Training Intervention Analysis

Ryan Maguire, id=18464854

11 November 2019

## Context: Celtic Study introduced in class.

Scaffold for the analysis when the primary response variable is VO2 max. You need to rerun the analysis using the Squat variables (i.e. Squat\_Pre, Squat\_Post) to see if there has been any improvemnt on average and provide a confidence interval for the likely average improvement of players in the populaiton of interest.

Therefore, the key question of interest is to see if there has been any improvemnt on average in the population of interest.

* State the appropriate null and alternative hypotheses?
* null: u > 130.24
* Alternate: u <= 130.24
* Define a Type I and Type II error and discuss the implication of making these errors in this application.

Type I Error: this occurs when the null hypothesis is rejected when it is true. For this data set, this would be accepting the null hypothesis and believing there was improvment, even though there was no noticable improvement

Type II Error: Occurs when the null hypothesis is false but is incorrectly accepted. For this data set, this would be saying there was no noticable improvment although there actually was.

## Read in the training intervention data

Read in the data and have a look at the variable names and structure of the data.

train.df <- read.csv("training.csv")  
glimpse(train.df)

## Observations: 18  
## Variables: 6  
## $ X <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...  
## $ ID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...  
## $ VO2.max\_Pre <dbl> 66.4, 70.9, 64.9, 68.6, 76.7, 75.6, 78.1, 73.1, 7...  
## $ VO2.max\_Post <dbl> 67.8, 81.7, 70.1, 73.0, 84.5, 78.4, 80.5, 76.0, 7...  
## $ Squat\_Pre <dbl> 120.80, 119.43, 129.84, 130.80, 110.55, 130.35, 1...  
## $ Squat\_Post <dbl> 140.47, 149.17, 158.67, 159.77, 139.86, 160.60, 1...

## Focus on the VO2 max response variables.

## Summary Statistics

train.df %>% select(Squat\_Pre,Squat\_Post) %>% summary()

## Squat\_Pre Squat\_Post   
## Min. : 99.15 Min. :129.2   
## 1st Qu.:121.17 1st Qu.:142.6   
## Median :130.09 Median :160.2   
## Mean :130.24 Mean :159.5   
## 3rd Qu.:140.07 3rd Qu.:170.8   
## Max. :159.61 Max. :189.7

## Mean and Standard Deviation

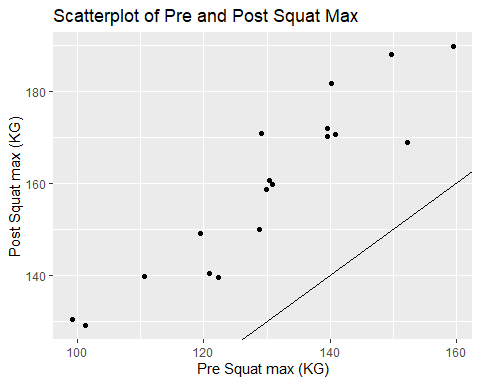
train.df %>% select(Squat\_Pre,Squat\_Post) %>%  
 summarize(Pre\_Mean=mean(Squat\_Pre), Pre\_SD= sd(Squat\_Pre),  
 Post\_Mean=mean(Squat\_Post), Post\_SD= sd(Squat\_Post))

## Pre\_Mean Pre\_SD Post\_Mean Post\_SD  
## 1 130.2361 16.43567 159.4506 18.67436

The Mean has clearly increased which could show that the training has in General improved their Squat max considerably, but the standard deviation has also increased only marginally but this means that the Squat values are spread further away from the mean.

## Scatterplot of Pre and Post with line of equality

train.df %>% ggplot(aes(x = Squat\_Pre, y = Squat\_Post)) +  
 geom\_point() +   
 ggtitle("Scatterplot of Pre and Post Squat Max") +  
 ylab("Post Squat max (KG)") +  
 xlab("Pre Squat max (KG)") +  
 geom\_abline(slope=1, intercept=0)



This scatter plot clearly outlines a positive correlation between the Post and Pre Squat max, this correlation seems very strong shown by the 0.9256596 correlation co-efficient, but the line of best fit is below alot of our data points this could suggest their is an outlier in the data that is causing this to occur. But the graph clearly shows the higher your pre-Squat the higher your post Squat would be in most cases after the training.

