

IST772– Problem Set 5

Sathish Kumar Rajendiran

Attribution statement: 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:

```
warpbreaks$breaks[warpbreaks$tension=="L"]
warpbreaks$breaks[warpbreaks$tension=="M"]

summary(warpbreaks)

summary(warpbreaks$breaks[warpbreaks$tension=="L"])

summary(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 to compare the means of medium and high tension in the warpbreaks data

- From the t-test above,
 - critical value t^* is 1.6199.
 - degrees of freedom is 33.74. It suggests the number of degree of freedom lost from each sample as a result of calculating the sameple mean.
 - p-value is 0.1146; p-value suggests that the probability of obtaining a value of t atleast as high as what was observed.
 - $t(33.74)=1.6199$, p-value =0.1146 is greater than alpha value (0.05) - it shows that the difference is not statistically significant and we fail to reject the null hypothesis.However, it doesnt mean that we accept the null hypothesis ,rather than we have no good evidence either way.
 - 95 percent confidence interval is between -1.203597 (lower band) and 10.648042 (higher band) which includes 0

- 95% confidence suggests that if we ran this experiment a large number of times (each time generating a new sample), then 95% of the confidence intervals constructed from those repetitions would actually contain the population mean difference.

```
# cat("Summary of breaks for Medium tension:\n")
# summary(warpbreaks$breaks[warpbreaks$tension=="M"])
# cat("\nSummary of breaks for High tension:\n")
# summary(warpbreaks$breaks[warpbreaks$tension=="H"])

#t-test
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
```

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).

BESTmcmc() to compare the means of medium and high tension in the warpbreaks data

- Please find details from R Console

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

## Loading required package: HDInterval

warpbreaksBest <- BESTmcmc(warpbreaks$breaks[warpbreaks$tension=="M"],warpbreaks$breaks[warpbreaks$tension=="H"])

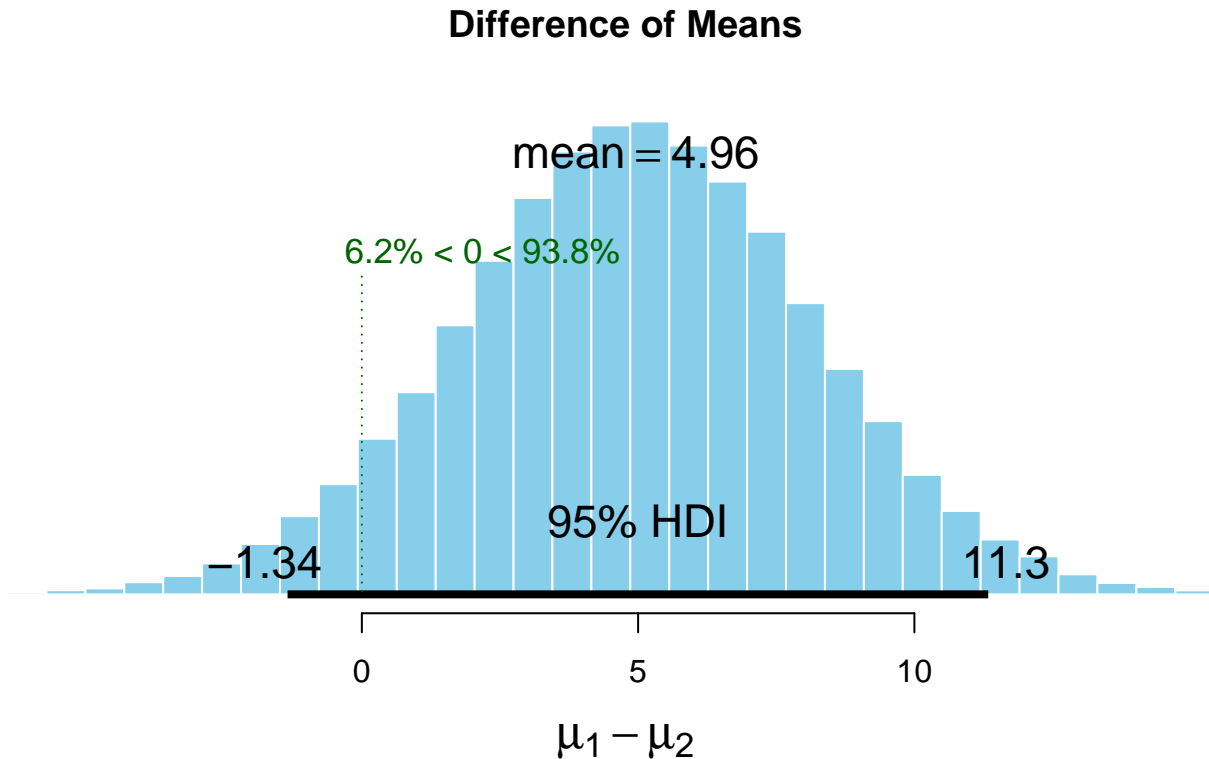
## Waiting for parallel processing to complete...

## done.
```

```
warpbreaksBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##      mean      sd median HDIlo HDIup  Rhat n.eff
## mu1    26.284  2.423 26.287 21.544 31.08 1.000 58868
## mu2    21.327  2.125 21.307 17.124 25.53 1.000 55247
## nu     34.961 29.724 26.286  1.653 94.25 1.001 18762
## sigma1  9.650  1.920  9.391  6.362 13.53 1.000 37879
## sigma2  8.477  1.796  8.252  5.296 12.15 1.000 36096
##
## '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.
```

```
#Bayesian probability distribution of the differences between two means
plot(warpbreaksBest)
```



```
diffBest <- warpbreaksBest$mu1 - warpbreaksBest$mu2
# summary(diffBest)
quantile(diffBest, c(0.025, 0.5, 0.975))
```

```
##      2.5%      50%      97.5%
## -1.419042  4.971410 11.269376
```

