Quinn Helmig and Sol Benishay

Introduction to Statistics

Professor Hoerl

3/15/19

Union Lacrosse Process Improvement Project

Executive Summary:

When first approached with this project, Quinn and I immediately knew that lacrosse would be a relevant and fun topic for us to do it on. With both of us having a deep knowledge of the game and Quinn being a member of the varsity team here, we knew we had the ability to execute a sound project on the sport. With the men's varsity team having a mediocre season, despite having the coaches and talent to do a lot better, we thought it would be a great idea to figure out what the team's issues were and how we can improve them. Our goal is to make the team more efficient. For the sake of this project, we are defining efficiency as our goal differential. The formula for goal differential (GD) is GD = goals forward-goals against. Positive differentials indicate a Union win and negative ones indicate an opponent winning. The goal is a larger positive goal differential. To increase this, we first needed to identify the main drivers behind GD. We were able to download the statistics from the team's 2018 season off the website and through film and using a multivariate scatterplot, we identified which explanatory variables (statistics) correlated most highly with goal differential. We concluded that Shots Against (SA), Assists (A), Ground balls (GB), Shots on Goal (SOG), and Assists Against (AA) were the most important explanatory variables behind goal differential. In the end we narrowed down down to SA and A as our variables for a designed experiment. Using the output from the regression analysis we identified the predicting formula for goal differential as GD= 18.390 + .091A - .655SA. Using high/low assists data points and high/ low shots against data points in the regression equation, we were able to conclude low shots against and high assists were most important for a high goal differential. We are hoping that we will be able to use this information to implement new drills and refocus the team on this variables to lead to success. We

believe that successful implementation of remedies to these factors will win the team a lot more games and hopefully championships.

Purpose:

With Quinn being a member of the varsity team here and both of us having lacrosse experience, we were hoping we would be able to use our knowledge to improve the teams success. We both have always been interested in what leads to a teams success in any sport and this project will allow us to learn how to do so and implement solutions on our own program. Our project will be using data from the 2018 lacrosse season to analyze and find how each variable affects our goals forward and goals against, and consequently our efficiency (goal differential). Using the data that was taken game-by-game as well as looking over the film from each game, we hoped we would be able to gather a good amount of important variables for the teams success. With access to Hudl, we were able to verify and had the prospect to implement new statistics outside the traditional ones. Hopefully our sports minded brains allow us come up with solutions towards bettering the Union men's lacrosse team success after analyzing the data. We would like to find which variables should be improved to increase the goals forward and decrease the goals against, thereby increasing goal differential and efficiency. Through understanding this, we will be able to talk about what concepts we should be working on more in practice to translate to more success in games.

Process Improvement Framework:

Through the process improvement framework, we were able to organize our ideas out of a large pool of lacrosse knowledge and figure out how we could make the team more successful. The flowchart allowed us to identify the things that need to happen in order for the team to win and have a high goal differential in a game. With access to the teams statistics through the website and game film access on Hudl, we were able to collect data on key variables to the team's success. We defined the teams success or efficiency as their goal differential, equal to goals forward-goals against. Using run and control charts from the game, we were able to identify outliers and any systematic variation, with those being the quality of team they were playing and the teams focus throughout the season. With a histogram of the goal differential, we found that most games lie between around -5 to 10. This seemed about right to us as most competitive college lacrosse games are in this range. Following this, we used multivariate regression analysis and scatterplots to narrow down to 5 factors which affect goal differential the most, picking those with the greatest magnitude correlations. In the end we executed a designed experiment using a regression equation of the 2 most important factors, Shots Against and Assists, and high and low data points of those from games as plug in data. In the end we concluded that by lowering shots against and increasing assists we could maximize our goal differential. Proper implementation of drills to improve these two variables will be vital to help maximize the team's performance and efficiency.

SIPOC:

Suppliers- Union Men's Lacrosse Players and Coaches

Inputs- Ball, sticks, goals, field, equipment.

