IST772 Problem Set 5

Abhijith Anil Vamadev

The homework for week five is based on exercises 6 through 10 on pages 86 and 87, but with changes as noted in the text in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 1. I did this homework by myself, with help from the book and the professor

# Chapter 5, Exercise 6

*The built‐in warpbreaks data set (see “? warpbreaks” for documentation) contains data for the number of warp thread breaks per loom with different amounts of tension (we will not consider the type of wool). The tensions are labelled “L”, “M” or “H” for low, medium and high tension. As a reminder, these subsetting statement accesses the breaks data for the low and medium tension:*

lower <- warpbreaks$breaks[warpbreaks$tension=="L"]  
med <- warpbreaks$breaks[warpbreaks$tension=="M"]  
high <- warpbreaks$breaks[warpbreaks$tension=="H"]

*Run a t‐test to compare the means of the breaks for medium tension (“M”) and breaks for high tension (“H”) in the warpbreaks data. Report the observed value of t, the degrees of freedom, and the p‐value associated with the observed value (1 pt). Assuming an alpha threshold of .05, decide whether you should reject the null hypothesis that the means are equal or fail to reject the null hypothesis (1 pt). In addition, report the upper and lower bound of the confidence interval (1 pt).*

t.test(warpbreaks$breaks[warpbreaks$tension=="M"], warpbreaks$breaks[warpbreaks$tension=="H"])

##   
## Welch Two Sample t-test  
##   
## data: warpbreaks$breaks[warpbreaks$tension == "M"] and warpbreaks$breaks[warpbreaks$tension == "H"]  
## t = 1.6199, df = 33.74, p-value = 0.1146  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.203597 10.648042  
## sample estimates:  
## mean of x mean of y   
## 26.38889 21.66667

* The t value is 1.619 and the degrees of freedom is 33.74, with a p-value of 0.1146. As the p-value is greater than the alpha (0.05) we fail to reject the null hypothesis. The upper and lower bound of the confidence interval are 10.68 and -1.203.

# Chapter 5, Exercise 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 warp breaks for medium tension (“M”) to breaks for high tension (“H”) (1 pt). Plot the result and document the boundary values that BESTmcmc() calculated for the HDI (1 pt). Write a brief definition of the meaning of the HDI and interpret the results from this comparison (1 pt).*

#The BESTmcmc wasn't working so, I did it online.

* HDI there is a 95% probability that the population mean difference between This HDI has a simple and intuitive interpretation: there is a 95% probability that the population mean difference between the two groups falls within this range.
* Medium tension has more breaks.

# Chapter 5, Exercise 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 tensions. Explain what information each test provides about the comparison of the breaks for medium tension and for high tension (1 pt for each report and interpretation).* \* means are equal - Null hypothesis. If the null hypothesized value is found in our confidence interval, then that would mean we have a bad confidence interval and our p-value would be high. Typically our null hypothesized value will be 0 (point of no difference), and if we find 0 in our confidence interval then that would mean we have a good chance of actually finding NO DIFFERENCE, which is typically the opposite of what we want. \* This is a statement about confidence intervals over the long run, not the particular one that we calculated. In contrast, the HDI is built up gradually from more than 100,000 steps in our Markov chain Monte Carlo process, with each step depicting a possible combination of the population parameters. A key distinction to reinforce:

* the BEST HDI directly models the population parameters of interest and shows us probability distributions for those parameters, whereas the confidence interval uses sample data to compute one and only one example of an upper and lower bound for the population mean.

# Chapter 5, Exercise 9

*Using the same warpbreaks data set, compare the breaks for medium tension to the breaks for low tension (“L”). Use all of the methods described earlier (t‐test, confidence interval, and Bayesian method) and explain all of the results (1 pt for each report and interpretation).*

t.test(warpbreaks$breaks[warpbreaks$tension=="M"], warpbreaks$breaks[warpbreaks$tension=="L"])

##   
## Welch Two Sample t-test  
##   
## data: warpbreaks$breaks[warpbreaks$tension == "M"] and warpbreaks$breaks[warpbreaks$tension == "L"]  
## t = -2.256, df = 26.554, p-value = 0.03252  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -19.1023204 -0.8976796  
## sample estimates:  
## mean of x mean of y   
## 26.38889 36.38889

* lower is better.
* P-value is 0.0325, since it is lower than alpha, we reject the null-hypothesis.
* The confidence interval is -19.103 and -0.897, which does not include 0, meaning we have a good chance of finding a difference between the two means.
* HDI is -10.3 and -0.776, with 100% < 0 < 0.00%, there is a 95% probability that the population mean difference falls between the HDI values.
* the effect size HDI does not overlap with zero, this is suggestive of a credible difference between means.

# Chapter 5, Exercise 10

*Consider this t‐test, which compares two datasets of n = 100,000 observations each:*

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 = -6.4981, df = 2e+05, p-value = 8.15e-11  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.14357363 -0.07703387  
## sample estimates:  
## mean of x mean of y   
## 17.09272 17.20303

*For rnorm() command was used to generate a random normal distribution of observations. The only difference between the two is that in the first rnorm() call, the mean is set to 17.1 and in the second it is set to 17.2. I think you would agree that this is a negligible difference, if we are discussing breaks. Run this line of code and comment on the results of the t‐test (1 pt). What are the implications in terms of using the NHST on very large data sets (e.g., is it a good idea)? (1 pt)*

*The t-value is high and the p-value is a lot smaller than the alpha, hence we reject the null hypothesis. The Confidence interval at 95% are -0.15 and -0.09, 0 doesn’t fall in the range of the confidence interval and there a chances of finding a difference between the two means.*  Increasing sample size, decreases the p-value. The larger the size of the sample, larger the difference between the conditions and larger the t-statistic and smaller the p-value. The t-statistic will grow larger with more sample size, which will reduce the p-value. Hence, no it isn’t a good idea to use NHST on large data sets.