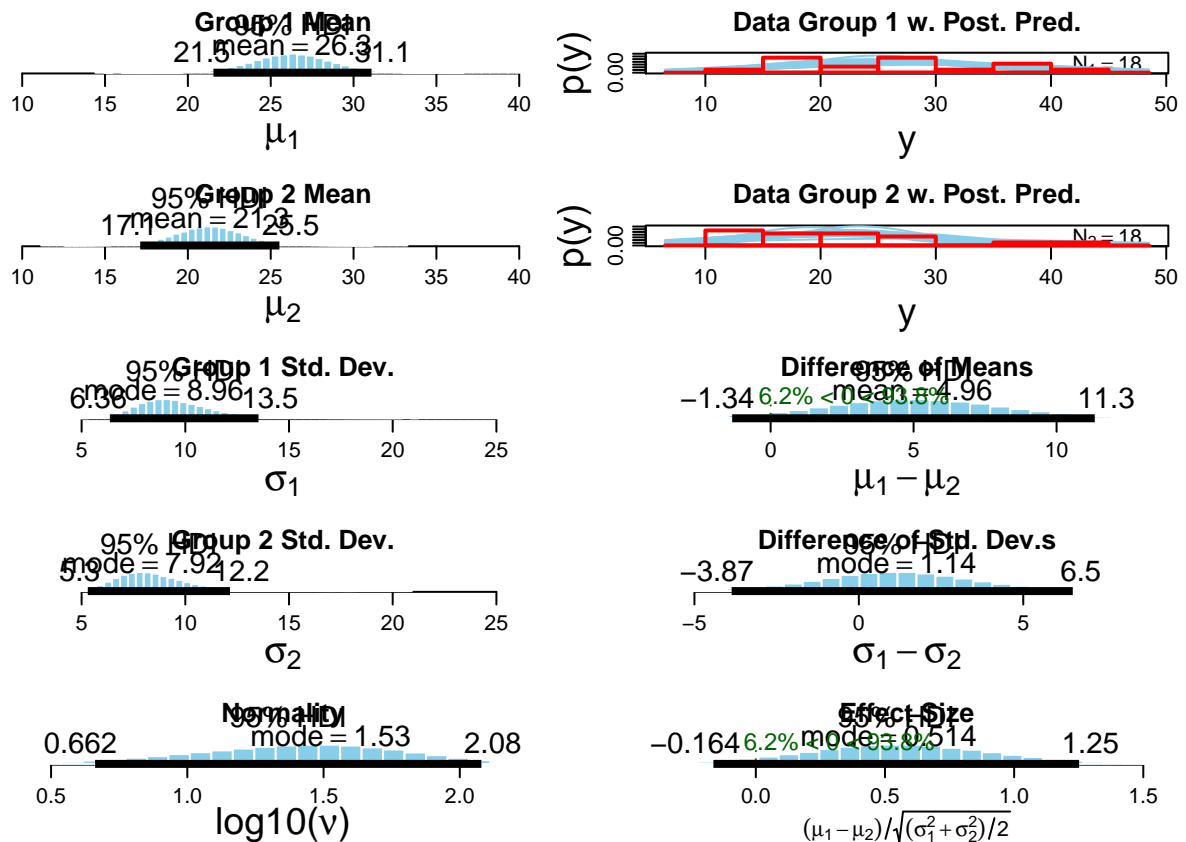
From the BESTmcmc() plot above, * HDI
* 95% Highest density interval (HDI) lie in the bell-shaped area between -1.388279 (HDIlo) and 11.286366 (HDIup); having mean value as 4.981 with 50% quantile (median) as 4.993

