IST772– Problem Set 6

Sathish Kumar Rajendiran

Attribution statement: 1. I did this homework by myself, with help from the book and the professor.

Chapter 6, Exercise 1

The datasets package (installed in R by default) contains a dataset called warpbreaks (see "? warpbreaks" for documentation) that shows the number of warp thread breaks per loom for different tensions (again, we will not consider the type of wool). For each tension, n = 18 observations were conducted. In this experiment, what is the dependent variable (outcome) and what is the independent variable? What is the total number of observations? (1 pt)

1) Dependent Variable:

- Dependent or Outcome variable is the "breaks" - The number of breaks

2) Independent Variable:

- Independent or predictor variable is the "tension" with categorical values as Low (L), Medium (M) and High (H)

3) Total Number of Observations:

- There were 18 observations for each "tension" category and overall 54 observations.

Please find more details from the R code below,

```
?warpbreaks

# This data set gives the number of warp breaks per loom, where a loom corresponds to a fixed length of

# A data frame with 54 observations on 3 variables.
#
# [,1]  breaks   numeric The number of breaks
# [,2]  wool     factor  The type of wool (A or B)
# [,3]  tension factor  The level of tension (L, M, H)
# There are measurements on 9 looms for each of the six types of warp (AL, AM, AH, BL, BM, BH).
```

```r
# head(warpbreaks)
#
# View(warpbreaks)
# dim(warpbreaks)

#Values across 3 "tension" categories
warpL <- warpbreaks$breaks[warpbreaks$tension=="L"]
cat("Number of Observations for tension factor L:", length(warpL))
```

```
## Number of Observations for tension factor L: 18
```

```r
# cat("Number of breaks with tension factor L", warpL)
warpM <- warpbreaks$breaks[warpbreaks$tension=="M"]
cat("\nNumber of Observations for tension factor M:", length(warpM))
```

```
##
## Number of Observations for tension factor M: 18
```

```r
warpH <- warpbreaks$breaks[warpbreaks$tension=="H"]
cat("\nNumber of Observations for tension factor H:", length(warpH))
```

```
##
## Number of Observations for tension factor H: 18
```

```r
cat("\n\nSummary:\n")
```

```
##
##
## Summary:
```

```r
summary(warpbreaks)
```
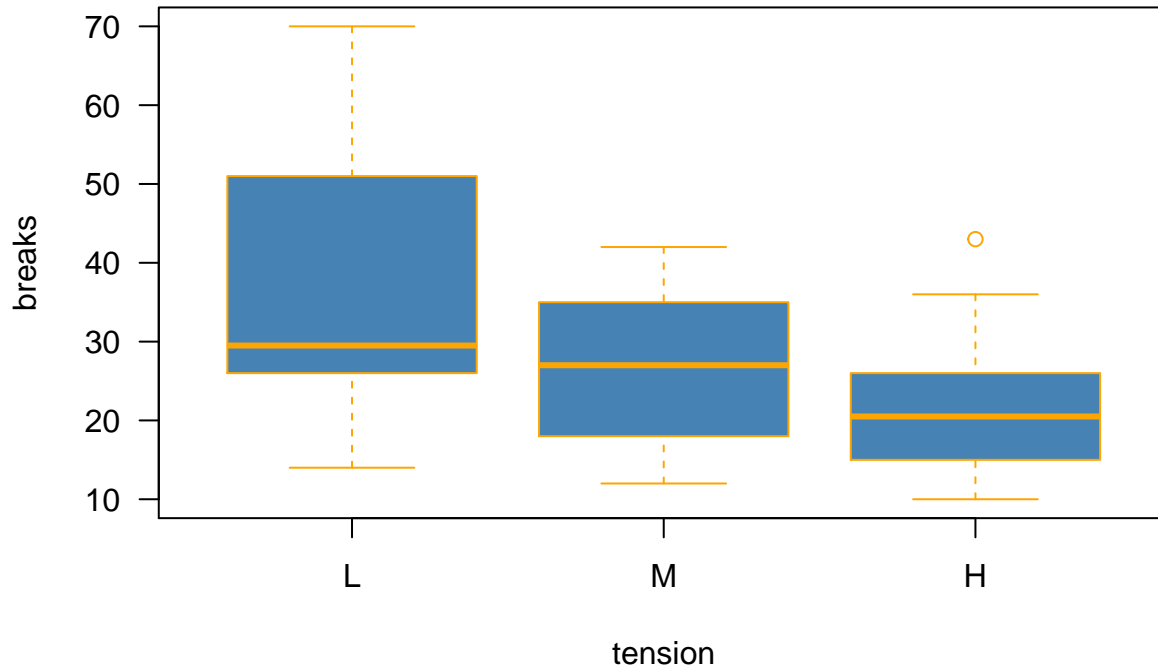
```
##      breaks        wool    tension
##  Min.    :10.00   A:27   L:18
##  1st Qu.:18.25   B:27   M:18
##  Median :26.00          H:18
##  Mean    :28.15
##  3rd Qu.:34.00
##  Max.    :70.00
```

```r
#box plot to compare the tensions
boxplot(breaks~tension, data=warpbreaks,
        border="orange",
        col="steelblue",
        freq=FALSE,
        las=1,
        # breaks=10,
        # notch = TRUE,
        horizontal = FALSE ,main=" Distribution of tenson vs breaks")
```

