[redacted]

Intro to Probability and Statistics 1223

Professor Jacob

8 June 2020

NFL Football Players 2019 Statistics Analysis

The data that I used for this project originated from the website https://www.pro-football-reference.com. This data appealed to me because I recently started playing fantasy football with my dad. I have been trying to learn more about the players, so I decided that this was the perfect opportunity!

To collect the information that I needed I searched through various players from different teams. I chose 25 players and got their statistics from the 2019 season. I chose these specific players because in my past I have either had them on my team or they are potential prospects for next season. Generally, when I am trying to choose a new player for my team, I consider all the factors that I listed in my table because that is what get the most points. I also look at how they played in previous seasons to try to look ahead into the next one.

Although there is a multitude of information the variables that I believe will be relevant to my project are the player's age, position, how many games they have played, and their total team wins. The categorical variables include the player's position and their age. The quantitative variables are how many games they played throughout the season and their teams total wins.

Using this data, I hope to find a relationship between a player's position and their age with the amount of games they played and how many wins their team had. I think that finding this relationship will be very interesting and will give me a new outlook while watching the games! Also, I find the player's age to be very interesting. There is a considerable gap including some up to 20 years!

I will be able to determine this relationship because players with different positions and amount of games played typically take on different amounts of bodily harm therefore resulting in differences to the age they continue to play the game and how many wins their team had over the season. Below is a chart containing the variables I am focusing on.

Player	Position	Age	Games Played	Team Wins
Larry Fitzgerald	Wide Receiver	36	16	5
Tom Brady	Quaterback	42	16	12
Davante Adams	Wide Reciever	27	12	13
Todd Gurley	Running Back	25	15	7
Drew Brees	Quaterback	41	11	13
Calais Campbell	Defensive End	33	16	6
Adam Vinatieri	Kicker	47	12	7
Matt Ryan	Quaterback	35	15	7
Jason Witten	Tight End	38	16	8
Maurice Canady	Corner Back	26	5	14
Matt Paradis	Center	30	16	5
Aaron Rodgers	Quarterback	36	16	13
Emmanuel Sand	Wide Receiver	33	10	13
Ryan Nall	Running Back	24	8	8
Joe Haden	Defensive Back	31	16	8
Antonio Brown	Wide Receiver	31	1	12
Stephen Gostkov	Kicker	36	4	12
Damarious Rand	Safety	27	11	6
Eli Manning	Quarterback	39	4	4
Greg Olsen	Tight End	35	14	5
Russell Wilson	Quarterback	31	16	11
Frank Gore	Running Back	36	16	10
Justin Tucker	Kicker	30	16	14
Jameis Winston	Quaterback	25	16	7
Stephon Gilmore	Corner Back	29	16	12

For the second part of the project, I examined the frequency and relative frequency of the player's position. Below is a table containing the data.

Player's Position	Frequency	Relative Frequency
Quarterback	7	28%
Wide Receiver	4	16%
Running Back	3	12%
Defensive End	1	4%
Kicker	3	12%
Tight End	2	8%
Corner Back	2	8%
Center	1	4%
Defensive Back	1	4%
Safety	1	4%
Total	25	100%

For the players that I chose 28% were quarterbacks. This is due to the fact that in a fantasy league quarterbacks typically score the most points each week making them very important! You can also see that out of the 10 positions that I chose from 6 made up less than 10%. So over half of the positions are usually low scoring.

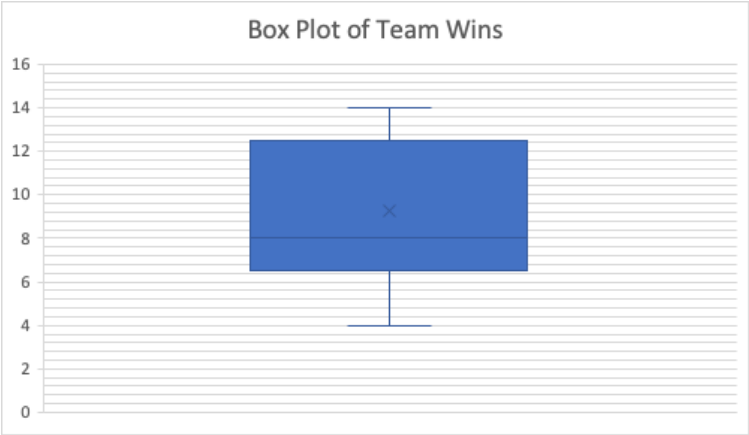
Next I compared a player's position to their age which is shown below.

	Ages 21-25	Ages 26-30	Ages 31-35	Ages 36-40	Ages 41-45	Ages 46-50
Quarterback	1	0	2	2	2	0
Wide Receiver	0	1	2	1	0	0
Running Back	2	0	0	1	0	0
Defensive End	0	0	1	0	0	0
Kicker	0	1	0	1	0	1
Tight End	0	0	1	1	0	0
Corner Back	0	2	0	0	0	0
Center	0	1	0	0	0	0
Defensive Back	0	0	1	0	0	0
Safety	0	1	0	0	0	0

This demonstrated that the majority of my chosen players were between the ages of 31 and 35, with the ages of 26 through 30 not being far behind. You can also see a relationship between the two variables that as players got older there was basically none in positions that are more physically daunting. It was interesting to

me that that the majority of players occurred in the middle of all the age ranges versus towards the younger end. You would think that younger players would be more athletic but this chart shows that as they grow older they develop more skill and increase their prominence within the game.

