Alyssa Vanderbeek (amv2187)
12 October 2018
HW 3

Problem 2: Cigarette smoking continues to be a public health problem with major consequences on heart and lung diseases. Less is actually known about the consequences of quitting smoking. A recent study selected a group of 10 women working at a small medical practice, ages 50-64, that had smoked at least 1 pack/day and quit for at least 6 years (data "HeavySmoke.csv").

1. The first question is to assess if their body mass index (BMI) has changed 6 years after quitting smoking. Perform an appropriate hypothesis test and interpret your findings.

Because we are looking at the difference in outcomes over time within the same individuals, we perform a paired t test with hypotheses

$$H_0 : \Delta = \mu_{6\text{yrs}} - \mu_{\text{baseline}} = 0$$
$$H_1 : \Delta = \mu_{6\text{yrs}} - \mu_{\text{baseline}} \neq 0$$

In plain english, we are testing whether there is (H_1) or is not (H_0) a difference in BMI in women who quit smoking 6 years prior. Note that we are interested in the change in BMI within each individual. For that reason, we can perform a paired t-test.

We know sample size $n = 10$, and the test statistic is given by
$t = \frac{\bar{\Delta}-0}{s/\sqrt{n}}$, where $\bar{\Delta} = \frac{\sum_{i=1}^{n} \Delta_i}{n}$ and $s = \sqrt{\frac{\sum_{i=1}^{n} (\Delta_i - \bar{\Delta})^2}{n-1}}$.

Table 1: Change in BMI in women who quit smoking

ID	BMI at baseline	BMI at 6 years	Change in BMI
11	25.6	31.1	5.5
12	24.4	27.6	3.2
13	31.0	36.6	5.6
14	20.4	20.8	0.4
15	22.3	23.2	0.9
16	22.2	23.8	1.6
17	20.8	26.1	5.3
18	23.5	31.0	7.5
19	26.6	29.2	2.6
20	23.0	24.0	1.0

Given our data, we find that

$$\bar{\Delta} = \frac{5.5 + 3.2 + ... + 1.0}{10} = 3.36,$$
$$s = \sqrt{\frac{(5.5 - 3.36)^2 + (3.2 - 3.36)^2 + ...(1.0 - 3.36)^2}{10 - 1}} = 2.46, \text{and}$$
$$t = \frac{3.36}{2.46/\sqrt{10}} = 4.31.$$

At a 5% significance level, we reject the null hypothesis that there is no change in BMI if the test statistic $t > t_{1-\alpha/2,n-1}$, where $t_{1-\alpha/2,n-1} = t_{0.975,9} = 2.2621572$. And since $4.31 > 2.26$, we reject the null in favor of

the alternative hypothesis and conclude that women who quit smoking after years of heavy consumption experience a change in their BMI after 6 years.

2. The investigators suspected an overall change in weight over the years, so they decided to enroll a control group of 50-64 years of age that never smoked (data NeverSmoke.csv). Perform an appropriate test to compare the BMI changes between women that quit smoking and women who never smoked. Interpret the findings.

$$H_0 : \Delta_{\text{smokers}} - \Delta_{\text{not smokers}} = 0$$
$$H_1 : \Delta_{\text{smokers}} - \Delta_{\text{not smokers}} \neq 0$$

Now we are interested in testing whether change in BMI in women ages 50-64 over a 6 year period is different between previously heavy smokers and those who never smoked. Because we have two independent sample (i.e. women who were heavy smokers cannot be said to have never smoked, and vice versa), we perform a two-sample t-test of means.

Table 2: 6-year change in BMI for female former smokers vs. never smokers

Heavy smokers	Never smoked
5.5	2.8
3.2	-0.9
5.6	-2.1
0.4	3.9
0.9	1.7
1.6	4.8
5.3	3.5
7.5	1.8
2.6	-0.8
1.0	0.8

From part 1 above, we know that in the smokers group, $\bar{\Delta}_1 = 3.36$ and $s_1 = 2.46$. To calculate these values for the non-smoking group ($n_2 = 10$), we use the same formulas as given in (1) and find that $\bar{\Delta}_2 = \frac{2.8-0.9+\ldots+0.8}{10} = 1.55$ and $s_2 = \sqrt{\frac{(2.8-1.55)^2+(-0.9-1.55)^2+\ldots(0.8-1.55)^2}{10-1}} = 2.28$.