Distribution of tenson vs breaks

Chapter 6, Exercise 2

After running the aov() procedure on the warpbreaks data set, the "Mean Sq" for tension is 1017.1 and the "Mean Sq" for Residuals is 141.1 (1 pt). Which one of these is the between-groups variance and which one is the within-groups variance? Explain your answers briefly in your own words. (1 pt)

1) AOV() :

- ANOVA table derived from aov() returns the following summary with "breaks" as dependent (outcome) variable and "tension" as independent (predictor) variable,

-        `Df Sum Sq Mean Sq F value  Pr(>F)`

- tension 2 2034 1017.1 7.206 0.00175 **

- Residuals 51 7199 141.1

- Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 '.' 0.1 ' ' 1

2) Between-groups variance:

- "tension" row with values as below is between-groups variance degrees of freedom (df) = 2 (k-1; with k = 3; among three group means only two can vary freely) Sum of Squares = 2034; between groups sum of squares. A raw initial calculation of variability. Mean Square value = 1017.1; between groups sum of squares divided by degrees of freedom. aka between-groups variance.that is 2034/2 = 1017.1

3) Within-groups (aka residuals) variance:

- "residuals" row with values as below is within-groups variance degrees of freedom (df) = 51 ; Out of 54 observation - we lose 1 degree of freedom for calculating the grand mean leaving 51 degrees of freedom within groups (aka residuals)) Sum of Squares = 7199; within groups sum of squares. A raw initial calculation of variability. Mean Square value = 141.1; within groups sum of squares divided by degrees of freedom. aka within-groups variance.that is $7199/51 = 141.1$

4) F-value: 7.206; ratio of the between-groups mean squares and within-groups mean squares ($1017.1/141.1 = 7.206$)

5) Pr(>F):0.00175- probability of larger F-ratio.i.e. the probability of finding F-value atleast this high. in other words 175 out of 10,000 would yeild F value as high as 7.206

Please find more details from the R code below,

```
# Analyze warpbreaks data that contains multiple groups with some mean differences
warpOut <- aov(breaks ~ tension, data=warpbreaks)
summary(warpOut)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## tension       2   2034  1017.1   7.206 0.00175 **
## Residuals    51   7199   141.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 6, Exercise 3

Based on the information in question 2 and your response to that question, calculate an F-ratio by hand or using a calculator. Given everything you have earned about F-ratios, what do you think of this one? Hint: If you had all the information you needed for a Null Hypothesis Significance Test, would you reject the null? Why or why not? (1 pt)

1) F-value:

- F-value: 7.206; ratio of the between-groups mean squares and within-groups mean squares.
- between-groups mean squares from the first line divided by within-groups mean squares second line
- $1017.1/141.1 = 7.206$

2) Interpretation on NHST: Null hypothesis - there is no difference between the means or all three groups were sampled from the same population.

- with $Pr(>F):0.00175$ - probability of larger F-ratio.i.e. the probability of finding F-value atleast this high. in other words 175 out of 10,000 would yeild F value as high as 7.206
- So, the results show that p-value on the F-tests has surpassed the traditional alpha level (0.05); p-value is less than the alpha level. Suggesting as to reject the Null hypothesis.

Please find more details from the R code below,

```
#
fvalue <- 1017.1/141.1
fvalue
```

```
## [1] 7.208363
```

Chapter 6, Exercise 4

Continuing with the warpbreaks example, there are three groups where each one has n = 18 observations. Calculate the degrees of freedom between groups and the degrees of freedom within groups. Explain why the sum of these two values adds up to one less than the total number of observations in the data set. (1 pt)

1) Degrees of freedom between-groups:

- degrees of freedom between-groups is (df) = 2 with k = 3; among three group means only two can vary freely.If we know the grand mean and means of two of the three groups then the third mean is no longer free to vary.So, we could calculate it from the other information we have. Hence, the degrees of freedom between-groups is k-1. that is 3-1 = 2 df.

2) Degrees of freedom within-groups:

- degrees of freedom (df) within-groups is 51
- for 3 groups with 18 observations each; total number of observations for within groups is 3 * 18 = 54. However, as we lose 1 degree of freedom for calculating the grand mean leaving 51 degrees of freedom within groups (aka residuals)).

3) Total Degrees of Freedom:

- Together, between-groups df and within-groups df always add up to total df.that is 2+ 51 = 53; Total number of observations is 54. As we can see total number of degrees of freedom is 1 less than the toal number of observations (54-1 = 53) it is because the F-ditribution depends on the both between-groups and within-groups degrees of freedom.

Please find more details from the R code below,

```r
# Analyze warpbreaks data that contains multiple groups with some mean differences
warpbreakResults <- aov(breaks ~ tension, data=warpbreaks)
summary(warpbreakResults)
```

```
##            Df Sum Sq Mean Sq F value  Pr(>F)
## tension     2   2034  1017.1   7.206 0.00175 **
## Residuals  51   7199   141.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 6, Exercise 5

Use R or R-Studio to run the aov() command on the warpbreaks data set. You will have to specify the model correctly using the "~" character to separate the dependent variable from the independent variable. Place the results of the aov() command into a new object called warpbreakResults. Run the summary() command on warpbreakResults and interpret the results briefly in your own words. As a matter of good practice, you should state the null hypothesis, the alternative hypothesis, and what the results of the null hypothesis significance test lead you to conclude. (1 pt) (Here's a summary of the APA recommended format for reporting statistical tests https://depts.washington.edu/psych/files/writing_center/stats.pdf.)

