bm_hw_3

Problem 2

1

```r
smoke_data = read_csv("./HeavySmoke.csv") %>%
  janitor::clean_names() %>%
  mutate(diff = bmi_base - bmi_6yrs)
```

```
## Parsed with column specification:
## cols(
##   ID = col_integer(),
##   BMI_base = col_double(),
##   BMI_6yrs = col_double()
## )
```

```r
diff_mean = mean(smoke_data$diff)
diff_sd = sd(smoke_data$diff)
n = 10
t = (diff_mean - 0)/(diff_sd/sqrt(n))
```

```r
qt(0.975, n-1)
```

```
## [1] 2.262157
```

```r
t.test(smoke_data$bmi_base, smoke_data$bmi_6yrs, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  smoke_data$bmi_base and smoke_data$bmi_6yrs
## t = -4.3145, df = 9, p-value = 0.001949
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.121709 -1.598291
## sample estimates:
## mean of the differences
##                   -3.36
```

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 > 0$$

$$\bar{d} = \sum_{i=1}^{n} \frac{d_i}{n} = -3.36$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} = 2.4627$$

$$n = 10$$

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = -4.3145$$

$$t_{n-1,1-\alpha/2} = 2.262157$$

$$|t| = 4.3145$$

$$For \ \ |t| > t_{n-1,1-\alpha/2}, \ reject \ H_0$$

Intepretation: We use paired t-test to test whether those 10 women's BMI has changed over 6 years after quitting smoking. According to the solutions listed above, we should reject the null, which means their BMI has changed significantly over 6 years.

2

```r
nonsmoke_data = read_csv("./NeverSmoke.csv") %>%
  janitor::clean_names()
```

```
## Parsed with column specification:
## cols(
##   ID = col_integer(),
##   BMI_base = col_double(),
##   BMI_6yrs = col_double()
## )
```

```r
n1=10
n2=10
qf(0.975, n1-1, n2-1)
```

```
## [1] 4.025994
```

```r
#test equality for variances
var.test(nonsmoke_data$bmi_base, nonsmoke_data$bmi_6yrs, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  nonsmoke_data$bmi_base and nonsmoke_data$bmi_6yrs
## F = 0.94826, num df = 9, denom df = 9, p-value = 0.9382
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2355353 3.8177044
## sample estimates:
## ratio of variances
##          0.9482638
```

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_0 : \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} = 0.94826$$

$$F_{n_1-1,n_2-1} = 4.025994$$

$$For\ F < F_{n_1-1,n_2-1},\ fail\ to\ reject\ H_0,\ \sigma_1^2 = \sigma_2^2$$

```
qt(0.975, n1+n2-2)
```

```
## [1] 2.100922
```

```
t.test(nonsmoke_data$bmi_base, nonsmoke_data$bmi_6yrs, var.equal = TRUE, paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  nonsmoke_data$bmi_base and nonsmoke_data$bmi_6yrs
## t = -0.69101, df = 18, p-value = 0.4984
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.262569  3.162569
## sample estimates:
## mean of x mean of y
##     28.86     30.41
```

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 25.15739$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} = -0.69101$$

$$|t| = 0.69101$$

$$t_{n_1+n_2-2,1-\alpha/2} = 2.100922$$

$$For\ \ |t| < t_{n_1+n_2-2,1-\alpha/2},\ fail\ to\ reject\ H_0,\ \mu_1 = \mu_2$$

Intepretation: First, we use F-test to test the equality of variances. The result shows that the variances of two groups are equal. Then we use t-test to test the equality of mean. The result shows that the means of two groups are equal. So there is no significant BMI changes between women who quit smoking and women who never smoked.

3

Show the corresponding 95% CI associated with part 2. Interpret it in the context of the problem.

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2,1-\alpha/2} s\sqrt{1/n_1 + 1/n_2} \leq \mu \leq (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2,1-\alpha/2} s\sqrt{1/n_1 + 1/n_2}$$

```
t = qt(0.975, 18)
s = sqrt(25.15739)
CIL = 28.86 - 30.41 - (t * s * sqrt(2/10))
CIR = 28.86 - 30.41 + (t * s * sqrt(2/10))
```

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2,1-\alpha/2} s\sqrt{1/n_1 + 1/n_2} = -6.262569$$
$$(\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2,1-\alpha/2} s\sqrt{1/n_1 + 1/n_2} = 3.162569$$
$$-6.262569 \leq \mu \leq 3.162569$$

Intepretation: The 95% CI for these two samples are (-6.262569, 3.162569). This CI means that we are 95% confidence that the true population mean difference between women that quit smoking and women who never smoked lies between the lower and upper limits of the interval.

4

a

For this new study, I would choose 50 women who never smoked and 50 women who quit smoking. To build the counterfactual, we should make sure that these two groups are comparable, which means except exposure, other conditions of women in each group should be the same (e.g health condition, age). Then, recording the BMI of each group. The possible bias in this study should be avoided is that 1) we should have sufficient sample size. Greater sample size can better represent the population. If the sample size is too small, the result might be inaccurate; 2) make sure there is no loss to follow up.

b

$$n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

Smoke sample size

Power	0.8	0.9
α		
0 .25	4.224461	5.516092
0.5	3.488391	4.669966

Never-Smoke sample size

Power	0.8	0.9
α		
0.25	7.400115	9.662704
0.5	6.11072	8.18052

Problem 3

A rehabilitation center is interested in examining the relationship between physical status before therapy and the time (days) required in physical therapy until successful rehabilitation. Records from patients 18-30 years old were collected and provided to you for statistical analysis (data "Knee.csv").

