Data 602 Project Report

Peter Baran, Janny Tam, Joel Penner

Introduction

Hockey is a dynamic sport that lends itself well to investigating emergent properties that are hidden in-game data. When compared to other professional sports, hockey seems to be tough to pin down with advanced analytics due to the nature of the game itself. With various factors including puck luck, the mental and physical state of players and staff, and is a much less predictable sport compared to other major sports.

The focus of our study is to explore the average penalty minutes a team takes by exploring The Wideman effect, where a hockey player had hit an NHL referee, and in doing so, allegedly creating a bias against the team with the league's referees. We will be looking at the 2015/2016 season for the Calgary Flames, along with comparing the average penalty minutes with other teams.

Secondly, we will be looking to see if there is a strong relationship between the goal differentials (goals for versus goals against), and the winning percentage for a team. In theory, the more goals that a team would have, the higher their winning percentage should be. We will be exploring with every Hockey team in the NHL for the 2010/2011 to the 2018/2019 season.

Dataset

The dataset we will be using is 'NHL Game Data' by Martin Ellis on Kaggle (Ellis, 2019). We obtained permission to use this data by conforming to the Terms of Use of Kaggle. This dataset contains game data from the 2010-2011 season to the 2018-2019 season. Having 9 seasons worth of data, there is more than enough data to discover trends in the modern era of hockey. We will be using the data frames 'game_team_stats.csv', 'game_data.csv' and 'team_info.csv'. The columns of the 'game_team_stats.csv' and 'games.csv' data frames contain information on a particular game while the rows are games from 2010 to 2019. The 'team_info.csv' data frame includes information on every NHL team that played in 2010 to 2019. Since each data frame includes the columns game_id and team_id, we can use these unique keys to merge the data and filter out columns to only get the information we need.

Guiding Questions

The Wideman Effect

During a hockey game on January 27, 2016, involving the Calgary Flames and the Nashville Predators, Dennis Wideman, a defenceman for the Calgary Flames, had cross-checked a linesman in what was stated to be an accident. The linesman had to undergo neck surgery and has not returned to work. Wideman had been suspended by the league for 20 games, but it was decided by an independent that a 10 game suspension was justified. This incident will hereafter be referred to as the Wideman hit.

From this hit, hockey analysts, the Calgary Flames team, and their fanbase believe that due to this incident, the NHL referees have been biased against the Calgary Flames, and issuing more penalties during hockey games against the team. We will be looking to see if there is a bias towards the Calgary Flames, by comparing the difference in means of penalty minutes, along with running permutation tests, and testing the four following null hypothesis:

1. Is there a statistical difference between the average amount of penalty minutes for the Calgary Flames before and after the Wideman hit in the 2015/2016 season
2. Is there a statistical difference between the average amount of penalty minutes for the Calgary Flames before and after the Wideman hit during the 2010/2011 and 2018/2019 season

To compare our findings, we are curious to see how the remaining NHL teams average penalty minutes had changed before and after the Wideman hit. We will remove the Calgary Flames games out of our 2015/2016 season dataframe, and will perform the same hypothesis and permutation test for the difference in average penalty minutes experienced by the league.

Goal Differentials Relationship with Winning Percentage

It would be expected that teams with a positive, and a higher goal differential (goals for - goals against) would be more likely to have a higher win percentage, as you would tend to have more goals scored against your opponent. With making this statement, we take a look at our data and experience the following:

In the 2017-2018 season, the Colorado Avalanche had a +13 goal differential, with a 51.13 win percentage. While in the same season, the New Jersey Devils had a -2 goal differential with a 51.72 win percentage.

One would expect that the Colorado Avalanche should have experienced a greater winning percentage in comparison to the New Jersey Devils. We will be building a linear model to see how strong the relationship between Goal Differential and Winning Percentage is, if it can be expressed as a linear function, and performing a bootstrap to confirm our findings.

The Wideman Effect

In order to analyze the Wideman effect, we cleaned the data and created a data frame that included the date of the game, season, team, penalty minutes and if the game happened before the Wideman hit on January 27, 2016 or after. For each question, we will be filtering from this data frame to create smaller samples.

