Preliminary Report: Home Team Advantage

Kiernan McKeegan

Background

This project aims to put a measure on a seemingly unproved observation. Sports are observed having a better atmosphere, locker rooms, game day preparation, and other local advantages for home teams. While making an argument for this to state that it is more beneficial for a team to play at home is seemingly easy, it can be difficult to prove with data. Especially in the NHL there are a multitude of factors that could go into the outcome of a game rather than just who is the home or away team. Therefore I will be determining if there is a statistically significant effect on the outcome of a game based on these terms. Understanding a possible advantage provides insight for teams on how to capitalize on their performance in order to maximize winnings. This will help me put a measure on travel, psychological, and environmental factors that influence the final result.

My motivation for this project is rooted in a long personal and career interest in hockey. Throughout grade school I played for a local team each year dedicating lots of time to understand the game and its many factors. I seemingly noticed that my team had more luck playing in our home arena versus when we were on the road. This piqued my interest into being something quantifiable. In my career I would like to pursue a Sports Statistician position. It would go to say my preferred sport is hockey but sports statistics in general are fascinating.

Data

The data I am using for this analysis is from Kaggle. This source contains lots of data sets with a multitude of variables I will use in my analysis. In my data I will be using information from 11,000 games that outlines the game outcomes, home team or away team win, shots, goals, and goal differentials. This data is spread out through several CSV files that will require data cleaning before any statistical analysis can conclude.

The data will be useful to analyze if there is a pattern between win rate and home team, as well as what variables could be influencing this trend the most. Possible variables that should be excluded are games with too few goals, and data that can be considered for further analysis are games with small differentials. The expectation of games won by the home team should be more than 50% to demonstrate that it is not due to random chance. I am cleaning the data and grouping them into dataframes where possible in order to gain an understanding of a rough visualization of the outcomes. In most of the sources I have observed I found that the home team does have a slightly higher chance of winning over the away team. However, my statistical analysis of these data sets will determine if the home team advantage does exist, if the data sample is distinct enough not to be due to random chance, and if simulated again could the same results appear.

Methods

In order to evaluate the presence of a home game advantage in the NHL, I applied several statistical method procedures. First from the total sum of all games I found what the total home team win percentage was. I then formulated a null and alternative hypothesis test to find out if this win percentage was due to random chance or not. My alternative hypothesis was that the average home win percentage would be greater than 0.5. My null hypothesis is that the average home win percentage would be equal to 0.5. After calculating, I graphed the result to visually observe the difference (Appendix A).

I then computed a confidence interval for each game based on the shots for each team. The confidence interval would tell me the range of the goal differential for each team that game.

This allowed me to do two things, first I found all the games that had a confidence interval differential encompassing 0 (lower end of interval was negative). I then threw out all of those ‘uncertain outcome’ games and recalculated the home team win percentage to observe games that stayed positive. Secondly, the interval allowed me to find the percentage of games that could have had flipped outcomes (lower end of interval from winning team intersecting upper end interval from losing team). I then calculated the percentage of home team wins that could have had flipped outcomes and away team wins that could have been flipped as well. This information gave me a rough overview of what percentage of games could have been flipped due to environmental factors, such as a home team advantage.

Finally, I used a permutation test to calculate if the home team win percentage was statistically significant. I went back to the same hypotheses from my original hypothesis test (H0=0.5, H1>0.5), however I generated my own null distribution. In order to generate this null distribution through a permutation test, I fixed the number of games and number of wins/losses in the dataset. However, I swapped the labels of the games (Labels ‘home win’ or ‘away win ') regardless of outcome (50% chance of being swapped). Under the null hypothesis there is no advantage for the home team to win, so this distribution should give me 50% of wins were from the home team and 50% were from the away team. I generated this null distribution 10,000 times and graphed the result (Appendix B).

Results

When I calculated the home team win percentage from the sum of all games, I found that the home team won 54.5% of the games. This percentage is very close to the null hypothesis, so I calculated a hypothesis test. From this I found a p-value <0.05, so I was able to reject the null hypothesis and continue my analysis.

I then calculated a confidence interval based on the shots which told me 94.4% of games had uncertain outcomes. Excluding these games, I recalculated the home team win percentage and found that the new percentage was 64.6%. This is over a 10% jump from the win percentage calculated with all of the games. I also found how many home wins and away wins had close enough goal differentials to have flipped outcomes under the interval. 45.8% of home wins could have had flipped outcomes and 48.6% of away wins could have had flipped outcomes. This means a larger percentage of away wins could have been flipped to home wins outcomes, suggesting external factors.

Finally I calculated a null distribution to re-calculate my hypothesis test. In my permutation test to generate the null distribution, I found that the average home team win rate was about 0.5 with a standard deviation of about 0.0047. I also found the largest home team win percentage under the null distribution to be 0.518. This means that no simulation from the null had a higher win percentage than the observed home win rate. I then recalculated my hypothesis test and found a p-value < 0.0001. This allowed me to reject the null hypothesis again, effectively stating it is very unlikely the home team win percentage observed is due to random chance.