Results:

- Dependent and Independent variables

  - Independent or predictor variable is the "tension" with categorical values as Low (L), Medium (M) and High (H). Total 3 major groups.
  - Dependent or Outcome variable is the "breaks" - The number of breaks. Total 54 observations with 18 observations for each independent variable groups.

- aov(breaks ~ tension, data=warpbreaks)

  - ANOVA table derived from aov() returns the following summary with "breaks" as dependent (outcome) variable and "tension" as independent (predictor) variable,
  -       `Df Sum Sq Mean Sq F value  Pr(>F)`
  - tension 2 2034 1017.1 7.206 0.00175 **
  - Residuals 51 7199 141.1
  - Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 '.' 0.1 ' ' 1

- Between-groups variance

  - "tension" row with values as below is between-groups variance
  - between-groups degrees of freedom (df) = 2 (k-1; with k = 3; among three group means only two can vary freely)
  - Sum of Squares = 2034; between groups sum of squares. A raw initial calculation of variability.
  - Mean Square value = 1017.1; between groups sum of squares divided by degrees of freedom. aka between-groups variance. that is 2034/2 = 1017.1

- within-groups variance

  - "residuals" row with values as below is within-groups variance
  - within-groups degrees of freedom (df) = 51 ; Out of 54 observation - we lose 1 degree of freedom for calculating the grand mean leaving 51 degrees of freedom within groups (aka residuals))
  - Sum of Squares = 7199; within groups sum of squares. A raw initial calculation of variability.
  - Mean Square value = 141.1; within groups sum of squares divided by degrees of freedom. aka within-groups variance.that is 7199/51 = 141.1

- F-value

  - F-value: 7.206; ratio of the between-groups mean squares and within-groups mean squares.
  - between-groups mean squares from the first line divided by within-groups mean squares second line
  - 1017.1/141.1 = 7.206

- Null hypothesis - there is no difference between the means or all three groups were sampled from the same population.

- Alternate Hypothesis - There is difference between the means or samples were drawn from differnt groups

- Conclusion on significance test

  - with Pr(>F):0.00175 - probability of larger F-ratio.i.e. the probability of finding F-value atleast this high. in other words 175 out of 10,000 would yeild F value as high as 7.206
  - So, the results show that p-value on the F-tests has surpassed the traditional alpha level (0.05); p-value is less than the alpha level. Suggesting us to reject the Null hypothesis.

Please find more details from the R code below,

```
# Analyze warpbreaks data that contains multiple groups with some mean differences
warpbreakResults <- aov(breaks ~ tension, data=warpbreaks)
summary(warpbreakResults)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## tension       2   2034  1017.1   7.206 0.00175 **
## Residuals    51   7199   141.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 6, Exercise 6

Load the BayesFactor package and run the anovaBF() command on the warpbreaks data set. You will have to specify the model correctly using the "~" character to separate the dependent variable from the independent variable. Produce posterior distributions with the posterior() command and display the resulting HDIs. (1 pt) Interpret the results briefly in your own words, including an interpretation of the BayesFactor produced by the grouping variable. (1 pt) As a matter of good practice, you should state the two hypotheses that are being compared. Using the rules of thumb offered by Kass and Raftery (1995), what is the strength of this result? (1 pt)

Results:

- anovaBF(breaks ~ tension, data=warpbreaks) - this produces bayesian approach to ANOVA
- posterior() distribution
  - posterior(bayesOut,iterations=10000) - generates posterior distribution of the Bayesian output
- Null hypothesis - there is no difference between the means or all three groups were sampled from the same population.
- Alternate Hypothesis - There is difference between the means or samples were drawn from differnt groups
- BayesFactor
  - [1] tension : 21.451 ±0.01%
  - Against denominator:
  - Intercept only

- Interpretation
  - Output of the Baysian posterior is organized into two groups.
    * Empirical mean and standard deviation for each variable -
      · each of these means is the mean value of the posterior distribution of the respective population parameter across 10,000 samples.
      · It calculated the grand mean and each groups deviation from the grand mean
      · Naive SE is the standard error; simply dividing the SD by the square root of 10,000.
  - Quantiles for each variable * It contains the upper and lower bounds of 95% of the HDI for each parameter * 2.5% is the lowest HDI bound and 97.5% is the upper bound for the parameters. * tension L and H doesnt straddle 0; Where as M tension straddle 0.
  - BayesFactor with Kass and Raftery rule of thumb * According to the rule of thumb privided by Kass and Raftery any odds above of 20:1 is considered strong evidence. This result with 21:1 ratio are strong evidence for the favored hypothesis. * ie. we can reject the null hypothesis

Please find more details from the R code below,