Process-

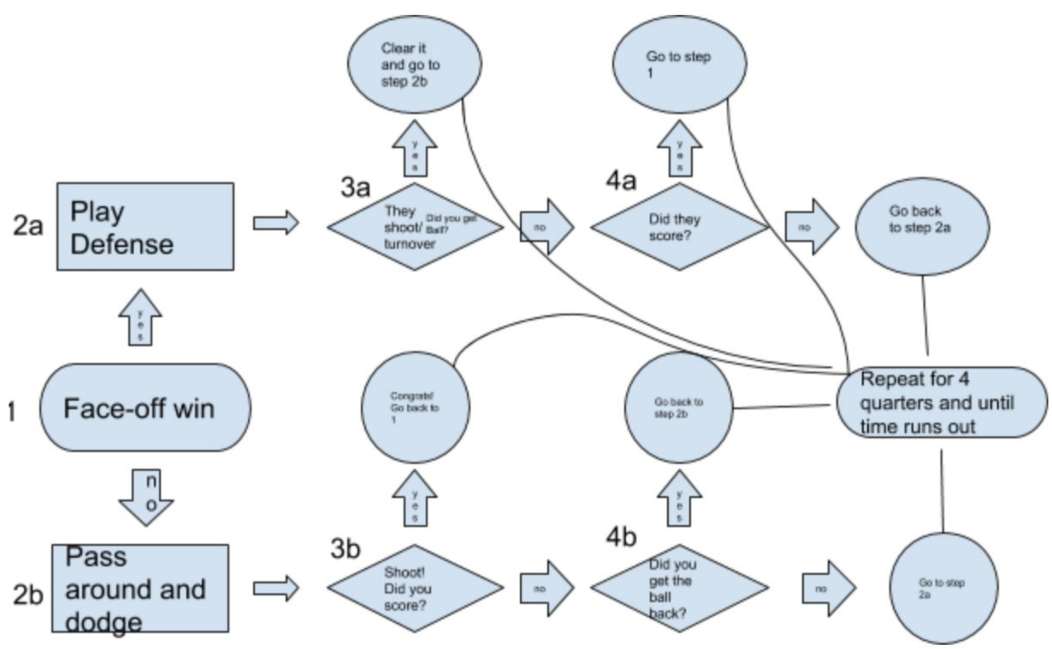
1. Win possession on face-off
2. Retain possession and get ball into box
3. Find open man
4. Shoot again
5. If it doesn't go in get possession back

6. Shoot again
7. Other team has ball
8. Play defense
9. They shoot, we get ball back or they score and take another faceoff

Outputs- Goals for the team (GF) and goals against the team (GA). Main output will be efficiency (E). E = GF-GA

Customers- Union Men’s Lacrosse Players, Coaches, and Fans

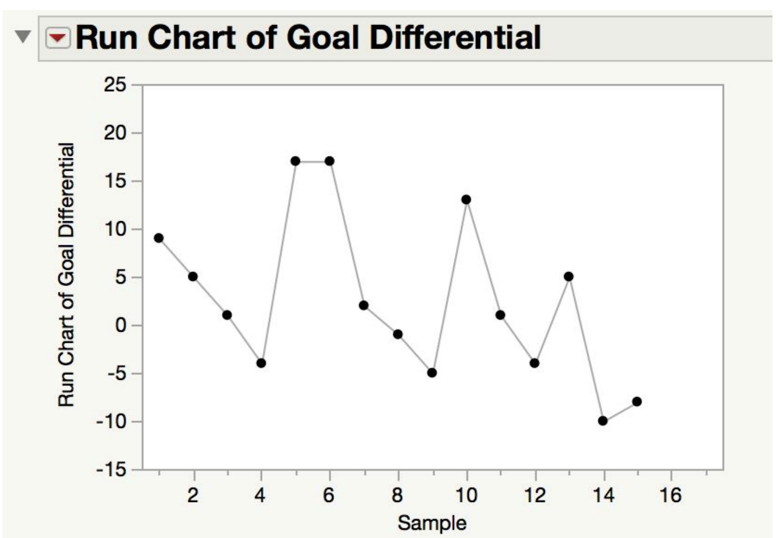
Flowchart:



