Midterm 2 Lab Portion Solutions - R Studio

Dwight Wynne

November 5, 2018

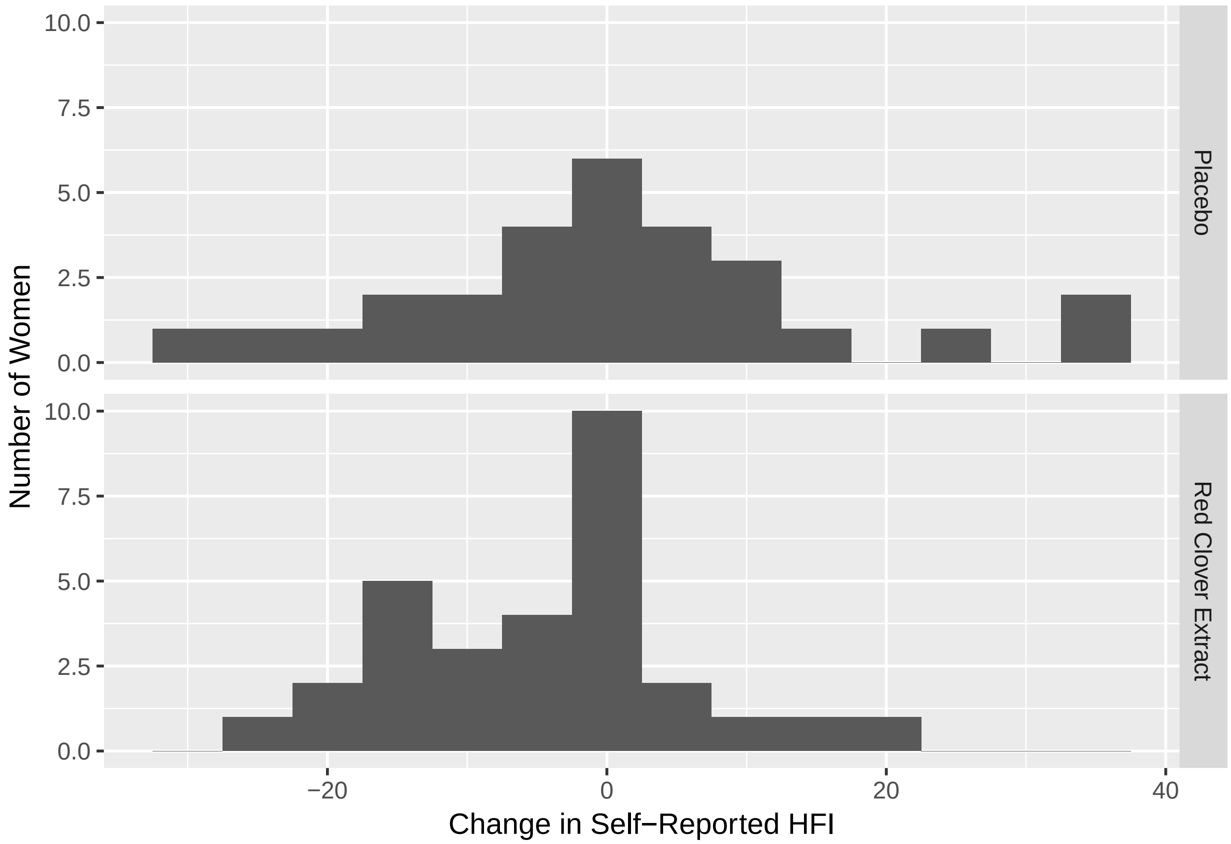
Problem 1

Post-menopausal women often develop sudden feelings of heat, skin redness, and sweating collectively known as hot flashes. Lambert and colleagues (2018) were interested in using Red Clover Extract to decrease the frequency and severity of these hot flashes in post-menopausal women. The rhfs.csv file on Titanium contains the self-reported change in hot flash intensity (HFI) over 12 weeks for their sample of 58 post-menopausal women.

Is there a significantly different effect of Red Clover Extract on hot flash intensity as compared to Placebo? Write 1-2 short paragraphs to answer this question. Support your answer by including and referring to software output.