The first step in testing our hypotheses is to first test whether the variances are equal between groups. We compute an F-statistic $F = \frac{s_1^2}{s_2^2} = \frac{2.46^2}{2.28^2} = 1.16$, and we reject the null hypothesis that variances are equal between groups and conclude that they are different if $F > F_{n_1-1,n_2-1}$, where $F_{n_1-1,n_2-1} = F_{9,9} = 4.03$. Since $1.16 < 4.03$, we do not reject the stated null hypothesis, and conclude that variances between groups are not different.

Since we can assume that variances are equal, we know our test statistic is given by

$$t = \frac{\bar{\Delta}_1 - \bar{\Delta}_2 - 0}{s_{pooled}/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Then

$$s_{pooled} = \sqrt{\frac{(10-1)(2.46)^2 + (10-1)(2.28)^2}{10+10-2}} = 1.37, \text{and}$$
$$t = \frac{3.36 - 1.55}{1.37/\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.704.$$

We reject the null in favor of the alternative if $t > t_{n_1+n_2-2,1-\frac{\alpha}{2}}$, where $t_{n_1+n_2-2,1-\frac{\alpha}{2}} = t_{18,0.975} = 2.1$. Since $1.704 < 2.1$, we do not reject the null hypothesis, and conclude that there is no difference in BMI change between former smokers and never smokers.

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

The 95% CI is calculated as

$$(\bar{\Delta}_1 - \bar{\Delta}_2 - (t_{n_1+n_2-2,1-\frac{\alpha}{2}})(s_{pooled})\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \bar{\Delta}_1 - \bar{\Delta}_2 + (t_{n_1+n_2-2,1-\frac{\alpha}{2}})(s_{pooled})\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$$

$$= (3.36 - 1.55 - (1.704)(2.37)\sqrt{\frac{1}{10}+\frac{1}{10}}, 3.36 - 1.55 + (1.704)(2.37)\sqrt{\frac{1}{10}+\frac{1}{10}})$$

$$= (-0.421, 4.041)$$

Therefore, we are 95% confident that the average difference in BMI change between former smokers and never smokers is between -0.42 and 4.04 BMI units. Since the null hypothesis (a difference of 0) is contained in this interval, we conclude that there is no difference in BMI change over time between groups. However, the lower limit of the CI is very close to 0, and so we may want to conduct a larger study to confirm or debunk the results seen here.

4. Suppose the researchers want to launch into a larger study to prove that a difference does exist between the two groups with respect to BMI changes.

(a) How would you design the new study? Comment on elements of study design such as randomization, possible causes of bias that should be avoided, etc.

There are two approaches to designing a study that tests this hypothesis, both of which are observational in nature: (1) designing a retrospective study that selects a group women who quit at least 6 years ago versus a group of women with similar demographics who never smoked, and (2) designing a prospective study that follows for 6 years women who are currently quitting smoking versus women who never smoked. Let's assume that we are designing the former (retrospective), and that we have access to past medical records for the women to be enrolled.

We need not worry about randomization here, since we already know the exposure (smoking vs. not). As a result, we must be extra vigilant about potential sources of bias. With regards to blinding, the participants are obviously not blinded, as is the case for observational studies. But it may be beneficial to blind the researchers. I would suggest that the statisticians who perform the analysis are not the same people who collect the data. In that way, they are aware only of groups "1" and "2", and not of which group is the which.

(b) Calculate the sample size for the new study. Assuming a two-sided test, create a table showing sample size estimates for 80% vs 90% power, 2.5% vs 5% significance level, using the following information: the true mean increase for smokers is 3.0 kg/m2, with a standard deviation of 2.0 kg/m2; for never-smokers the true mean increase is 1.7 kg/m2, with a standard deviation of 1.5 kg/m2.

For this study we know

$$\Delta_1 = 3.0, \sigma_1^2 = 2.0 \text{ for the former smokers, and}$$

$$\Delta_2 = 1.7, \sigma_2^2 = 1.5 \text{ for the never smokers.}$$

.

