Final Project Report

NHL Penalty Kill Prediction and Team Penalty Analysis

Gordon Liang, Fariha Babar, Ethan Kakavetsis, Peter Nguyen

Group 61

University of California, Berkeley
Data 102 — Data, Inferences, Decisions
Professor Strang, Professor Sridharan
May 6, 2024

Contents

List of Figures

List of Tables

1 Data Overview

Our data consisted of two datasets, Penalty Kills (2015-2024).csv for penalty kill probabilities and 2023-24 NHL Penalty Counts.csv for penalty distributions, both extracted from the National Hockey League's (NHL) API, and consist of the population from the 2015-2016 season to the 2023-2024 season. To get the data for the penalty kill set, we called used request.get on the URI, https://api.nhle.com/stats/rest/en/team/penaltykill?limit=-1, and turned the returned json into a pandas Dataframe and saved it as a csv after a little bit of cleaning.

Table 1: First 5 rows of Penalty Kills (2015-2024).csv

Game ID	Team	Penalties Committed	Penalties Drawn	Penalty Diff
2023020001	NSH	6	4	2
2023020001	TBL	4	6	-2
2023020002	PIT	4	2	2
2023020002	CHI	2	4	-2
2023020003	SEA	4	4	0

The process for getting penalty distribution data was more complicated. Using the URI, https://api-web.nhle.com/v1/club-schedule-season/{team}/now' with each team's abbreviations, we were able to get the game ID for every game played this season and then call https://api-web.nhle.com/v1/gamecenter/gameID/play-by-play on each game. The returned API gives a row of each play and the event so from there we'd group by game and calculate the number of times a penalty was called to get the penalty difference for each team in each game.

Table 2: First 5 rows of 2023-24 NHL Penalty Counts.csv

teamFullName	seasonId	penaltyKillPct	ppGoalsAgainst	shGoalsFor	timesShorthanded	Penalties Killed
Buffalo Sabres	20172018	0.77872	52	9	235	183
Detroit Red Wings	20192020	0.74336	58	4	226	168
Chicago Blackhawks	20152016	0.79681	51	11	251	200
Washington Capitals	20212022	0.81893	44	8	243	199
Montréal Canadiens	20232024	0.76706	58	9	235	183

Our data represents a sample because we only see 82 games worth of a team's penalty killing. And because certain teams committed less penalties, our sample size varies by team. Though, because an 82 game sample is fairly large, we expect our sample to be a fairly good estimate of the population.

The players involved in the data were fully aware that these are tracked by the league. For the first dataset, each row represents a team for a specific season. For the latter dataset, each row represents a game for a specific team in the 2023-2024 season, so each game has two rows, one for each team.

Selection bias is potentially a concern in our penalty kill set because teams that commit more penalties and are thus more shorthanded will have more opportunities to lower or raise their penalty kill rate (PK%). Neither measurement error nor convenience sampling were concerns.

We didn't modify the dataset for differential privacy as it was unneeded. While the "Penalties Killed" column wasn't in the original data given, we got it by multiplying the given "timesShorthanded" and "penaltyKillRate" columns together. It would've been nice to get a column on the amount of minor power plays and amount of major power plays each team had to kill but an overwhelming amount of them are minor power plays so it was fine to not have it. Otherwise, there was no concern for missing data and our cleaning was limited to adding the extra column and filtering the relevant columns.

2 Research Questions

In our first question we aimed to estimate each team's true probability of killing a power play. This gives a better idea of which teams are better at penalty killing with sample size considered. Teams can then use this data to identify successful penalty killing teams and search for features that may be causing this. To achieve this we perform Bayesian Hierarchical Modeling because our prior will have a heavier effect on teams that weren't shorthanded as often compared to teams that committed a lot of penalties. A limitation of this method is having our prior overshadow the likelihood. If our prior is too strong, every team will look like they have the same penalty kill distribution. And conversely, a really weak prior will just show our observed data as is.