Game_ID <dbl>	Home_or_Away <chr>	Date_Time <date>	Season <dbl>	Team <chr>	PIM <dbl>	Before_or_After <chr>
2011030221	away	2012-04-29	20112012	New Jersey Devils	12	Before
2011030221	home	2012-04-29	20112012	Philadelphia Flyers	6	Before
2011030222	away	2012-05-01	20112012	New Jersey Devils	12	Before
2011030222	home	2012-05-01	20112012	Philadelphia Flyers	32	Before
2011030223	away	2012-05-03	20112012	Philadelphia Flyers	4	Before
2011030223	home	2012-05-03	20112012	New Jersey Devils	10	Before
2011030224	away	2012-05-06	20112012	Philadelphia Flyers	10	Before
2011030224	home	2012-05-06	20112012	New Jersey Devils	4	Before
2011030225	away	2012-05-08	20112012	New Jersey Devils	2	Before
2011030225	home	2012-05-08	20112012	Philadelphia Flyers	8	Before

1-10 of 22,868 rows   Previous 1 2 3 4 5 6 ... 100 Next

We first looked at if the Wideman effect was apparent for the Calgary Flames in the 2015-2016 season by using a hypothesis test. With our dataframe, we took out all the games played by the Calgary Flames in the 2015-2016 season. In this season, they played 82 games and of those games 47 were before the Wideman hit and 35 were after.

H_0:

$$\mu_{Before\ the\ Wideman\ Hit} - \mu_{After\ the\ Wideman\ Hit} \geq 0 \equiv \mu_{Before\ the\ Wideman\ Hit} \geq \mu_{After\ the\ Wideman\ Hit}$$

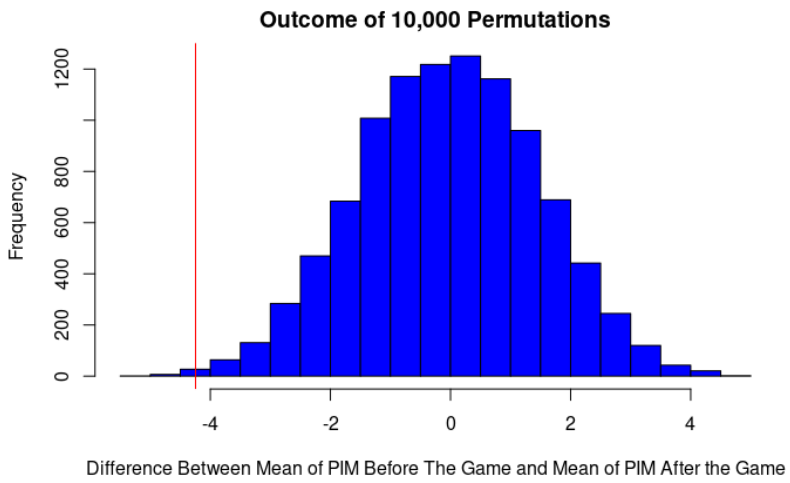
H_A:

$$\mu_{Before\ the\ Wideman\ Hit} - \mu_{After\ Wideman\ Hit} < 0 \equiv \mu_{Before\ the\ Wideman\ Hit} < \mu_{After\ the\ Wideman\ Hit}$$

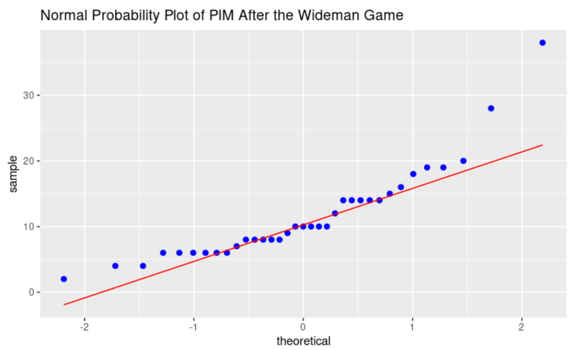
The null hypothesis states that the average penalty minutes taken by the Calgary Flames before the Wideman hit is greater than the average penalty minutes taken by the Calgary Flames after. This means that the Calgary Flames did not experience the Wideman effect in the 2015-2016 season. Meanwhile, the alternative hypothesis states that the Wideman effect does exist because the average penalty minutes taken by the Calgary Flames after the Wideman hit is greater than the average penalty minutes taken before the game. We first performed a permutation test to test this hypothesis.