Table 3 shows the sample size requirements for 80% and 90% power, and 5% and 2.5% significance using R calculations. Below is the R code for sample size calculations, where sample size estimates are rounded up to the nearest integer.

```
a0.05_p0.8 = ceiling(power.t.test(d = 3.0 - 1.7,
                                  sd = 2^2 + 1.5^2,
                                  power = 0.8)$n)
a0.05_p0.9 = ceiling(power.t.test(d = 3.0 - 1.7,
                                  sd = 2^2 + 1.5^2,
                                  power = 0.9)$n)
a0.025_p0.8 = ceiling(power.t.test(d = 3.0 - 1.7,
                                   sd = 2^2 + 1.5^2,
                                   power = 0.8,
                                   sig.level = 0.025)$n)
a0.025_p0.9 = ceiling(power.t.test(d = 3.0 - 1.7,
                                   sd = 2^2 + 1.5^2,
                                   power = 0.9,
                                   sig.level = 0.025)$n)
```

Table 3: Total sample size requirements for the specified power and significance levels

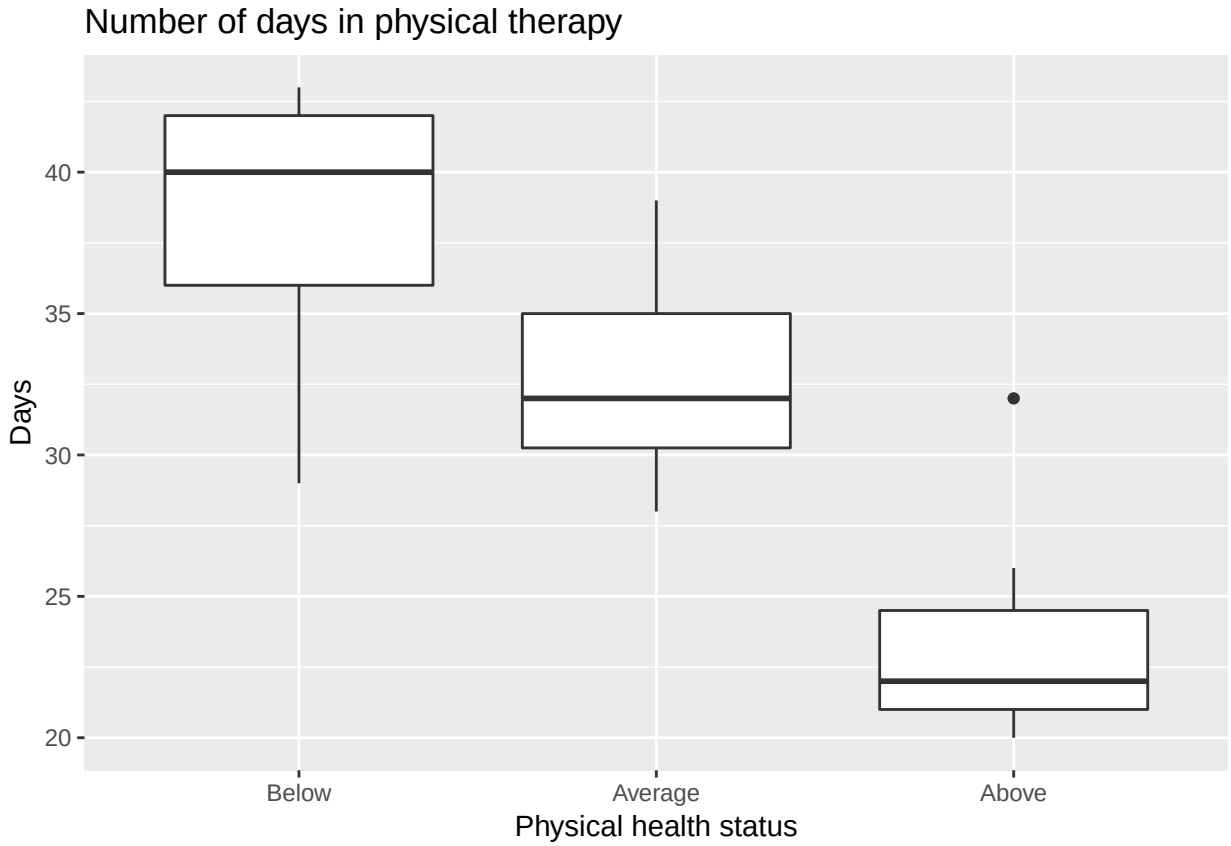
	80% power	90% power
2.5% significance	441	576
5% significance	364	487

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").

