A Statistical Framework for Hockey Analytics: Advice and observations from time my working with the Kingston Frontenacs

Joshua Pohlkamp-Hartt

Queen's University

January 16, 2016

The Role of a Statistician and Their Place in Hockey

Statisticians are experts in the following:

- *Producing trustworthy data*
- *Analyzing data to make their meaning clear*
- *Drawing practical conclusions from data*

*(*American Statistical Society*)*

Within the context of hockey this means:

- Creating and monitoring data streams
- Analyzing and modeling the data to answer management questions
- Provide sensible and easily digestible results

Interpreting the conversation

All projects will start with a discussion and from this a question concerning our team's performance for us to analyze should arise. We can then begin to think about the following questions:

- Are there existing methods?
- What factors/co-factors could we use to observe this problem? What types of data are they?
- Is the data readily available or do we need to collection it?
- What biases/barriers may affect our data collection?
- What methods might we use if no existing methods apply
- What assumptions are we making?
- How best can we report the results of our analysis?

Data Collection

Data collection for hockey is often achieved in three ways:

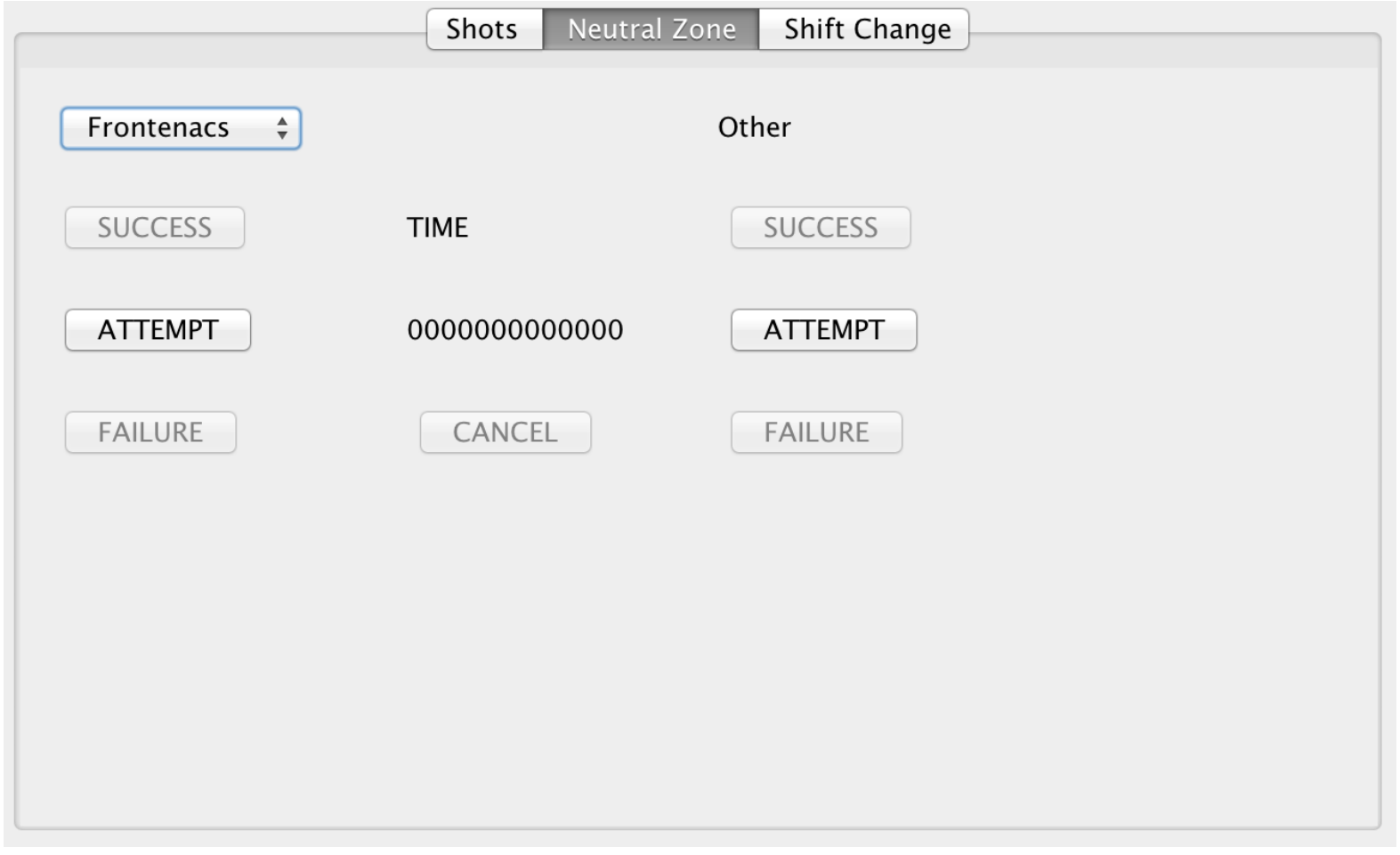
1. Through aggregation/scraping of league data
   - Low cost
   - Low collection bias
   - Lots of data
   - Data may not describe the question you are investigating
2. Through purchasing third-party information
   - Moderate to high cost
   - Repeated costs for more data
   - Low collection bias
   - Data may describe the question but may not be ideal
3. Through direct recording of in game events
   - Get the data you want!
   - Much higher cost
   - Must create strict guidelines to ensure data quality
   - Potential collection bias

