Parin Patel

Homework 5

11/03/2019

# Attribution Statement:

1. Homework 4 by Parin Patel: I did this homework by myself, with help from the book and the professor. In addition, I used the following websites to help with a conversion from scientific notation to decimal :
   1. <http://www.easysurf.cc/scintd.htm>

# Exercises:

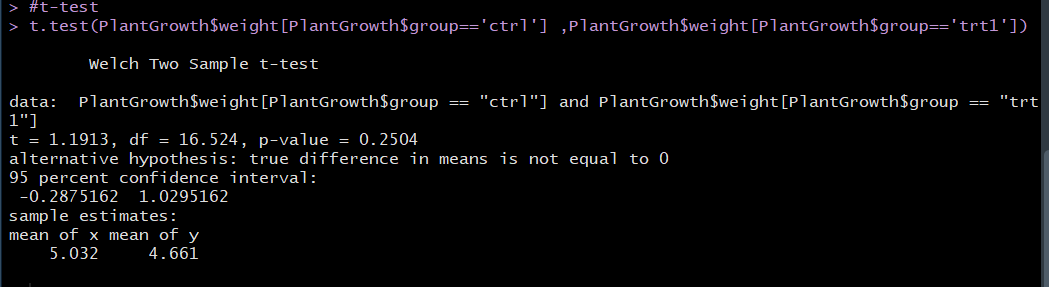
**6. The PlantGrowth data contains three different groups, with each representing various plant food diets. The group labeled “ctrl” is the control group, while “trt1” and “trt2” are different types of experimental treatment. As a reminder, this subsetting statement accesses the weight data for the control group:**

**PlantGrowth$weight[PlantGrowth$group=='ctrl']**

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

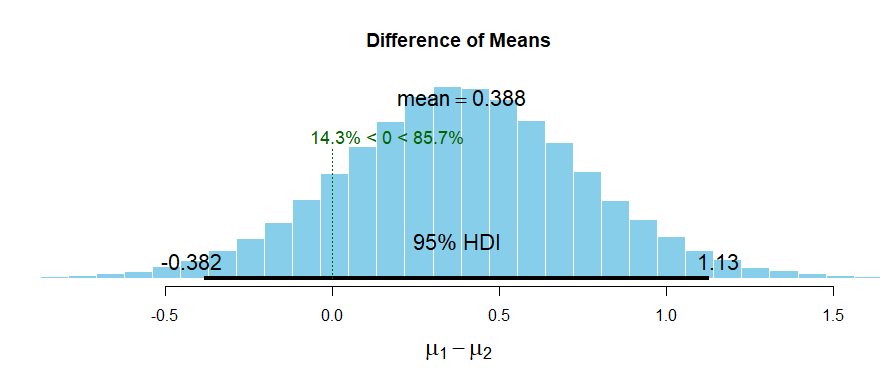
**PlantGrowth$weight[PlantGrowth$group=='trt1']**

**Run a t-test to compare the means of the control group (“ctrl”) and the treatment groups 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 the alpha threshold of 0.05 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.**



Assuming that the “alpha level” probability is 0.05 , the calculated p-value is 0.25 and the t test value is 1.913. Since the p-value is greater than the alpha level of 0.05, we fail to reject the null hypothesis. Our degrees of freedom is 16.5. The confidence interval has a min of -0.288 and as max of as 1.030.

**7. Install and library () the BEST package. Not 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.**



The BESTmcmc() calculated the boundary values for the highest density variable to range from -0.382 and 1.13. HDI, is the highest density interval. It is used to create an estimate on the population mean difference between two groups. Based on our graph, we can infer that there is a 95% probability that the population mean difference between the two groups lie in the bell-shaped area between -0.382 and 1.13. There is a greater likelihood that the population mean difference is somewhere in the region between 0.2 and 0.5; near a weight of 0.388. The likelihood of the population mean difference being 0 or higher is 85.7%.