## Calculate the Improvement

Calculate a new variable and have a look at the data frame to see that it has been created. High vlaues of VO2 max are good to Post-Pre is a better measure than Pre-Post to capture this.

train.df <- train.df %>% mutate(Improvement = Squat\_Post-Squat\_Pre) %>%  
 glimpse()

## Observations: 18  
## Variables: 7  
## $ X <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...  
## $ ID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...  
## $ VO2.max\_Pre <dbl> 66.4, 70.9, 64.9, 68.6, 76.7, 75.6, 78.1, 73.1, 7...  
## $ VO2.max\_Post <dbl> 67.8, 81.7, 70.1, 73.0, 84.5, 78.4, 80.5, 76.0, 7...  
## $ Squat\_Pre <dbl> 120.80, 119.43, 129.84, 130.80, 110.55, 130.35, 1...  
## $ Squat\_Post <dbl> 140.47, 149.17, 158.67, 159.77, 139.86, 160.60, 1...  
## $ Improvement <dbl> 19.67, 29.74, 28.83, 28.97, 29.31, 30.25, 32.34, ...

## Mean and Standard Deviation of Improvement

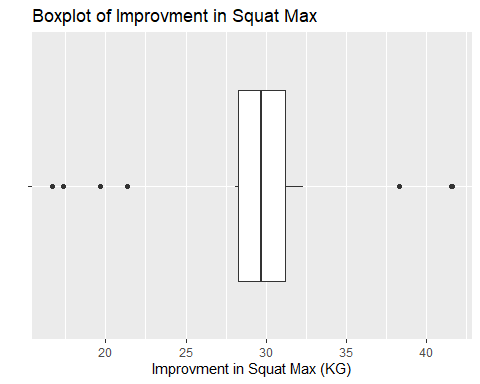
train.df %>% select(Improvement) %>%  
 summarize(Imp\_Mean=mean(Improvement), Imp\_SD= sd(Improvement))

## Imp\_Mean Imp\_SD  
## 1 29.21444 7.116578

The mean shows there are clear improvements but if we look at the data set we can clearly see some severe outliers that could be skewing the mean like the 41.5 improvement and the 17.33 improvement are quite far apart from the 29.21 mean that was calculated, the standard deviation though shows that their is a bit of spread in our data from the mean this is again likely because of certain outliers in the data set. Interpret!

## Boxplot of Improvement

train.df %>% ggplot(aes(x = "", y = Improvement)) +  
 geom\_boxplot() +   
 ggtitle("Boxplot of Improvment in Squat Max") +  
 ylab("Improvment in Squat Max (KG)") +  
 xlab("") +  
 coord\_flip()

 The box plot is relatively compact showing that our values when we take out the outliers are not very spread apart, also this boxplot shows us that there are alot of outliers in our data set shown by the 7 dots outside from the boxplot, the boxplot shows us that the median is a very good estimate of what an average player would improve as the median is pretty much dead center of the boxplot, there’s no heavy skewing to the left or right.

## 95% Confidence Interval Using the t.test function

train.df %>% select(Improvement) %>% t.test()

##   
## One Sample t-test  
##   
## data: .  
## t = 17.417, df = 17, p-value = 2.832e-12  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 25.67545 32.75344  
## sample estimates:  
## mean of x   
## 29.21444

Based on the output given answer the following questions:

* What is the likely average improvement in the population of interest? Interpret the relevant 95% Confidence Interval carefully. This confidence interval shows that we can be 95% confident that any individual can improve their Squat max somewhere between 25.68kg and 32.75kg, so because the mean falls within this range I would expect an individual to maybe improve by around 29.2kg.
* Use the relevant interval estimate and p-value to decide whether you think there is sufficient evidence in the samples provided to claim that there is any improvement in the population of interest!
* The p value is tiny and is less than 0.05 so therefore we can reject the null hypothesis and accept our alternate hypothesis that there is an improvement in the mean Squat maxes.
* What are the assumptions underlying the one sample t-test presented?

Firstly, we presume that the scale of measurement follows a continuous scale.

Secondly, the collected data was from a randomly selected portion of the total population.

Thirdly, when we plot the data, we will almost certainly find a normal distribution curve.

Again, a sample of size 19 is used, this again suggests that were dealing with a normal distribution curve.