Data Collection - Example

We agreed to provide information on neutral-zone play by reporting each player's 5v5 neutral-zone success rates (attack and defense). The OHL does not record neutral-zone data or any potential proxy. We were restricted by a small budget making large scale data collection impossible. Due to this we designed a recording/scraping hybrid data collection process.

- We developed a JAVA program to collect player ice-time and neutral-zone data.
- Since neutral-zone events are not traditionally recorded, strict guidelines and training were needed to ensure consistency in our data and address all potential situations.
- We used rvest to scrape the OHL website to obtain special teams data.
- Data merging can be a huge headache.
- Moderate costs for data collection and programming.

Data Collection - Java



Free million dollar idea: Someone develop an app to do this.

Data Analysis - Preliminary Reporting

Once we collect the data, we are able to report preliminary statistics directly. When deciding on these preliminary statistics to report there are several things to consider.

- Statistics should be easy to understand by non-statisticians
- Keep the analysis streamlined
- Ensure assumptions made are valid
- Should lead well to future analysis
- Create a narrative for the analysis
- Develop/refine the lines of communication

Data Analysis - Preliminary Reporting Example

We considered each attempt to get through the neutral-zone as a realization of a Bernoulli random variable. For each player's two roles (offensive and defensive) the course of a game the amount of success can be modeled as a pair of binomial random variables. We reported the estimated probabilities for these binomial distributions to describe a players skills in the neutral-zone.

The most obvious way to estimate a players success rate is $\hat{p} = \frac{successes}{attempts}$. This is ideal in several ways: easy to understand for decision makers, is the MLE, and has a well defined distribution.

We provided weekly reports of the estimates, $\hat{p}_O$, $\hat{p}_D$, for each player at even strength with a narrative when possible.

Data Analysis - Estimated Neutral Zone Differential Formula

To provide a singular neutral-zone statistic, we developed the Estimated Neutral Zone Differential (END). We referenced a player's success rate to the average rate for similar players. For a player X and reference group G, we have

$$END(X, G) = (\hat{p}_O(X) - \hat{p}_O(G)) + (\hat{p}_D(X) - \hat{p}_D(G)). \quad (1)$$

- Pro - Single descriptive result
- Pro - Statistically shown to model to perceived player quality
- Pro - Cool name
- Con - Not easily understood
- Con - No well defined distribution under our assumptions

Including Covariates into the analysis

Our previous analysis considered the variables to be independent of the other players and game situation. This is a strong assumption that is unlikely true. Including these variables in our analysis may give more clarity about the variables we are analyzing.

- Regression analysis is a common method to include covariates.
- Many varieties of regression models exist, use depends on our data and assumptions.
- Care must be had when selecting covariates to ensure data quality and independence.
- More difficult to describe, relating our results back to terms from our previous analysis and clear figures are useful.
- Most every numerical computer program can perform regression.
- Model selections is important and care must be taken to ensure our resulting model is valid and fits with our objectives.

