Homework 11

Miles Tweed

12/3/2020

Problem 1

a) The population being inferred about is the collection of all New York AirBnB's. For this test, $H_0 : \mu = \$150$ and $H_a : \mu \neq \$150$. Using a significance level of 0.05, the p-value suggests that we should fail to reject H_0 and the confidence interval agrees with this conclusion because in includes the value \$150. The small Cohen's d score suggests that the effect size is minimal («0.2) which would be expected when we fail to reject to null hypothesis.

```
airbnb <- read.csv(file.path(PROJ,DATA,'listings.csv'))
test <- t.test(airbnb$price, mu=150, alternative = 'two.sided')
print(test)
```

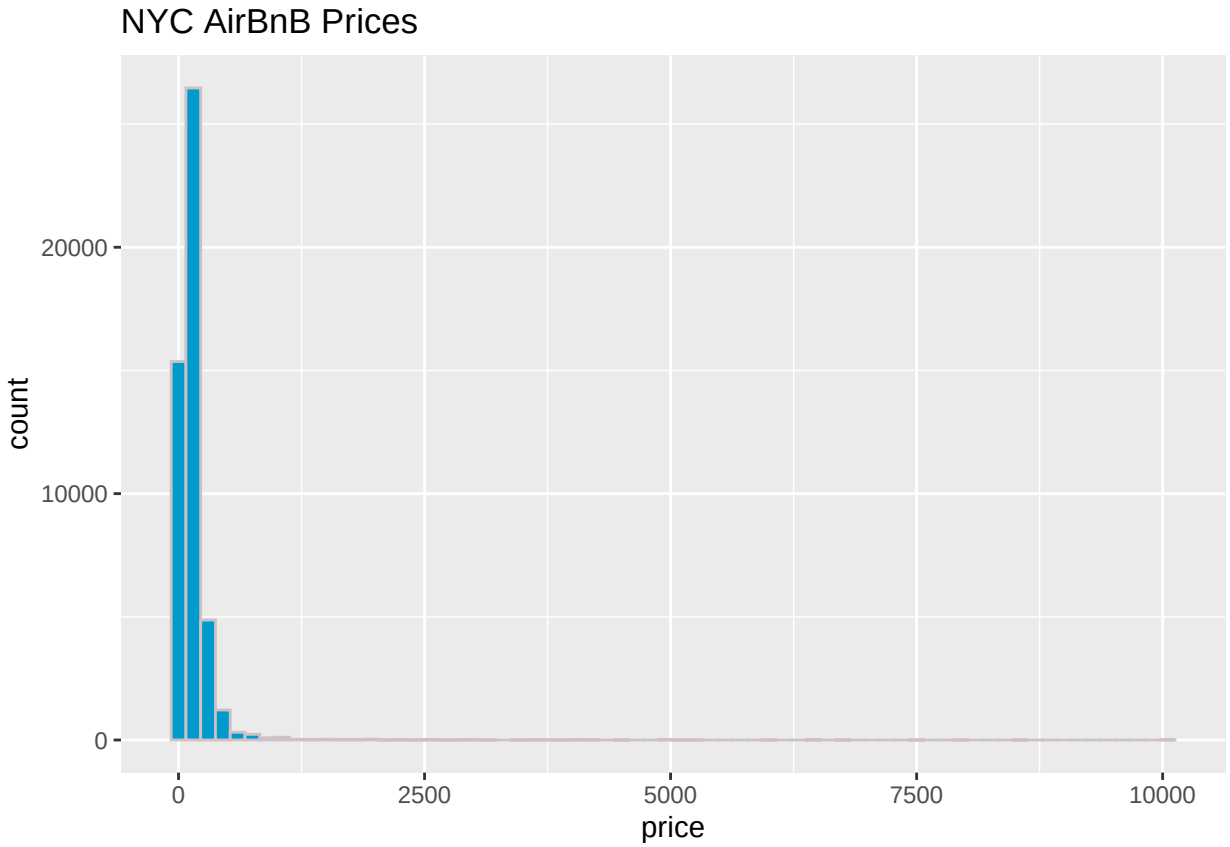
```
##
##  One Sample t-test
##
## data:  airbnb$price
## t = 1.3578, df = 48863, p-value = 0.1745
## alternative hypothesis: true mean is not equal to 150
## 95 percent confidence interval:
##  149.3554 153.5509
## sample estimates:
## mean of x
##  151.4532
```

```
mu <- 150
x.bar <- mean(airbnb$price)
s <- sd(airbnb$price)
d = abs(x.bar-mu)/s
print(paste("The Cohen's d score is", round(d,3)))
```

```
## [1] "The Cohen's d score is 0.006"
```