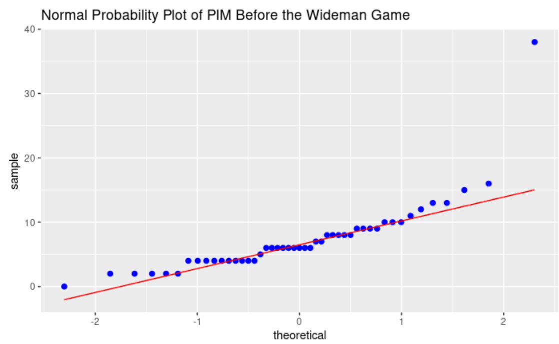
From our observed data, we found the average penalty minutes of the Calgary Flames in the 2015-2016 season before the Wideman hit was 7.3829 and the average penalty minutes of the Calgary Flames in the 2015-2016 season afterwards was 11.6286.

From the graph below, we can see that the red line is the observed difference in the penalty minutes is calculated by subtracting the penalty minutes before the Wideman hit by the penalty minutes after the Wideman hit. The p-value we calculated is 0.0025. With a p-value that is smaller than 0.05, we can reject the null hypothesis because it is not likely that we will see a more extreme value than our observed value. Therefore, we can conclude that the average penalty minutes after the Wideman hit is greater than the average penalty minutes before the Wideman hit. The Calgary Flames did in fact experience the Wideman effect in the 2015-2016 season.



We also used a student T test to validate our permutation test.





Since the penalty minutes variable is a discrete number, it is hard to show that the samples can be modeled by the Normal Distribution. From the t.test we obtained a p-value of 0.002738 similar to the p-value calculated in the permutation test. Therefore, the decisions from the t-test is similar to the outcome from the Permutation test. We reject the null hypothesis and conclude that on average, the penalty minutes of the Calgary Flames in the 2015-2016 season before the Wideman hit is less than the penalty minutes after. Ultimately, the Wideman effect exists for the Calgary Flames in the 2015-2016 season.

To ensure that we do not misappropriate this result as evidence of discriminatory officiating, we must check that the Wideman effect did not occur for all teams. In comparing the Calgary Flames to the remainder of the NHL league, we will be computing the following hypothesis test:

H_0:

$$\mu_{Before\ the\ Wideman\ Hit} - \mu_{After\ the\ Wideman\ Hit} \geq 0 \equiv \mu_{Before\ the\ Wideman\ Hit} \geq \mu_{After\ the\ Wideman\ Hit}$$

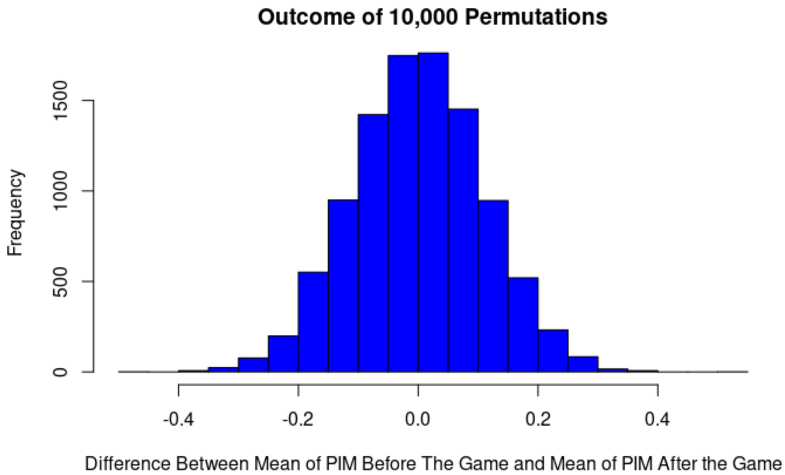
H_A:

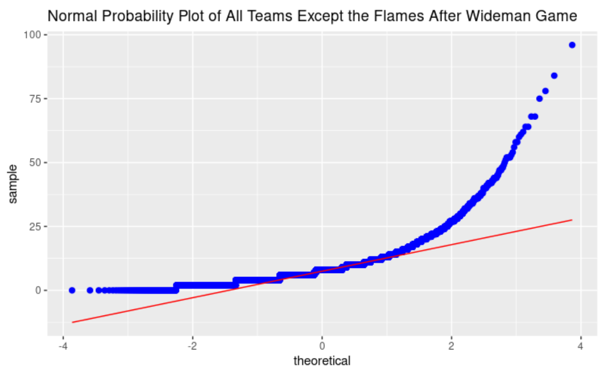
$$\mu_{Before\ the\ Wideman\ Hit} - \mu_{After\ the\ Wideman\ Hit} < 0 \equiv \mu_{Before\ the\ Wideman\ Hit} < \mu_{After\ the\ Wideman\ Hit}$$