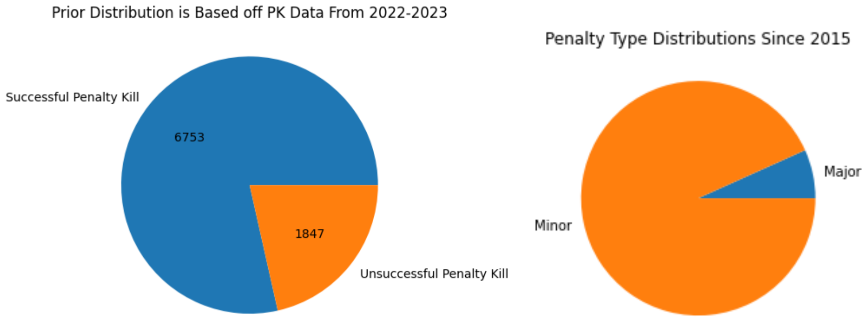
Our second question was multi-faceted: which teams commit significantly more/less penalties than other teams and which teams consistently commit more penalties than their opponents? To answer the first part, we performed A/B testing where our test statistic was the difference in penalties per game between the tested team and the rest of the league. To answer the second part, we perform a Likelihood Ratio Test where the null hypothesis is that a team's penalty difference follows a rounded normal distribution centered at 0 and the alternate hypothesis is that the rounded normal distribution is centered at 1.

This allows teams to identify which teams are more or less disciplined and game plan accordingly. If the penalty difference for a team isn't normally distributed, the likelihood ratio test might not perform well. We might reject the null hypothesis for a team when in reality the distribution skewed the mean. Likewise, for the A/B Test, because a team plays 82 games leaving 480 games in the season they're not involved in, the team's penalties per game which is part of the test statistic calculation is more vulnerable to outliers which can lead to a misleading discovery.

3 EDA

Categorical Variables

Figure 1(a): From 2015-2023, 80.01% of power plays have been killed. We use this to formulate our prior distribution for our Bayesian Hierarchical Model.



(a) Penalty kill successes and failures

(b) Penalty types

Figure 1: Categorical variable pie charts

Figure 1(b): While this might not be directly coded into our model, major penalties are five minutes long whereas minor penalties are two minutes long which influences the chances of a power play getting killed. Due to the heavy majority of penalties being minor penalties along with fighting penalties being a major penalty for both players (and thus not resulting in a power play), we don't factor penalty type into our model.

Numerical Variables

Figure 2: This bar plot depicts the distribution of the penalty difference for the Carolina Hurricanes, showing the distribution to be approximately normal. This motivates our second question's assumption that penalty difference for any given team is normally distributed and makes it easier to conclude which teams are more likely to commit more penalties consistently.

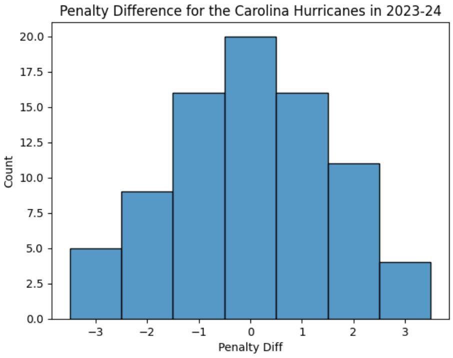


Figure 2: Bar chart plot of penalty difference across all teams

Figure 3: This violin plot graph shows the empirical distribution and the mean/variance of each NHL team's penalty difference across each game of the 2023-2024 season. This is relevant because we can see that, from our observed data, penalty difference across teams is approximately normally distributed. This motivates the assumption for our second research question of which teams commit more penalties consistently that penalty difference for any given team is normally distributed.

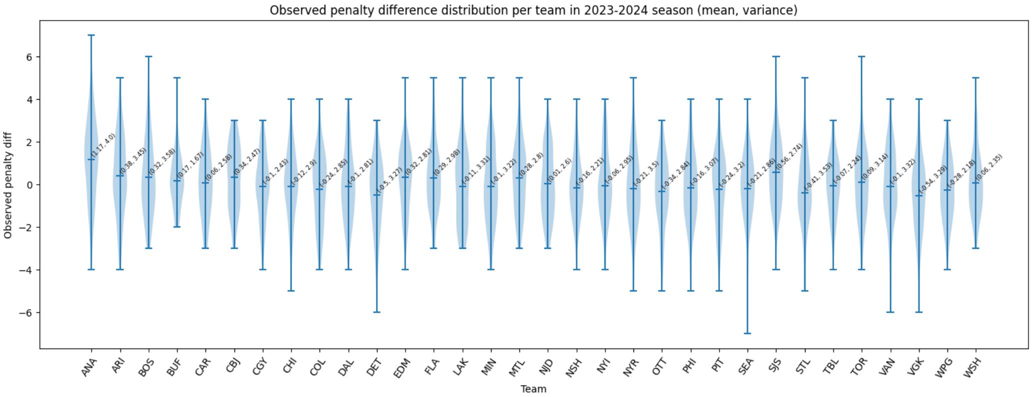


Figure 3: Observed Penalty Distribution per team in 2013-2014 season (mean, variance)

Figure 4: This bar plot compares the amount of penalties committed to the amount of penalties drawn by each team. This is crucial as we can use this for the second research question where we determine which teams commit a significant amount of penalties in the context of the whole league. The observed information here can be used in the multiple hypothesis testing method we use to answer this question for observed data to compare our test statistic too.

