Homework 3

Jaisal Amin

10/5/2018

Problem 2

1.)

```r
heavy_smoke = read_csv("HeavySmoke.csv")
```

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

```r
BMI_start = heavy_smoke$BMI_base
BMI_end = heavy_smoke$BMI_6yrs
```

The p-value is < 0.05 so we reject the null hypothesis that the mean BMI at the start is different from the mean BMI at the end.

2.)

```r
never_smoke = read_csv("NeverSmoke.csv")
```

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

```r
heavy_diff = BMI_start - BMI_end
never_diff = never_smoke$BMI_base - never_smoke$BMI_6yrs

var.test(heavy_diff, never_diff)
```

```
##
##  F test to compare two variances
##
## data:  heavy_diff and never_diff
## F = 1.1627, num df = 9, denom df = 9, p-value = 0.826
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2888038 4.6811133
## sample estimates:
## ratio of variances
##          1.162722
```

Because the p-value is > 0.05 we fail to reject the null hypothesis that the variances are equal.

3.)

95% CI of the difference: (0.2888038, 4.6811133). We are 95% confident that the true difference of variance is between 0.2888038 and 4.6811133.

4a.)

I would design a cohort study to follow the BMI of smokers who recently (within the last 3 months) quit and non-smokers. Since you cannot assign the treatment (smoker/non-smoker) you cannot randomize treatment but you can and should randomize across demographic factors such as age, sex, ethnicity, socioeconomic status, etc. The representative sample should be large enough to comfortably conduct multiple tests and should be followed for at least 6 years.

4b.)

```
smokers_mean = 3.0
never_mean = 1.7
smokers_sd = 2.0
never_sd = 1.5
sd_pooled = (((smokers_sd^2) + (never_sd^2))/2)^0.5
##Calculating Cohen's d in order to use power function
cd = (smokers_mean - never_mean)/sd_pooled
```

80% vs. 90% power

```
pwr.t.test(d = cd, sig.level = 0.05, power = 0.8, type = c("two.sample"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 30.01813
##               d = 0.7353911
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
pwr.t.test(d = cd, sig.level = 0.05, power = 0.9, type = c("two.sample"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 39.84411
##               d = 0.7353911
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

2.5% vs. 5% significance level

```
pwr.t.test(d = cd, sig.level = 0.025, power = 0.9, type = c("two.sample"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 47.1809
##               d = 0.7353911
##       sig.level = 0.025
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
pwr.t.test(d = cd, sig.level = 0.05, power = 0.9, type = c("two.sample"))
```

```
##
##      Two-sample t test power calculation
##
##               n = 39.84411
##               d = 0.7353911
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Problem 3

```r
knee = read_csv("Knee.csv")
```

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

1.)

Descriptive statistics

```r
knee %>%
  summary
```

```
##      Below        Average           Above
##  Min.   :29   Min.   :28.00   Min.   :20.00
##  1st Qu.:36   1st Qu.:30.25   1st Qu.:21.00
##  Median :40   Median :32.00   Median :22.00
##  Mean   :38   Mean   :33.00   Mean   :23.57
##  3rd Qu.:42   3rd Qu.:35.00   3rd Qu.:24.50
##  Max.   :43   Max.   :39.00   Max.   :32.00
##  NA's   :2                    NA's   :3
```

Looking at the summary data, the mean and median decrease across groups, with medians between groups varying slightly more than means. The IQR of the "average" and "above" groups are similar however the below group has a higher IQR suggesting greater variability or outliers within that group.

2.)

```r
knee_data = knee %>%
  gather(key = "level", value = "recovery_days", Below:Above, na.rm = TRUE)

anova(lm(recovery_days~factor(level), data = knee_data))
```

```
## Analysis of Variance Table
##
## Response: recovery_days
##               Df Sum Sq Mean Sq F value    Pr(>F)
## factor(level)  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
```

The hypothesis for this ANOVA test is that the mean between all 3 levels ("Below", "Average", "Above") was equal. We reject the null if F > F(k-1, n-k, 1-a) and we fail to reject if it is less than or equal to. In this case the p-value is far below the 0.01 significance level so we reject the null hypothesis that the means are equal. Because we reject the null, we can proceed to pairwise comparisons:

3.)

```r
pairwise.t.test(knee_data$recovery_days, knee_data$level, p.adj = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  knee_data$recovery_days and knee_data$level
##
##         Above   Average
## Average 0.0011  -
## Below   1.1e-05 0.0898
##
## P value adjustment method: bonferroni
```

```r
knee_aov = aov(recovery_days~factor(level), data = knee_data)
TukeyHSD(knee_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = recovery_days ~ factor(level), data = knee_data)
##
## $`factor(level)`
##                    diff        lwr      upr     p adj
## Average-Above  9.428571  3.8066356 15.05051 0.0010053
## Below-Above   14.428571  8.5243579 20.33278 0.0000102
## Below-Average  5.000000 -0.4113011 10.41130 0.0736833
```

```r
DunnettTest(recovery_days ~ level, data = knee_data, control = "Below")
```

4.)

```r
admissions = as_tibble(UCBAdmissions)

admissions %>%
  filter(Gender == "Female") %>%
  group_by(Admit) %>%
  summarize(sum(n))
```

```
## # A tibble: 2 x 2
##   Admit    `sum(n)`
##   <chr>       <dbl>
## 1 Admitted      557
## 2 Rejected     1278
```

```r
admissions %>%
  filter(Gender == "Male") %>%
  group_by(Admit) %>%
  summarize(sum(n))
```

```
## # A tibble: 2 x 2
##   Admit    `sum(n)`
##   <chr>       <dbl>
## 1 Admitted     1198
## 2 Rejected     1493
```

```r
female_prop = 557 / (1278 + 557)
male_prop = 1198 / (1198 + 1493)
```

Point estimate of female student admittance is 0.3035422 and point estimate for male admittance is 0.4451877.

```r
sort_ad = spread(admissions, key = Admit, value = n) %>%
  mutate(sum = Admitted + Rejected) %>%
  mutate(prop_admit = Admitted / sum) %>%
  select(c(Gender, prop_admit))

female = sort_ad %>%
  filter(Gender == "Female")

male = sort_ad %>%
  filter(Gender == "Male")

f = female$prop_admit
m = male$prop_admit

t.test(f, m, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  f and m
## t = 0.24772, df = 9.3979, p-value = 0.8097
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2906653  0.3626712
## sample estimates:
```

```
## mean of x mean of y
## 0.4172692 0.3812662
```