Mohammad Javad Ranjbar

810101173

Homework 4

Statistical Inference, Fall 2022

Question #1:

a) False. The bootstrap distribution is created by resampling from an original sample with a replacement.
b) True. The data provide convincing evidence that at least one pair of population means differ (but we can't tell which one) using ANOVA. But the pairwise analysis will identify at least one pair of significantly different means.
c) False. The resamples should be the same sample size as the original sample for it to be a representative distribution.
d) True. For the paired analysis. First, we have to calculate the difference between each group's values. Therefore, we only have one group of variables and we use these differences to make inferences.
e) False. You can think of the within-group variance as the background noise that can obscure a difference between means.
f) False. We need to calculate the difference for each item then we start doing interference based on that.
g) True. For the paired analysis. First, we have to calculate the difference between each group's values. Therefore, we only have one group of variables and we use these differences to make inferences.
h) False. Type-ll error is equal to β.

Question #2:

a)

$$\bar{x} = \frac{76 + 84 + 69 + 92 + 58 + 89 + 73 + 97 + 85 + 77}{10} = \frac{800}{10} = 80$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

$$= \frac{(76-80)^2}{9} + \frac{(84-80)^2}{9} + \frac{(69-80)^2}{9} + \frac{(92-80)^2}{9} + \frac{(58-80)^2}{9}$$
$$+ \frac{(89-80)^2}{9} + \frac{(73-80)^2}{9} + \frac{(97-80)^2}{9} + \frac{(85-80)^2}{9} + \frac{(77-80)^2}{9}$$
$$= \frac{1234}{9} = 137.11 \Rightarrow s = 11.71$$

$$SE = \frac{s}{\sqrt{n}} = \frac{11.71}{\sqrt{10}} = 3.7$$

b) Margine of error $= t^* SE$
The students have been chosen at random and they are independent. Also, scores usually have a not extremely skewed distribution similar to the normal distribution. The number of students that have been chosen is less than 10% of all the students in that university but it is smaller than 30. Therefore, we use the Student's ***t***-Distribution for this question.
df = n-1 = 9
If we consider $\alpha = 0.1$ we have $t^* = 1.83$(`> qt(0.05,df=9)` `[1] -1.833113`)

$ME = 1.83^*3.9 = 6.77$

c) $\bar{x} - ME < \mu < \bar{x} + ME$

$$80 - 6.77 < \mu < 80 + 6.77 \Rightarrow 73.23 < \mu < 86.77$$

90% of random samples of 10 students will yield confidence intervals containing the true mean score of students.

Question #3:

a)

$$H_0: \mu = 8 \text{ hourse(NewYorkers sleep 8 hours a night on average)}$$

$$H_A: \mu < 8 \text{ hourse(NewYorkers sleep less than 8 hours a night on average)}$$

a)

Independence

- Sampled observations are 25 random independent New Yorkers.
- The sampling is without replacement, and 25 is less than 10% of the population of New York.

Sample size/skew:

- The amount of sleep that people get per day is a natural trait that is normally distributed. For this size sample, slight skew is acceptable, and the min/max suggests there is not much skew in the data.
- The size of the sample is small and not larger than 30 (n<30).

Therefore, we use Student's ***t***-Distribution.

b) $SE = \frac{s}{\sqrt{n}} = \frac{0.77}{5} = 0.154$

$$T = \frac{\text{Obeservation} - \text{Null}}{SE} = \frac{7.73 - 8}{0.154} = -1.75$$

$$df = n - 1 = 25 - 1 = 24$$

c) p-value=0.0464 (`> pt(-1.75,df=24)` `[1] 0.04644754`). If we consider α =0.05 we can conclude that $p - value < \alpha$. Therefore, the data provide enough evidence to reject the H_0 meaning that the data suggests that New Yorkers sleep less than 8 hours a night on average.

d) Based on the definition we know. P-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true (p-value $= P(\text{observed or more extreme outcome} \mid H_0 \text{ true})$).
In this context observing a sample with $\bar{x} = 7.73$ if the average hours of sleep for New Yorkers is 8 hours. (p-value $= P(\bar{x} = 7.73 \mid H_0 \text{ true})$)
Based on the calculated p-value and its comparison with α we understand that H_0 is rejected. Means that it would be very unlikely to for our sample to have $\bar{x} = 7.73$ (New Yorkers sleep 7.73 hours on average) if the null hypothesis were true, and hence reject H_0.

e) No, using p-value we concluded that H_0 has been rejected meaning that 8 will not be present in our CI.