1. Generate descriptive statistics for each group and comment on the differences observed

Table 4: Summary statistics across

Physical Health	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Below	29	36.00	40	38.00000	42.0	43	2
Average	28	30.25	32	33.00000	35.0	39	NA
Above	20	21.00	22	23.57143	24.5	32	3



2. Using a type I error of 0.01, obtain the ANOVA table. State the hypotheses, decision rule and conclusion.

H_0 = Physical status has no effect on the amount of physical therapy needed for rehabilitation. (Duration of physical therapy is the same in all physical status categories)

H_1 = Physical status has some effect on the amount of physical therapy needed for rehabilitation. (Duration of physical therapy is different in at least one physical status category)

With a sample size of 30 participants (10 per group) and $\alpha = 0.01$, we will reject the null hypothesis in favor of the alternative if the computed F statistic F is greater than the critical value $F_{3-1,30-1,1-\alpha} = F_{2,29,0.99} = 5.420445$.

As seen in the below ANOVA table, our F-statistic for the test is greater than 19. Since $19 > 5.420445$, we reject the null in favor of the alternative that physical status affects the amount of physical therapy needed for rehabilitation.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(physical_status)	2	795.2457	397.62286	19.2802	1.45e-05
Residuals	22	453.7143	20.62338	NA	NA

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.

```
# multiple testing adjustments
pairwise.t.test(knee_data$days, knee_data$physical_status, p.adj = 'bonferroni')
```

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

```
TukeyHSD(aov(days~factor(physical_status), data = knee_data))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = days ~ factor(physical_status), data = knee_data)
##
## $`factor(physical_status)`
##                      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
```

```
summary(multcomp::glht(aov(days~factor(physical_status), data = knee_data),
               lincft = mcp(method = "Dunnett")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = days ~ factor(physical_status), data = knee_data)
##
## Linear Hypotheses:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0                      38.000      1.606  23.667   <0.001
## factor(physical_status)Average == 0   -5.000      2.154  -2.321   0.0681
## factor(physical_status)Above == 0    -14.429      2.350  -6.139   <0.001
##
## (Intercept) == 0                     ***
## factor(physical_status)Average == 0 .
## factor(physical_status)Above == 0    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

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

Using this data, we are interested in knowing whether someone's overall physical health status (e.g. below, at, or above average) impacts the amount of time needed in physical therapy in order to fully recover. Since we have three categories of physical fitness, we can use an ANOVA test to understand whether at least one of these groups has a significantly different distribution of physical therapy duration. According to this test, and at a 5% significant level, we conclude that in fact physical fitness does have an impact on the duration of physical therapy. Upon further examination, people in the 'above average' category spend a significantly less amount of time (number of days) in physical therapy, whereas there was no difference in the amount of time spent in physical therapy for those of average or below average health. Statisical results are provided above, as well as a figure depicting the distribution of physical therapy duration in each group.

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.

The below table gives the observed proportion of acceptances for men and women, where each proportion is calculated as $\hat{p} = \frac{\text{total number of acceptances}}{\text{total number of applications}}$.

Gender	Acceptances	Proportion of total applications
Female	557	0.304
Male	1198	0.445

To compute a 95% CI for the proportion estimate $\hat{p}$ in each group, we need to first establish that the the data follows a normal distribution ($np(1-p) \geq 5$). This is clearly true for men ($(n_m)(p_m)(1-p_m) = (2691)(0.445)(1-0.445) = 664.61 > 5$) and women ($(n_w)(p_w)(1-p_w) = (1835)(0.304)(1-0.304) = 388.26 > 5$), and so we can assume a normal distributions. Therefore, the 95% for any one group is given by

$$(\hat{p} - z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

For men, this is equal to

$$(0.445 - z_{0.975}\sqrt{\frac{0.445(1-0.445)}{2691}}, 0.445 + z_{0.975}\sqrt{\frac{0.445(1-0.445)}{2691}}) = (0.426, 0.464)$$

Thus we are 95% confident that between 42.6% and 46.4% of male applicants are admitted to UC Berkeley across departments.

For women, we have

$$(0.304 - z_{0.975}\sqrt{\frac{0.304(1-0.304)}{1835}}, 0.304 + z_{0.975}\sqrt{\frac{0.304(1-0.304)}{1835}}) = (0.283, 0.325)$$

And we are 95% confident that between 28.3% and 32.5% of female applicants are admitted.

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.

Because we're dealing with proportions in two populations (male and female), we can perform a two-sample test of proportions. Our null hypothesis is that the proportions of accepted men and women across all departments are the same, and our alternative hypothesis is that they are different;

$$H_0 : p_{men} - p_{women} = 0$$

$$H_1 : p_{men} - p_{women} \neq 0$$

```
## $t
## [1] -9.558269
##
## $p.val
## [1] 0
```