Matthew L. Pergolski
IST 772
Dr. Block
11/8/2021

Homework 5

Beginning Statement

"I produced the material below with no assistance [direct quote from IST 772 class syllabus]." Note: Homework questions from the book may have been copied/pasted into the document for both the student and viewer's convenience.

The homework for week five is exercises 6 through 10 on pages 86 and 87.

Homework Question 6

Question:

6. The PlantGrowth data set contains three different groups, with each representing various plant food diets (you may need to type data(PlantGrowth) to activate it). The group labeled "ctrl" is the control group, while "trt1" and "trt2" are different types of experimental treatment. As a reminder, this subsetting statement accesses the weight data for the control group:

PlantGrowth$weight[PlantGrowth$group=="ctrl"]

and this subsetting statement accesses the weight data for treatment group 1:

PlantGrowth$weight[PlantGrowth$group=="trt1"]

Run a t-test to compare the means of the control group ("ctrl") and treatment group 1 ("trt1") in the PlantGrowth data. Report the observed value of t, the degrees of freedom, and the p-value associated with the observed value. Assuming an alpha threshold of .05, decide whether you should reject the null hypothesis or fail to reject the null hypothesis. In addition, report the upper and lower bound of the confidence interval.

Answer/Student Response:

The following code was generated for this problem:

```
#data
data(PlantGrowth)

#subset - control group
ctrl <- PlantGrowth$weight[PlantGrowth$group=="ctrl"]
ctrl

#subset - treatment group 1
trt1 <- PlantGrowth$weight[PlantGrowth$group=="trt1"]
trt1

#t.test
t.test(ctrl, trt1)
```

The following was observed as output from the t.test:

```
>     t.test(ctrl, trt1)

        Welch Two Sample t-test

data:  ctrl and trt1
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875162  1.0295162
sample estimates:
mean of x mean of y
    5.032     4.661
```

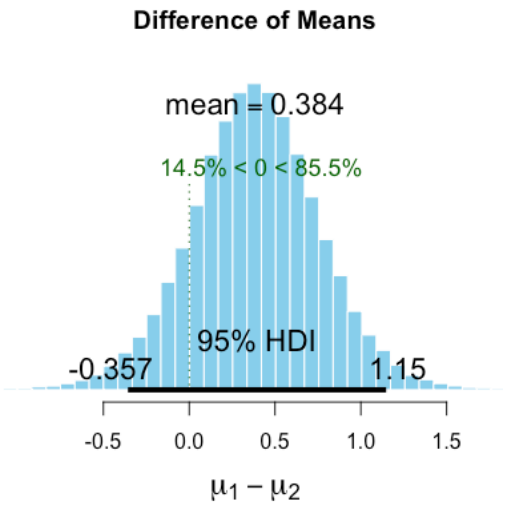
The observed value of t comes out to be 1.1913 while the degrees of freedom correspond to 16.524; the p-value equates to 0.2504. Based on our alpha value of 0.05, we fail to reject the null hypothesis that there is no difference between the two groups, since our p-value (0.2504) is greater than our alpha value (0.05). The upper and lower bound of the confidence interval is the following: -0.2875162 through 1.0295162.

Homework Question 7

Question:
7. Install and library() the BEST package. Note that you may need to install a program called JAGS onto your computer before you try to install the BEST package inside of R. Use BESTmcmc() to compare the PlantGrowth control group ("ctrl") to treatment group 1 ("trt1"). Plot the result and document the boundary values that BESTmcmc() calculated for the HDI. Write a brief definition of the meaning of the HDI and interpret the results from this comparison.

Answer/Student Response:
The following was observed as the output of the BESTmcmc() function (i.e., data and plot):

```
>     BESTmcmc(ctrl, trt1)
Waiting for parallel processing to complete...done.
MCMC fit results for BEST analysis:
100002 simulations saved.
         mean      sd  median  HDIlo  HDIup  Rhat n.eff
mu1     5.0282  0.2235  5.0286 4.5745  5.461 1.000 52111
mu2     4.6418  0.3060  4.6386 4.0423  5.261 1.000 52681
nu     34.3035 29.6046 25.7322 1.0361 92.525 1.001 20535
sigma1  0.6603  0.2041  0.6230 0.3465  1.064 1.002 24806
sigma2  0.8938  0.2764  0.8428 0.4539  1.449 1.001 25690

'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.
'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).
'n.eff' is a crude measure of effective sample size.
```



The HDI boundary values are -0.357 and 1.15. The HDI conveys to the reader that the population mean has a 95% probability of existing between the interval that is shown (-0.357 through 1.15). The point estimate, or 0.384, is our most likely candidate value for the population mean. We can see that, from the graph, that about 14.5 percent of values from this distribution are shown to be on the left hand side of the zero value (i.e., negative) while 85.5% of the values are in the positive realm.