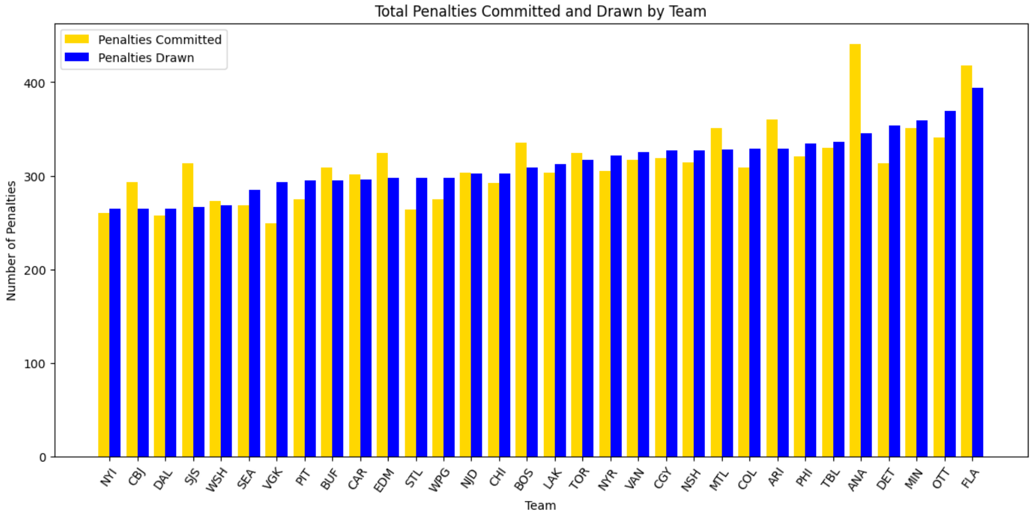


Figure 4: Bar chart plot of total penalties committed and drawn by each team

Figure 5: This bar plot displays the difference in average penalties by each team, highlighting which teams have higher penalties committed (higher positive numbers, more to the right of the x-axis) and which teams tend to have higher penalties drawn on average (higher negative numbers, more to the left of the x-axis). These figures can also lead to observed data statistics (mean difference) to compare to a test statistic during multiple hypothesis testing for the second research question.