- HDI Interpretation
 - HDI is built from Markov chain Monte Carlo process with more than 100,000 simulations of mean difference between medium and high tension warpbreaks data
 - Lower and upper limits of 95% HDI suggests that the likelihood of population mean difference of 0 or larger is between 6% to 94%.
 - Difference between $\mu_1 - \mu_2$ is 4.98

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).

```
plotAll(warpbreaksBest)
```



- Null hypothesis significance test(NHST)
 - Based on the comparison between the means for medium and high tension, $t(33.74) = 1.6199$, $p\text{-value} = 0.1146$ is greater than alpha value (0.05) - it shows that the difference is not statistically significant and we fail to reject the null hypothesis. However, it doesn't mean that we accept the null hypothesis, rather than we have no good evidence either way.
- Confidence Interval (CI) -

- 95% confidence interval is between -1.203597 (lower band) and 10.648042 (higher band) which includes 0.
- 95% confidence suggests that if we ran this experiment a large number of times (each time generating a new sample), then 95% of the confidence intervals constructed from those repetition would actually contain the population mean difference.
- Bayesian distribution- Highest density interval (HDI)
 - BESTmcmc() way to compare the warp breaks for medium tension (“M”) to breaks for high tension (“H”) ran with 100,002 simulations suggests that the highest density interval (HDI) lie in the bell-shaped area between -1.388279 (HDIlo) and 11.286366 (HDIup). It includes 0. Range is slightly higher than the confidence interval.
 - mean value from HDI plot is 4.98
 - From plotAll(warpbreaksBest), we can see the Group 2 has better distribution and denser packaged standard deviation.

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 to compare the means for medium and low tension
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

#Bayesian distribution (HDI)
warpbreaksLBest <- BESTmcmc(warpbreaks$breaks[warpbreaks$tension=="M"],warpbreaks$breaks[warpbreaks$tension=="L"])