For the third part of the project I created two graphs to display one of my quantitative variables, the amount of team wins!



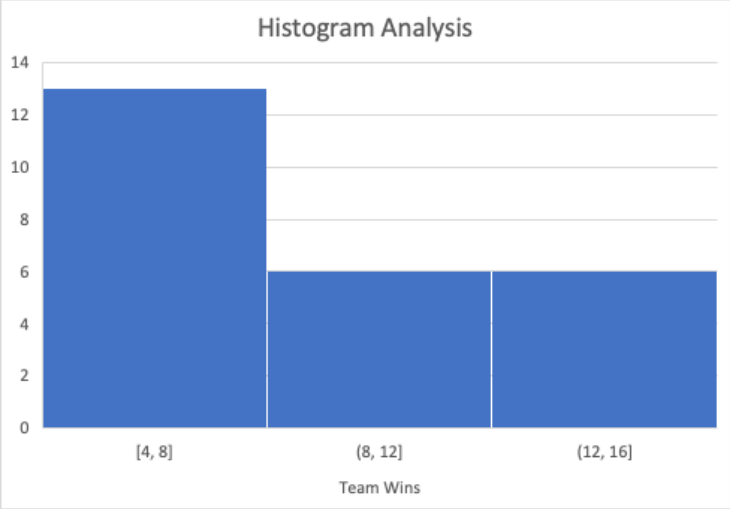
By looking at these graphs you can gather a variety of data. According to my calculations I've listed the five number summary, mean, and standard deviation below.

Five number Summary:
Minimum: 4
1st Quartile: 6.5.
Median: 8
3rd Quartile: 12.5
Maximum: 14

Mean: 9.28

Standard Deviation: 3.310589071

There are no outliers contained within this data being that the numbers range from a minimum of 4 to a maximum of 14 making all the numbers relatively close.



You can see that the distribution of this histogram is skewed right.

When analyzing the data above you can see that the greatest number of team's wins was between 4 and 8, meaning of the selected player's teams more had less wins. It is also interesting how there was the same number of wins in the 8 through 12 and 12 through 16 categories. This shows that the player's team wins with more wins were generally around the same.

For the fourth part of the project I created a hypothesis test for a quantitative and categorical variable!

$H0 : \mu = 8$

$Ha : \mu > 8$

The quantitative variable that I created the hypothesis test for was team wins. The teams play 16 games a season, so half is 8 and if you win more than 8 you have a good season! My null hypothesis states that the player's team wins are

equal to 8 games while my alternative hypothesis shows the wins over 8 games. I wanted to do a hypothesis test for the team wins of the players because I wanted to see how many had a successful season!

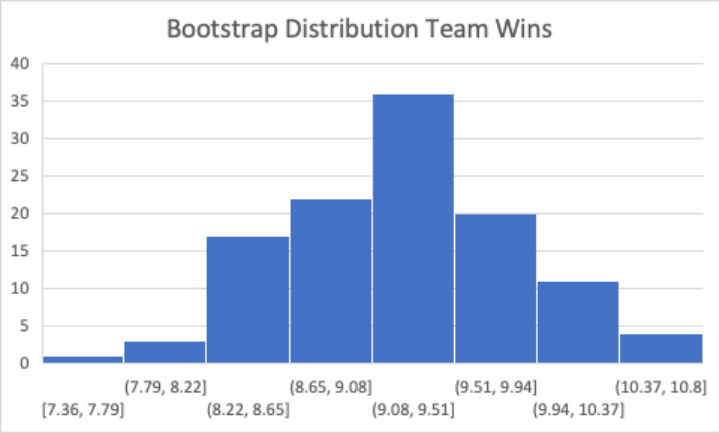
Next is the hypothesis test for a categorical variable, which in this case is the player's position.

H0: $P_{QB} = 0.1$

Ha: $P_{QB} > 0.1$

Out of my selected players I had 10 different positions. I consider the quarterback to be the most valuable player and most consistent scorer, so I decided to create a null and alternative hypothesis to see if quarterbacks are 1/10 of the players on the fantasy roster or greater than that. Considering that quarterbacks typically gather the most points I usually implement one as my flex player to get more points each week therefore making the alternative hypothesis true in my case.

Now onto the fifth part of the project!