Finally, we can presume that equal variance exists when the samples are almost equal.

* Explain why or why not the assumptions look justified based on the output provided.

I believe the assumptions I outlined above can’t be validated due to the relative size of our data set which contains 19 players which is not really enough to get a general sense of the total population in my opinion.

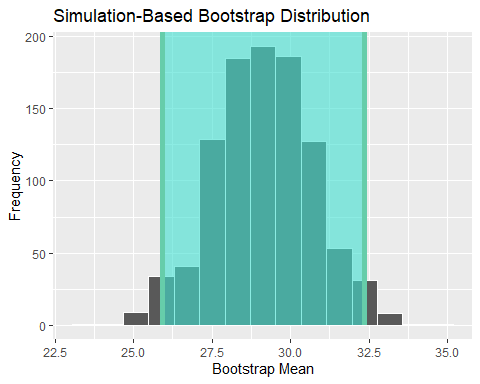
## 95% Bootstrap CI for the mean

boot <- train.df %>%  
 specify(response = Improvement) %>%  
 generate(reps = 1000, type = "bootstrap") %>%  
 calculate(stat = "mean")  
  
percentile\_ci <- get\_ci(boot)  
round(percentile\_ci,2)

## # A tibble: 1 x 2  
## `2.5%` `97.5%`  
## <dbl> <dbl>  
## 1 25.9 32.3

boot %>% visualize(endpoints = percentile\_ci, direction = "between") +  
 xlab("Bootstrap Mean") + ylab("Frequency")

## Warning: `visualize()` should no longer be used to plot a confidence  
## interval. Arguments `endpoints`, `endpoints\_color`, and `ci\_fill` are  
## deprecated. Use `shade\_confidence\_interval()` instead.



Well, the green shows our confidence Interval for the mean and we can see that we have a normal distribution graph after a thousand resamples and most of these values fall within our confidence interval which means our confidence interval is validated, there are a couple outliers but their influence is more than likely minimal to say the least and they seem balance each other out.

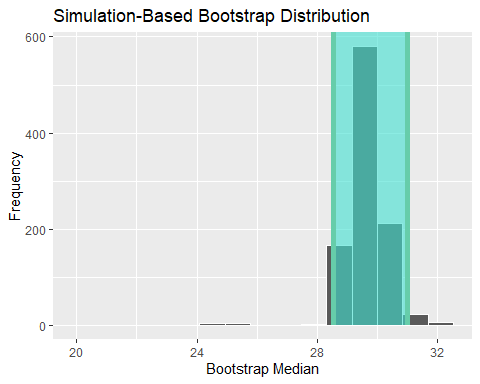
## 95% Bootstrap CI for the median

boot.median <- train.df %>%  
 specify(response = Improvement) %>%  
 generate(reps = 1000, type = "bootstrap") %>%  
 calculate(stat = "median")  
  
percentile\_ci\_median <- get\_ci(boot.median)  
round(percentile\_ci\_median,2)

## # A tibble: 1 x 2  
## `2.5%` `97.5%`  
## <dbl> <dbl>  
## 1 28.5 31

boot.median %>% visualize(endpoints = percentile\_ci\_median, direction = "between") +  
 xlab("Bootstrap Median") + ylab("Frequency")

## Warning: `visualize()` should no longer be used to plot a confidence  
## interval. Arguments `endpoints`, `endpoints\_color`, and `ci\_fill` are  
## deprecated. Use `shade\_confidence\_interval()` instead.



Again the green shade shows us the interval of our confidence interval of the median and like before alot of the values fall within our interval which again validates our interval, again there are a few outliers with a couple of sever outliers off to the left but their frequency is tiny, this may skew the results slightly but relatively little as has been seen.