Assuming that data are normally distributed, answer the questions below: 1. Generate descriptive statistics for each group and comment on the differences observed (R only). (4p)

1

```
knee_data = read_csv("./Knee.csv") %>%
  janitor::clean_names()
```

```
## Parsed with column specification:
## cols(
##   Below = col_integer(),
##   Average = col_integer(),
##   Above = col_integer()
## )
```

```
summary(knee_data$below)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      29      36      40      38      42      43       2
```

```
sd(knee_data$below, na.rm = T)
```

```
## [1] 5.477226
```

```
summary(knee_data$average)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   30.25   32.00   33.00   35.00   39.00
```

```
sd(knee_data$average, na.rm = T)
```

```
## [1] 3.91578
```

```
summary(knee_data$above)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   20.00   21.00   22.00   23.57   24.50   32.00       3
```

```
sd(knee_data$above, na.rm = T)
```

```
## [1] 4.197505
```

Group	mean	sd	3rd Qu.
Below	38	5.477226	42
Average	33	3.91578	35
Above	23.57	4.197505	24.5

As we can see above, the average days required in physical therapy until successful rehabilitation is largest in Below group and smallest in Above group. The number of 3rd Qu. is largest in the Below group and smallest in the Above group. These two findings are conform with our common sense. If the physical status before therapy is relatively better, the days required should be relatively shorter.

2

```r
below <- knee_data$below
average <- knee_data$average
above <- knee_data$above

knee_reshape <- c(below, average, above)
ind<-c(rep(3,length(below)),rep(2,length(average)),rep(1,length(above)))
new_data_knee <- as.data.frame(cbind(knee_reshape,ind))

res<-lm(knee_reshape~factor(ind), data=new_data_knee)
anova(res)
```

```
## Analysis of Variance Table
##
## Response: knee_reshape
##             Df Sum Sq Mean Sq F value    Pr(>F)
## factor(ind)  2 795.25  397.62   19.28 1.454e-05 ***
## Residuals   22 453.71   20.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H_0: \ \mu_1 = \mu_2 = \mu_3 = \ \ldots \ = \mu_k H_1: \ at\ least\ two\ means\ are\ not\ equal$$
$$F = \frac{Between\ SS/(k-1)}{With\ SS/(n-k)} \sim F_{k-1,n-k}\ distribution\ under\ H_0 F = 19.28$$

```r
qf(1-0.01, 2, 22)
```

```
## [1] 5.719022
```

$$F_{k-1,n-k,1-\alpha} = 5.719022 F > F_{k-1,n-k,1-\alpha},\ reject\ H_0$$

3 (R only). (5p)

Bonferroni

```r
pairwise.t.test(new_data_knee$knee_reshape, new_data_knee$ind, p.adj='bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  new_data_knee$knee_reshape and new_data_knee$ind
##
##   1       2
## 2 0.0011  -
## 3 1.1e-05 0.0898
##
## P value adjustment method: bonferroni
```

```r
qt( 1-((0.01/3)/2), 22 )
```

```
## [1] 3.290888
```

As we can see, according to the bonferroni adjustment, there are no mean differences between each group, which means:

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

Tukey

```r
res1<-aov(knee_reshape~factor(ind), data=new_data_knee)
summary(res1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(ind)  2  795.2   397.6   19.28 1.45e-05 ***
## Residuals   22  453.7    20.6                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

```r
TukeyHSD(res1, conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
## 
## Fit: aov(formula = knee_reshape ~ factor(ind), data = new_data_knee)
## 
## $`factor(ind)`
##          diff       lwr      upr     p adj
## 2-1  9.428571  2.168498 16.68864 0.0010053
## 3-1 14.428571  6.803969 22.05317 0.0000102
## 3-2  5.000000 -1.988063 11.98806 0.0736833
```

According to the Tukey method, we can see that the mean between below and above and the mean between average and above are different. Tukey method is less conservative than Bonferroni.

Dunnett

```r
library(DescTools)
```

```r
x <- c(29,42,38,40,43,40,30,42)
y <- c(30,35,39,28,31,31,29,35,39,33)
z <- c(26,32,21,20,23,22,21)
dunn_knee <- c(x,y,z)
g <- factor(rep(1:3, c(8, 10, 7)),
            labels = c("below",
                       "average",
                       "above"))
DunnettTest(dunn_knee, g, control = "above", conf.level = 0.99)
```

```
## 
##   Dunnett's test for comparing several treatments with a control :  
##     99% family-wise confidence level
## 
```

```
## $above
##                     diff   lwr.ci   upr.ci    pval
## below-above   14.428571 7.173453 21.68369 6.9e-06 ***
## average-above  9.428571 2.520317 16.33683 0.00069 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the Dunnett method, we can see that the mean between below and above group, and the mean between average and above group are different. This conclusion is consistent with the result using Tukey's method.

4. Write a short paragraph summarizing your results as if you were presenting to the rehabilitation center director.(1p)

Problem 4

For this problem you will use the built-in R data called "UCBAdmissions" (library 'datasets'), an example of sex bias in admission practices. You are interested in comparing the proportions of women vs men admitted at Berkeley (over all departments).

1. Provide point estimates and 95% CIs for the overall proportions of men and women admitted at Berkeley. Briefly comment on the values. (5p)

```r
library(datasets)
ucb_ad = UCBAdmissions
```

2. Perform a hypothesis test to assess if the two proportions in 1) are significantly different. Report the results including the test statistic and p-value and an overall conclusion of your findings. This part should contain both 'hand' and R calculations. For the latter, feel free to use built-in functions or to create your own. (5p)