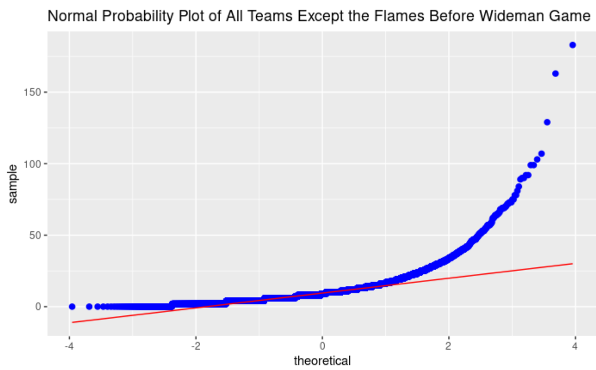
Where our Null Hypothesis states that the average penalty minutes taken by the league, except for the Calgary Flames, before the Wideman hit, is greater than or equal to the average of penalty minutes taken by the league, except the Calgary Flames, after the Wideman hit. This would conclude that the Wideman Effect did not exist. Our Alternative Hypothesis states that the average penalty minutes taken by the league, except for the Calgary Flames, is less than the average penalty minutes taken by the league, except for the Calgary Flames, after the Wideman hit. This would suggest that the Wideman Effect does exist.

From our dataset we were able to find that the average number of penalty minutes before the Wideman hit served by the NHL excluding the Calgary Flames is 11.02 minutes, while the average amount of penalty minutes served after the hit, excluding the Calgary Flames is 8.93 minutes. This would suggest that after the Wideman hit, the league took less penalty minutes after the hit than before.

From the graph below, we can see the outcome of 10,000 permutation tests. From these tests, the observed penalty minute difference is 2.0879, and the probability of observing our test result is as extreme as our observed penalty minutes difference is 1.0.







We also ran a t-test on our data, and can assume from the plots that our data is approximately normal. From the t-test, we observed a P-Value of 1.0, and we fail to reject the null hypothesis.

From the calculated P-Value of 1.0 with our permutation test, we fail to reject the null hypothesis and we conclude that the rest of the NHL league, excluding the Calgary Flames, did not experience an increase in penalty minutes in the 2015/2016 season after the Wideman hit.

With the tests above, we were able to confirm that the Calgary Flames did experience the Wideman effect while the other teams did not in the 2015-2016 season. However, do we see this effect for the Calgary Flames when we are looking at all 9 seasons?

H_0:

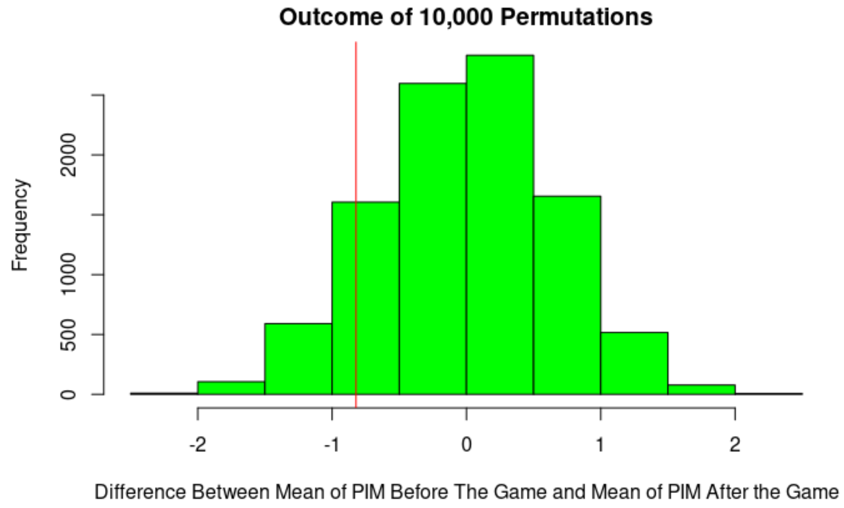
$$\mu_{Before\ the\ Wideman\ Hit} - \mu_{After\ the\ Wideman\ Hit} \geq 0 \equiv \mu_{Before\ the\ Wideman\ Hit} \geq \mu_{After\ the\ Wideman\ Hit}$$

