STAT 516 Course Project

Jackson MacDonald

May 11, 2024

1 A brief overview:

In this project, I will be using data from the 2023-2024 National Football League (NFL) Season via NFL Next-Gen Stats to compare the average expected yards after catch (YAC) of wide receivers to tight ends. Expected YAC (xYAC) is calculated by the NFL Next Gen Stats team using advanced Machine Learning models, and analytics. According to NFL Next Gen Stats *"Expected Yards After Catch (xYAC) leverages Next Gen Stats player tracking data to differentiate between a receiver's actual yards gained after the catch and their expectations at the moment of the catch. Similar to our Completion Probability model, the xYAC model takes into account several in-play factors such as the speed, direction, and acceleration of the receiver, the distance between the receiver and the nearest defender, the number of defenders in the path of the receiver, the number of blockers in the path of the receiver, among several other metrics."* We will use this metric from 86 WRs and 29 TEs in the 2024 season. We want to compare the differences between the distributions of xYAC between the two main positions of pass catchers in football Wide Receivers and Tight Ends. We want to see if there is statistical evidence to show that one position group has a higher average xYAC than the other and make a statistical inference on why this may be. We will do a further analysis of these distributions below, as we will prove their normality and IID samples of these data points.

2 Parameters of Interest

In this study, we will have 2 main parameters of interest and take their difference. The average xYAC for wide receivers and tight ends is shown in the Histograms in Figure 1. We will run the Shapiro-Wilk normality test on each distribution to ensure the normality of these distributions. We use an $\alpha = 0.05$ in this study, so we want our p-value to be greater than 0.05 to reject the null hypothesis and assume normality for these distributions. Running the Shapiro-Wilk normality test on the WR data gets us W = 0.97949, p-value = 0.1886, and for the TE data we get W = 0.94171, p-value = 0.1112. Therefore since p-value(s) = 0.1886 and 0.1112 $> \alpha$, we can assume normality for both distributions. The estimator that we will be using will be $\bar{X} - \bar{Y}$, where $\bar{X}$ is the sample mean for WR xYAC and $\bar{Y}$ is the sample mean for TE xYAC. We want to calculate the basic properties of the estimator (i.e. bias, variance, MSE, consistency). We know that the bias of this estimator will be 0 because of the Weak Law of Large Numbers. When we are using the sample mean we know that it will converge in probability to the population mean, and therefore it will be unbiased. This also means that the MSE($\bar{X} - \bar{Y}$) = Var($\bar{X} - \bar{Y}$). The Var($\bar{X} - \bar{Y}$) = 1.749134, as calculated using RStudio. Taking the limit as $\lim_{n\to\infty}$ MSE($\bar{X} - \bar{Y}$) = Var($\bar{X} - \bar{Y}$) = $\frac{1}{n_x^2}Var(X) + \frac{1}{n_y^2}Var(Y) = 0$ so this is also a consistent estimator. Therefore we will use $\bar{X} - \bar{Y}$ as our estimator because it is unbiased and consistent. This estimator uses $\bar{X}$ = 3.813448, and $\bar{Y}$ = 4.359600, in which the mean will be -0.546152. So, the interpretation of our estimator is that we expect on average the xYAC of WRs to be 0.546152 less than the xYAC of TEs.

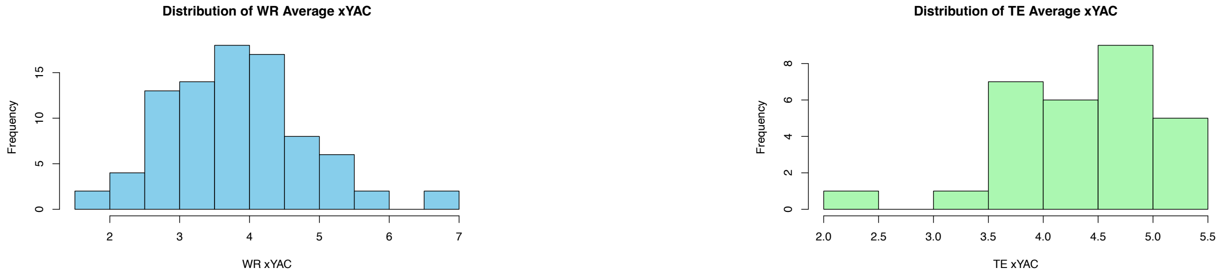


Figure 1: WR and TE xYAC Histograms

3 Confidence Interval

The 95% confidence interval for the difference of mean xYAC in WRs and TEs is created by taking the difference in means and assuming nonequal variances for our normal data. We will use the differences in mean plus and minus the appropriate t value times the root of the sample sizes. This will end up being $t_{\alpha/2,n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $t_{0.05/2,86+29-2}\sqrt{\frac{1}{86}+\frac{1}{29}} = t_{0.025,113}\sqrt{\frac{1}{86}+\frac{1}{29}}$. This will yield a 95 percent confidence interval: [-0.8924129 -0.1998909] given the sample estimates: mean of x mean of y as 3.813448 and 4.359600 respectively. The interval will have asymptotically correct coverage as we constructed this test with a sample of 86 WRs and 29 TEs from a single season of data. My interpretation of this confidence interval is that if we were to repeat this process many times with a lot of data, then 95% of the time the interval -0.8924129 to -0.1998909 would contain the true difference in the means of xYAC in WRs and TEs when our test statistic is (WR xYAC) - (TE xYAC), meaning TE xYAC is greater than WR xYAC in these samples.