### Overall Conclusion ??

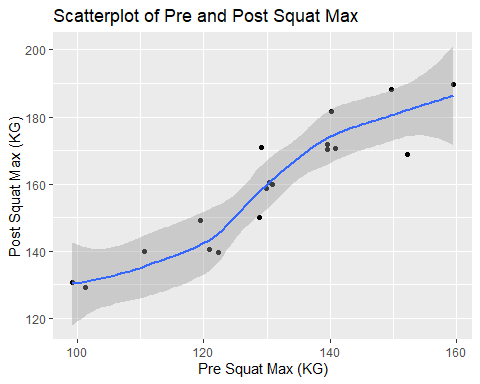
From all the data above we fail to reject the null hypothesis outlined and accept the fact that the players had increased their Squat max from their previous Squat max, the only part im skeptical on is whether we could fully apply this to the population because our data set is relatively small with only 19 players , im also concerned how they were sampled, are the players randomly selected? Are the players all from the one club, these are concerns that could lead us to an error in our assumptions of what this data means in regards to the entire population, maybe in fact most people would not improve their overall squat as a result of the training, also we have no data or information on the training these players are taking.

## Assessing the relationship between pre and post VO2 max The aim here is to see if there is any relationship between pre and post VO2 max. In other words, we are going to find if we can predict the value of post VO2 max for a player based on his pre VO2 max value. To answer the question, a scatter plot along with a smoother is displayed as below.

## Scatterplot of Pre and Post with line of equality

train.df %>% ggplot(aes(x = Squat\_Pre, y = Squat\_Post)) +  
 geom\_point() +   
 ggtitle("Scatterplot of Pre and Post Squat Max") +  
 geom\_smooth()+  
 ylab("Post Squat Max (KG)") +  
 xlab("Pre Squat Max (KG)")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



## Correlation coefficient

cor(train.df$Squat\_Pre,train.df$Squat\_Post)

## [1] 0.9256596

Change the analysis done for VO2.max to Squat and answer the following questions accordingly.

* What is the correlation between the pre and post Squats? Interpret it.

The correlation as shown is 0.9256595, this shows a very strong positive relationship because it is quite close to one, suggesting that the training has improved the players squats.

* Would you consider the above scatter plot (i.e. for the pre and post squats) to contain outliers? If so, identify the relevant outliers.

I believe there are a couple of outliers, for example the one where the players pre squat was 130 and he improved all the way up to 170 which is abnormal given the data set.

* What does the smoother suggest regarding the suitability of a simple linear regression model in this context?

The smoother suggests that there is wide range of possible values in the lower ranges and the upper ranges but in the middle it gets fairly tight suggesting that there is little spread between the values

### Fit the model

Assume that a model of the form (for i = 1, …, 50), together with the usual assumptions is appropriate in this context. Using the provided output from a regression analysis carried out on these data answer the following questions:

### Fit a regression model (standard output)

model <- lm(Squat\_Post~Squat\_Pre,data=train.df)  
model

##   
## Call:  
## lm(formula = Squat\_Post ~ Squat\_Pre, data = train.df)  
##   
## Coefficients:  
## (Intercept) Squat\_Pre   
## 22.476 1.052

According to the output given, the regression model can be described as .

Change the analysis done for VO2.max to Squat ans see if there is any relationship between player’s pre and post squats and answer the following questions. Answer the following questions:

* Write down the equation of line of best fit.

The equation: y = 1.052x + 22.476

* Provide an interpretation of the slope and intercept of the corresponding line of best fit.

The slope of the scatterplot is 0.9505703422, This suggests that there is a strong positive linear correlation because of its closeness to 1.

* Predict the post Squat of a player whose pre Squat is 110.

The answer is 138.196 by using the equation from the previous parts

* Why is the prediction in the previous part is different from 140, i.e. the Post Squat for player 5.

The line of best fit is only an approximation of the value, not a definite value, there can be variation from the graph and the calculated value.

train.df[5,]

## X ID VO2.max\_Pre VO2.max\_Post Squat\_Pre Squat\_Post Improvement  
## 5 5 5 76.7 84.5 110.55 139.86 29.31

* Predict the post Squat of a player whose pre Squat is 220. Explain if you have any concern related to this prediction?  
  Answer is 253.92 My concern with this prediction is that the value of 220 for a pre squat max is a very large value compared to every other player, this would make this value a huge outlier in our data set, if this value were in our data set it would alter alot of our calculations and skew our conclusions.

### Overall Conclusion??

The overall conclusion after studying this data set is that we can accept our original null hypothesis and say that yes the players who underwent this training did indeed improve their squat max, Again my only issue is the size of the data set is relatively small and also we don’t know how the players were sampled, if they were all from the same club for instance we couldn’t really be certain our data applies to the whole population.