HW5_Videtti

##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.

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

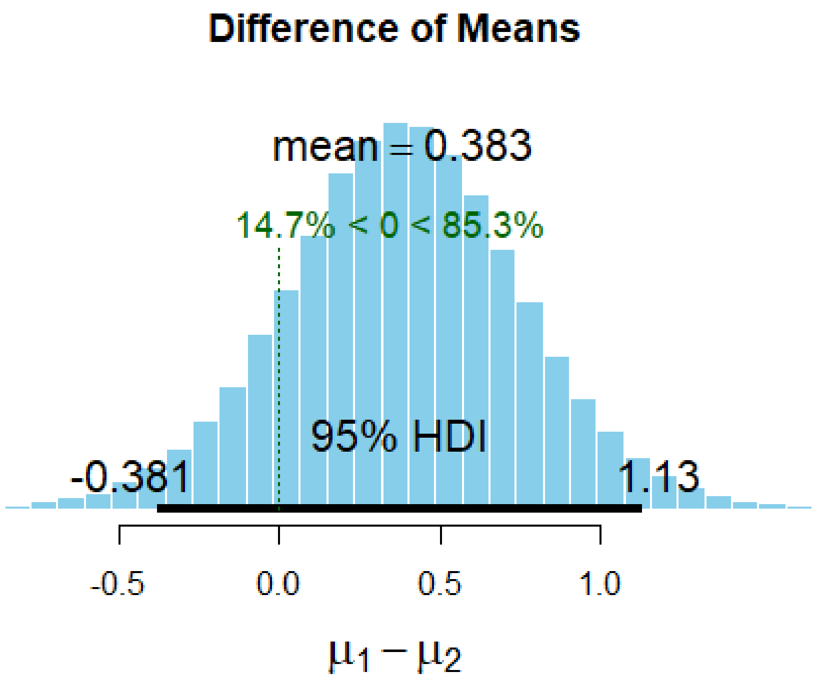
```
##
##  Welch Two Sample t-test
##
## data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and
PlantGrowth$weight[PlantGrowth$group == "trt1"]
## t = 1.1913, df = 16.524, p-value = 0.2504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2875162  1.0295162
## sample estimates:
## mean of x mean of y
##     5.032     4.661
```

#We See that the observed value of t is 1.9193, the degrees of freedom are 16.524, and the p-value is 0.2504. Given an alpha threshold of .05, there is not statistically significant evidence to reject the null hypothesis, thus we will fail to reject the null hypothesis. Lastly, the upper and lower bounds of the 95% confidence interval are 1.0295162 and -0.2875162, respectively.

##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.

```
#install.packages("BEST")
library(BEST)

## Loading required package: HDInterval

BESTmcmc(PlantGrowth$weight[PlantGrowth$group=='ctrl'],
PlantGrowth$weight[PlantGrowth$group=='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      5.0264  0.2253  5.0267 4.5728   5.471 1.000 57532
## mu2      4.6410  0.3069  4.6394 4.0330   5.250 1.000 53141
## nu      34.6493 30.1590 25.7515 1.1181  94.749 1.000 19787
## sigma1   0.6602  0.2011  0.6226 0.3392   1.059 1.001 27655
## sigma2   0.8988  0.2876  0.8455 0.4576   1.462 1.000 21399
##
## '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.

plot(BESTmcmc(PlantGrowth$weight[PlantGrowth$group=='ctrl'],
PlantGrowth$weight[PlantGrowth$group=='trt1']))

## Waiting for parallel processing to complete...done.
```



#We see that the boundary values for the 95% HDI are -0.381 and 1.13. HDI refers to the highest density interval, which gives us a range of values that have a specific chance of containing the population parameter, which in this case is the difference in population mean weights between the control group and the group receiving treatment 1. Putting that all together, we can conclude from our results in this Exercise that there is a 95% chance that the difference in the population mean weights between plants with no treatment and plants that receive treatment 1 is between -0.358 and 1.15

##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").

#The null hypothesis test told us we should fail to reject the null hypothesis. That is, we do not have statistically significant evidence to conclude that there is any difference in the population mean weights between plants with no treatment and plants that receive treatment 1.
#The 95% confidence interval tells us about uncertainty in any given estimate of the difference in the population mean weights between plants with no treatment and plants that receive treatment 1. Our 95% confidence interval was -0.2875162 to 1.0295162. In the long run, 95% of tests for the difference

in the population mean weights between plants with no treatment and plants that receive treatment 1 will have a 95% confidence interval that contains the difference in the population mean weights between plants with no treatment and plants that receive treatment 1.
#The 95% HDI gives us a range of values that the difference in the population mean weights between plants with no treatment and plants that receive treatment 1 is 95% likely to fall in. This means that the difference in the population mean weights between plants with no treatment and plants that receive treatment 1 is 95% likely to fall between -0.358 and 1.15. We see a larger proportion of values in the 95% HDI above 0, which indicates that we are likely to find that the mean weight among the control group is higher than the mean weight of the plants receiving treatment 1.

##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.

