# IMT 573: Problem Set 6

## Statistical Learning

Ali Qazi

Due: July 31 2022

**Collaborators:** Akeel Qazi, Anthony Mercado

**Instructions:** Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Download the `06_ps_statlearn.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `06_ps_statlearn.Rmd`.
2. Replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps6_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

**Setup:** In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(stringr)
library(gridExtra)
```

### Problem 1: Are sons taller than fathers?

Here we analyze the dataset of fathers' and sons' height, used by Pearson and which we saw in the last problem set. It contains two variables, fathers' height and sons' height. If you take a simple mean, you see that in average sons are taller than fathers. But can this difference just be due to chance? Let's find out.

```
father_son_data <- read.csv("fatherson.csv")
str(father_son_data)
```

**(a) To begin load the `fatherson.csv.bz2` data. Create density plots of both heights on the same figure. Comment the plots. What do they look like? What do they suggest in terms of fathers' and sons' relative height?**

```
## 'data.frame':    1078 obs. of  1 variable:
##  $ fheight.sheight: chr  "165.2\t151.8" "160.7\t160.6" "165\t160.9" "167\t159.5" ...
```

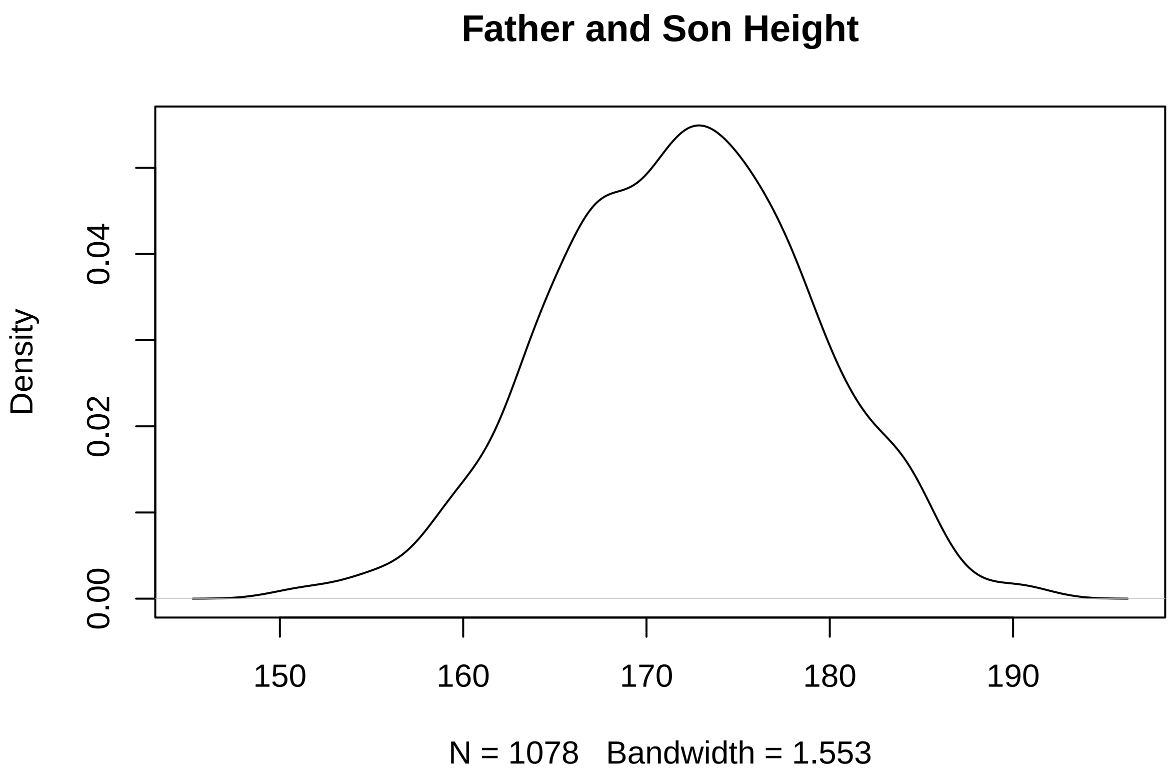
```
fs_data <- str_split_fixed(father_son_data$fheight.sheight, "\t", 2)
fsdata.df <- as.data.frame(fs_data)
str(fsdata.df)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ V1: chr  "165.2" "160.7" "165" "167" ...
##  $ V2: chr  "151.8" "160.6" "160.9" "159.5" ...
```