4 Hypothesis Testing:

In this experiment, the null hypothesis is that the true mean of WR xYAC is equal to the mean of TE xYAC, $H_o : \mu_1 = \mu_2$, and the alternative is that the true difference in means is not equal to 0 $H_A : \mu_1 \neq \mu_2$. The formula for the Test Statistic is $t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1}{n_1}+\frac{s_2}{n_2}}}$. Under the null hypothesis, it is t distribution with 62.57 4 degrees of freedom. The test has an exact type I error control asymptotically because in finite samples we might not. This is due to the volatile nature of sports analytics, specifically in the NFL. In the future, and more samples there may be a shift in data to provide a different conclusion and a type I error in a finite sample. From the RStudio t-test, the value we got was t = -3.1524, df = 62.574, and p-value = 0.002486. Since the p-value of 0.002486 < is our predefined $\alpha = 0.05$, we reject our null hypothesis. There is no substantial evidence from our test to accept our null that the true mean of WR xYAC is equal to the mean of TE xYAC.

5 Conclusion:

In this experiment, we found that the 95% confidence interval between the xYAC of WRs and TEs is [-0.8924129 -0.1998909]. From this, we can say that if we were to repeat this process many times with a lot of data, then 95% of the time the interval -0.8924129 to -0.1998909 would contain the true difference in the means of xYAC in WRs and TEs. Our hypothesis testing conclusion also tells us that there is no substantial evidence from our test to accept our null hypothesis that the true mean of WR xYAC is equal to the mean of TE xYAC. This tells us that since the confidence interval does not include the null value of 0, then we conclude that there is a statistically significant difference between the WRs and TEs. Furthermore, the data shows that the xYAC of TEs is greater than the xYAC of WRs. To speculate why this may be, it could be because of sheer athletic difference between WRs, and TEs as TEs are much bigger and stronger on average, while WRs are faster and more elusive. The difference could also be caused by inconsistency in the calculation of the xYAC stat by NFL Next Gen stats. There are many factors within the sports analytics world and to pinpoint the cause of this a more intensive test would need to be done on the xYAC stat. Our conclusion from this study though is that there is a statistically significant difference between the xYAC in the WRs and TEs in favor of TEs.

Stats516FinalProject.R

Jackson MacDonald

2024-05-11

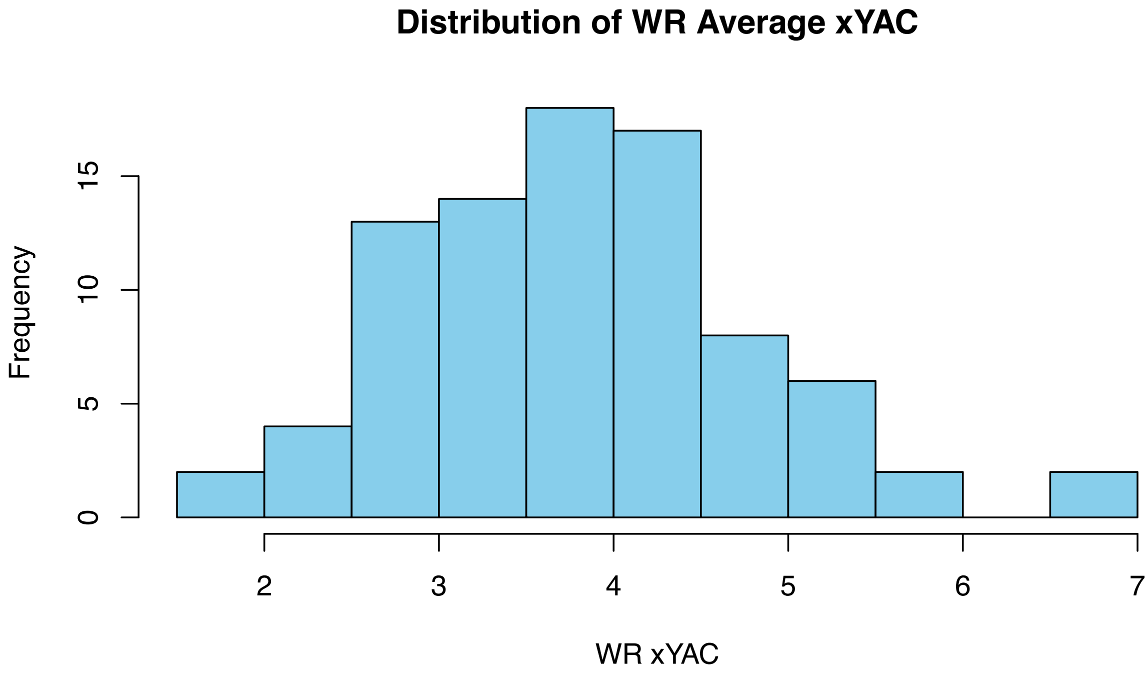
```r
data <- readRDS("/Users/jacksonmacdonald/Downloads/ngs_2023_receiving.rds")

# Filter data for TE and WR
wr_data <- subset(data, player_position == "WR" & week == 0)
te_data <- subset(data, player_position == "TE"& week == 0)
wr_yac <- wr_data$avg_expected_yac
te_yac <- te_data$avg_expected_yac
difference <- wr_yac - te_yac
```