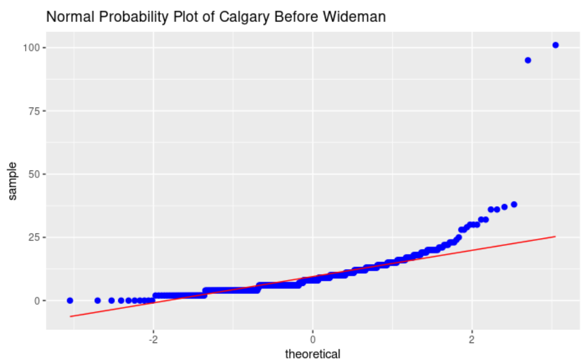
H_A:

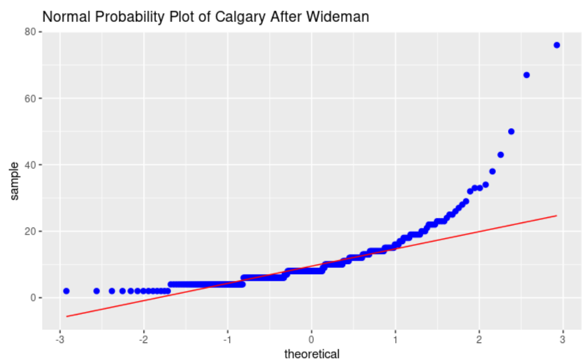
$$\mu_{Before\ the\ Wideman\ Hit} - \mu_{After\ the\ Wideman\ Hit} < 0 \equiv \mu_{Before\ the\ Wideman\ Hit} < \mu_{After\ the\ Wideman\ Hit}$$

Using the same hypothesis tests above, we have the null hypothesis stating that the average penalty minutes taken by the Calgary Flames from 2010 up to the Wideman hit in 2016 is greater than the average penalty minutes taken by the Calgary Flames from after the Wideman hit to 2019. This means that the Wideman effect did not exist for the Calgary Flames when we look at the 9 seasons. While the alternative hypothesis states that the Wideman effect does exist as the average penalty minutes after the Wideman hit is greater than the average penalty minutes before the game.

From our sample, the average penalty minutes before the Nashville game is 9.864 and the average penalty minutes after the Nashville game is 10.686. Using a permutation test, we can see that the observed difference in penalty minutes is -0.822, which we can see is the red line on the graph, and the p-value is 0.117. The probability of observing a test value that is just as extreme as the observed difference in penalty minutes is 0.117. Since this is greater than 0.05, we can fail to reject the null hypothesis and can conclude that the Calgary Flames' average penalty minutes before the Wideman hit is greater than the average penalty minutes afterwards. Therefore, the Calgary Flames did not experience the Wideman effect when we look at all 9 seasons in the dataframe.







We will use a student T test on the samples. From the Normal Probability plots, we will assume that both the samples are approximately Normally Distributed. The p-value obtained from the t.test was 0.1075, similar to the p-value found in the permutation test. Therefore, using the t.test we fail to reject the null hypothesis and conclude that there was not a statistically significant increase in penalty minutes awarded to the Calgary Flames after the Wideman hit. So the Calgary Flames did not experience the Wideman effect throughout the 9 seasons. Using the permutation model, our 95% confidence interval is $-1.2997 \le \mu_{Before} - \mu_{After} \le 1.2717$ does include the value of zero, which would suggest that with 95% confidence, we can statistically infer that there would be no increase in penalty minutes after the Wideman hit for the remainder of the 2015-2016 season to the 2018-2019 season.

Goal Differentials and the relationship with Winning Percentage

A goal differential in hockey is calculated as (Goals For - Goals Against). In each game, the team who scores more goals is the winner. A team's season goal differential is further calculated by summing their goal differential from every game in a season. As a measure of team success, 'win percentage' is also calculated as (Wins / Games Played). Using our dataframe, we will calculate the goal differential by subtracting home_goals by away_goals if the team is home and subtracting away_goals by home_goals if the team is away.