Description of Data: The main output will be efficiency or goal differential, defined as goals forward minus goals against. The main input variables are assists for/ against, shots for/ against, shots on goal for/ against, shots on goals percentage for/ against, ground balls for/ against, turnovers for/ against, face-off percentage for/ against, and penalties in minutes for/ against. These are most of the variables available on the Union

College Athletics website. Later on, we narrow down our variables to the 5 most important ones by using subject matter knowledge and t-ratios. There is some concern over whether the statistics are correct on the website as well as on Hudl. On Hudl, I can organize clips by what happened, eg. I can organize it to see only turnovers. This is done by someone who works at Hudl, so their statistics might not match up with the same as the person who took them at the game.

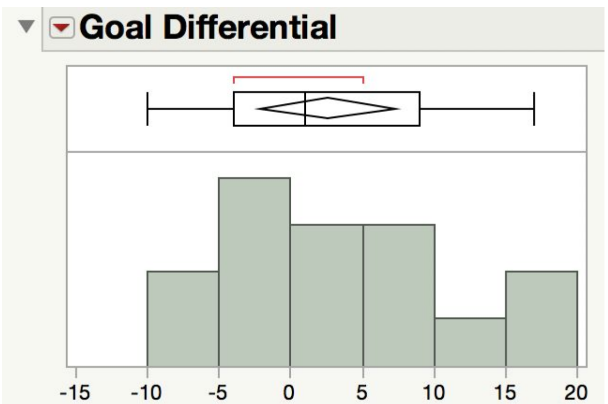
Stability of Data: It seems as though the data is relatively stable with a possible high and low outlier or two by samples 5,6 and 14,15. I see very little evidence of significant special causes except for the possibility of the teams being really good or really bad. The team likes to consider every opponent as equal so we will not strongly identify those games as special cause outliers. Other than these possible outliers due to the quality of team that was played, we conclude that the run chart output is stable with no specific pattern in the chart.



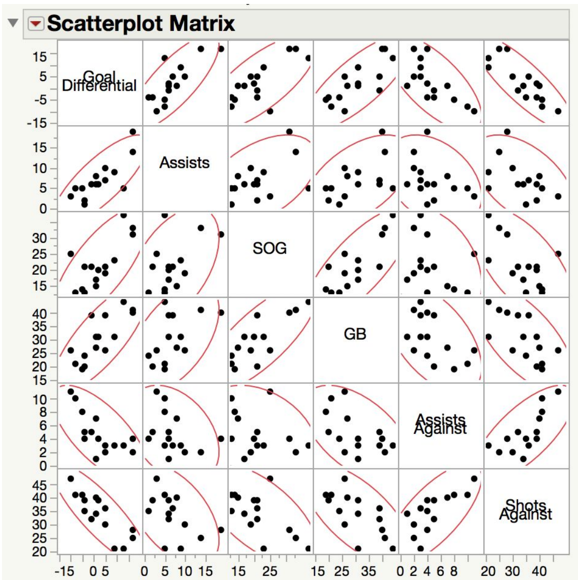
Application of Five Tools:

-Data Based

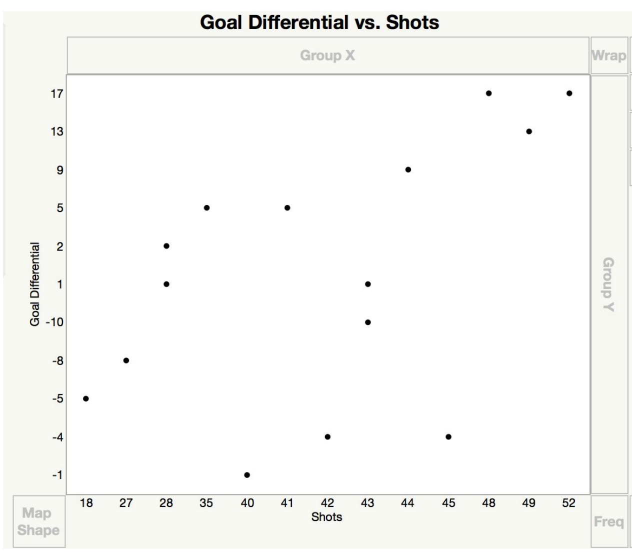
- **Histogram:** We can conclude that in this histogram, most of our goal differential falls between -5 and 10. We would prefer this to be as high and as a positive number as possible. This would mean that we were scoring more goals than the other team. For the distribution, we have a median of 1, mean of 2.5, and standard deviation of 8.49. We expected the curve to look like this because it represents how the team did each game. As evident from the graph and the season, they had a lot of close games, and some that we either won by a lot or lost by a lot. The graphs shape right now is unimodal skewed left.