We have a single numerical variable measured in two different, seemingly independent samples. We suspect that two-sample procedures are appropriate. Based on the figure below, or based on the combined sample size of 58, we suspect t-procedures are appropriate. There appear to be a few outliers in the Placebo group but nothing too extreme.

```r
library (ggplot2)
rhfs_hist <- ggplot(rhfs, mapping = aes(x = Change_in_HFI)) +
  geom_histogram(center = 5, binwidth = 5)
rhfs_hist_labeled <- rhfs_hist +
  labs(x = "Change in Self-Reported HFI", y = "Number of Women") +
  facet_grid(Group~.)  # histograms one on top of the other
print(rhfs_hist_labeled)
```



H_0: the two groups have the same population mean, $\mu_{RCE} = \mu_{placebo}$; or, the red clover extract has no effect on the mean change in hot flash intensity (compared to placebo)

H_a: the two groups have different population means; $\mu_{RCE} \neq \mu_{placebo}$; or, the red clover extract has an effect on the mean change in hot flash intensity (compared to placebo)

Significance level: $\alpha = 0.05$

We perform a two-sample t test:

```
t.test(Change_in_HFI ~ Group, data = rhfs)
```

```
##
##  Welch Two Sample t-test
##
## data:  Change_in_HFI by Group
## t = 1.3547, df = 46.8, p-value = 0.182
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.235166 11.449451
## sample estimates:
##           mean in group Placebo mean in group Red Clover Extract
##                       0.1071429                       -4.5000000
```

Because the p-value of $0.182 > 0.05$, we fail to reject the null hypothesis.

Conclusion: we did not find evidence of a significantly different effect of red clover extract on hot flash intensity as compared to placebo

Problem 2

Suppose a gas station chain is interested in showing that a competitor's gas prices are higher. One morning, they decide to send people out to record the Regular Unleaded gasoline price at 50 randomly selected competitor's stations throughout Southern California. The company knows the following:

- On this day, the company's own stations in Southern California sell Regular Unleaded for, on average, \$3.80 per gallon
- Previous studies have suggested that a sample standard deviation of \$0.15 per gallon is a reasonable assumption

Is 50 a large enough number of gas stations to show that the competitor's prices are higher, if the competitor's true average price is \$3.85 per gallon? Write a short report (1-2 sentences) answering this question. Support your answer by including and referring to software output.

We perform a one-sample t power analysis. The alternative hypothesis here is $\mu > 3.8$ since the gas station only cares if the competitor has higher prices.

```r
power.t.test(n = 50, delta = 3.85 - 3.80, sd = 0.15, type = "one.sample", alternative = "one.sided")
```

```
## 
##      One-sample t test power calculation 
## 
##               n = 50
##           delta = 0.05
##              sd = 0.15
##       sig.level = 0.05
##           power = 0.7515644
##     alternative = one.sided
```