Game_ID <dbl>	Season <dbl>	Team <chr>	Home_Or_Away <chr>	Won <lgl>	Home_Goals <dbl>	Away_Goals <dbl>	Shots <dbl>
2011030221	20112012	New Jersey Devils	away	FALSE	4	3	26
2011030221	20112012	Philadelphia Flyers	home	TRUE	4	3	36
2011030222	20112012	New Jersey Devils	away	TRUE	1	4	35
2011030222	20112012	Philadelphia Flyers	home	FALSE	1	4	20
2011030223	20112012	Philadelphia Flyers	away	FALSE	4	3	28
2011030223	20112012	New Jersey Devils	home	TRUE	4	3	31
2011030224	20112012	Philadelphia Flyers	away	FALSE	4	2	22
2011030224	20112012	New Jersey Devils	home	TRUE	4	2	43
2011030225	20112012	New Jersey Devils	away	TRUE	1	3	30
2011030225	20112012	Philadelphia Flyers	home	FALSE	1	3	28

1-10 of 22,868 rows Previous 1 2 3 4 5 6 ... 100 Next

With some data wrangling, we were able to get each team's win percentage and their goal differential for each season. This will be our sample data.

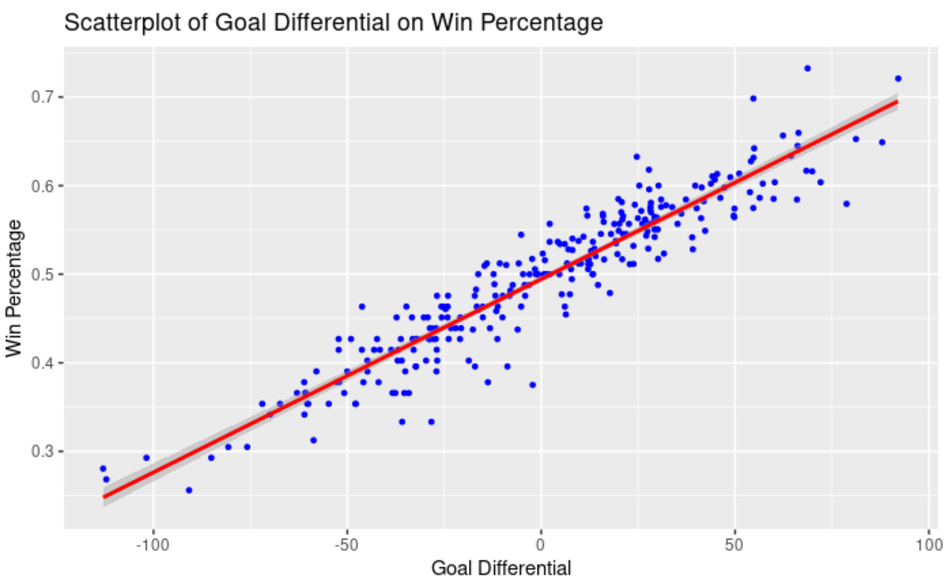
Team <chr>	Goal_Diff <dbl>	Win_Perc <dbl>
Anaheim Ducks	2	0.5568182
Atlanta Thrashers	-46	0.4146341
Boston Bruins	79	0.5794393
Buffalo Sabres	12	0.5168539
Calgary Flames	13	0.5000000
Carolina Hurricanes	-3	0.4878049
Chicago Blackhawks	39	0.5280899
Colorado Avalanche	-61	0.3658537
Columbus Blue Jackets	-43	0.4146341
Dallas Stars	-6	0.5121951

1-10 of 272 rows

Given the relatively straightforward relationship between out-scoring your opponent and winning games, we intend to investigate if one can be expressed as a linear function of the other.