```
library(BayesFactor)
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```
bayesOut <- anovaBF(breaks ~ tension, data=warpbreaks) # Calc Bayes Factors
mcmcOut <- posterior(bayesOut,iterations=10000)  # Run mcmc iterations
boxplot(as.matrix(mcmcOut[,2:4]), main=NULL, border="orange", col="steelblue",)
```
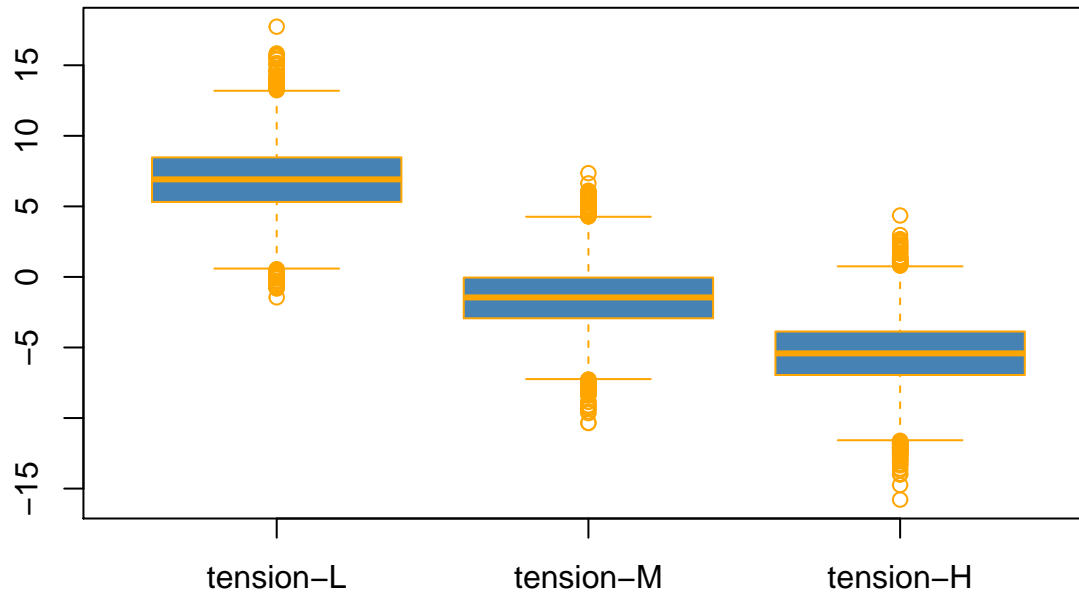
```r
summary(mcmcOut)
```

```
## 
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
## 
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
## 
##               Mean     SD Naive SE Time-series SE
## mu          28.157  1.661  0.01661        0.01661
## tension-L    6.913  2.367  0.02367        0.02786
## tension-M   -1.484  2.183  0.02183        0.02183
## tension-H   -5.428  2.307  0.02307        0.02577
## sig2       148.254 30.139  0.30139        0.33039
## g_tension    1.054 11.708  0.11708        0.11708
## 
## 2. Quantiles for each variable:
## 
##                 2.5%      25%      50%      75%    97.5%
## mu          24.87882  27.0357 28.1336 29.27907  31.413
## tension-L    2.30916   5.3138  6.9161  8.47130  11.593
## tension-M   -5.81399  -2.9316 -1.4531 -0.04341   2.780
## tension-H  -10.06825  -6.9505 -5.4173 -3.86920  -1.031
## sig2       100.49031 126.6506 144.6022 165.68296 215.816
## g_tension    0.06378   0.1923  0.3639  0.75618   4.305
```

