Training Intervention Analysis

Protim Ganguly

28 October 2018

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 tolerance interval for teh likely improvement for 95% of players in the populaiton of interest (with 95% confidence).

```
library(infer)
library(tidyverse)

## -- Attaching packages ----------------------------------------------------
---------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts -------------------------------------------------------------
---------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tolerance)
```

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_intervention_data.csv")
glimpse(train.df)

## Observations: 18
## Variables: 5
## $ 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    <int> 120, 120, 130, 130, 110, 130, 140, 120, 140, 100,...
## $ Squat_Post   <int> 140, 150, 160, 160, 140, 160, 170, 140, 170, 130,...
```

Focus on the VO2 max response variables.

Summary Statistics

```
train.df %>% select(Squat_Pre,Squat_Post) %>% summary()
```

```
##    Squat_Pre     Squat_Post   
##  Min.   :100   Min.   :130.0  
##  1st Qu.:120   1st Qu.:142.5  
##  Median :130   Median :160.0  
##  Mean   :130   Mean   :159.4  
##  3rd Qu.:140   3rd Qu.:170.0  
##  Max.   :160   Max.   :190.0
```

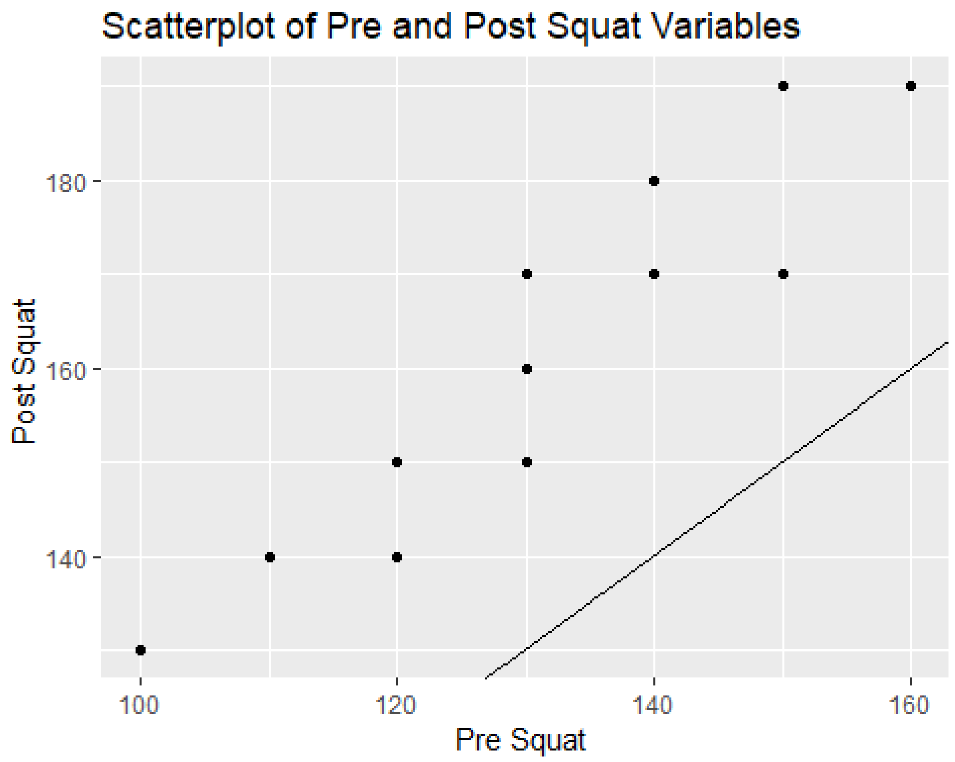
Mean and Standard Deviation

```
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 16.44957  159.4444 18.62074
```

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 Variables") + 
  ylab("Post Squat ") + 
  xlab("Pre Squat ") + 
  geom_abline(slope=1, intercept=0)
```