**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 HDI 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 1 (“trt1”).**

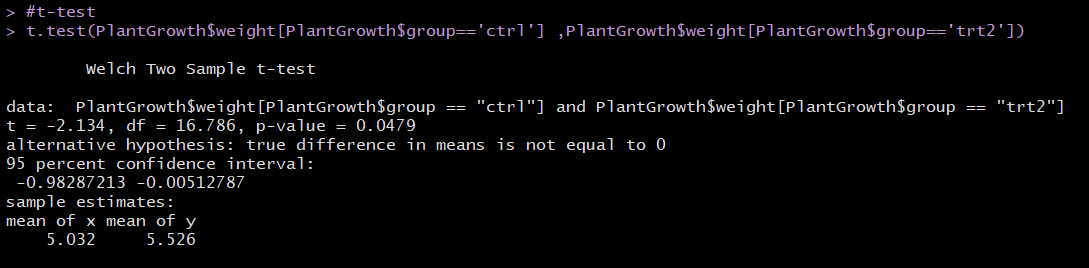
From the analysis we conducted in exercise 6, the 95 percent confidence interval ranged from 0.288 to a 1.030 lbs. The definition of a 95% confidence interval is that if we ran an experiment a large number of times (each time generating a new sample) then around 95% of the confidence intervals we created from the repetitions would actually contain the population mean difference. This statement is with regards to the long run possibilities of the confidence interval, and not about the accuracy of it. Additionally, in exercise 6, we calculated the T value to be 1.1913. The confidence interval assumes that the t-distribution makes a good model of difference between the population means estimated from the two independent samples. And while this is true for normal samples with no major outliers or other problems, it can be a bad model for samples with high skew, unequal variances, and other issues. Additionally, it’s important to note that we utilized Welch’s t-test for our confidence interval calculations, which does not require the assumption of equal variance between the populations, and therefore is an appropriate model to use for our dataset. However, if the sample did have an unequal variance between populations, it would be best to use the BESTmcmc() procedure. This is because the formula makes no special assumption about the shape of the posterior distribution. The result will be a stretched out HDI with raised tails that signal greater uncertainty in the samples.

The HDI stands for the highest density interval. It used to create an estimate on the population mean difference between two groups. Based on our graph n exercise 7, we can infer that there is a 95% probability that the population mean difference between the two groups lie in the bell-shaped area between -0.382 and 1.13. There is a greater likelihood that the population mean difference is somewhere in the region between 0.2 and 0.5; near a weight of 0.388. The likelihood of the population mean difference being 0 or higher is 85.7%.

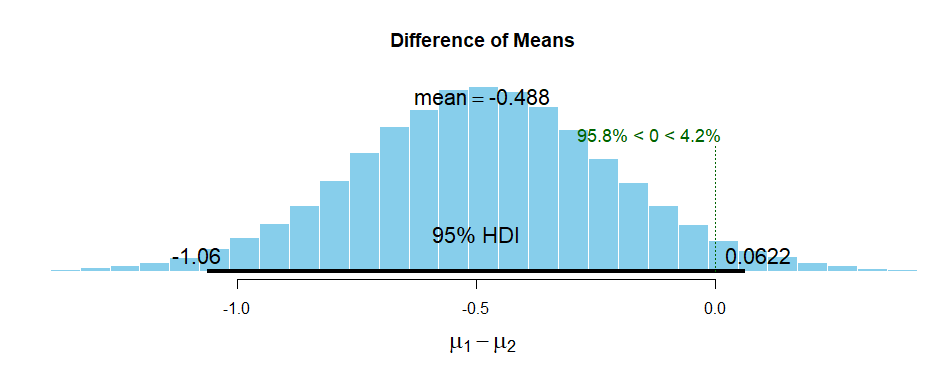
Unlike the confidence interval which uses the sample data to compute only one example of an upper and lower bound for the population mean, the BEST HDI model the population parameters of interest and shows us a probability distribution for those parameters. By showing a range of values, the HDI allows for more detail to be shown.

Finally, from exercise 6, we can utilize the the null hypothesis significance test (NHST) to understand if there’re is a mean difference between the mean of the two independent groups. By setting our alpha value to 0.05, we calculated our p-value to be 0.25. Sine our p is greater than the alpha level, we fail to reject the null hypothesis. This means that we cannot accept the null. But the p-value does not inform us on the likelihood of the null hypothesis occurring. Therefore, the test has not identified a consequential relationship between the two-independent means. This means that by rejecting the null, we are can express support for the alternative hypothesis of “true difference in means is not equal to 0”. And therefore, the difference in weight between the control and treatment 1 groups is not equal to zero.

**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.**



Assuming that the “alpha level” probability is 0.05 , the calculated p-value is 0.048 and the t test value is -2.134. Since the p-value is less than the alpha level of 0.05, we are able to reject the null hypothesis. Our degrees of freedom is 16.8. The confidence interval has a range from -0 0.98287213 to -0.00512787. Based on the confidence interval, we are able to view the range of values. Since we rejected the null, there is a definite likelihood of a consequential relationship between the two means. Therefore, we do not express support for the alternative hypothesis and instead can express that the difference in weight between the control and treatment 2 group is equal to zero.