```r
plot(mcmcOut)
```
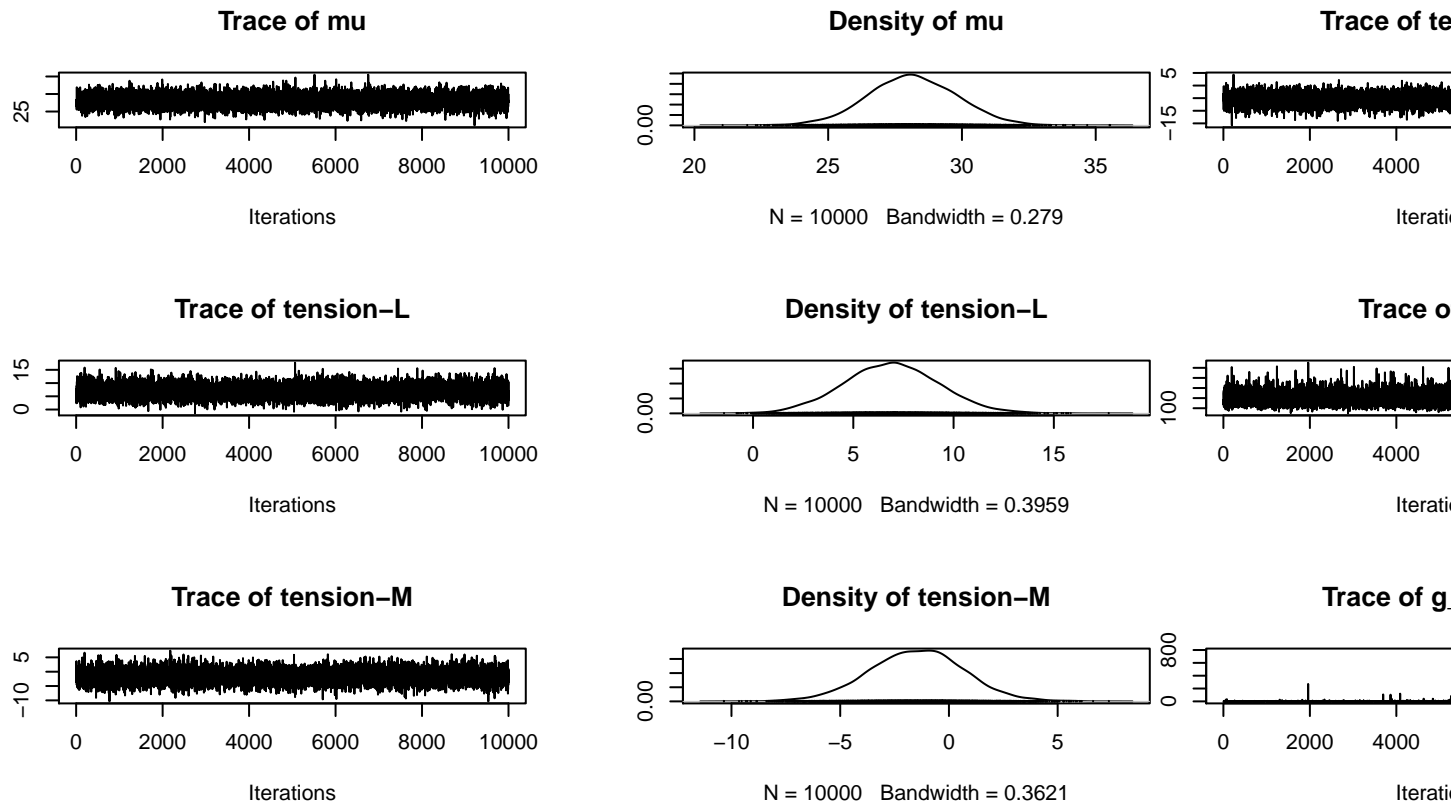


```r
par(mfcol=c(1,1))
hist(mcmcOut[,"tension-L"], main ="tension-L HDI of the Posterior distribution")
#mark the upper and lower boundaries of the 95% HDI with vertical lines
abline(v=quantile(mcmcOut[,"tension-L"],c(0.025)),col="blue")
abline(v=quantile(mcmcOut[,"tension-L"],c(0.975)),col="green")
```

tension−L HDI of the Posterior distribution
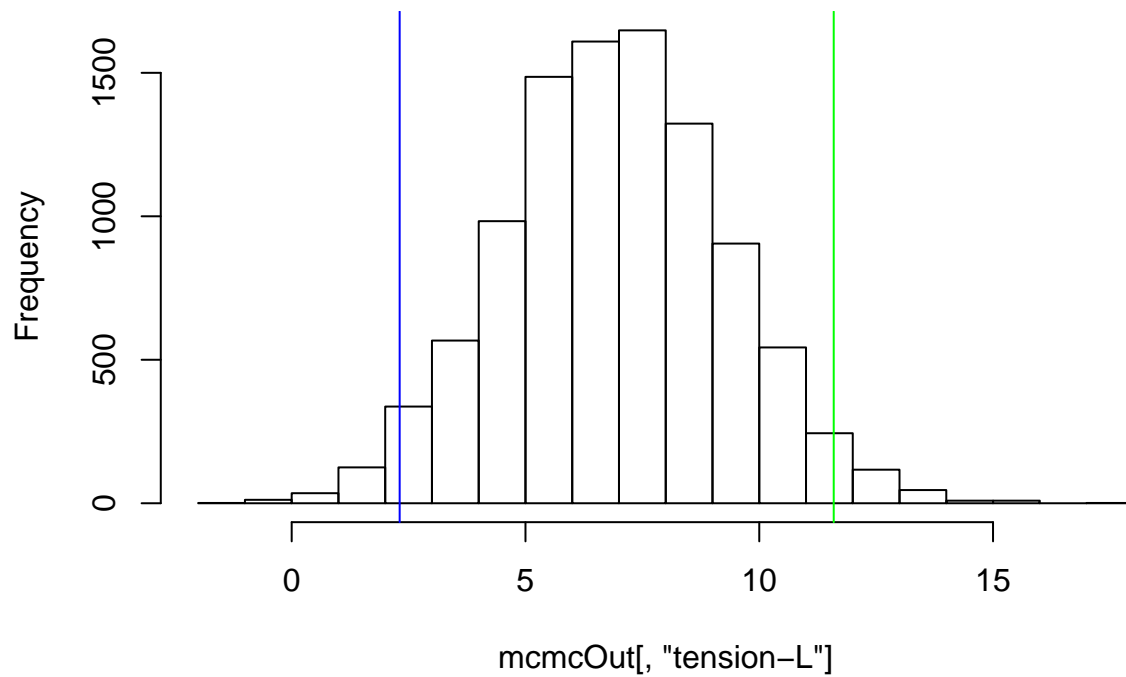

