percentage and perform the same test on height instead to see if the results become more conclusive.

We could also have more inference testing related to frontcourt positions, since we focused on a lot more than intended on the backcourt for this lab report. Many famous players right now such LeBron James or Kevin Durant are all frontcourt players, so it would be interesting to explore that particular aspect of the NBA.

References

Cirtautas, J. (2020, March 08). NBA Players: Biometric, biographic and basic box score features from 1996 to 2019 season. Retrieved October 9, 2020, from https://www.kaggle.com/justinas/nba-players-data

Goldsberry, K. (2020, April 30). Seven ways the NBA has changed since Michael Jordan's Bulls. Retrieved October 9, 2020, from https://www.espn.com/nba/story/_/id/29113310/seven-ways-nba-changed-michael-jordan-bulls