Conclusion

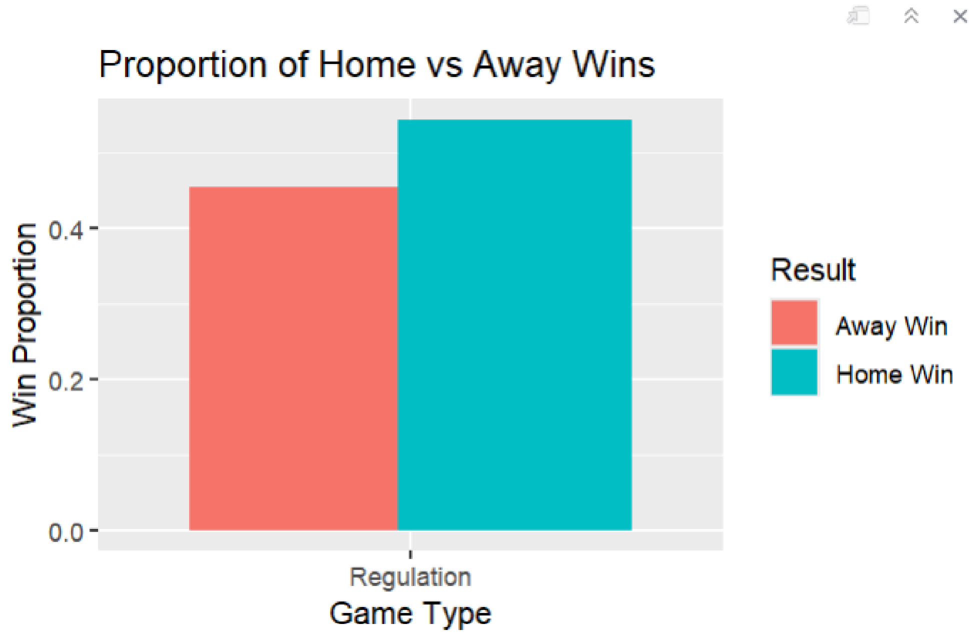
From my analysis I was able to conclude that a home team advantage does exist in the NHL. The initial percentage showed that home teams win 54.5% of games, already exceeding the null of 50% under which no advantage exists. Formal hypothesis testing confirmed that this result is statistically significant, leading to further analysis. Accounting for uncertainty through confidence intervals strengthened my finding. I found the home team won 64.6% of games when I excluded uncertain outcomes. As well, nearly half of all away wins were suggested to be close enough in their goal differential that they could have been a home win instead, due to external factors. The permutation test was the strongest evidence of the home team advantage. By simulating 10,000 datasets under the assumption of no home team advantage, the null distribution told me I should expect a win rate of 50%. However there was no win rate larger than what I did observe in the data. The p-value (<0.0001) confirms the observed advantage is not due to random variation.

Discussion

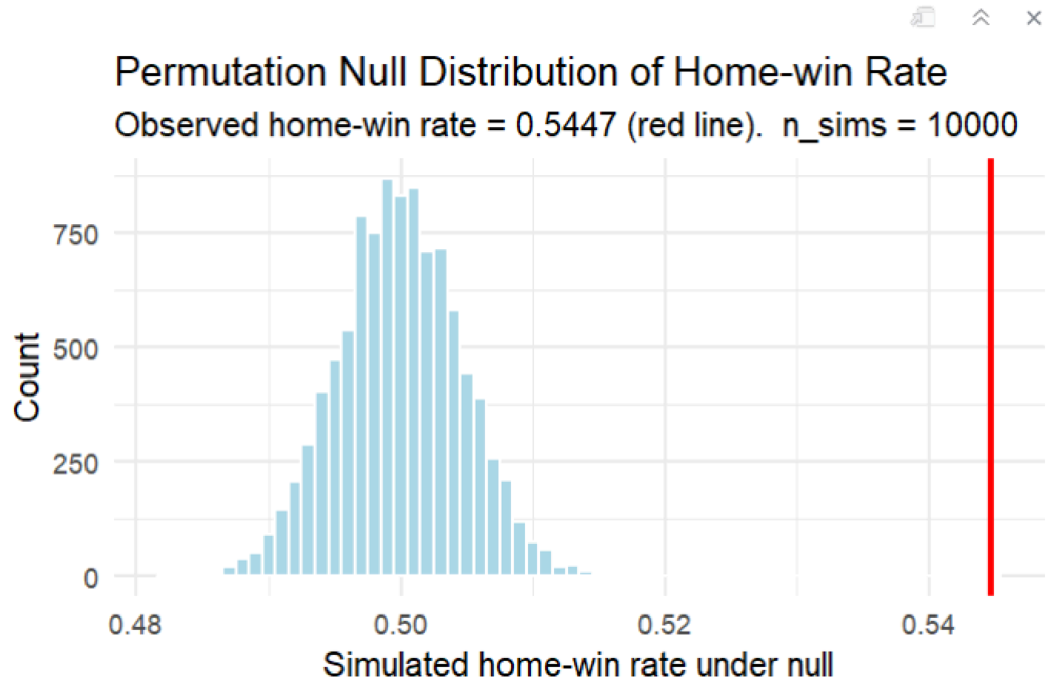
A potential bias to address was the calculation of the confidence interval. I found each one through the shots per game to find the goal differential range in that game. However this can be flawed since a typical game may have 40-50 shots but only 3 goals, making the standard error high and the interval wide. To account for this I multiplied the standard error by the probability of actually scoring a goal on a shot, however it is an imperfect model. As observed by the 94.4% of uncertain game outcomes, the interval still encompassed a large number of the games, and only games that had some higher scoring and smaller standard error were considered certain. This suggests that using a confidence interval, while conceptually useful for identifying stray outcomes, may overestimate uncertainty and therefore understate the true effect of the home ice advantage.

Appendix

A:


Proportion of Home vs Away Wins
Win Proportion
0.4
0.2
0.0
Regulation
Game Type
Result
🟥 Away Win
🟦 Home Win

B:


Permutation Null Distribution of Home-win Rate
Observed home-win rate = 0.5447 (red line). n_sims = 10000
Count
750
500
250
0
0.48
0.50
0.52
0.54
Simulated home-win rate under null