bm_hw_3

Problem 2

1

```
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()
## )
```

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

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

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

```
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

```
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()
## )
```

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

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

```
#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.

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

```r
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

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

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

2. Using a type I error of 0.01, obtain the ANOVA table. State the hypotheses, decision rule and conclusion (R only). (5p)-
3. Based on your response in part 3, perform pairwise comparisons with the appropriate adjustments (Bonferroni, Tukey, and Dunnett – 'below average' as reference). Report your findings and comment on the differences/similarities between these three methods (R only). (5p)
4. Write a short paragraph summarizing your results as if you were presenting to the rehabilitation center director.(1p)