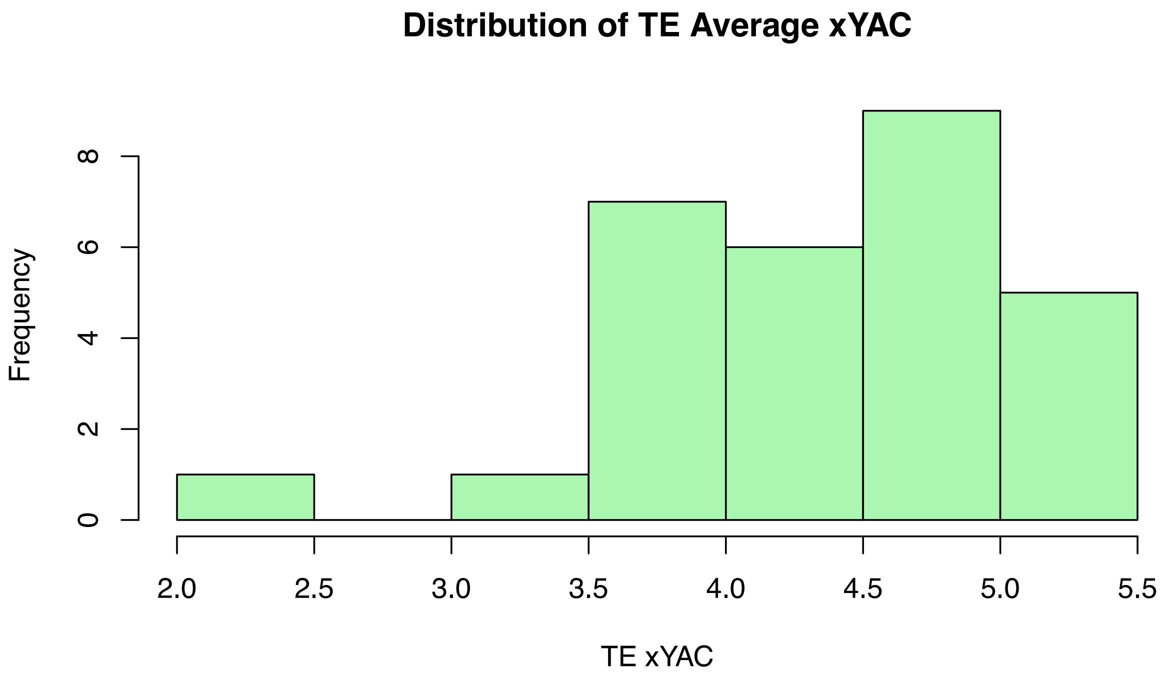
```r
var(difference)
```

```
## [1] 1.749134
```

```r
hist(wr_yac,
     main = "Distribution of WR Average xYAC",
     xlab = "WR xYAC",
     ylab = "Frequency",
     col = "skyblue",
     border = "black")
```



```
hist(te_yac,
     breaks = 5,
     main = "Distribution of TE Average xYAC",
     xlab = "TE xYAC",
     ylab = "Frequency",
     col = "#abf7b1",
     border = "black")
```



```
# Calculate the CI
t.test(x = wr_yac, y = te_yac, conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wr_yac and te_yac
## t = -3.1524, df = 62.574, p-value = 0.002486
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8924129 -0.1998909
## sample estimates:
## mean of x mean of y
##  3.813448  4.359600
```

```
var(te_yac)
```

```
## [1] 0.5530495
```

```
var(wr_yac)
```

```
## [1] 0.9412949
```

```
shapiro.test(wr_yac)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wr_yac
## W = 0.97949, p-value = 0.1886
```

```
shapiro.test(te_yac)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  te_yac
## W = 0.94171, p-value = 0.1112
```