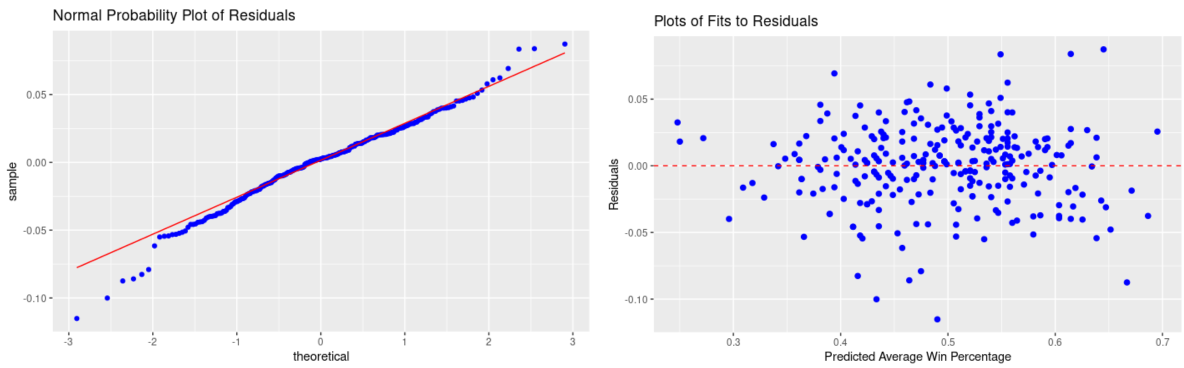
H_0: A team's win percentage cannot be expressed as a linear function of their goal differential.
H_A: A team's win percentage can be expressed as a linear function of their goal differential.

To explore these hypotheses, we first processed game data for every team, for every season. A data frame was constructed where each row represented a certain team's goal differential and winning percentage for a single season. This data was then used to build a linear model of *Winning Percentage ~ Goal Differential*



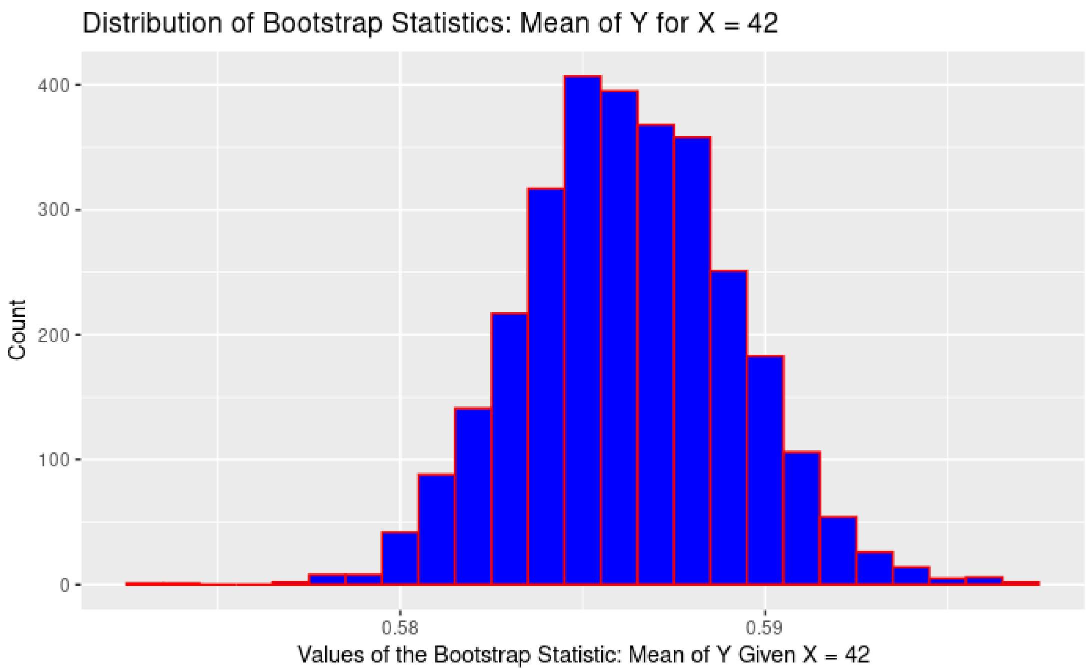
$$Win\,\% \;=\; 0.4945 \;+\; (0.002182 \;*\; Goal\,Differential)$$

The above figure illustrates the strong linear relationship between these two variables. The correlation coefficient for this model sits at a strong *0.9357.* Before drawing conclusions, we must first verify the model for normality of residuals and homoscedasticity conditions.



As shown above, the win percentage variable is normally distributed, and the predicted average win percentage is homoscedastic. Thus we can conclude that the model is statistically sound. With these conditions holding, we can evaluate our hypotheses. An F-test was conducted and the results were F_{Obs} = 1682, P-Value = 0.0000000000000002. Therefore we can reject the null hypothesis and conclude that a team's win percentage can be expressed as a linear function of a team's season goal differential.