b) The normality assumption is not satisfied and the distribution is highly right skewed. This is because there are some extreme outliers on the upper end. We should be concerned about the legitimacy of the result because of the effect of these outliers on sample statistics such as sample mean and sample standard deviation.

```
airbnb %>% ggplot(aes(x = price, color = I('lavenderblush3'), fill = I('deepskyblue3'))) +
  geom_histogram(binwidth = 150) + labs(title = 'NYC AirBnB Prices')
```




```
fn <- fivenum(airbnb$price)
fQ <- fn[2]
tQ <- fn[4]

out.low <- fQ - (1.5 * IQR(airbnb$price))
out.high <- tQ + (1.5 * IQR(airbnb$price))

# Number of upper end outliers
airbnb %>% filter(price > out.high) %>% select(price) %>% count()
```

```
##      n
## 1 2873
```

```
# Number of lower end outliers
airbnb %>% filter(price < out.low) %>% select(price) %>% count()
```

```
##   n
## 1 0
```

c) The percentage of times that the p-value was less than 0.05 was more than three times what I would expect. I would only expect this to happen 5% of the time if the results of the initial t-test were correct and the correct decision was to fail to reject the null hypothesis.

```
our_pop <- airbnb$price
mu_pop <- mean(our_pop)

n.sim <- 10000
p.values <- numeric(n.sim)
```

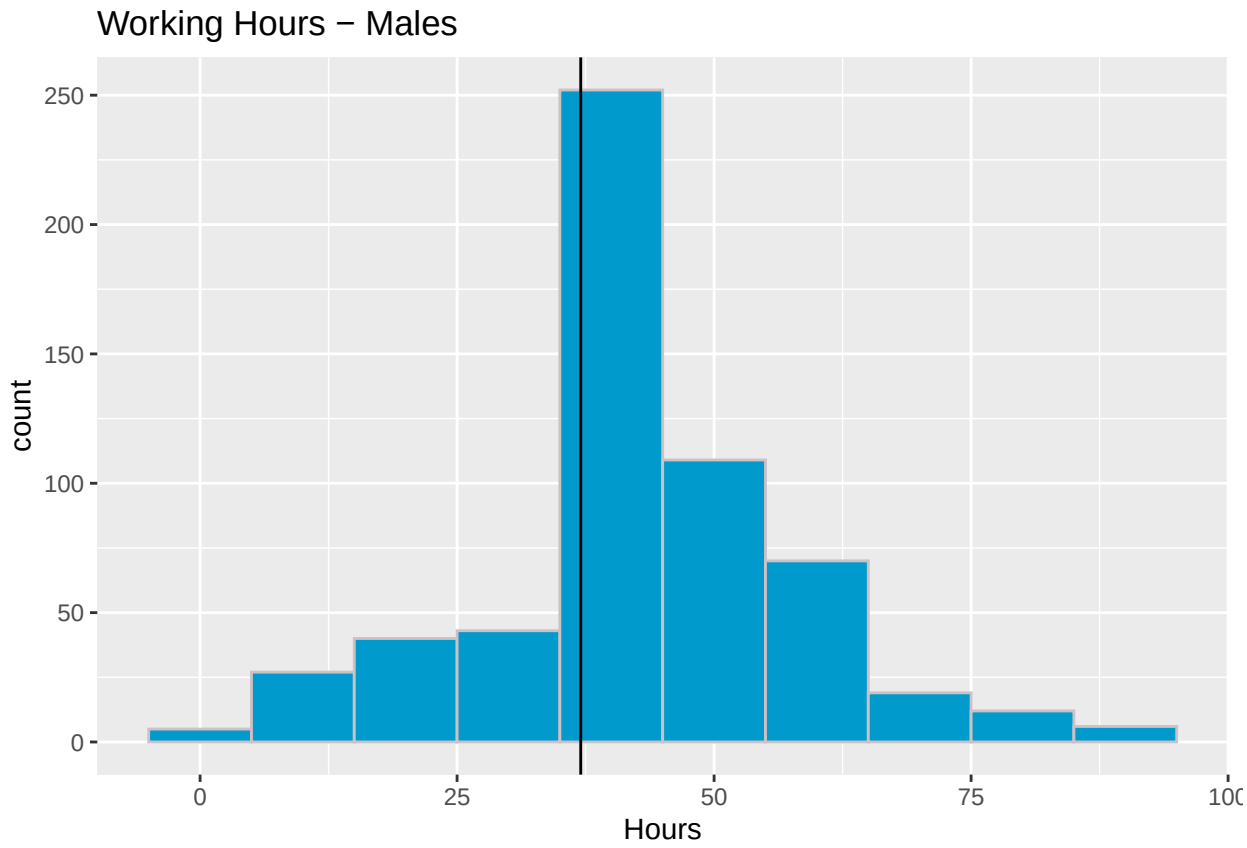
```r
samp.size <- 30

for (j in 1:n.sim){
  x <- sample(our_pop, samp.size)
  p.values[j] <- t.test(x, mu=mu_pop)$p.value
}

mean(p.values < 0.05)
```

```
## [1] 0.1642
```