Homework Question 8

Question:
8. Compare and contrast the results of Exercise 6 and Exercise 7. You have three types of evidence: the results of the null hypothesis test, the confidence interval, and the HD from the BESTmcmc() procedure. Each one adds something, in turn, to the understanding of the difference between groups. Explain what information each test provides about the comparison of the control group ("ctrl") and the treatment group ("trt1").

Answer/Student Response:
Within the null hypothesis test, NHT, we can determine if the difference between the two groups is significant. Our alpha value of 0.05 was provided and we observed a p-value, after the t.test was ran, of 0.2504. Because of this observation, we cannot reject the null hypothesis since our p-value is not less than the alpha value. As a side note, the null hypothesis is always assumed to conclude that no difference lies between the two comparison groups.

The confidence interval informs us of a possible mean difference over the long run; say, if the test was run 100 times, the population mean would likely show within the interval of 95 of those tests. This interval was determined to be -0.2875162 through 1.0295162.

Lastly, the BESmcmc() function allows us to have a clearer, more conclusive probability table/illustration of where the population mean lies. From this specific test, we saw that about 85% of values were above zero, and a similar HDI slot of -0.357 through 1.15. The likely population mean equates to 0.384, resulting in a small difference.

It's important to know that we cannot prove for sure the population mean difference is within the interval produced by these functions, but it is our best estimation. With that said, however, since we determined that our results were not statistically significant using the NHT, we can conclude that we fail to reject the null hypothesis that there is no (meaningful) difference between the two groups.

Homework Question 9

Question:
9. Using the same PlantGrowth data set, compare the "ctrl" group to the "trt2" group. Use all of the methods described earlier (t-test, confidence interval, and Bayesian method) and explain all of the results.

Answer/Student Response:
The following code and results were developed and observed:

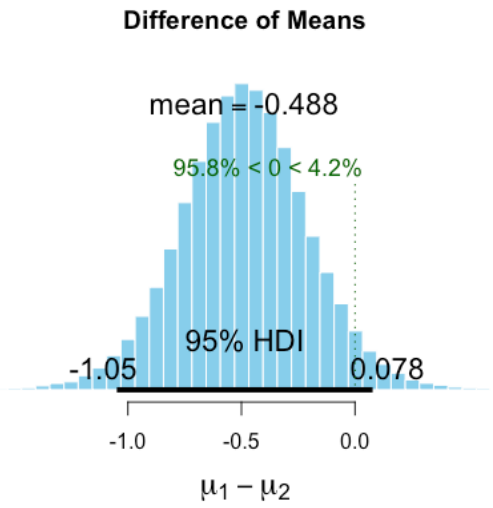
```
#ctrl group
ctrl

#trt2 group
trt2 <- PlantGrowth$weight[PlantGrowth$group=="trt2"]
trt2

#t.test
t.test(ctrl, trt2)
```

```
>     t.test(ctrl, trt2)

        Welch Two Sample t-test

data:  ctrl and trt2
t = -2.134, df = 16.786, p-value = 0.0479
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.98287213 -0.00512787
sample estimates:
mean of x mean of y
    5.032     5.526
```

We have a t value of t = -2.134 and a degrees of freedom value corresponding to df = 16.786; we also have a p-value of 0.0479, which is smaller -- barely -- than the alpha value of 0.05, meaning that our results are statistically significant and we can reject the null hypothesis that there is no difference between the two groups. Our confidence interval has a range of -0.98287213 and -0.00512787, meaning that -- over the long run -- our population mean would be estimated to correspond to a place within this range 95 out of 100 times.

From the Bayesian perspective, the following was developed and observed:

```
#bayesian method
BESTmcmc(ctrl, trt2)
plot(BESTmcmc(ctrl, trt2))
```



Within our distribution, we can conclude, with 95% confidence, that the true population mean difference lies within the interval range of -1.04 and 0.0755. From this, we see that 95.9% of the data within the distribution is negative, and that the most likely mean equates to -.0487, per the graph.

Homework Question 10

Question:
10. Consider this t-test, which compares two groups of n = 100,000 observations each:

t.test/(rnorm(100000,mean=17.1,sd=3.8),rnorm(100000,mean=17.2,sd=3.8))

For each of the groups, the rnorm() command was used to generate a random normal distribution of observations similar to those for the automatic transmission group in the mtcars database (compare the programmed standard deviation for the random normal data to the actual mtcars data). The only difference between the two groups is that in the first norm () call, the mean is set to 17.1 mpg and in the second it is set to 17.2 mpg. I think you would agree that this is a negligible difference, if we are discussing fuel economy. Run this line of code and comment on the results of the t-test. What are the implications in terms of using the NHST on very large data sets?

Answer/Student Response:
We seem to have a t value of t = -5.8156, degrees of freedom of df = 2e+05, and p-value of p-value = 6.05e-09 (meaning that our resutls are statistically significant and we reject the null hypothesis). We have a confidence interval of -0.13244696 -0.06567583, which is a relatively small range. We can conclude that, with large datasets, we will see statistically significant results for even the smallest of differences between the data will cause the t.tester to reject the null hypothesis that the two groups have no (meaningful) difference, which can be seen as unreliable.