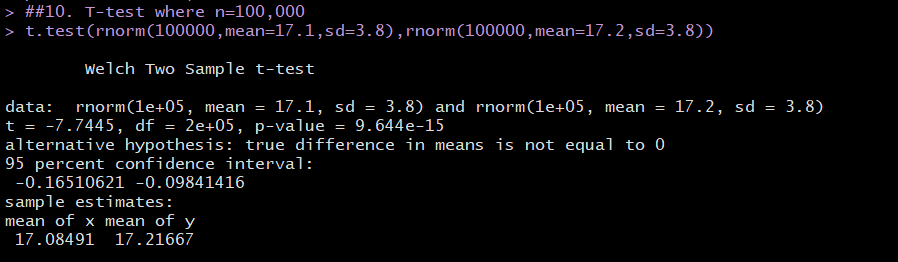
The BESTmcmc() calculated the boundary values for the highest density variable to range between -1.06 and 0.0622. Based on our graph, we can infer that there is a 95% probability that the population mean difference between the two groups lie in the bell-shaped area between -1.06 and 0.0622. There is a greater likelihood that the population mean difference is somewhere in the region between -0.4 and -0.5 near a weight of -0.488. The likelihood of the population mean difference being 0 or lower is 95.8%

This analysis is in line with our hypothesis of the difference in weight being equal to zero and the range of the confidence intervals falls around the HDI range.

**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 rnorm() 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 discuss‑ ing 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?**



Assuming that the “alpha level” probability is 0.05 , the calculated p-value is 9.644e-15 (or 0.000000000000009664) and the t test value is 7.7445. Since the p-value is less than the alpha level of 0.05, we are able to reject the null hypothesis. Our degrees of freedom are very high at 2e+05 (or 200,000). The confidence interval has a range from -0.16510621 to -0.09841416. Based on the confidence interval, we are able to view the range of values. Since we rejected the null, there is a definite likelihood of a consequential relationship between the two means. Therefore, we do not express support for the alternative hypothesis and instead can express that the difference in mpg between the control and treatment 2 group is equal to zero.

NHST, with regards to large datasets, may not always result in useful outcomes because large datasets where n is near or greater than 1,000 observations can lead to every difference becoming statistically significant, no matter how small the difference may be. Overall, the NHST uses a model that assumes zero difference between the two populations and looks for extreme outcomes. This makes assumptions highly questionable and is an issue for large datasets.

# Appendix A: Final Script

install.packages("animation")

install.packages("ggplot2")

library("animation")

library("ggplot2")

data("PlantGrowth")

PlantGrowth$weight[PlantGrowth$group=='ctrl'] #subsetting for weight data ctrl

PlantGrowth$weight[PlantGrowth$group=='trt1'] #subsetting for weight data trt1

##6. t-test to compare means of ctrl1 and trt1

#sample mean calc

CtrlMean<-mean( sample(PlantGrowth$weight[PlantGrowth$group=='ctrl'],size=10,replace=TRUE) ) #ctrl

Trt1Mean<-mean( sample(PlantGrowth$weight[PlantGrowth$group=='trt1'],size=10,replace=TRUE) ) #trt1

#mean difference

CtrlMean - Trt1Mean

#t-test

t.test(PlantGrowth$weight[PlantGrowth$group=='ctrl'] ,PlantGrowth$weight[PlantGrowth$group=='trt1'])

##7: use best library to cimoare the ctrl to trt1 group.

install.packages("BEST")

library("BEST")

PlantsBest <- BESTmcmc(PlantGrowth$weight[PlantGrowth$group=='ctrl'] ,PlantGrowth$weight[PlantGrowth$group=='trt1'])

#plot

plot(PlantsBest, main=NULL)

##9. compare the Ctrl and the trt2 group

#sample mean calc

CtrlMean<-mean( sample(PlantGrowth$weight[PlantGrowth$group=='ctrl'],size=10,replace=TRUE) ) #ctrl

Trt2Mean<-mean( sample(PlantGrowth$weight[PlantGrowth$group=='trt2'],size=10,replace=TRUE) ) #trt2

#mean difference

CtrlMean - Trt2Mean

#t-test

t.test(PlantGrowth$weight[PlantGrowth$group=='ctrl'] ,PlantGrowth$weight[PlantGrowth$group=='trt2'])

PlantsBest2 <- BESTmcmc(PlantGrowth$weight[PlantGrowth$group=='ctrl'] ,PlantGrowth$weight[PlantGrowth$group=='trt2'])

#plot

plot(PlantsBest2, main=NULL)

##10. T-test where n=100,000

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