Problem 2

Here is the **code chunk** to do the calculation in Problem 2

1

```
heavysmoke_df <- read_csv(file = "./data/HeavySmoke.csv")
heavysmoke_df = janitor::clean_names(heavysmoke_df) %>%
  mutate(diff = bmi_6yrs-bmi_base) %>%
  mutate(mean = sum(diff)/length(bmi_6yrs)) %>%
  mutate(sd = (mean - diff)^2)

s_d_1= sqrt(sum((heavysmoke_df$sd))/(length(heavysmoke_df$id)-1))
t_statistics = 3.36 / (s_d_1/sqrt(9))
t_critical = qt(0.975,9)
t.test(heavysmoke_df$bmi_6yrs,heavysmoke_df$bmi_base,paired = T)
```

2

```
neversmoke_df <- read_csv(file = "./data/NeverSmoke.csv") %>%
  janitor::clean_names() %>%
  mutate(diff = bmi_6yrs-bmi_base) %>%
  mutate(mean = sum(diff)/length(bmi_6yrs)) %>%
  mutate(sd = (mean - diff)^2)

s_d_2 = sqrt(sum((neversmoke_df$sd))/(length(neversmoke_df$id)-1))
s_sqr = (9*s_d_1^2+9*s_d_2^2)/(18)
t_stat = (3.36 - 1.55) / (sqrt(s_sqr)*sqrt(1/9+1/9))
t_crit = qt(0.975,18)
diff_heavy = heavysmoke_df$bmi_6yrs - heavysmoke_df$bmi_base
diff_never = neversmoke_df$bmi_6yrs - neversmoke_df$bmi_base
f_crit=qf(0.975,9,9)
var.test(diff_heavy,diff_never,alternative = "two.sided")
res = t.test(diff_heavy,diff_never,var.equal = FALSE, paired = FALSE)
```

4

```
power.t.test(power = .90, delta = 3.0, sd=2.0, sig.level = 0.05, alternative = c("two.sided"))
power.t.test(power = .80, delta = 3.0, sd=2.0, sig.level = 0.05, alternative = c("two.sided"))
power.t.test(power = .90, delta = 3.0, sd=2.0, sig.level = 0.025, alternative = c("two.sided"))
power.t.test(power = .80, delta = 3.0, sd=2.0, sig.level = 0.025, alternative = c("two.sided"))
power.t.test(power = .90, delta = 1.7, sd=1.5, sig.level = 0.05, alternative = c("two.sided"))
power.t.test(power = .80, delta = 1.7, sd=1.5, sig.level = 0.05, alternative = c("two.sided"))
power.t.test(power = .90, delta = 1.7, sd=1.5, sig.level = 0.025, alternative = c("two.sided"))
power.t.test(power = .80, delta = 1.7, sd=1.5, sig.level = 0.025, alternative = c("two.sided"))
```

Problem 3

Here is the **code chunk** to read the files.

```
knee_df = read_csv(file = "./data/Knee.csv") %>%
  janitor::clean_names()
```

1

Here is the **code chunk** to generate descriptive statistics for each group.

```
summary(knee_df$below)
```

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

```
sd(!is.na(knee_df$below))
```

```
## [1] 0.421637
```

```
summary(knee_df$average)
```

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

```
sd(!is.na(knee_df$average))
```

```
## [1] 0
```

```
summary(knee_df$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(!is.na(knee_df$above))
```

```
## [1] 0.4830459
```

From the vaue below, we could see the "below" group has the highest median and mean, while the "above" group has the slowest median and mean. The "above" group has the largest standard deviation 0.48 while the average has the standard deviation 0.

2

Here is the **code chunk** to obtain the ANOVA table. The null hypothesis is all means of different groups are equal: H0: u1 = u2 = u3 The corresponding alternative hypothesis is that not all means are equal:

```
knee_df = gather(knee_df, key = type, value = value, below:above) %>%
  filter(!is.na(value))
knee_df$type=as_factor(knee_df$type)
res <- aov(value~type,data = knee_df)
summary(res)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## type         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
```

```
qf(0.01, df1 = 2, df2 = 22)
```

```
## [1] 0.01005493
```

From the table, we are able to find that the p - value is 1.45e-05, which is significantly lower than 0.05. We reject the null hypothesis and conclude that there is enough evidence to show that at least two of the means are different.

