Homework 5

Kwasi Mensah

2022-11-09

Author: Kwasi Mensah Homwork\_Number: 5 Due Date: “2022-12-11” Output: pdf\_document Attribution statement:

I did the homework by myself, with help from the book and the professor #R Markdown #Run these three functions to get a clean test of homework code

dev.off() #Clear the graph window

## null device   
## 1

cat('\014') #Clear the console

rm(list = ls()) #Clear user objects from the environment

#R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: ##############################################################

1. 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 “trtl” and “trt2” are different types of experimental treatment. As a reminder, this subsetting statement accesses the weight data for the control group:

PlantGrowthgroup==’ctrl’]

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

PlantGrowthgroup==’ctrl’]

Run a t-test to compare the means of the control group (“ctrl”) and treatment group 1 (“trtl”) 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.

# Load the Plant Growth data set  
data(PlantGrowth)  
?PlantGrowth

## starting httpd help server ... done

# Break the weights down by group  
ctrl.weights = PlantGrowth$weight[PlantGrowth$group == 'ctrl']  
trt1.weights = PlantGrowth$weight[PlantGrowth$group == 'trt1']  
trt2.weights = PlantGrowth$weight[PlantGrowth$group == 'trt2']  
  
  
# T-test between treatment1 and the control group.  
t.test(ctrl.weights,trt1.weights)

##   
## Welch Two Sample t-test  
##   
## data: ctrl.weights and trt1.weights  
## 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 t value is 1.1913, the degree of freedom is 16.524, the p-value is .2504. When assuming there is a alpha threshold of .05, we would decide to fail to reject the null hypothesis. This is due to the p-value being .2504, which is not less than or equal to .05, so it is not statically significant. The confidence interval is -.2875162 to 1.0295162.

1. 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 (“trtl”). Plot the result and document the boundary values that BESTmcmc() calculated for the HDI. Write a brief definition of the meaning of the HD’ and interpret the results from this comparison.

library(rjags)

## Warning: package 'rjags' was built under R version 4.2.2

## Loading required package: coda

## Linked to JAGS 4.3.1

## Loaded modules: basemod,bugs

library(BEST)

## Warning: package 'BEST' was built under R version 4.2.2

## Loading required package: HDInterval

## Warning: package 'HDInterval' was built under R version 4.2.2

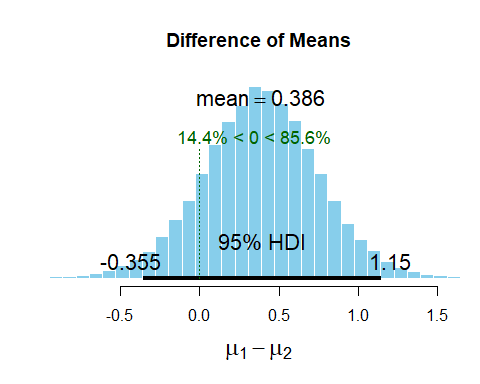
PGBest <- BESTmcmc(ctrl.weights,trt1.weights)

## Compiling model graph  
## Resolving undeclared variables  
## Allocating nodes  
## Graph information:  
## Observed stochastic nodes: 20  
## Unobserved stochastic nodes: 5  
## Total graph size: 59  
##   
## Initializing model  
##   
##   
## Sampling from the posterior distributions:

PGBest

## MCMC fit results for BEST analysis:  
## 100002 simulations saved.  
## mean sd median HDIlo HDIup Rhat n.eff  
## mu1 5.0274 0.2262 5.0274 4.5687 5.465 1.000 51521  
## mu2 4.6419 0.3054 4.6394 4.0456 5.256 1.000 52281  
## nu 34.6328 29.7718 25.9807 1.1699 94.143 1.000 19559  
## sigma1 0.6615 0.2045 0.6223 0.3381 1.066 1.001 24898  
## sigma2 0.8943 0.2782 0.8440 0.4533 1.441 1.000 26107  
##   
## '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(PGBest)



# The boundary values calculated of the 95% Highest Density Interval or HDI ranges are from -.355 to 1.15. There is a 95% probability that the population mean difference between the two groups falls within the given range. # The Highest Density Interval

1. 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 (“trtl”).

#The null hypothesis test has t-value of 1.1913 and a p-value of .2504, which is the probability of the t-value will occur. Also due to the p-value not being less than or equal to .05 we will fail to reject the null hypothesis. The confidence interval is -0.2875162 and 1.0295162, which tells us that is we do 100 replications we are 95 % sure that the population mean may have a chance of falling between the specified range and 5% sure that wouldn't fall within the range. The confidence interval passes 0 which means that the results are not statically significant. Regarding the confidence interval its results show the long run possibilities, not about the accuracy of this particular confidence interval. The Highest Density interval result show that there is a 95 % probability that the population mean will fall between the specified range in the graph. The range of the population mean is between -.355 and 1.15, which also passes 0. The population mean is most likely to be .386. The population value could be 0, and there is probability that there is no difference between the means of the control group and the treatment1 group.

1. 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.

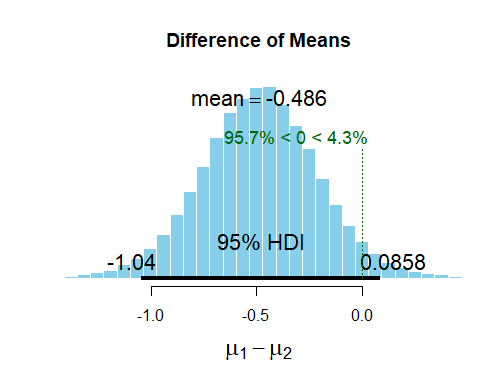
t.test(ctrl.weights,trt2.weights)

##   
## Welch Two Sample t-test  
##   
## data: ctrl.weights and trt2.weights  
## 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

PGBest2 <- BESTmcmc(ctrl.weights,trt2.weights)

## Compiling model graph  
## Resolving undeclared variables  
## Allocating nodes  
## Graph information:  
## Observed stochastic nodes: 20  
## Unobserved stochastic nodes: 5  
## Total graph size: 59  
##   
## Initializing model  
##   
##   
## Sampling from the posterior distributions:

plot(PGBest2)



# The null hypothesis test, the t-value is -2.134. The degree of freedom is 16.786. Our analysis has a p-value of .0479 meaning that we should reject the null hypothesis, and. If we do 100 replications, we are 95% confident that the population mean may be between the range of, -0.98287213 and -0.00512787. The boundary values calculated of the 95% Highest Density Interval or HDI ranges are from -1.04 to .0858. There is a 95% probability that the population mean difference between the two groups falls within the given range. The histogram shows that most likely the value of the mean is -.486. although the results show that the analysis is statically significant, it is best to still run multiple iterations of the experiment but with larger sample sizes. This will help us have a stronger statistically significant results and help to reduce the size of the confidence interval ranges.

1. 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?

set.seed(111)  
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 = -6.4831, df = 199991, p-value = 9.004e-11  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.14388842 -0.07708418  
## sample estimates:  
## mean of x mean of y   
## 17.08820 17.19869

# When we use larger sample sizes most differences between groups whether large or small, will result as being statistically significant. The p-values of the two analyses were both less than .05 and awfully close to 0. The confidence interval range is between -0.14388842 and -0.07708418. Although the means are not known we could assume that they would be different between the two groups of n=100,000 obersavtions. We believe that there might be a chance that the the population mean value may fall between the confidence interval given range.