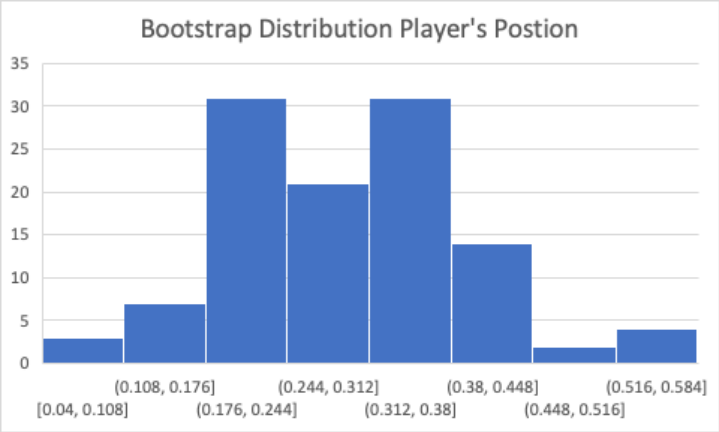
Above is a histogram showing the bootstrap distribution for my quantitative variable above, team wins.

The standard error for this data set is 0.61.

The 95% confidence interval is between 8.03 and 10.48.

Based on this information I can reject my previously stated null hypothesis that teams' wins are equal to 8 because the lowest part of the confidence interval is over 8.

Below is a histogram showing my categorical variable, player's position.



The standard error for this data set is 0.09.
The 95% confidence interval is between 10.57% and 48.41%

My null and alternative hypothesis for the categorical variable above represents the number of quarterbacks on a fantasy team, whether its 1/10 or greater than that. So, to create my bootstrap distribution I took the various position of my 25 players and counted the quarterbacks as 1 and all the other positions as 0. This allowed me to find my confidence interval above which is 10.57% and 48.41% or 0.1056 and 0.4840. Considering that the lowest part of my confidence interval is over 0.1, I have to reject my null hypothesis.

Looking at both of these bootstraps and their conclusions you can tell that if you want your fantasy team to have more wins you need more quarterbacks!

The sixth part of the project we are going to repeat my hypothesis test for my categorical variable and see how that compares to the previous one with bootstrapping!

H0: $P_{QB} = 0.1$

Ha: $P_{QB} > 0.1$

My fantasy team has 10 different positions and I consider the quarterback to be the most valuable one therefore I want to test to see if the quarterbacks were equal to 1/10 of the players on the roster or greater than that!

The 95% confidence interval that I calculated without bootstrapping was between 10.04% and 45.96%. This is very similar to my bootstrapping results which gave a 95% confidence interval of 10.57% and 48.41%. This still allows me to reject my null hypothesis because 10.47% or 0.1004 is still greater than 0.1. Furthermore, this rejection leads to the conclusion that more quarterbacks would lead to a better season and more points for your team!

It was interesting to see how my original data set's confidence interval came out so close to the bootstrap's. Originally, going into calculating this part of the project I assumed that they would be very different. However, after seeing my results and thinking about how bootstraps are just resampled data with the original size as the data set it made more sense that they would be so close!

The seventh part of the project is repeating the hypothesis test for my quantitative variable, team wins. My hypothesis test is below. In fantasy football the teams play a total of 16 games so if you win more than 8 you have a successful season. So my null hypothesis is the mean equal to 8 wins and the alternative hypothesis is greater than 8.

$H_0 : \mu = 8$

$H_a : \mu > 8$

Below I made a chart to compare these results from the bootstrapping!

	Bootstrapping Results	**Repeat Test Results**
Mean	9.30	9.28
Standard Error	0.61	0.57
95% Confidence Interval	Between 8.03 and 10.48	Between 8.14 and 10.42

The mean and standard error of both of the hypothesis tests are very similar, which is what you would expect using the same dataset just without bootstrapping! Looking at this data what I want to compare the most is the 95% confidence intervals. They are very close, and I can still reject my previously stated null hypothesis because the lowest part of the confidence interval is not equal to 8, it is over that! This shows that the team wins of the players were over 8 and therefore had a successful season!

The final part of the project consists of creating conditional probabilities. The data that I used comes from a two-way table that I created in the second part of this project. This table compares a player's position to their age.

	Ages 21-25	Ages 26-30	Ages 31-35	Ages 36-40	Ages 41-45	Ages 46-50
Quarterback	1	0	2	2	2	0
Wide Receiver	0	1	2	1	0	0
Running Back	2	0	0	1	0	0
Defensive End	0	0	1	0	0	0
Kicker	0	1	0	1	0	1
Tight End	0	0	1	1	0	0
Corner Back	0	2	0	0	0	0
Center	0	1	0	0	0	0
Defensive Back	0	0	1	0	0	0
Safety	0	1	0	0	0	0

Here I will be using this table to find the probability that the older the player's age the fewer positions played. Using this formula, I will be able to find the probability of this occurrence.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Here A refers to the ages of the players while B refers to their position.

So for example, lets consider that A is ages 41-45 and B is Quarterback. This allows for P(A|B) = 1/6 because there is 6 different age categories and the probability of falling into one is 1/6. P(B) is 1/10 because there is 10 different positions and the probability of failing into one is 1/10. For $P(A \cap B)$ for the intersection of a and b with the specific categories of ages 41-45 and quarterback that is 2. So to compute the conditional probability use (2/25)/0.1 because there are 25 players altogether. This gives me a conditional probability of 0.8!

So the older the player the fewer positions played!