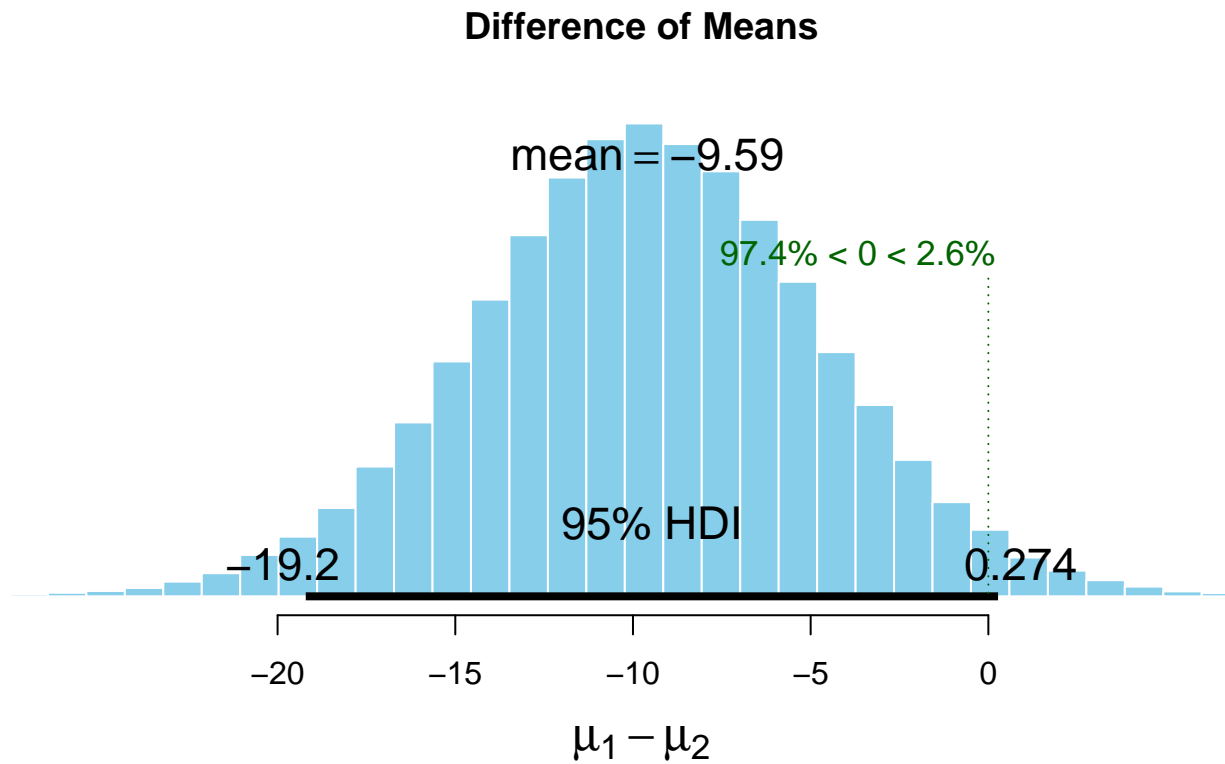
## Waiting for parallel processing to complete...done.

warpbreaksLBest

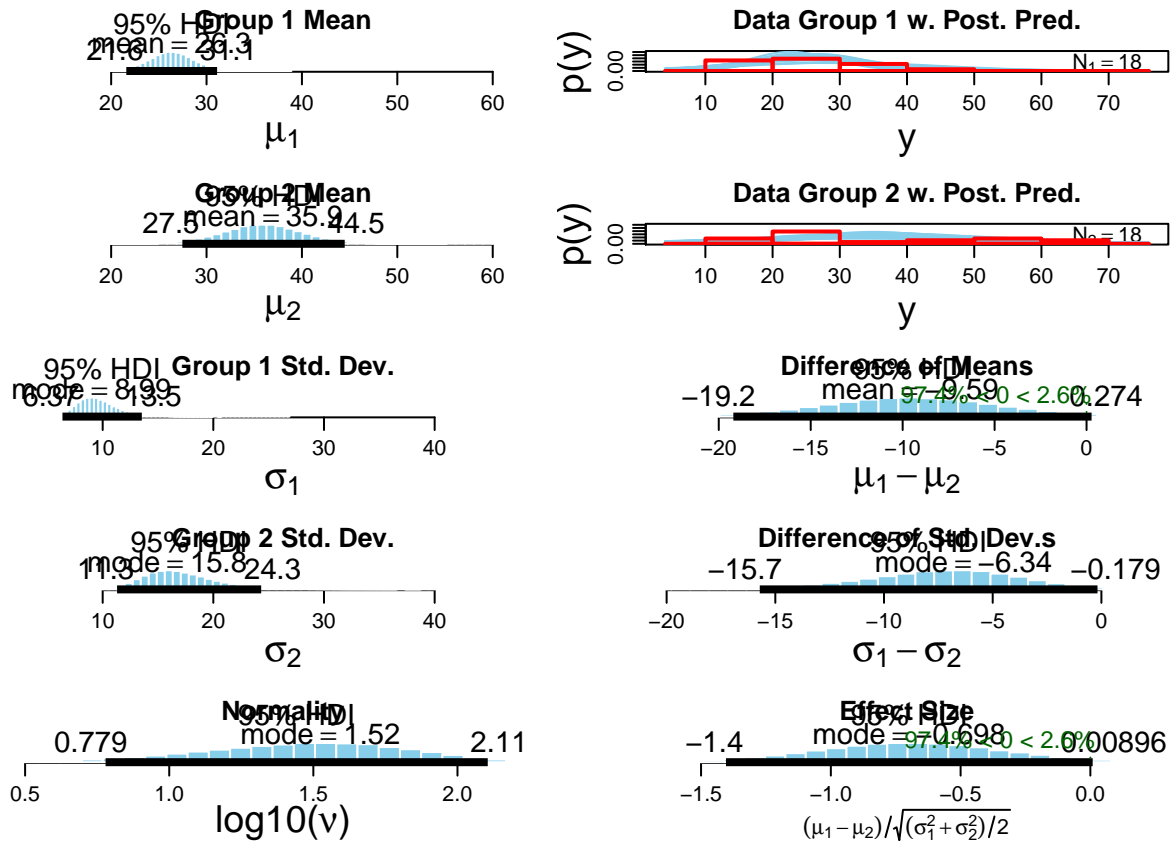
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##      mean      sd median HDIlo HDIup Rhat n.eff
## mu1   26.321  2.407 26.314 21.587 31.11   1 59875
## mu2   35.915  4.305 35.889 27.465 44.48   1 57080
## nu    38.938 30.467 30.631  2.030 99.60   1 22412
## sigma1 9.684  1.918  9.416  6.368 13.54   1 40165
```