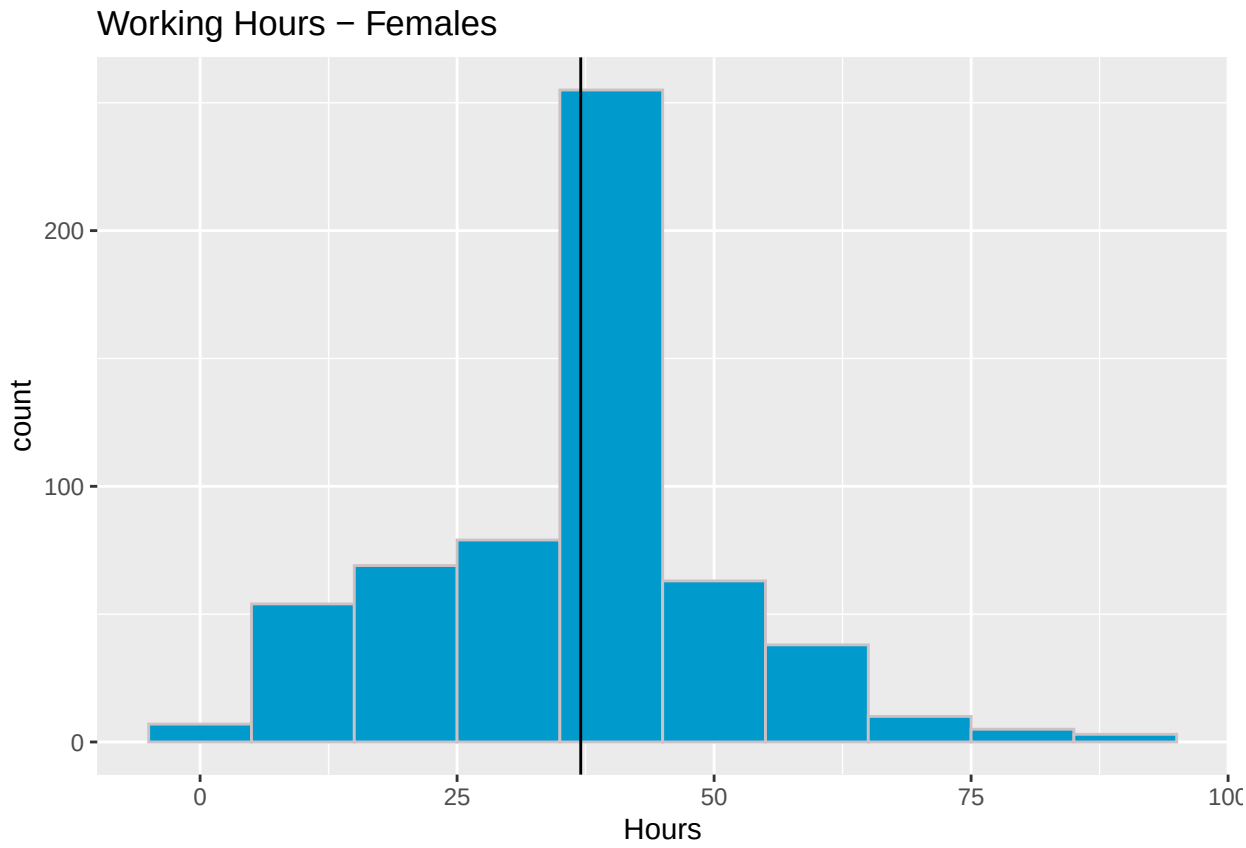
Problem 2

a) The normality assumptions seem to be satisfied with both the male and female samples. Both distributions have several outliers according to the $1.5 * IQR$ condition. However, since both distribution are similar, they should be comparable. Additionally, the sample size is adequately large so that the central limit theorem could apply.