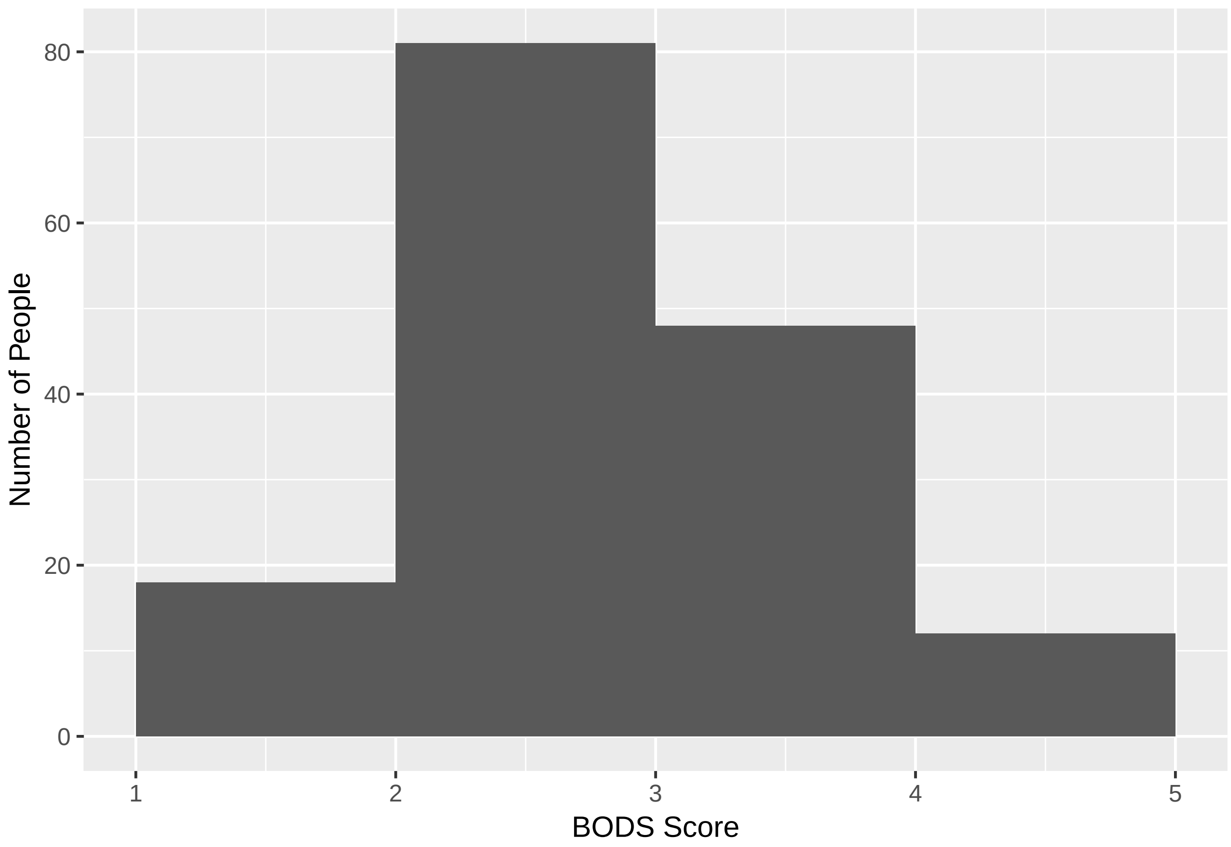
Given a significance level of 0.05, the power is 0.752, which is less than 0.8. We conclude that 50 stations is not a sufficiently large sample for the gas station to detect if the competitor is actually charging, on average, \$0.05 more.

Problem 3

They'll measure anything these days. The disgust.csv contains scores of 159 Americans on various scales that measure how disgusting they find germs, pathogens, and similar things. The variable BODS in the dataset represents people's disgust with body odor, normalized from 1 (not disgusting at all) to 5 (completely disgusted by it). Estimate with 95% confidence the mean score on this scale of body odor disgust if all Americans were to take the researchers' survey. Write a short sentence interpreting your estimate.

Although we have many different variables, we are only interested in one numerical variable, BODS. This suggests we are in the one-mean inference domain. With a sample size of 159, we are plenty good to use t-confidence intervals, but just in case you wanted to check:

```r
bods_hist <- ggplot(disgust, mapping = aes(x = BODS)) +
  geom_histogram(center = 1.5, binwidth = 1)
bods_hist_labeled <- bods_hist +
  labs(x = "BODS Score", y = "Number of People")
print(bods_hist_labeled)
```



The sample is skewed right, but not horribly so, and there are not any obvious outliers.

We perform the one-sample t confidence interval:

```
t.test(disgust$BODS, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  disgust$BODS
## t = 48.776, df = 158, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.767545 3.001135
## sample estimates:
## mean of x
##   2.88434
```

The 95% CI is (2.77, 3.00).

Interpretation: We are 95% confident that the population mean BODS score is between 2.77 and 3.00.

Alternative interpretation: We estimate the population mean BODS score to be 2.88. We are 95% confident that our estimate is within 0.059 of the actual population mean value.