- **Multivariate:** The multivariate chart below, or a scatterplot matrix, shows the different correlations that the variables have with each other. We used this to help cut down our explanatory variables from the 18 we already had. Although we end up using only 2 explanatory variables for the designed experiment, we felt as if it was important to identify at least 5 for as we started with 16. We concluded that those 5 variables with the highest magnitude correlations with goal differential as the lead drivers behind goal differential. Those 5 were: Shots Against (SA), Assists (A), Ground Balls Forward (GB), Shots on Goal (SOG), and Assists Against (AA). Through this, we can conclude that assists and shots on goal against are the 2 most important variables to goal differential. For this, we limited the variables to put on it. Many of the variables can be counted as duplicates because as we get one it decreases the other teams.

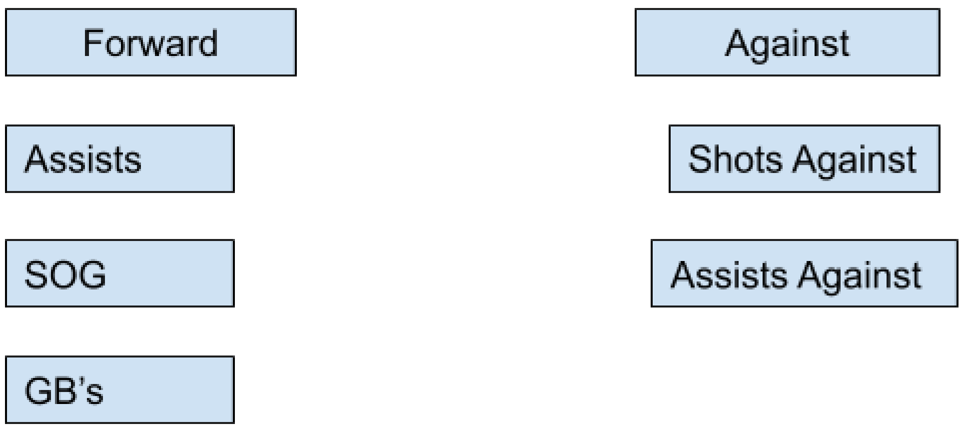


- **Scatterplot:** The scatterplot illustrates shots forward on the X axis and goal differential on the y axis. From this plot, I see a moderately strong positive correlation between Shots and Goal Differential. In the majority of cases on this plot, shooting more led to a higher, more wanted goal differential. Although there are some clear outlier, most notably towards the bottom, the graph keeps a general upward trend.

