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

```
## [1] 2.4627
```

```
## [1] 3.36
```

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.

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

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

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.

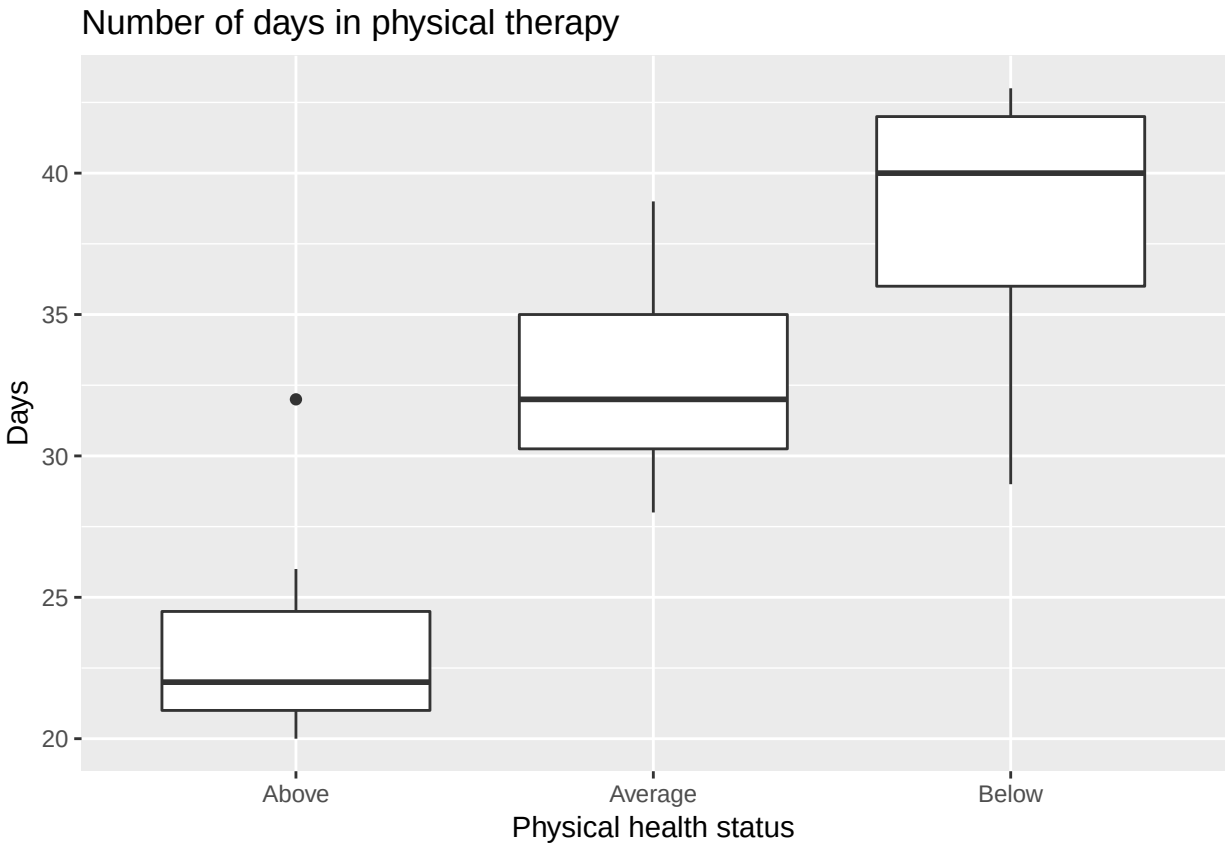
(b) Calculate the sample size for the new study. Assuming a two-sided test, create a table showing sample size estimates for 80% vs 90% power, 2.5% vs 5% significance level, using the following information: the true mean increase for smokers is 3.0 kg/m2, with a standard deviation of 2.0 kg/m2; for never-smokers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

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 1: 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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
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
##
##         Above   Average
## Average 0.0011  -
## Below   1.1e-05 0.0898
##
## 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-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
```

```
# need Dunnett test!!!
```

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.