For 90% confidence interval we have $t^* = 1.71$ (
```
> qt(0.95,df=24)
[1] 1.710882
```
).

$$\bar{x} - t^*SE < \mu < \bar{x} + t^*SE$$

$$7.73 - 1.71 * 0.154 < \mu < 7.73 + 1.71 * 0.154 \Rightarrow 7.4605 < \mu < 7.9995$$

8 is not in 90% confidence interval confirming the p-values findings. Therefore, the H_0 hypothesis will be rejected.

Question #4:

We need to calculate the difference between mean of two groups. Then, we start the interference process.

$$\mu_{intensive} = \mu_{Paced} \Rightarrow \mu_{intensive} - \mu_{Paced} = 0$$

Our hypothesizes would be as followed:

$$H_0: \mu_{intensive} - \mu_{Paced} = 0 (\text{There is no difference})$$

$$H_A: \mu_{intensive} - \mu_{Paced} \neq 0 (\text{There is difference})$$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{41.4736}{12} + \frac{56.5504}{10}} = 3.02$$

For calculating degree of freedom, we can use the approximate method as followed:

$$df = \min(n_1 - 1, n_2 - 2) = 9$$

We can calculate the exact value:

$$C = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{\frac{41.4736}{12}}{\frac{41.4736}{12} + \frac{56.5504}{10}} = 0.38$$

$$df = \frac{(n_1 - 1) * (n_2 - 2)}{(n_2 - 2) * C^2 + (1 - C)^2 * (n_1 - 1)} = \frac{9 * 11}{11 * (0.38)^2 + (1 - 0.38)^2 * 9} = 19.6043972258 \approx 20$$

$$T = \frac{\text{Observed} - \text{Null}}{SE} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{3.02} = \frac{46.31 - 42.79}{3.02} = 1.1655$$

$p - value = 0.25$ (
```
> 2*pt(1.165,df=20,lower.tail = FALSE)
[1] 0.2577267
```
)

if we consider $\alpha = 0.05$. We conclude that, $p - value > \alpha$ and we can not reject the null hypothesis.

Question #5:

effect size $= \delta = \bar{x} - \mu = 0.5, \beta = 0.2, \alpha = 0.05$
```
> qnorm(0.975)> qnorm(0.8)
[1] 1.959964  [1] 0.8416212
```
$, sd = 2.2, n = ?$

$$z = \frac{\text{Observation-null}}{SE} \Rightarrow SE = \frac{\delta}{z} = \frac{0.5}{(0.84 + 1.96)} = 0.1785714$$

$$n = \frac{s_1^2 + s_2^2}{SE} = \frac{s_1^2 + s_2^2}{SE} = \frac{4.84 + 4.84}{0.1785714} = 55$$

We need 55 new enrollees.

Question #6:

First we need to calculate the difference between each group.

Before	After	Difference
25	27	25-27=-2
25	29	-4
27	37	-10
44	56	-12
30	46	-16
67	82	-15
53	57	-4
52	61	-9
53	80	-27
60	59	1
28	43	-15

$$H_0: \mu_{Diff} = 0 (\text{There is no difference})$$

$$H_A: \mu_{Diff} \neq 0 (\text{There is difference})$$

$$\bar{x} = \frac{-2-4-10-12-16-15-4-9-27+1-15}{11} = -10.27$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(-2+10.27)^2}{10} + \frac{(-4+10.27)^2}{10} + \frac{(-10+10.27)^2}{10} + \frac{(-12+10.27)^2}{10}$$
$$+ \frac{(-16+10.27)^2}{10} + \frac{(-15+10.27)^2}{10} + \frac{(-4+10.27)^2}{10} + \frac{(-9+10.27)^2}{10}$$
$$+ \frac{(-27+10.27)^2}{10} + \frac{(1+10.27)^2}{10} + \frac{(-15+10.27)^2}{10} = \frac{636.1819}{10} = 63.62$$

$$s = \sqrt{63.62} = 7.97$$

$$df = n - 1 = 11 - 1 = 10$$

$$SE = \frac{s}{\sqrt{n}} = \frac{7.97}{\sqrt{11}} = 2.40$$

$$T = \frac{\text{Observed} - \text{Null}}{SE} = \frac{-10.27 - 0}{2.40} = 4.28$$