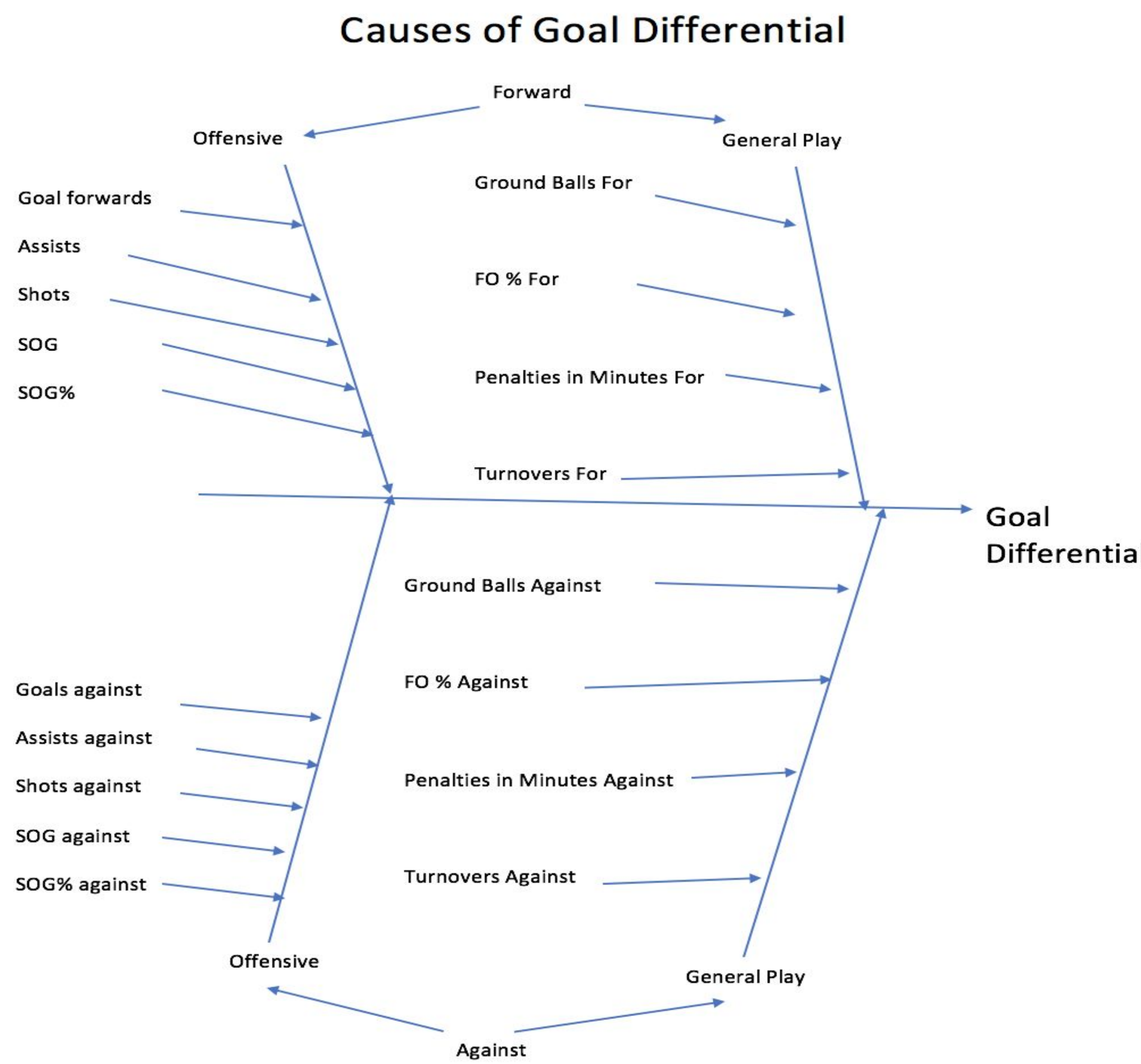


-Knowledge Based

- **Affinity Diagram:** We were able to make an affinity diagram with the 5 variables that are the most important towards goal differential. This helps to split up everything into things that increase goal differential and things that decrease goal differential. From this, we are able to understand what is helping our goal differential go up and down.



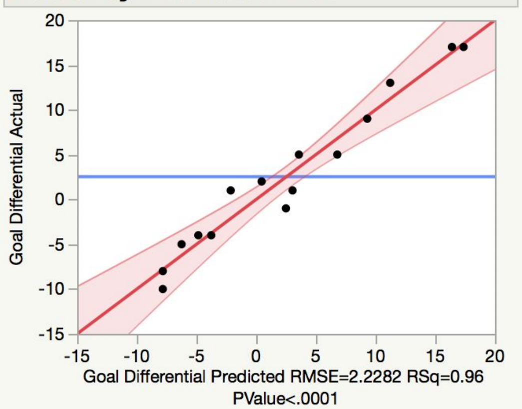
- **Cause and Effect Diagram (Fishbone):** We thought a cause and effect diagram of all 18 explanatory variables we had would be a great way to organize the variables and types of variables that affect goal differential. I first segmented the variables by whether they were for our team or against our team and then by offensive or general play statistics. As can be seen, the top and bottom of the fishbone are opposite of each other and all these factors lead to goal differential.