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: 6
## $ 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    <int> 120, 120, 130, 130, 110, 130, 140, 120, 140, 100,...
## $ Squat_Post   <int> 140, 150, 160, 160, 140, 160, 170, 140, 170, 130,...
## $ Improvement  <int> 20, 30, 30, 30, 30, 30, 30, 20, 30, 30, 40, 40, 2...
```

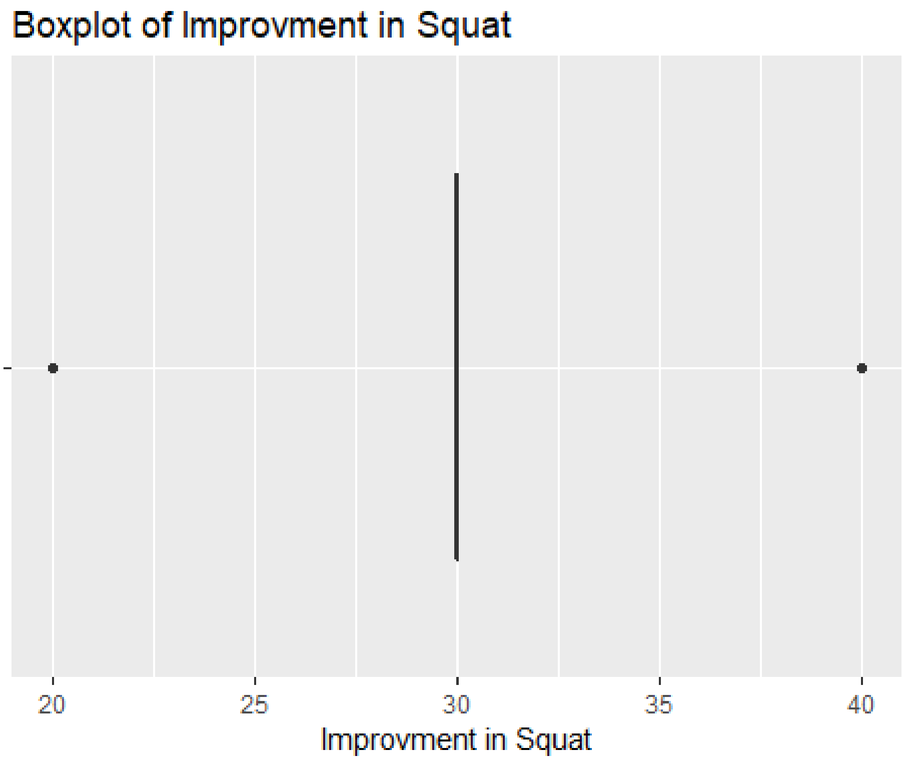
Mean and Standard Deviation of Improvement

```
train.df %>% select(Improvement) %>%
           summarize(Imp_Mean=mean(Improvement), Imp_SD= sd(Improvement))

##   Imp_Mean   Imp_SD
## 1 29.44444 6.391375
```

Boxplot of Improvement

```
train.df %>% ggplot(aes(x = "", y = Improvement)) +
        geom_boxplot() +
  ggtitle("Boxplot of Improvment in Squat") +
  ylab("Improvment in Squat") +
  xlab("") +
  coord_flip()
```



95% Confidence Interval

Using the t.test function

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

## 
##  One Sample t-test
## 
## data:  .
## t = 19.545, df = 17, p-value = 4.356e-13
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  26.26609 32.62280
## sample estimates:
## mean of x 
##  29.44444
```

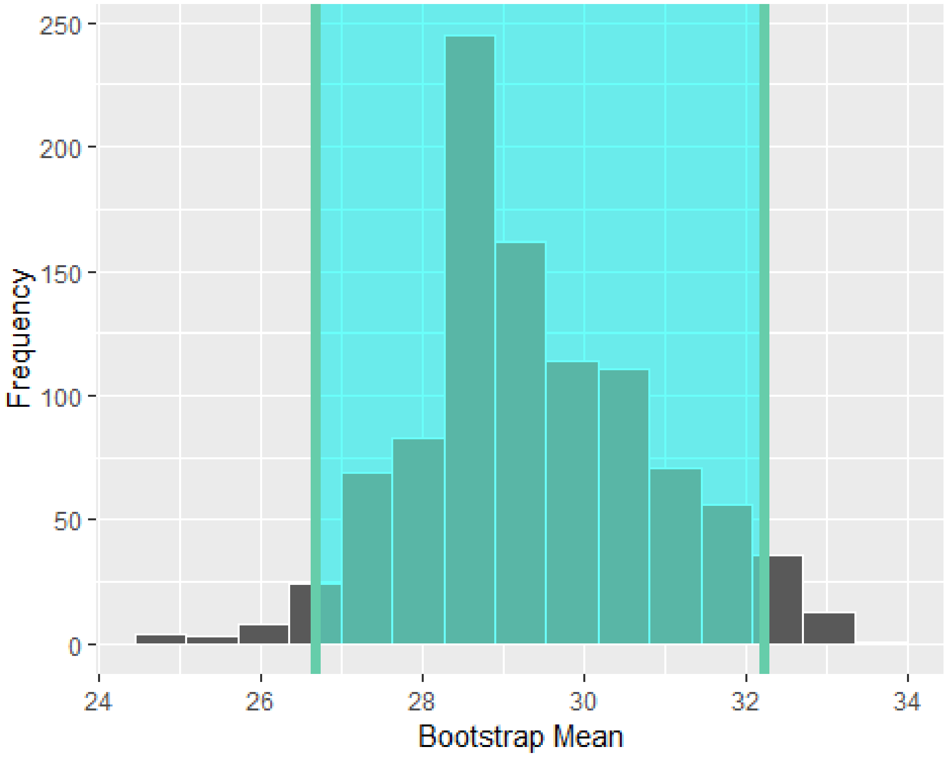
95% Bootstrap CI for the mean

```
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   26.7    32.2
```

```
boot %>% visualize(endpoints = percentile_ci, direction = "between") +
                   xlab("Bootstrap Mean") + ylab("Frequency")
```