```
#Lower bound HDI
quantile(mcmcOut[,"tension-L"],c(0.025))
```
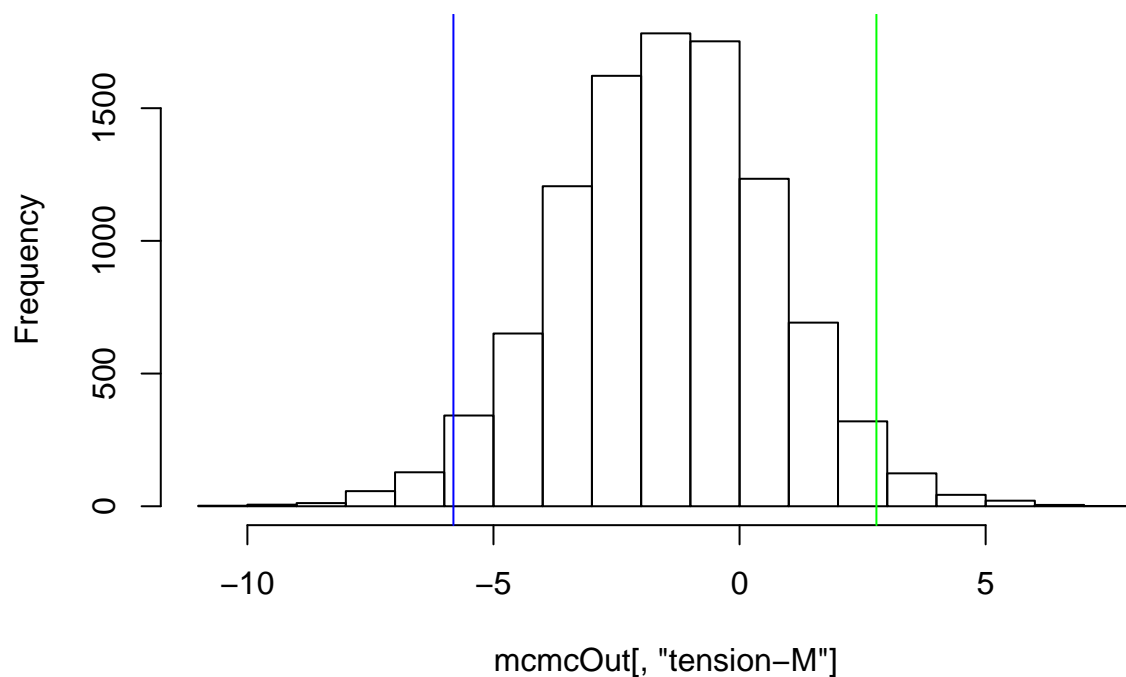
```
##     2.5%
## 2.309164
```

```
#Upper bound HDI
quantile(mcmcOut[,"tension-L"],c(0.975))
```

```
##    97.5%
## 11.59252
```

```
par(mfcol=c(1,1))
hist(mcmcOut[,"tension-M"], main ="tension-M HDI of the Posterior distribution")
#mark the upper and lower boundaries of the 95% HDI with vertical lines
abline(v=quantile(mcmcOut[,"tension-M"],c(0.025)),col="blue")
abline(v=quantile(mcmcOut[,"tension-M"],c(0.975)),col="green")
```

tension−M HDI of the Posterior distribution


```r
#Lower bound HDI
quantile(mcmcOut[,"tension-M"],c(0.025))
```

```
##      2.5%
## -5.813992
```

```r
#Upper bound HDI
quantile(mcmcOut[,"tension-M"],c(0.975))
```

```
##     97.5%
## 2.780341
```

```r
par(mfcol=c(1,1))
hist(mcmcOut[,"tension-H"], main ="tension-H HDI of the Posterior distribution")
#mark the upper and lower boundaries of the 95% HDI with vertical lines
abline(v=quantile(mcmcOut[,"tension-H"],c(0.025)),col="blue")
abline(v=quantile(mcmcOut[,"tension-H"],c(0.975)),col="green")
```

tension–H HDI of the Posterior distribution
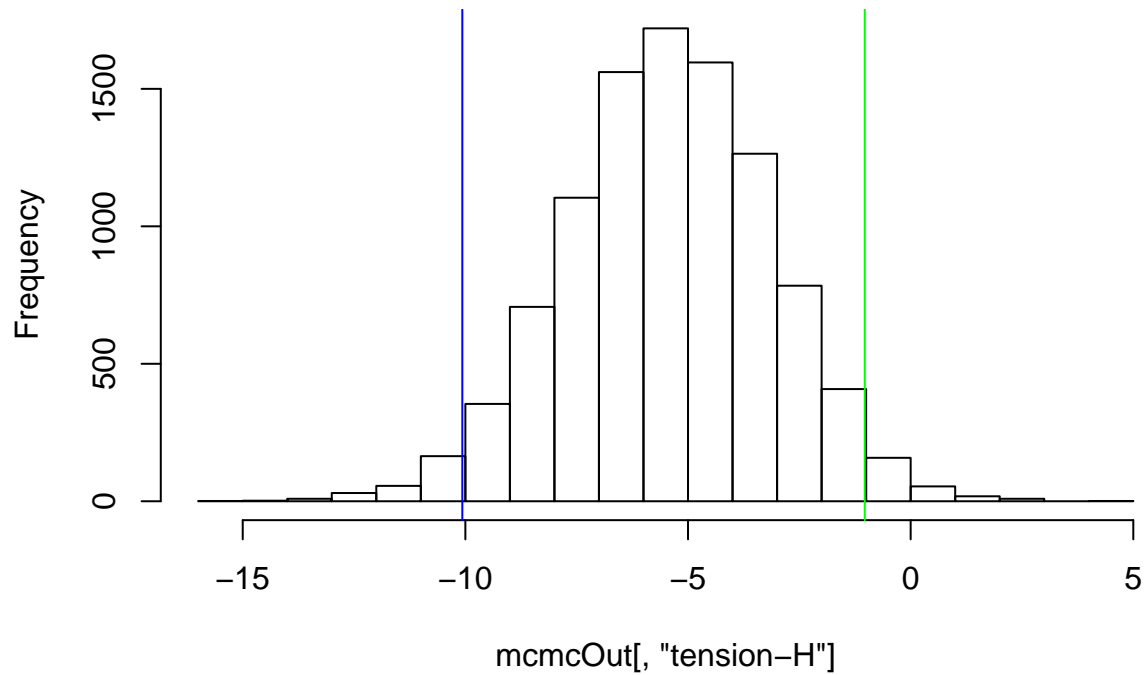
mcmcOut[, "tension−H"]