Causes of Goal Differential
Forward
Offensive
General Play
Goal forwards
Assists
Shots
SOG
SOG%
Ground Balls For
FO % For
Penalties in Minutes For
Turnovers For
Goal Differential
Ground Balls Against
FO % Against
Penalties in Minutes Against
Turnovers Against
Goals against
Assists against
Shots against
SOG against
SOG% against
Offensive
General Play
Against

Regression Analysis: Using all of our variables, we can make a very accurate model of goal differential. However, using all of our 16 variables is extremely unrealistic. Using our subject matter knowledge and understanding redundancy within the variables, we were able to narrow it down and choose the variables that correlated directly to goal differential the best. Many of the 16 variables that we started with are duplicates because one will increase for one team and decrease for the other team. The others did not have a strong enough impact to even use in our model. Once we narrowed it down, the variables we came upon were assists, ground balls, shots on goal, assists against, and shots against.

Response Goal Differential

Actual by Predicted Plot



Effect Summary

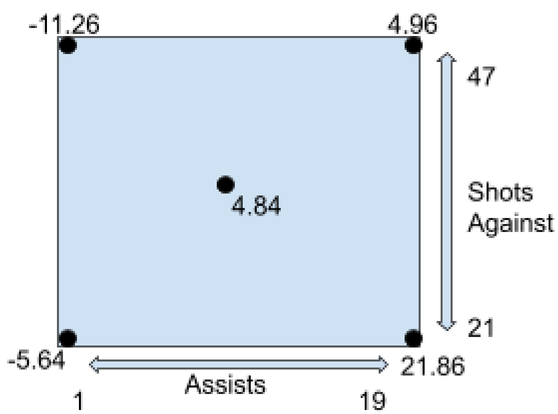
Source	LogWorth	PValue
Assists	3.067	0.00086
Assists Against	1.250	0.05629
SOG	1.161	0.06900
Shots Against	1.049	0.08939
GB	0.199	0.63171

Remove Add Edit FDR

Summary of Fit

RSquare	0.955746
RSquare Adj	0.931161
Root Mean Square Error	2.228216
Mean of Response	2.533333
Observations (or Sum Wgts)	15

Designed Experiment: After initially narrowing down our experiment to the 5 variables used in the regression analysis, we had to further narrow it down 2 variables to make a designed experiment out of this. Since most of our experiment was observational to this point, it was tough to pick just 2 variables that could represent a whole lacrosse game. In the end, we moved to assists and shots against as our 2 variables. We chose these as they have the highest absolute value t-ratio's, thus the most important variables in the games. They also contradict each other as assists go hand in hand with goals forward and shots against is the most important variable for goals against. As you can see on the square, you have the highest goal differential when you have more assists and fewer shots against. You have the lowest goal differential when you have fewer assists and more shots against.



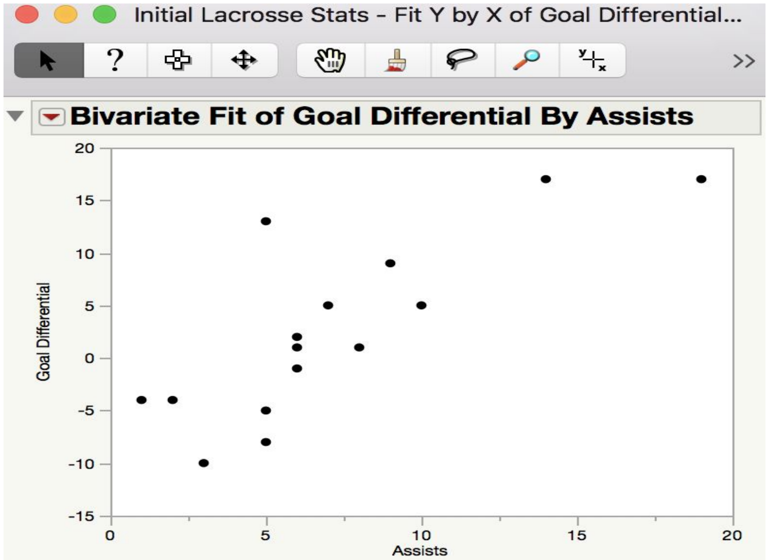
Hypothesis Testing:

Null hypothesis: Assists do not have an affect on goal differential

Alternative hypothesis: Assist do have an affect on goal differential

$\alpha = .05$

JMP Output with confidence interval for correlation between assists and goal differential



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.794439	0.475814	0.92874	0.0004*
Covariance	31.17619			
Count	15			

Variable	Mean	Std Dev
Assists	7.066667	4.620864
Goal Differential	2.533333	8.492574

Conclusion: As can be seen in the jmp output above, the significance probability if .0004. With our alpha value set at .05 and the significance probability way below that, there is strong evidence to reject the null hypothesis and conclude that assists do in fact have a significant effect on goal differential. If the lower and upper bounds contained 0, we would have less evidence to reject the null hypothesis.

Confidence Intervals: The first set of confidence intervals seen below is the CI's based on the histogram of goal differential. Based on our data, we should see that 95% of the population should come in between -2.16 and 7.24. As well, the standard deviation of the population should be between 6.21 and 13.39. The second figure shows the effects that the variables have on goal differential and their matching 95% confidence intervals.

Since assists have 2 positive signs, that means that it has a positive effect on the goal differential. Then, the shots against has 2 negative signs which means that each shot against has a negative effect on the goal differential.

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	2.533333	-2.1697	7.236364	0.950
Std Dev	8.492574	6.217639	13.39363	0.950

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	18.394044	5.051529	3.64	0.0034*	7.3877085	29.400379
Assists	0.9013851	0.202109	4.46	0.0008*	0.4610281	1.3417421
Shots Against	-0.655123	0.121079	-5.41	0.0002*	-0.91893	-0.391315

The Story: We first knew that we needed to identify the five most important key variables that affect goal differential in order to find solutions. Using regression analysis, we took the 5 factors with the highest magnitude correlations and focused on them for the majority of the project, until we had to come down to 2 for the experiment. Because our data was previously collected, we needed to find a natural experiment in the data that would help us identify root variables. From the regression equation for goal differential calculated by JMP from assists and shots against, we plugged in high and low natural occurrences of those 2 most important variables to design our experiment. In the end we concluded that the optimal situation for a high goal differential would be a game with a lot of assists and not a lot of shots against.

Future Steps: With the realization that assists for and shots against are the two most important factors to the team's efficiency, we believe that getting the team to understand this and be able to implement better will be the solution. Through the stressing of passing, teamwork when scoring, and looking for the open man by coaches and the running of drills to simulate assists, the team should be able to increase their numbers. For shots against, emphasis of defensive techniques and not letting opponents get shots of should be stressed in practice. The team should increase their defensive focus to help this and lower the goals scored on them. By properly focusing and improving on these two variables, the team should be able to increase their success a lot.

Summary: With the results from the JMP output and conclusions we have made of it, it can be clearly seen that many factors affect goal differential in the game, but some are much more important than others. After coming down to 5 with the highest magnitude correlations, we identified assists forward and shots against as the two most important factors to the team's efficiency and goal differential. Using the predictive regression equation for goal differential from JMP, we simulated what would happen in different situations with the variables. In the end as we suspected, more assists for the team and less shots against would reap a higher goal differential and more wins. This goes to reaffirm much of the cliche lessons we have learned from coaches in the past; always emphasizing team offense and statements like "defense wins championships". Although we have yet to prove the latter theory statistically, this project gave us the baseline and experience to be able to think statistically and execute inferences and experiments on other processes. The process improvement framework taught us to sequentially and systematically think about causes and solutions to different issues in processes. Doing everything in the step by step order and using the different statistical and knowledge based tools, allowed us to gain a better understanding statistical thinkings application to process improvement. I believe after this, we will be able to apply these new insights from our statistical inferences to improve on the field performance. More importantly, we will now have to foundation to apply this framework to other cases in need of process improvement.