Having established the validity of this relationship as a linear model, the model can be used to make predictions of winning percentage. Doing an application on our linear model, suppose that the Calgary Flames had a goal differential of 42. Without applying any numbers, we would assume that if Calgary Flames scored 42 more times than their opponents they should have a higher win percentage. Using the linear model, we are 95% confident that a teams win percentage when they have a 42 goal differential is between $0.5807 \;\le \mu_{y\,|\,x\,=\,42} \le\; 0.5916$. Therefore, their win percentage would be between 58.07% and 59.16% when they have a 42 goal differential. We can support this confidence interval by doing a bootstrap of 3000 iterations. We get an $\overline{a} \;=\; 0.4957$ and $\overline{b} \;=\; 0.002216$. Using the bootstrap method, our estimate of the linear model will be $Win\,Percentage \;=\; 0.4946 \;+\; 0.00221 \;*\; Goal\,Differential$.Therefore we get the following distribution when the goal differential is 42.



The 95% bootstrap confidence interval is $0.5807 \leq \mu_{y\,|\,x=42} \leq 0.5919$. Therefore, using the bootstrap confidence interval, all teams with a 42 goal differential will have a win percentage of 58.07% to 59.19%. Comparing the two models, the win percentages are very close. Thus confirming that our 95% confidence interval is compelling.

Team <chr>	Goal_Diff <dbl>	Win_Perc <dbl>
Chicago Blackhawks	39	0.5280899
Washington Capitals	39	0.5416667
Boston Bruins	40	0.6000000
NY Rangers	40	0.5744681
Detroit Red Wings	41	0.5632184
NY Rangers	41	0.5980392
Los Angeles Kings	42	0.5490196
St Louis Blues	42	0.5824176
Pittsburgh Penguins	44	0.6105263
St Louis Blues	44	0.6022727

231-240 of 272 rows

Looking at our dataframe, we can see that the teams with the 42 goal differential had a 0.549 and 0.582 win percentage. We can see that the Los Angeles Kings win percentage does not fall in our 95% confidence interval while the St Louis Blues win percentage did.

Summary

From our Hypothesis testing and Permutation tests, we were able to prove statistically that the Calgary Flames did have on average more penalty minutes issued against them after the Dennis Wideman hit, proving the Wideman effect does exist for the 2015/2016 season. However, having stated our findings, we also were able to conclude that the league did not experience the same change in average penalty minutes taken before and after the hit, strengthening our hypothesis that the Wideman effect exists for the Calgary Flames. Lastly, we were able to see that throughout our data set, nine seasons from 2010 to 2019, the Calgary Flames did not see a statistical significant increase in penalty minutes after the Wideman effect. This suggests that the bias against the Calgary Flames was only for the remainder of the 2015/2016 season.

From our exploration of a relationship between Goal Differentials and Winning Percentage, we were able to see that there is a very strong linear relationship between the two variables with a strong positive correlation. We have estimated a model where for every goal differential increase, a team will experience an increase to their predicted Winning Percentage by 0.22 %.

We had run a bootstrap simulation with our findings, and have found a 95% confidence interval of 0.587 to 0.5919, with an assumption of a 42 goal differential. While comparing our two models we can see that the estimated win percentage is very close, and does confirm that our confidence interval has statistical support.

Further investigation into this topic could include adapting the linear model to predict goal differential needed to qualify for the playoffs. Alternatively, we could create a multivariable regression model to see how other factors would correlate with a team's Winning Percentage.

References

Cosh, C., 2017. *Colby Cosh On Vigilante Justice And The Wideman Effect: When The Laws Of Hockey Go Wrong* | *National Post*. [online] Nationalpost.com. Available at: <https://nationalpost.com/opinion/colby-cosh-on-vigilante-justice-and-the-wideman-effect-when-the-laws-of-hockey-go-wrong> [Accessed 5 October 2020].

Ellis, M., 2019. NHL Game Data. kaggle.com. Available at:https://www.kaggle.com/martinellis/nhl-game-data/activity [Accessed 27 September 2020].

SPORTSNET, 2016. Gotta See It: Wideman plows into referee after scary hit. YouTube. https://youtu.be/Nj4PoDrqv-E

Appendix - R Code