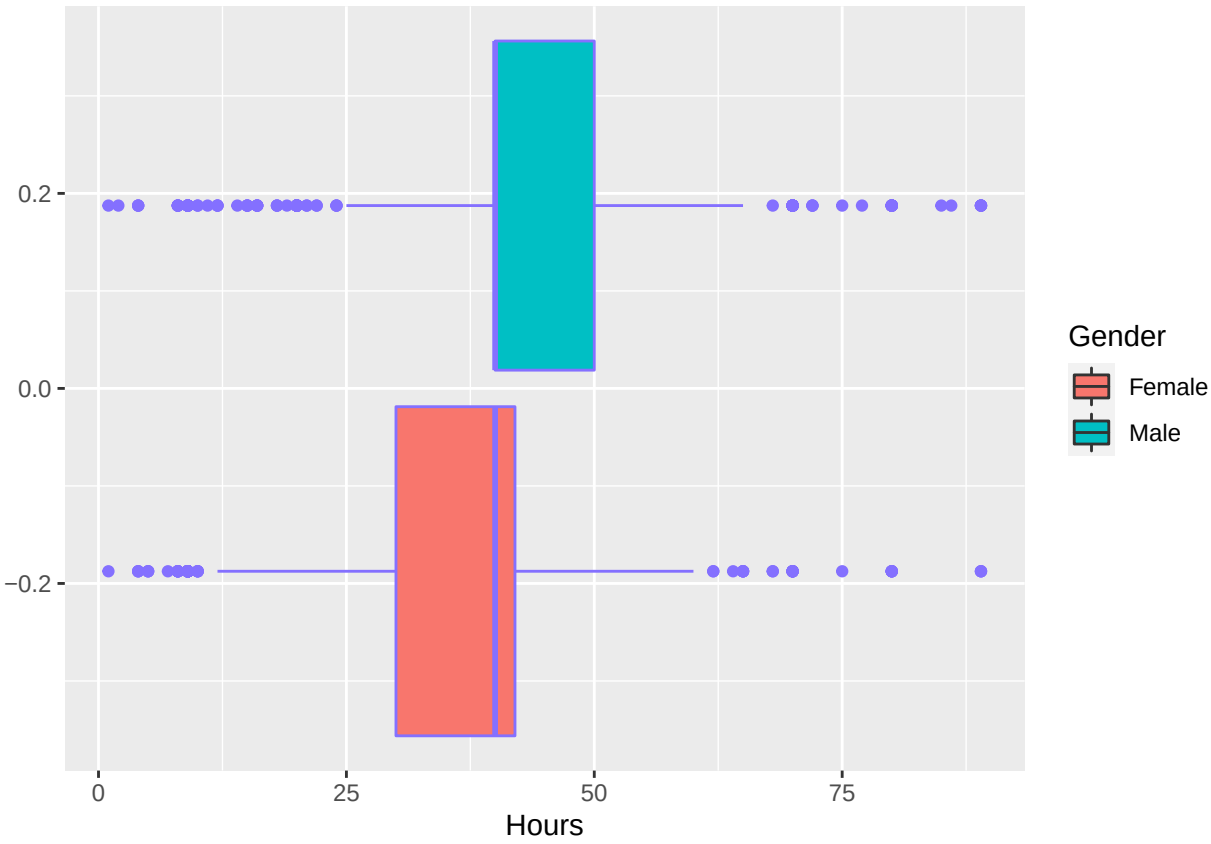
```r
# DATA
work <- read.csv(file.path(PROJ,DATA,'workweek2012.csv'))

# Histogram of working hours for males
work.mmean <-
  work %>% filter(Gender=='Female') %>%
  select(Hours) %>%
  summarise(Avg = mean(Hours))

work %>% filter(Gender == 'Male') %>%  ggplot(aes(x = Hours, color = I('lavenderblush3'),
                                                  fill = I('deepskyblue3'))) +
  geom_histogram(binwidth = 10) + labs(title = 'Working Hours - Males') +
  geom_vline(xintercept = work.mmean$Avg)
```




```
# Histogram of working hours for females
work.fmean <-
  work %>% filter(Gender=='Female') %>%
  select(Hours) %>%
  summarise(Avg = mean(Hours))

work %>% filter(Gender == 'Female') %>%  ggplot() +
  geom_histogram(aes(x = Hours, color = I('lavenderblush3'), fill = I('deepskyblue3')), binwidth = 10) +
  labs(title = 'Working Hours - Females') +
  geom_vline(xintercept = work.fmean$Avg)
```




```
work %>%
  ggplot() + geom_boxplot(aes(x=Hours, fill = Gender, color = I('lightslateblue')))
```




```
# outliers male
male.hours <- work %>% filter(Gender == 'Male') %>% select(Hours)
fn <- fivenum(male.hours$Hours)
fQ <- fn[2]
tQ <- fn[4]

out.low <- fQ - (1.5 * IQR(male.hours$Hours))
out.high <- tQ + (1.5 * IQR(male.hours$Hours))

# Number of upper end outliers (male)
male.hours %>% filter(Hours > out.high) %>% select(Hours) %>% count()
```

```
##    n
## 1 37
```

```
#Number of lower end outliers (male)
male.hours %>% filter(Hours < out.low) %>% select(Hours) %>% count()
```

```
##    n
## 1 67
```

```
# outliers female
female.hours <- work %>% filter(Gender == 'Female') %>% select(Hours)
fn <- fivenum(female.hours$Hours)
fQ <- fn[2]
tQ <- fn[4]

out.low <- fQ - (1.5 * IQR(female.hours$Hours))
out.high <- tQ + (1.5 * IQR(female.hours$Hours))
```

```r
# Number of upper end outliers (female)
female.hours %>% filter(Hours > out.high) %>% select(Hours) %>% count()
```

```
##    n
## 1 26
```

```r
#Number of lower end outliers (female)
female.hours %>% filter(Hours < out.low) %>% select(Hours) %>% count()
```

```
##    n
## 1 48
```