```
fsdata.df <- as.data.frame(lapply(fsdata.df, as.numeric))
str(fsdata.df)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ V1: num  165 161 165 167 155 ...
##  $ V2: num  152 161 161 160 163 ...
```

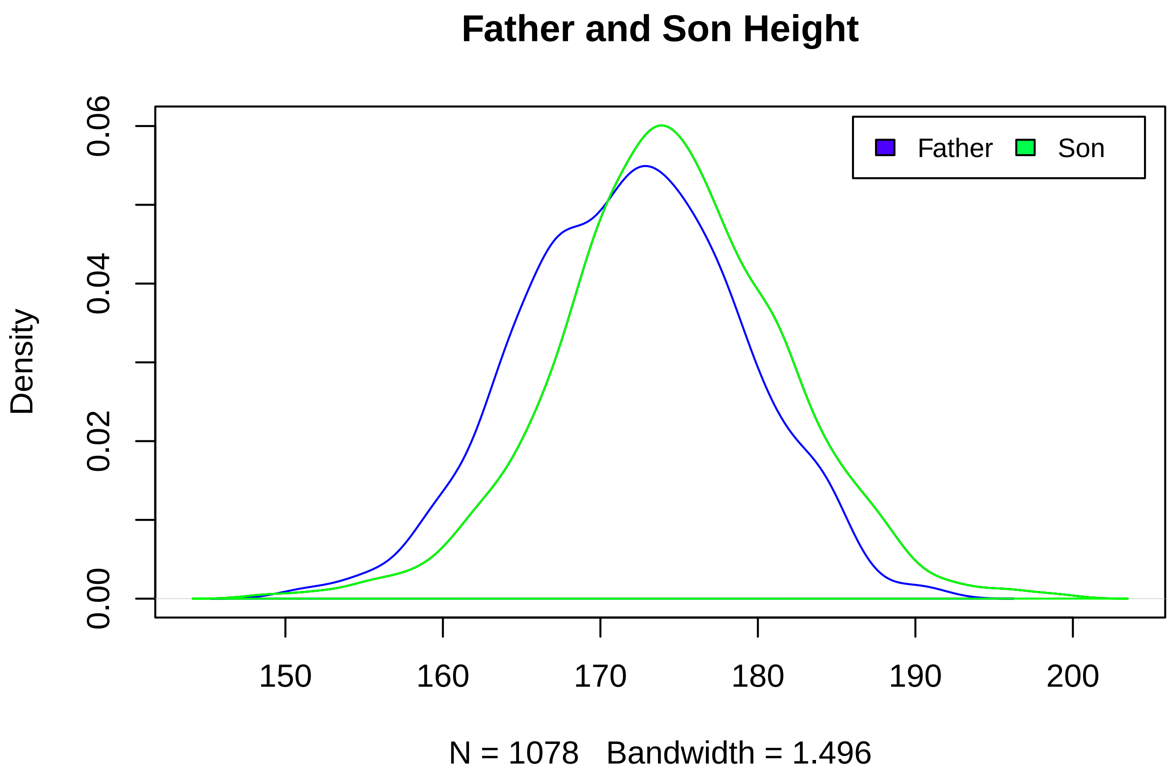
```
d <- density(fsdata.df$V1) # returns father density data
p <- density(fsdata.df$V2) # returns son density data
plot(d, main="Father and Son Height") # plot data
```



```
plot(p, main="Father and Son Height")

polygon(d, col = "transparent", border = "blue")
polygon(p, col = "transparent", border = "green")

legend("topright", inset=.02,
   c("Father","Son"), fill=topo.colors(3), horiz=TRUE, cex=0.8)
```



In this density we see that father, blue line, is smaller than than the son, red line.

```
sonmean <- mean(fsdata.df$V2)
fathermean <- mean(fsdata.df$V1)
standdev <- sd(fs_data2$`sapply(fs_data, as.numeric)`)
```

**(b) But is this difference statistically significant? Let's do a $t$-test. Here I ask you to *compute yourself the $t$-value*, do not use any pre-existing functions! What do you find? Why did you use/did you not use pooled standard deviations? Explain!**

```
## Error in is.data.frame(x): object 'fs_data2' not found
```

```
set <- length(fs_data2$`sapply(fs_data, as.numeric)`)
```

```
## Error in eval(expr, envir, enclos): object 'fs_data2' not found
```

```
tvalue <- (sonmean - fathermean) / (standdev/sqrt(set))
```

```
## Error in eval(expr, envir, enclos): object 'standdev' not found
```

```
tvalue
```

```
## Error in eval(expr, envir, enclos): object 'tvalue' not found
```

```
t.test(fsdata.df$V2, fsdata.df$V1)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  fsdata.df$V2 and fsdata.df$V1
## t = 8.3239, df = 2152.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.935475 3.128532
## sample estimates:
## mean of x mean of y
##  174.4572  171.9252
```