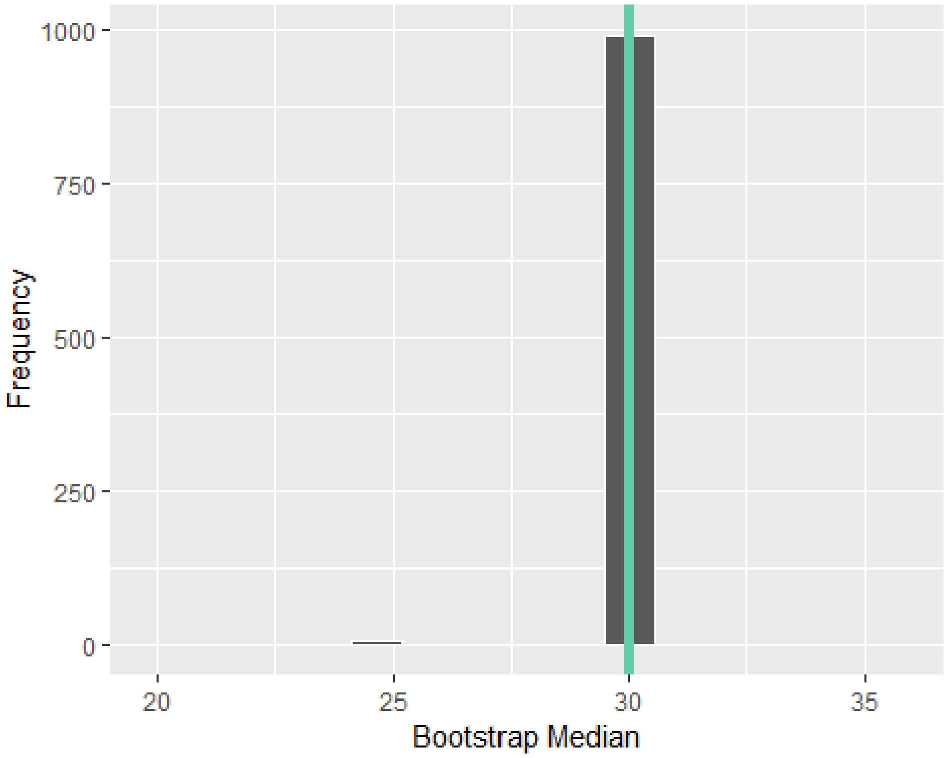
95% Bootstrap CI for the medan

```
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     30      30
```

```
boot.median %>% visualize(endpoints = percentile_ci_median, direction = "betw
een") +
                   xlab("Bootstrap Median") + ylab("Frequency")
```



95% Tolerance Interval

```
normtol.int(train.df$Improvement, alpha = 0.05, P = 0.95)
```

```
##   alpha    P    x.bar 1-sided.lower 1-sided.upper
## 1  0.05 0.95 29.44444      13.76674      45.12215
```

Conclusion -

From the initial analysis of the sample of 18 players we can see that there has been an improvement in average squats after they had gone through aerobic interval training. Previously the mean of squats was 130 and after the training it was 159.44. From the line of equality we see mostly post squat variables are above the line which also proves the average improvement in Squats.

However for our analysis its better to use a paired design of squat improvement over each individual player which is good since each person acts their own controller. Since all squat improvements are positive of each individual we can say that the training has indeed proved to be better on an individual level as well. If we compute the box plot of the improvements of all individual players we see that summary of statistics where the mean and upper quartile and lower quartile all have same value of 29.44.

Now we want to infer from the sample that there will be an improvement in average over the whole population if the training is provided for squat variables. Proceeding as outlined above, we would calculate 95% confidence limits in make an inference about the likely value of the population mean for squat improvement. The basis of our 95% confidence in the results from any one sample is that we are making an estimate of the population mean. The Central Limit Theorem states that if we take a large enough sample then the sample mean will vary from the population mean +/- the standard error. We compute standard error using Standard Deviation of the population however since we will never know that from beforehand we assume the SD of the sample taken. However for this assumption we need to modify our multiplier to Standard Deviation of sample if the sample size is less than 30 which we do in R using t.test. Now since our sample size is 18 which is less than 30 R does the job for us. So here we see that R gives us that we are 95% confident that the population mean of Improvement for Squat variables will lie between 26.26 and 32.62 for 1 sided t Test.

Another way for estimating population mean is bootstrapping. Here a random sample is taken 1000 times from the sample of 18 players with replacement and the mean of Improvement in squat variables is computed for each sample and then averaging out all the sample means give us an estimate of improvement over the population. get_ci returns upper and lower 95 percent confidence levels which basically gives that the population mean of Improvement for Squat variables will lie between 26.67 and 32.22 for bootstrapping. Also since many samples are taken from one sample it mostly assumes a normal distribution model as we can see in the graph. While estimating the midpoint (median) of Improvement in Squat Variables over the whole population we see that the upper and lower 95 percent confidence levels is of the same value 30 which shows that bootstrapping gives us a point estimate for the statistic(median) of the population.

Statistical tolerance limits are limits within which we expect a stated proportion of the population to lie. Here we are finding out the tolerance interval of 95% population for a 1 sided tolerance interval. Alpha is the difference between 100% and the confidence level 95% which is 0.05. We see that it returns the sample mean which is 29.44 and the lower tolerance bound 13.76 and upper tolerance bound of 45.12