$p - value = 0.0008056962$ (`> pt(4.28,df=10,lower.tail = FALSE)` `[1] 0.0008056962`)

Therefore, $p - value < \alpha$ and we reject the null hypothesis.

Question #7:

a)

H_0: The mean height is the same across all the mentioned countries.

$$\mu_{USA} = \mu_{UK} = \mu_{India}$$

H_A: The mean height differs between at least one pair of countries.

b)

		DF	Sum SQ	Mean SQ	F-value	Pr (>F)
Group	Class	2	1.75	0.875	0.02683461117	0.9735
Error	Residuals	69	684.75	32.607		
	Total	71	686.5			

$$\bar{y}_{usa} = \frac{180 + 183 + 172 + 178 + 169 + 179 + 178 + 180}{8} = 177.375$$
$$\bar{y}_{uk} = \frac{185 + 181 + 180 + 179 + 164 + 173 + 180 + 178}{8} = 177.5$$
$$\bar{y}_{uk} = \frac{170 + 183 + 180 + 175 + 181 + 183 + 176 + 167}{8} = 176.875$$
$$\bar{y} = \frac{176.875 + 177.5 + 177.375}{3} = 177.25$$

$$SSG = \sum_{j=1}^{k} n_j * (\bar{y}_j - \bar{y})^2$$
$$= 8 * (177.375 - 177.25)^2 + 8 * (177.5 - 177.25)^2 + 8 * (176.875 - 177.25)^2$$
$$= 1.75$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (180 - 177.25)^2 + (183 - 177.25)^2 + (172 - 177.25)^2 + (178 - 177.25)^2$$
$$+ (169 - 177.25)^2 + (179 - 177.25)^2 + (178 - 177.25)^2 + (180 - 177.25)^2$$
$$+ (185 - 177.25)^2 + (181 - 177.25)^2 + (180 - 177.25)^2 + (179 - 177.25)^2$$
$$+ (164 - 177.25)^2 + (173 - 177.25)^2 + (180 - 177.25)^2 + (178 - 177.25)^2$$
$$+ (170 - 177.25)^2 + (183 - 177.25)^2 + (180 - 177.25)^2 + (175 - 177.25)^2$$
$$+ (181 - 177.25)^2 + (173 - 177.25)^2 + (176 - 177.25)^2 + (167 - 177.25)^2$$
$$= 686.5$$

$SSE = SST - SSG = 686.5 - 1.75 = 684.75$

Degrees of freedom:

- Total: $df_T = n - 1 = 24 - 1 = 23$
- Group: $df_G = k - 1 = 3 - 1 = 2$
- Error: $df_E = df_T - df_G = 23 - 2 = 21$

Mean squares:

- Group: $MSG = \frac{SSG}{df_G} = \frac{1.75}{2} = 0.875$

- Error: $\text{MSG} = \frac{\text{SSE}}{\text{df}_E} = \frac{684.75}{21} = 32.607$

F-statics:

- $F = \frac{\text{MSG}}{\text{MSE}} = \frac{0.875}{32.607} = 0.02683461117$

Calculating p-value:

$p - \text{value} = 0.9735$(

```
> pf(0.02683461117, 2, 69, lower.tail = FALSE)
[1] 0.9735324
```

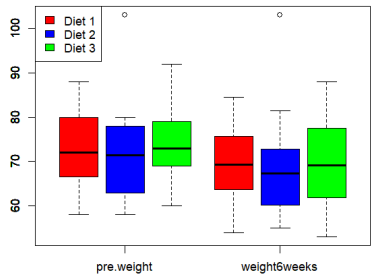
)

c) p-value is large($p - \text{value} > \alpha$), we fail to reject H_0. The data do not provide convincing evidence that at least one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

Question #8:

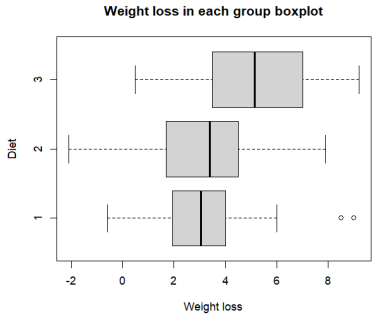
a)

- The weights boxplot before and after the diet are like the below figure:



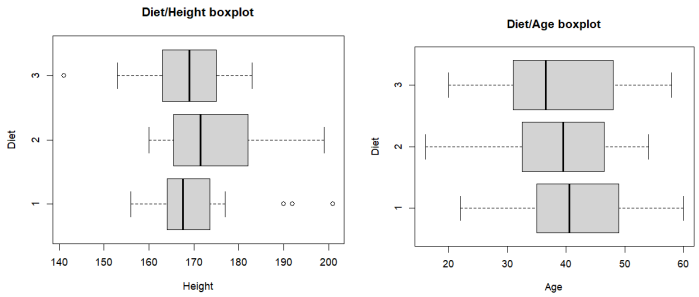

```
> boxplot(df[,c(5,7)], boxfill = NA, border = NA) #invisible boxes - only axes and plot area
> boxplot(df[df$Diet=="1", c(5,7)], xaxt = "n", add = TRUE, boxfill="red",
+         boxwex=0.25, at = 1:ncol(df[,c(5,7)]) - 0.3) #shift these left by -0.15
> boxplot(df[df$Diet=="2", c(5,7)], xaxt = "n", add = TRUE, boxfill="blue",
+         boxwex=0.25, at = 1:ncol(df[,c(5,7)])  ) #shift to the right by +0.15
> boxplot(df[df$Diet=="3", c(5,7)], xaxt = "n", add = TRUE, boxfill="green",
+         boxwex=0.25, at = 1:ncol(df[,c(5,7)]) + 0.3) #shift these left by -0.15
> legend(x = "topleft",legend=c("Diet 1", "Diet 2", "Diet 3"), fill = c("red","blue","green"))
```

- The weight loss in each group has a boxplot like this:




```
> df['difference']<-df$pre.weight-df$weight6weeks
> boxplot(df$difference ~ df$Diet ,main = "Weight loss in each group boxplot" ,xlab = "gender", ylab="Diet",horizontal = TRUE)
```

- The age and height boxplots are shown below:




```
> boxplot(df$Height ~ df$Diet ,main = "Diet/Height boxplot" ,xlab = "Height", ylab="Diet",horizontal = TRUE)
> boxplot(df$Age ~ df$Diet ,main = "Diet/Age boxplot" ,xlab = "Age", ylab="Diet",horizontal = TRUE)
```

b)

Hypothesis

H_0: The mean weight loss is the same across all the mentioned diet groups.

$$\mu_1 = \mu_2 = \mu_3$$

H_A: The mean weight loss differs between at least one pair of the mentioned diet groups.

Based on ANOVA test because p-value is small (less than α), reject H_0. Meaning that, there is a significant difference in the mean weight loss between groups.

c) The ANOVA table is shown here:

```
> result <- aov( difference~Diet, data = df)
> summary(result)
            Df Sum Sq Mean Sq F value  Pr(>F)
Diet         2   69.3   34.67    6.32 0.00284 **
Residuals   79  433.3    5.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is obvious that $p - value = 0.00284$. Therefore with α=0.05 we have $p - value < \alpha$. Therefore, with significance level of 0.05 we reject H_0. Therefore we can conclude that the mean weight loss differs between at least one pair of the mentioned diet groups.

d)

Now in order to find which groups caused this significant difference we have to do a pairwise comparison with the below hypothesis for each pair of groups:

H_0: The mean weight loss in diet group i and diet group j is the same.

$$\mu_i - \mu_j = 0$$

H_A: The mean weight loss in diet group i and diet group j is not the same.

$$\mu_i - \mu_j \neq 0$$

We check the above hypothesis for every two groups:

```
> pairwise.t.test(df$difference, df$Diet, p.adj = "none", pool.sd = FALSE)