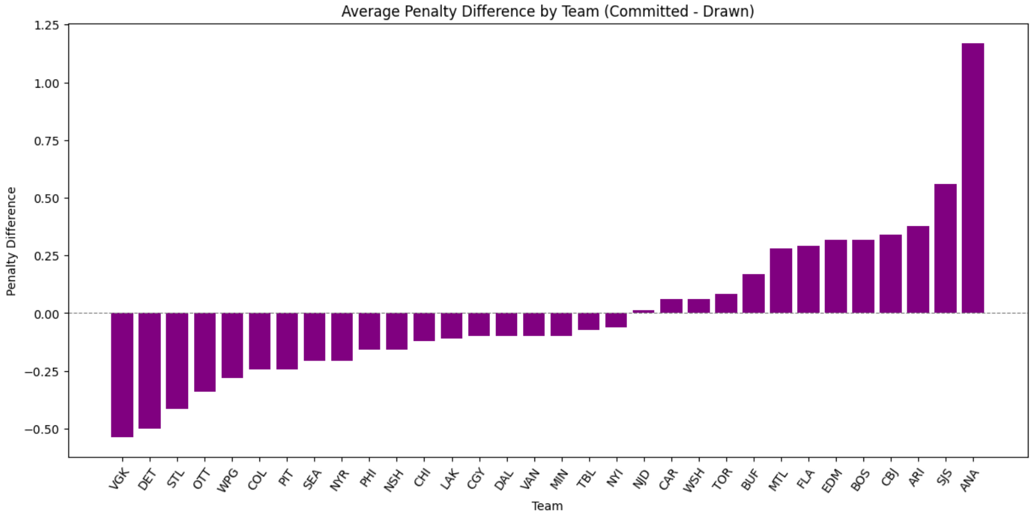


Figure 5: Bar chart plot of average penalty difference by each team

4 Bayesian Hierarchical Modeling

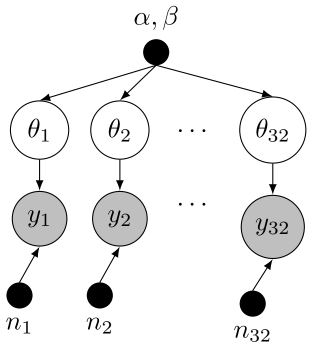
Research Question Explored

Research Question 1- For each team in the NHL, what is their true probability of killing a penalty?

The top three NHL teams in terms of penalty killing are all separated by less than two points in PK%. With hockey being a lengthy season of hot and cold streaks, two percentage points isn't enough to draw a conclusion. Especially considering the fact that certain teams go shorthanded more often than others and so the denominator varies for each team. In this section, we aim to quantify the uncertainty within each team's penalty killing and identify teams that may be better than their observed numbers.

Method

Each group is one of 32 NHL teams during the 2023-2024 season. A hierarchical model is useful due to the fact that each team has a different amount of power plays to kill. As we established in the previous part, some teams commit significantly more penalties and some significantly less penalties than the rest of the league. The parameters for our prior distributions were randomly sampled from a hyperprior uniform distribution. We chose these values based off grouped penalty kill numbers from the 2022-2023 season and floor divided by 32 so the prior wouldn't be too strong. We used a beta for our prior because it acted as a counter of the number of penalties killed and the numbers of penalties that were converted. We used a binomial distribution for our likelihood it fit the scheme of getting a probability of how many successes (killed penalties) we observe in n trials (power plays).



$$\alpha \sim \text{Unif}(0, 211)$$

$$211 = \text{No. of penalties killed (2022-23) // 32}$$

$$\beta \sim \text{Unif}(0, 57)$$

$$57 = \text{No. of power play goals (2022-23) // 32}$$

$$\theta_i \sim \text{Beta(a,b)}$$

$$y_i \sim \text{Binom}(\theta_i, n_i)$$

Results

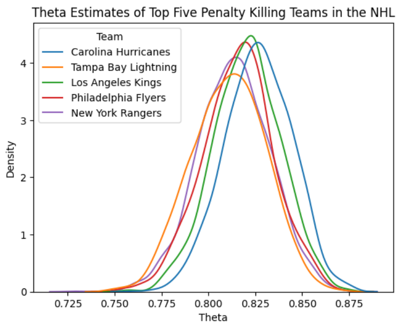


Figure 6: Estimated theta of top 5 penalty killing teams

For our prior hyperparameters, we chose to sample α uniformly from 0 to 211 (where 211 is chosen because that's the average amount of penalties that were killed off per team in the previous season) and β uniformly from 0 to 57 (the average amount of penalties converted per team in the previous season). We chose this method to add variance in how strong our prior is. After performing the sampling, we ranked our top five based on LMSE Estimation. While all of the top five teams remained the same, the Philadelphia Flyers moved up a spot and the New York Rangers down a spot. This makes sense as the Flyers were shorthanded 233 times this season whereas the Rangers were only shorthanded 212 times so this was one of the cases where a small difference in our observed data may have been due to a difference in sample size. The Carolina Hurricanes led the league in Penalty Kill Rate and also led in our LMSE estimation. They had a .95 Credible Interval of [.7899, .8579], meaning that the probability of the Carolina Hurricanes killing off a penalty has a 95% chance of being between .7899 and .8579.

Table 3: Top 5 Penalty Killing Teams

Team	.95 CI bottom bound	.95 CI upper bound	LMSE
Carolina Hurricanes	0.7898798673	0.8578567622	0.8265006678
Los Angeles Kings	0.7839926966	0.8548874404	0.8205868067
Philadelphia Flyers	0.7803263317	0.8547141907	0.8156711548
New York Rangers	0.775074224	0.8512979862	0.8135347014
Tampa Bay Lightning	0.7737665729	0.8484783118	0.8108019382