Regression Example - Line Optimization Formulation

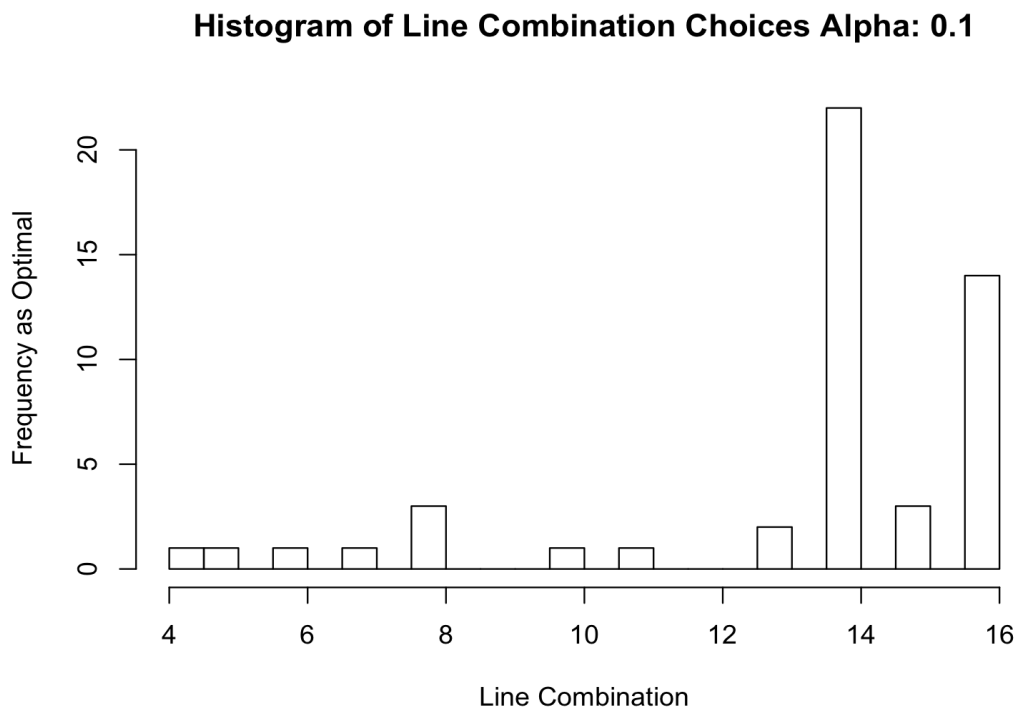
Another area that we were interested in analyzing was line combinations and there affect on scoring. We used logistical regression to model the effect of players and their lines (2 or 3 term interactions) on the probability of a goal being scored in the next minute. We evaluated the difference in cumulative pessimistic confidence bounds on the coefficients for models on goals for and against. That is for a line combination we get the metric,

$$\begin{aligned}\Delta_\alpha = &(\sum_{i=1}^{18}(\beta_{i,gf} + \Phi(\alpha)S(\beta_{i,gf})) + \sum_{j=1}^{3}(\gamma_{j,gf} + \Phi(\alpha)S(\gamma_{j,gf})) \\ &+ \sum_{k=1}^{4}(\zeta_{k,gf} + \Phi(\alpha)S(\zeta_{k,gf}))) - (\sum_{i=1}^{18}(\beta_{i,ga} + \Phi(1-\alpha)S(\beta_{i,ga}) \\ &+ \sum_{j=1}^{3}(\gamma_{j,ga} + \Phi(1-\alpha)S(\gamma_{j,ga})) + \sum_{k=1}^{4}(\zeta_{k,ga} + \Phi(1-\alpha)S(\zeta_{k,ga}))).\end{aligned} \tag{2}$$

Where α is how pessimistic we are.

Regression Example - Line Optimization Application

To avoid issues with unequal use of line combinations in the data, we employed a bagging algorithm. For each sampling from our data we selected the line combination with the largest Δ_α. Then we reported the optimal line combination to the combination selected most often across the samples.



Advanced Modeling - Upping the complexity

As the last example showed, we can take a relatively simple regression problem and complicate things pretty fast. Depending on our data or objectives we may need to use more advanced methods.

- There are a variety of fields for data types (spatial, time-series, etc.).
- We have to do our statistical due diligence (research on methods).
- Results can get more complex and difficult to report, remember to use simple/common terms and clear plots.
- Programs to perform the analysis are more sparse. R usually has a package, or you can write one.
- We shouldn't be shy about taking the road (methods) less traveled.
- "Done is better than perfect."