```
## sigma2 17.229  3.475 16.750 11.302 24.32    1 38513
##
## '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 HDI
plot(warpbreaksLBest)
```



```
plotAll(warpbreaksLBest)
```



```
diffLBest <- warpbreaksLBest$mu1 - warpbreaksLBest$mu2
# summary(diffLBest)
quantile(diffLBest, c(0.025, 0.5, 0.975))
```

```
##          2.5%          50%          97.5%
## -19.3844006  -9.5882363   0.1111023
```

From above to compare the means for medium and low tension,

- Null hypothesis significance test(NHST) (null hypothesis : there is no difference between the means for medium and low tension)
 - Based on the comparison between the means for medium and low tension, $t(26.554) = -2.256$, $p\text{-value} = 0.03252$ is less than alpha value (0.05) - it shows that the difference is statistically significant and we may reject the null hypothesis. However, if we lower the alpha level to 0.01 then the mean difference won't be statistically significant and that we may fail to reject null hypothesis.
- Confidence Interval (CI) -
 - 95% confidence interval is between -19.1023204 (lower band) and -0.8976796 (higher band) which doesn't include 0.
 - 95% confidence suggests that if we ran this experiment a large number of times (each time generating a new sample), then 95% of the confidence intervals constructed from those repetitions would actually contain the population mean difference.
- Bayesian distribution- Highest density interval (HDI)

- BESTmcmc() way to compare the warp breaks for medium tension (“M”) to breaks for high tension (“L”) ran with 100,002 simulations suggests that the highest density interval (HDI) lie in the bell-shaped area between -19.348758855 (HDIlo) and 0.002098449 (HDIup). It doesnt include 0. Range is slightly higher than the confidence interval.
- From plotAll(warpbreaksBest), we can see the Group 1 has better distribution and densed packaged standard deviation with mean value as 26.3

Chapter 5, Exercise 10

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

```
options(scipen=999) # to handle scientific notions
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(100000, mean = 17.1, sd = 3.8) and rnorm(100000, mean = 17.2, sd = 3.8)
## t = -5.6783, df = 199995, p-value = 0.00000001362
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12968287 -0.06312999
## sample estimates:
## mean of x mean of y
## 17.11313 17.20954
```

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? (1 pt)

- From the t-test above,
 - critical value t^* is -5.8389.
 - degrees of freedom is 199,996. It suggests the number of degree of freedom lost from each sample as a result of calculating the sameple mean.
 - p-value is 0.0000000002487; Assuming alpha level as 0.05, p-value is lot less than 0.05 and suggests that its statistically significant and probablity of obtaining a value of t atleast as high as what was observed.
 - As we dont know the purpose of the trial and alpha level - we can neither reject nor fail to reject the null hypothesis.
 - 95 percent confidence interval lie between -0.1405030 (lower band) and -0.0740502 (higher band).
 - 95% confidence suggests that if we ran this experiment a large number of times(each time generating a new sample), then 95% of the confidence intervals constructed from those repetition would actually contain the population mean difference