```
t.test(fsdata.df$V2 - fsdata.df$V1)
```

```
##
##  One Sample t-test
##
## data:  fsdata.df$V2 - fsdata.df$V1
## t = 11.783, df = 1077, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.110352 2.953656
## sample estimates:
## mean of x
##  2.532004
```

Two sample mean tests with the assumption that the standard deviation are not equal, tells us that pooled standard deviations are not used.

Hint: read OIS 7.3

```

x <- seq(-150, 200, by = 12.5)
dt(x, df = 5) # 5 degrees of freedom
```

**(c) Look up the $t$-distribution table. (Or compute the relevant quantiles). What is the likelihood that such a $t$ value happens just by random chance? Hint: be sure to consider the *degrees of freedom* in current case carefully!**

```
##  [1] 4.163004e-12 7.015896e-12 1.242702e-11 2.337836e-11 4.737973e-11
##  [6] 1.055225e-10 2.659002e-10 7.930437e-10 3.018705e-09 1.688232e-08
## [11] 1.897677e-07 1.131735e-05 3.796067e-01 1.131735e-05 1.897677e-07
## [16] 1.688232e-08 3.018705e-09 7.930437e-10 2.659002e-10 1.055225e-10
## [21] 4.737973e-11 2.337836e-11 1.242702e-11 7.015896e-12 4.163004e-12
## [26] 2.575592e-12 1.651212e-12 1.091568e-12 7.411414e-13
```

The likelihood that such a t values happens just by random chance is low. T values are a type of test statistic. The t value is 16.39008 which is similar to the manual t value that I got above.

**(d) Based on your above analysis, state clearly your conclusion to the question - are sons taller than fathers?** The conclusion through t test is that sons are taller than fathers on average. Height of sons are 2.45 inches away from the predicted value by the model on average.

### Problem 2: Fathers and Sons - the Monte Carlo approach

Next, let's re-visit the fathers and sons height, but this time by doing Monte Carlo analysis on a computer. You will proceed as follows: create two samples of random normals, similar to the data above, using the mean and standard deviation over both fathers and sons. Call one of these samples `fathers'' and the other`sons’ ’. What is the difference in their means? And now you repeat this exercise many times and see if you can get as big a difference as what you saw above in the data.

```
# overall mean and standard deviation of combined father and sons heights
fs_data2 %>%
  summarize(mean(`sapply(fs_data, as.numeric)`), sd(`sapply(fs_data, as.numeric)`))
```

**(a) First, compute the overall mean and standard deviation of combined fathers’ and sons’ heights. Now create two sets of normal random variables, both with the same mean and standard deviation that you just computed above. Call one of these `fathers` and the other `sons`. What is the father-son mean difference? Compare the result with that you found in the previous problem.**

```
## Error in summarize(., mean(`sapply(fs_data, as.numeric)`), sd(`sapply(fs_data, as.numeric)`)): objec
```

```
# normal random variables
fathers <- rnorm(y, mean = 173, sd = 7)
```

```
## Error in rnorm(y, mean = 173, sd = 7): object 'y' not found
```

```
sons <- rnorm(y, mean = 173, sd = 7)
```

```
## Error in rnorm(y, mean = 173, sd = 7): object 'y' not found
```

```
meandiff <- sonmean - fathermean
meandiff
```

```
## [1] 2.532004
```

The mean difference is 2.543 which is close to the height of sons that are 2.45 inches away from the predicted value by the model on average.

```
random.numbers = sample(x = meandiff , size = 1000, replace = TRUE)
rtimes <- random.numbers
mean(rtimes) # true difference in means is not equal to 0
```

**(b) Now repeat the previous question a large number of times $R$ (1000 or more). Each time store the difference, so you end up with $R$ different values for the difference. What is the mean of the difference values? Explain what do you get. What is it standard deviation? Compare it to that you computed in the previous problem for the difference in data (when doing $t$-test). What is the largest difference (in absolute value)?**

```
## [1] 1.499
```

```
sd(rtimes)
```

```
## [1] 0.5002492
```

```
max(rtimes)
```

```
## [1] 2
```

**(c) Find the 95% quantile of (the absolute value) your difference. Compare this number to the actual father-son difference you found in the data.** Hint: use the R function `quantile` for this.

```
quantile(rtimes, 0.95)
```

```
## 95%
##   2
```