Advanced Methods Example - Player Monitoring

To provide the coaches with a way to monitor and evaluate player performance within games we used exponentially weighted moving average quality control charts. For a player's neutral-zone offensive success rate score we have,

$$EWMA_{NeutralOffense}(t) = \sum_{i=1}^{t} \lambda(1-\lambda)^{i-1} \frac{T - \hat{p}_O(i)}{T}. \quad (3)$$

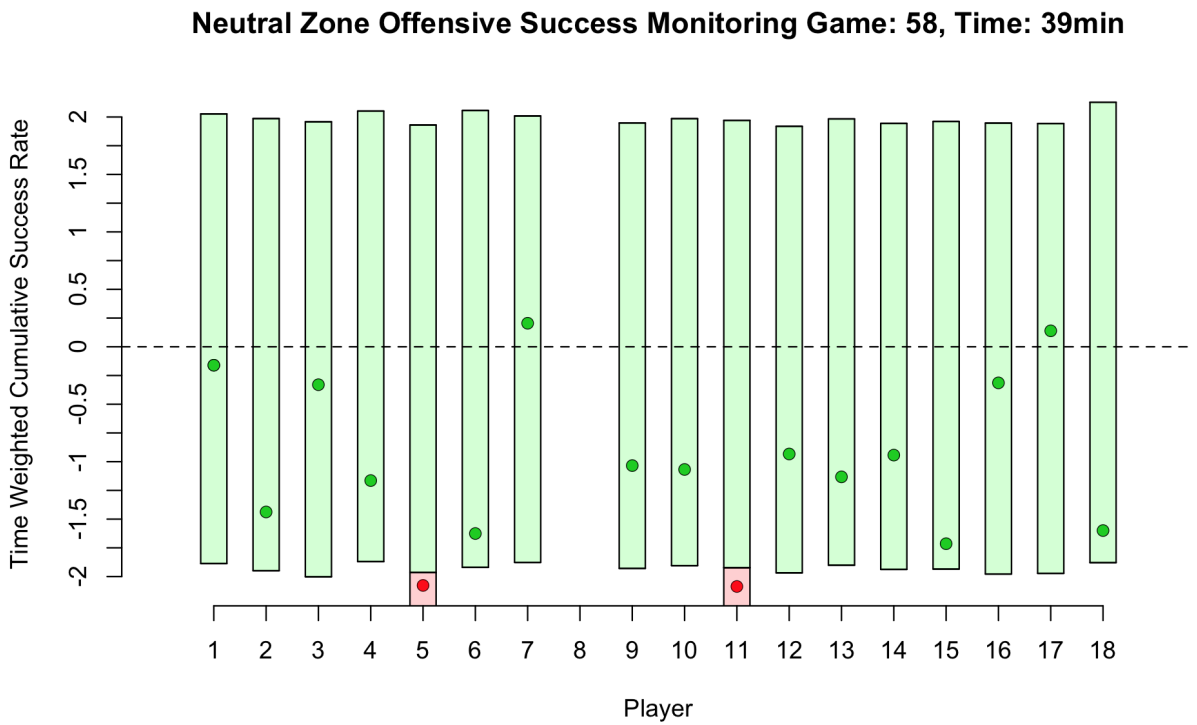
T being the target value for a player.
Under the null hypothesis that $H_0 : \mu_{NeutralOffense} = T$, we have

$$EWMA_{NeutralOffense}(t) \sim N(0, \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]}). \quad (4)$$

We then set significance bounds on this statistic and identified when a player is performing significantly different from their target value.

Data Analysis - Player Monitoring Online Application

We can use this method to give the status of the current entire team simultaneously as an in-game reporting tool. We standardized all player's scores to make the plot easier to read.



We can see here that players 5 and 11 are under performing at this time with respect to their average.

Concluding Remarks

- Data collection in any circumstance should not be impossible but we must be vigilant to not introduce avoidable bias.
- Start simple and build to complex.
- Create a narrative to explain the results.
- Plan ahead how for reporting results. (Figures go a long way!)

- Don't be afraid to go outside your statistical comfort zone. This is where the fun begins!
- Go Fronts Go!

Acknowledgments

This research is supported by the Kingston Frontenacs, Queen's University and my doctoral supervisors Dr. Takahara and Dr. Thomson.

If you want more info or a copy of my JAVA program email me.
pohlkamp.hartt@gmail.com

References


D. Freedman, *Bootstrapping regression models*, The Annals of Statistics **9** (1981), no. 6, 1218–1228.


D. Montgomery, *Introduction to statistical quality control*, John Wiley & Sons, 2007.


T Purdy, *Shots, fenwick and corsi*, February 2011.


M. Schuckers and J. Curro, *Total hockey rating (thor): A comprehensive statistical rating of national hockey league forwards and defensemen based upon all on-ice events*, Proceedings of the 2013 MIT Sloan Sports Analytics Conference, 2013.


S. Simmons, *Why hockey's trendy advanced stats are a numbers game*, May 2014.


R. Tibshirani T. Hastie and J. Friedman, *The elements of statistical learning*, vol. 1, Springer, 2001.


R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) **58** (1996), no. 1, 267–288.