        Pairwise comparisons using t tests with non-pooled SD

data:  df$difference and df$Diet

  1      2
2 0.7165 -
3 0.0062 0.0023

P value adjustment method: none
```

For the above pairwise test, we need to modify α based on Bonferroni correction therefore we have:

$$\alpha^* = \frac{\alpha}{3} = 0.0167$$

However we could just use the bulletin Bonferroni correction like this:

```
> pairwise.t.test(df$difference, df$Diet, p.adj = "bonf", pool.sd = FALSE)

        Pairwise comparisons using t tests with non-pooled SD

data:  df$difference and df$Diet

  1      2
2 1.0000 -
3 0.0186 0.0069

P value adjustment method: bonferroni
```

- Group 1 and 2:
  H_0: The mean weight loss in diet group 1 and diet group 2 is the same.
  $$\mu_1 - \mu_2 = 0$$
  H_A: The mean weight loss in diet group 1 and diet group 2 is not the same.
  $$\mu_1 - \mu_2 \neq 0$$

  p-value in comparison between group 1 and 2 is 0.72 which is bigger than $\alpha^* = 0.0167$ and we can not reject the H_0.

- Group 1 and 3:
  H_0: The mean weight loss in diet group 1 and diet group 3 is the same.
  $$\mu_1 - \mu_3 = 0$$
  H_A: The mean weight loss in diet group 1 and diet group 3 is not the same.
  $$\mu_1 - \mu_3 \neq 0$$

  p-value in comparison between group 1 and 3 is 0.0062 which is smaller than $\alpha^* = 0.0167$ and we can not reject the H_0.

- Group 3 and 2:
  H_0: The mean weight loss in diet group 3 and diet group 2 is the same.
  $$\mu_3 - \mu_2 = 0$$
  H_A: The mean weight loss in diet group 3 and diet group 2 is not the same.
  $$\mu_3 - \mu_2 \neq 0$$

  p-value in comparison between group 2 and 3 is 0.0023 which is smaller than $\alpha^* = 0.0167$. Therefore, we can reject the H_0.

Therefore the difference found between diets 1-3, and 2-3.

Based on the above tests we conclude that there is a significance difference in mean weight loss of diet groups. This difference has been caused by the difference of mean weight loss between diets 1-3, and 2-3.

Question #9:

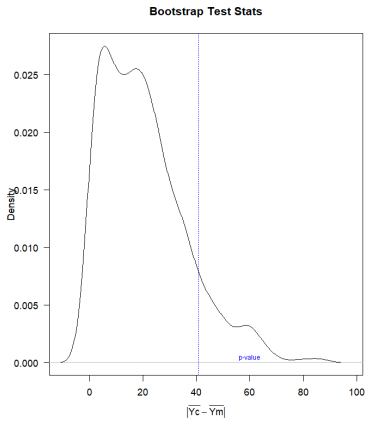
H_0: The mean area is the same in both groups.

$$\mu_{\text{London}} = \mu_{\text{Berlin}}$$

H_A: the mean area differs between groups.

$$\mu_{\text{London}} \neq \mu_{\text{Berlin}}$$

a) The bootstrap distribution is as followed. It is slightly skewed.



b) P-value

a. For the original sample:

```
> X <- df[df$City=="London",]  # London group
> Y <- df[df$City=="Berlin",]   # Berlin group
> ttest <- t.test(X$Area..Meter., Y$Area..Meter.,pool.sd = FALSE)
> ttest

        Welch Two Sample t-test

data:  X$Area..Meter. and Y$Area..Meter.
t = -1.7476, df = 24.954, p-value = 0.09283
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -89.047761   7.300508
sample estimates:
mean of x mean of y
 277.7692  318.6429
```

Because p-value>0.05 we fail to reject H_0,

b. For the bootstrap sample:

```
> #...calculate the p-value
> mean( Boot.test.stat1 >= test.stat1)
[1] 0.11
```

Because p-value>0.05 we fail to reject H_0

c) For the original sample:

```
95 percent confidence interval:
 -89.047761    7.300508
```

For the bootstrap sample:

```
> quantile(Boot.test.stat1, c(.975,.025))
     97.5%       2.5%
59.8327652  0.8094697
```