Discussion

Our results are significant as the top ranking teams continue to show a high number of penalty kills even after adjusting for the number of penalties faced. The intervals for these top teams are quite high and similar, with 0.95 confidence intervals bounded around 0.78 and 0.85, suggesting the model is quite accurate in capturing each team's ability. For individual tests, teams; improvement in their penalty kills could influence coaching or game strategy, as player roles can be changed during penalty kills for instance. In total, the distribution of penalty kills across teams can help set precedents or rules across the league for penalty kills (and say what is considered an appropriate amount). Our limitations lay largely in having a smaller data set and thus sample size to work with, along with the fact that each penalty committed/drawn is taken independently when these events are likely not independent. We avoided p-hacking largely by sticking to the same variables in our model and not adjusting them for the sake of training model accuracy. One additional test we could develop with more data, particularly time-series data to see how the teams penalty kill performance changes over time (the data set here only looks at the 2023-2024 season). Data on each teams opponent would also allow us to run tests to see how certain opposing teams may affect the number of penalty kills, if there is a correlation. Additionally, data on the context of the penalties would allow us to make better inferences and add more to our model. For example, a penalty at the end of the game likely won't last the whole duration so it'll be easier to kill off. Likewise if the team with a man advantage commits a penalty during the power play, their power play is now over and both teams had to only kill off part of a power play, making for an easier kill.

5 Multiple Hypothesis Testing

Research Question Explored

Research Question 2 - Which teams commit significantly more/less penalties than the rest of the league? Which teams commit more penalties than their opponents consistently?

There's a general consensus in the hockey world that penalties called should balance out. In other words, if your team is consistently getting more penalties called against them, they're either experiencing biased officiating or straight-up committing more penalties. To get a clearer perspective of this, test two null hypotheses for each team in the NHL: a given team commits the same amount of penalties per game as other teams and a given team's penalty differential for each game follows a normal distribution centered at zero. The alternate hypotheses for each team are that they don't commit the same amount of penalties (can be either more or less than other teams) and that their penalty difference distribution is centered at one rather than zero. For the latter hypothesis, we want to isolate teams that seemingly get called for more penalties than their opponent consistently so we set our specific alternate hypothesis to be that their penalty difference is distributed normally with a mean of one rather than zero.

Method

To test the first hypothesis, we perform A/B testing with our test statistic being the absolute difference in penalties per game between the team and the rest of the league. To construct the null hypothesis we pseudo-randomly shuffle team labels 1000 times per team and calculate the test statistic at each iteration (resulting in 32,000 total iterations for the 32 teams). Then, we applied both Benjamini-Hochberg and Bonferroni Correction to control False Discovery Proportion and Family-Wise Error Rate.

To test the second hypothesis test, we define our likelihood ratio to be the probability density function under the alternative hypothesis divided by the probability density function under the null hypothesis. Then we calculated the threshold with which our false positive rate would be lower than 0.05. We used the variance of a team's specific penalty difference as the variance for our hypotheses.

In order to test the null hypothesis that all teams have a similar distribution of penalties.

Alternative hypothesis: Certain teams commit more penalties than their opponents consistently.

Steps

1. After loading the datasets, setting the number of simulations to 1000 and desired alpha (TPR) of 0.05.
2. Set seed for reproducability and generate a certain amount of random states for each team for A/B testing
3. Store actual penalties/game for a team - mean penalties/game for the other teams as penalty differences
4. Random sample teams and get their penalty differences, use extreme differences to general p-values
5. Apply Benjamini-Hochberg procedure to to p-values (control for expected false discovery rate)

   (a) Sort p-values
   (b) Set threshold k*alpha/m (Find the largest sorted p-value that's still below the comparison value, use that p-value as the threshold - multiple testing)
   (c) For teams below threshold, label them as being more statistically significant (True) for having extreme penalty differences consistently (False otherwise)

6. Apply Bonferroni Correction (control for family wise error rate, allow less false positives)
   (a) Set threshold alpha/m
   (b) For teams below threshold, label them as being more statistically significant (True) for having extreme penalty differences consistently (False otherwise)
7. Likelihood Ratio Test
   (a) Set threshold for test statistic under the null hypothesis with a normal distribution and standard deviation scale
   (b) Use threshold and standard deviation to calculate gamma
   (c) Use standard deviation and threshold to calculate likelihood ratio for each team
   (d) Compare the likelihood ratio (LR) for each team to gamma, if the LR is greater than gamma, that means the test statistic is higher, that team rejects the null hypothesis, and label that team as True for rejecting the null hypothesis.