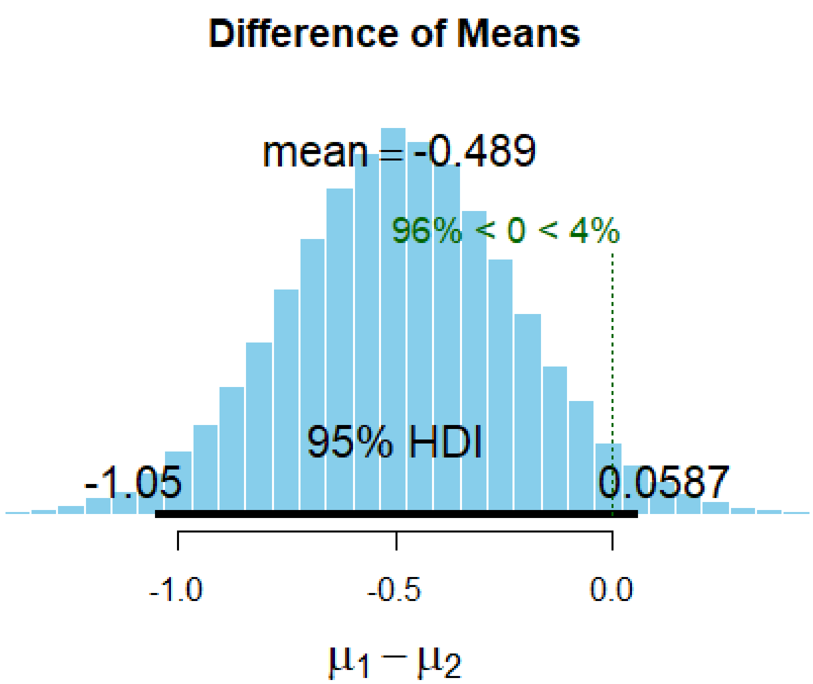
```
t.test(PlantGrowth$weight[PlantGrowth$group=='ctrl'],
PlantGrowth$weight[PlantGrowth$group=='trt2'])
```

```
##
##  Welch Two Sample t-test
##
## data:  PlantGrowth$weight[PlantGrowth$group == "ctrl"] and
PlantGrowth$weight[PlantGrowth$group == "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 See that the observed value of t is -2.134, the degrees of freedom are 16.786, and the p-value is 0.0479. Given an alpha threshold of .05, there is statistically significant evidence to reject the null hypothesis, thus we will indeed reject the null hypothesis. Lastly, the upper and lower bounds of the 95% confidence interval are -0.00512787 and -0.98287213, respectively.

#The null hypothesis test told us we should reject the null hypothesis. That is, we do have statistically significant evidence to conclude that there is a difference in the population mean weights between plants with no treatment and plants that receive treatment 2.
#The 95% confidence interval tells us about uncertainty in any given estimate of the difference in the population mean weights between plants with no treatment and plants that receive treatment 2. Our 95% confidence interval was -0.98287213 to -0.00512787. In the long run, 95% of tests for the

```
difference in the population mean weights between plants with no treatment
and plants that receive treatment 2 will have a 95% confidence interval that
contains the difference in the population mean weights between plants with no
treatment and plants that receive treatment 2.

plot(BESTmcmc(PlantGrowth$weight[PlantGrowth$group=='ctrl'],
PlantGrowth$weight[PlantGrowth$group=='trt2']))

## Waiting for parallel processing to complete...done.
```




```
#We see that the boundary values for the 95% HDI are -1.05 and 0.0587. HDI
refers to the highest density interval, which gives us a range of values that
have a specific chance of containing the population parameter, which in this
case is the difference in population mean weights between the control group
and the group receiving treatment 2. Putting that all together, we can
conclude from our results in this Exercise that there is a 95% chance that
the difference in the population mean weights between plants with no
treatment and plants that receive treatment 2 is between -1.05 and 0.0766. We
see a larger proportion of values in the 95% HDI below 0, which indicates
that we are likely to find that the mean weight among the control group is
lower than the mean weight of the plants receiving treatment 2.
```

```
##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 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?

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

##
##  Welch Two Sample t-test
##
## data:  rnorm(1e+05, mean = 17.1, sd = 3.8) and rnorm(1e+05, mean = 17.2,
sd = 3.8)
## t = -6.2547, df = 199991, p-value = 3.991e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.13946414 -0.07291339
## sample estimates:
## mean of x mean of y
##  17.08607  17.19226

#We See that the observed value of t is -6.2547, the degrees of freedom are
199991, and the p-value is 0.0000000003991. Given an alpha threshold of .05,
there is statistically significant evidence to reject the null hypothesis
that there is a difference in the population mean difference between the two
random samples. While there is technically a difference between the
population means, we agreed that the difference is negligible, so saying
there is a difference may be misleading. This is an implication of using the
NHST on very large data sets, where we find conclusions that may be
misleading, even if they are true. This is also a flaw of the NHST in
general, as it is simply a "go / no-go" test and does not give very specific
detail.
```