b) For this two sample t-test I will set $\bar{x}_1$ as the mean working hours for females and $\bar{x}_2$ as the mean working hours for males. Therefore, my null hypothesis is $H_0 : (\mu_1 - \mu_2) = 0$ or $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis will be $H_a : (\bar{x}_1 - \bar{x}_2) \neq 0$ or $H_0 : \mu_1 \neq \mu_2$. The p-value suggests that the difference is statistically significant ($\ll 0.0001$) and we should reject the null hypothesis. The result suggests that the average hours worked by females if less than that of males since the t-value and both ends of the confidence interval are negative. In other words, we could be 95% confident that a random sample of females work, on average, as much as 8.2 hours or as little as 4.7 hours less than males. The Cohen's d score suggests that the effect size is small bordering on medium. As stated earlier, the confidence interval agrees with this conclusion since it does not include zero and is entirely negative.

```r
t.test(female.hours$Hours, male.hours$Hours, alternative = 'two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  female.hours$Hours and male.hours$Hours
## t = -7.2908, df = 1163.6, p-value = 5.68e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.239410 -4.745153
## sample estimates:
## mean of x mean of y
##  37.02744  43.51973
```

```r
x.bar.1 <- mean(female.hours$Hours)
x.bar.2 <- mean(male.hours$Hours)
s1 <- sd(female.hours$Hours)
s2 <- sd(male.hours$Hours)
n1 <- length(female.hours$Hours)
n2 <- length(male.hours$Hours)
x.bar.diff <- (x.bar.1 - x.bar.2)
s.pooled <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2))
d = abs(x.bar-mu)/s.pooled
print(paste("The Cohen's d score is", round(d,3)))
```

```
## [1] "The Cohen's d score is 0.096"
```

Problem 3

9.56 This is misleading because, using a 0.05 statistical significance level, we would expect to reject the null hypothesis in 5% of the trials if the null hypothesis was, in fact, true. Three trials out of sixty is exactly 5% of the trials.

10.14

a) The groups are college students that consume alcohol with energy drinks added and those that consume alcohol without energy drinks added. The population means are the mean blood alcohol content (BAC) for those two categories post consumption over all college students who drink alcohol. The null hypothesis, if BAC post consumption of alcohol with energy drinks is μ_1 and BAC post consumption of alcohol without energy drinks is μ_2, is $H_0 : (\mu_1 - \mu 2 = 0)$ or $H_0 : \mu_1 = \mu 2$.

b) The confidence interval provides information about the direction of the relationship between the two variables. If the confidence interval is negative, it can be inferred that the group mean described by μ_1 is, on average, less than the group mean describe by μ_2 and is it is positive the opposite can be inferred. Also, it can provide a rang of values that the two mean will likely differ by (with 95% confidence).