```r
#Lower bound HDI
quantile(mcmcOut[,"tension-H"],c(0.025))
```

```
##      2.5%
## -10.06825
```

```r
#Upper bound HDI
quantile(mcmcOut[,"tension-H"],c(0.975))
```

```
##      97.5%
## -1.030974
```

```r
#Bayes factor analysis
bayesOut
```

```
## Bayes factor analysis
## --------------
## [1] tension : 21.451 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```r
#Bayesian distribution (HDI)
library(BEST)
```

```
## Loading required package: HDInterval
```

```
warpBest <- BESTmcmc(warpbreaks$breaks[warpbreaks$tension=="M"],warpbreaks$breaks[warpbreaks$tension=="
```

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

```
## done.
```

```
warpBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##           mean      sd median  HDIlo HDIup  Rhat n.eff
## mu1     26.312  2.419 26.326 21.467 31.01 1.000 59251
## mu2     35.902  4.281 35.879 27.493 44.45 1.000 56629
## nu      38.701 30.451 30.313  2.341 99.13 1.000 22152
## sigma1   9.684  1.898  9.427  6.391 13.52 1.000 41517
## sigma2  17.217  3.442 16.761 11.143 24.13 1.001 37781
##
## '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.
```
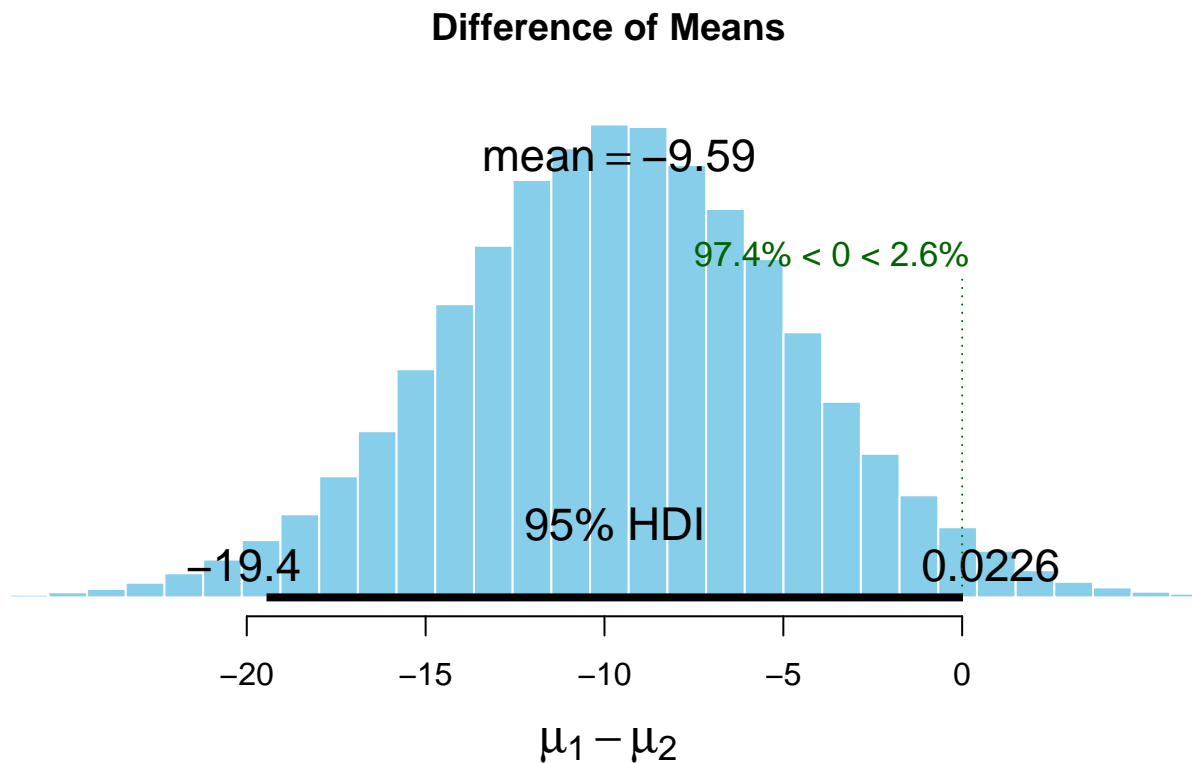
```
plot(warpBest)
```

Difference of Means

Chapter 6, Exercise 7

In situations where the alternative hypothesis for an ANOVA is supported and there are more than two groups, it is possible to do post-hoc testing to uncover which pairs of groups are substantially different from one another. Using the warpbreaks data, conduct a t-test to compare groups "M" and "L" (preferably a Bayesian t-test). Interpret the results of this t-test. (1 pt) In addition to the Bayesian t-test, feel free to run Tukey's HSD or another post-hoc procedure.

Results:

- Welch Two Sample t-test (95% confidence interval)

  - critical value t* is -2.256
  - degrees of freedom is 26.554. It suggests the number of degree of freedom lost from each sample as a result of calculating the sameple mean.
  - p-value is 0.03252; p-value suggests that the probablity of obtaining a value of t atleast as high as what was observed.
  - t(26.554)=-2.256, p-value =0.03252 is less than alpha value (0.05) - it shows that the differnce is statistically significant and we can reject the null hypothesis.
  - 95 percent confidence interval is between -19.1023204 (lower band) and -0.8976796(higher band) which doesnt 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 repetition would actually contain the population mean difference.

- Bayesian distribution (HDI)

  - From the BESTmcmc() plot above,
  - HDI
    * 95% Highest density interval (HDI) lie in the bell-shaped area between -19.5 (HDIlo) and 0.0123 (HDIup); having mean value as -9.65
  - 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 2.6%
    * Difference between mu1-mu2 is -9.65
    * plotAll(warpBest) - group 1(Medium tension value is significant than Low tension group)

Please find more details from the R code below,