3

Now we performed pairwise comparisons with the appropriate adjustments (Bonferroni, Tukey, and Dunnett – 'below average' as reference). Here is the **code chunk**:

```
pairwise.t.test(knee_df$value,knee_df$type,p.adjust.method = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  knee_df$value and knee_df$type
##
##         below   average
## average 0.0898  -
## above   1.1e-05 0.0011
##
## P value adjustment method: bonferroni
```

```
TukeyHSD(res)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = value ~ type, data = knee_df)
##
## $type
##                      diff       lwr        upr     p adj
## average-below  -5.000000 -10.41130  0.4113011 0.0736833
## above-below   -14.428571 -20.33278 -8.5243579 0.0000102
## above-average  -9.428571 -15.05051 -3.8066356 0.0010053
```

```
dunnetttest<-glht(res, linfct=mcp(type="Dunnett"))
summary(dunnetttest)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = value ~ type, data = knee_df)
##
## Linear Hypotheses:
##                       Estimate Std. Error t value Pr(>|t|)
## average - below == 0   -5.000      2.154  -2.321   0.0543 .
## above - below == 0    -14.429      2.350  -6.139 6.93e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Problem 4

Here is the **code chunk** to load the data file and conduct data cleaning:

```
library(datasets)
data("UCBAdmissions")

ucb_df <- as.data.frame(UCBAdmissions) %>%
  janitor::clean_names()
```

1

Here is the **code chunk** to provide point estimates and 95% CIs for the overall proportions of men and women admitted at Berkeley.

```
num_men_total = sum(filter(ucb_df,gender=="Male")$freq)
num_women_total = sum(filter(ucb_df,gender=="Female")$freq)
num_men_admitted = sum(filter(ucb_df,gender=="Male", admit=="Admitted")$freq)
num_women_admitted = sum(filter(ucb_df,gender=="Female", admit=="Admitted")$freq)

prop_men = num_men_admitted / num_men_total
prop_women = num_women_admitted / num_women_total

prop.test(num_men_admitted,num_men_total, p=0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  num_men_admitted out of num_men_total, null probability 0.5
## X-squared = 32.12, df = 1, p-value = 1.449e-08
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4263168 0.4642163
## sample estimates:
##         p
## 0.4451877
```

```
prop.test(num_women_admitted,num_women_total,p=0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  num_women_admitted out of num_women_total, null probability 0.5
## X-squared = 282.51, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2826669 0.3252490
## sample estimates:
##         p
## 0.3035422
```

From the data above, we could obtain that the point estimate for the overall proportion of men admitted at Berkeley is 0.45, the 95% confidence interval (0.42, 0.46). The point estimate for the overall proportion of women admitted at Berkey is 0.30, the 95% confidence interval is (0.28, 0.32). From the data above, we

observe that there is no overlap of the two confidence interval, and we could make a strong guess that indeed there exists sex bias in admission practices.

2

Here is **code chunk** to perform a hypothesis test to assess if the two proportions in previous example are significantly different. The null hypothesis H0 is that p1=p2. The coressponding alternative hypothesis H1 is that p1 != p2.

```
res <- prop.test(x = c(num_men_admitted, num_women_admitted), n = c(num_men_total,num_women_total))
res
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(num_men_admitted, num_women_admitted) out of c(num_men_total, num_women_total)
## X-squared = 91.61, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1129887 0.1703022
## sample estimates:
##    prop 1    prop 2
## 0.4451877 0.3035422
```

From the data above, we could obtain the p-value is 2.2e-16, which is way smaller than 0.05. We reject the null hypothesis and conclude that there is enough evidence to show that there is sex bias in admission practices.