10.24

a) If the population means were equal, the difference between the means would be within the range of the standard deviation. Because of this, the calculation of the t score would not deviate much higher than one.

```
x.bar.1 <- 2.9
x.bar.2 <- 0.1
n1 <- 237
n2 <- 95
s1 <- 3.6
s2 <- 0.5
se <- sqrt(s1^2/n1 + s2^2/n2)
t <- (x.bar.1-x.bar.2)/se

print(paste('The t statistic is: ',round(t,2)))
```

```
## [1] "The t statistic is:  11.7"
```

b) Since the p-value for a test statistic this large would be much less than 0.001, we could infer that we should reject the null hypothesis. The test statistic is positive which would imply that the mean score of those that inhaled was larger. This could be confirmed by looking at the confidence interval.

c) It is assumed that the samples were generated using randomization and that the distributions of the samples are roughly normal centered around the mean. Since the sample is adequately large it can be assumed that the central limit theorem applies in this case.

10.49

a) The before and after variables are dependant because they are a measure of the same individual over time. In other words, the success of the treatment is always relative to the status of the individual before the trial.

b) The values are related in that the difference of the before mean and the after mean is identical to the mean of the differences.

```
x.bar.before <- mean(c(150,165, 135))
print(paste('Mean before score:',x.bar.before))
```

```
## [1] "Mean before score: 150"
```

```
x.bar.after <- mean(c(130,140,120))
print(paste('Mean after score:',x.bar.after))
```

```
## [1] "Mean after score: 130"
```

```
x.bar.diff <- mean(c((150-130),(165-140),(135-120)))
print(paste('Mean score difference:',x.bar.diff))
```

```
## [1] "Mean score difference: 20"
```

```r
diff.mean <- x.bar.before - x.bar.after
print(paste('d = before - after:',diff.mean))
```

```
## [1] "d = before - after: 20"
```

c) The confidence intervals are overlapping which indicates that the difference is likely to not be statistically significant.

```r
sd.before <- sd(c(150,165, 135))
print(paste('Before standard deviation:',sd.before))
```

```
## [1] "Before standard deviation: 15"
```

```r
ci95.low.before <- x.bar.before - 1.96*sd.before
ci95.high.before <- x.bar.before + 1.96*sd.before
print(paste('95% confidence interval before:',ci95.low.before, 'to', ci95.high.before))
```

```
## [1] "95% confidence interval before: 120.6 to 179.4"
```

```r
sd.after <- sd(c(130,140,120))
print(paste('After standard deviation',sd.after))
```

```
## [1] "After standard deviation 10"
```

```r
ci95.low.after <- x.bar.after - 1.96*sd.after
ci95.high.after <- x.bar.after + 1.96*sd.after
print(paste('95% confidence interval after:',ci95.low.after, 'to', ci95.high.after))
```

```
## [1] "95% confidence interval after: 110.4 to 149.6"
```

```r
t.test
```

```
## function (x, ...)
## UseMethod("t.test")
## <bytecode: 0x55c9d69e8fc8>
## <environment: namespace:stats>
```

10.58

a)

```r
before.raw <- c(5.08,5.99, 5.32, 6.03, 5.44)
after.raw <- c(5.36, 5.98, 5.62, 6.26, 5.68)

check <- data.frame(before.raw,after.raw)
names(check) <- c('before', 'after')

check['diff'] <- check$after - check$before

print(paste('Average improvement:',mean(check$diff)))
```

```
## [1] "Average improvement: 0.208"
```

b) Since the p-value is below 0.05, the result is statistically significant given a 95% confidence level. This is confirmed by the fact that the 95% confidence interval does not contain 0. The confidence interval is negative with the before reading as $\bar{x}_1$ which indicates that we can be 95% confident that the before reading will be between 0.05 and 0.36 less than the after reading.

```r
t.test(check$before, check$after, paired = TRUE, alternative = 'two.sided')
```

```
##
```

```
##  Paired t-test
##
## data:  check$before and check$after
## t = -3.7155, df = 4, p-value = 0.02056
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.36343134 -0.05256866
## sample estimates:
## mean of the differences
##                  -0.208
```

c) This completely changes the analysis. Instead of analysing the mean of the differences it is simply comparing the means. The statistical significance disappears and the confidence interval contains zero indicating that there is no difference.

```
t.test(check$before, check$after, alternative = 'two.sided')
```

```
##
##  Welch Two Sample t-test
##
## data:  check$before and check$after
## t = -0.85276, df = 7.7236, p-value = 0.4194
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7739891  0.3579891
## sample estimates:
## mean of x mean of y
##     5.572     5.780
```