```
#t-test
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)
library(BEST)
warpBest <- BESTmcmc(warpbreaks$breaks[warpbreaks$tension=="M"],warpbreaks$breaks[warpbreaks$tension=="
```

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

```
warpBest
```

```
## MCMC fit results for BEST analysis:
## 100002 simulations saved.
##            mean      sd median  HDIlo  HDIup Rhat n.eff
## mu1      26.311   2.408 26.316 21.652  31.17    1 60550
## mu2      35.905   4.291 35.879 27.469  44.41    1 56811
## nu       39.035  30.760 30.522  2.095 100.43    1 22656
## sigma1   9.691   1.919  9.417  6.400  13.57    1 40842
## sigma2  17.225   3.463 16.768 11.232  24.25    1 39192
##
## '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.
```
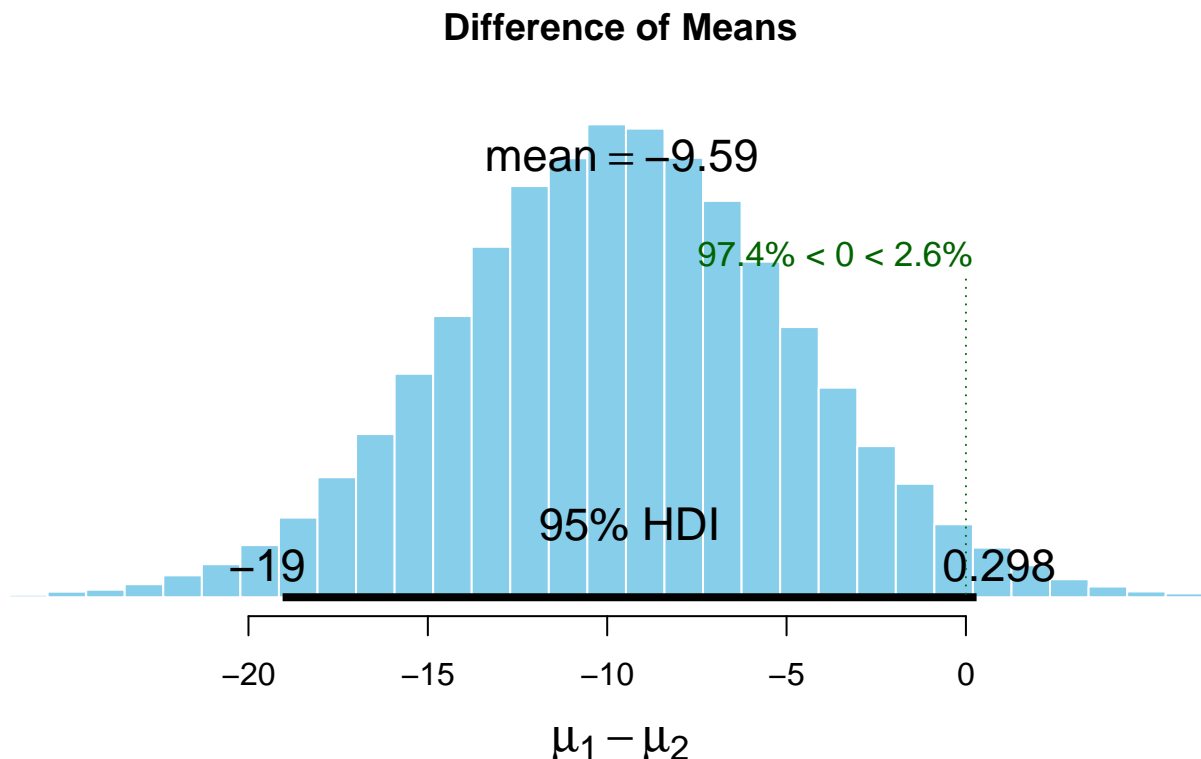
```
#Bayesian probability distribution of the differences between two means
plot(warpBest)
```

Difference of Means

```
plotAll(warpBest)
```


Group 1 Mean
95% HDI
mean = 26.3
21.9   31.2
μ_1

Data Group 1 w. Post. Pred.
p(y)
$N_1 = 18$
y

Group 2 Mean
95% HDI
mean = 35.9
27.5   44.4
μ_2

Data Group 2 w. Post. Pred.
p(y)
$N_2 = 18$
y

Group 1 Std. Dev.
95% HDI
mode = 8.91
6.4   13.6
σ_1

Difference of Means
95% HDI
mean = −9.59
−19   100% < 2.6% 0.298
$\mu_1 - \mu_2$

Group 2 Std. Dev.
95% HDI
mode = 16
11.2   24.3
σ_2

Difference of Std. Dev.s
95% HDI
mode = −6.93
−15.7   −0.266
$\sigma_1 - \sigma_2$

Normality
95% HDI
mode = 1.53
0.773   2.11
$\log10(\nu)$

Effect Size
95% HDI
mode = −0.712
−1.39   97.4% < 0 < 2.6% 0.0133
$(\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$