Results

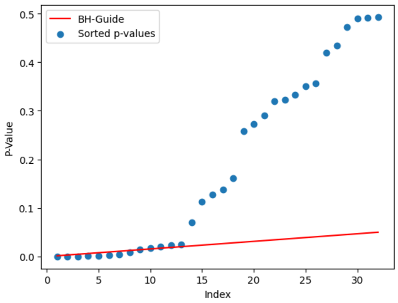


Figure 7: Bonferroni plot for multiple hypothesis testing

Using Benjamini-Hochberg (controlling for expected false discovery rate), we set our new p-value threshold for rejection at 0.011 and reject the null hypothesis for the following teams: Anaheim Ducks, Arizona Coyotes, Dallas Stars, Florida Panthers, New York Islanders, Seattle Kraken, St. Louis Blues and the Vegas Golden Knights (8 teams total). Using Bonferroni Correction (controlling for family wise error rate, allowing less false positives), our threshold lowered to 0.0016 which led us to reject the null hypothesis for: Anaheim Ducks, Dallas Stars, Florida Panthers and the Vegas Golden Knights (4 teams). Of the eight teams where we rejected the null hypothesis using either error control method, only the Anaheim Ducks, Arizona Coyotes and Florida Panthers committed more penalties per game than the rest of the league so while they committed significantly more penalties than the rest of the league, the other five teams committed significantly less penalties than the rest of the league. For the latter hypothesis, we only rejected the null hypothesis for two teams: the Anaheim Ducks and the San Jose Sharks (two teams).

Discussion

To start, the fact that all the teams identified by the Bonferroni correction (which is stricter) were also rejected by the Benjamini-Hochberg procedure (less strict) indicates that the discovery of these teams rejecting the null hypothesis is significant. For individual tests, teams identified as having significantly different penalty kill rates might lead to decisions regarding coaching and training strategies, as well as adjusting player roles. These decisions would be especially important if penalties are considered harmful to the team's performance. Teams that reject the null hypothesis under both correction methods would likely benefit the most from such decisions. Overall, these findings can lead to league-wide precedents and rules, especially in regards to penalties and how they are called in order to make games more fair. Possible limitations lie in our controls. The stricter Bonferroni correction, which controls the FWER (Family-Wise Error Rate), or the probability of making one or more false discoveries among all hypotheses tested, reduces false positives but does not quite decrease false negatives. The less stringent Benjamini-Hochberg procedure controls more for false negatives, but by its own nature is less strict with and so more likely to allow for false positives. Thus, we do not have the perfect test here to control for false positives and negatives. We avoided p-hacking largely by sticking to the same control methods in our model (Bonferroni and Benjamini-Hochberg), settling on them for a more honestly look at our model's accuracy. Additionally, for the likelihood ratio test, we made the assumption that each team's penalty difference is normally distributed. One additional test we could develop with more data, particularly time-series data over seasons to see how the teams' penalties trend over seasons. We could once again use data on each games opponent in order to run tests to see how certain opposing teams may affect a team's penalty trends, if there is a correlation.

For our calculations, less false rejections of the null hypothesis is most appropriate to more accurately predict which teams will consistently commit higher penalties.

6 Conclusion

From the Bayesian Hierarchical Modeling we were able to estimate a team's true probability of killing a penalty and get a clearer perspective of the league's top penalty killing teams. From hypothesis testing we found nine teams that commit significantly different amounts of penalties than the rest of the league and also identify teams that consistently commit more penalties than their opponent. We found that the Anaheim Ducks both commit more penalties in general than the league but also commit more penalties than their opponents so teams facing them can expect to focus more on strengthening their power play and, though it's frowned upon in the league and an embellishment penalty if noticed, work on selling the calls more because they're a team that simply plays more on the edge in terms of physicality.

We didn't need to merge different data sources but we did use two separate data sets from the same source. Our data didn't include the context of penalties that were or weren't killed which would affect penalty kill rates. Future studies can aim to add the probability of a power play being abridged (cut off by another penalty during the power play) or the probability of another penalty happening by the shorthanded team (resulting in a 5 on 3) into the hierarchical model.

References

"Chapter 9 - Six Enduring Controversies in Measurement and Statistics." Quantifying the User Experience: Practical Statistics for User Research, by Jeff Sauro and James R. Lewis, Elsevier Science, 2012, p. 241. ScienceDirect, https://www.sciencedirect.com/topics/computer-science/bonferroni-adjustment:∼:text=Applying

"False Discovery Rate." Columbia University Mailman School of Public Health, Columbia University, https://www.mailman.columbia.edu/research/population-health-methods/false-discovery-rate. Accessed 5 May 2024.