ANOVA & Experimental Groups Homework

Tim Hulak

I produced the material below with no assistance

Exercises 1-7 on pages 117 and 118 of *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*

1. The data sets package (installed in R by default) contains a data set called InsectSprays that shows the results of an experiment with six different kinds of insecticide. For each kind of insecticide, n = 12 observations were conducted. Each observation represented the count of insects killed by the spray. In this experiment, what is the dependent variable (outcome) and what is the independent variable? What is the total number of observations?

```
data(InsectSprays)
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

```
dim(InsectSprays)
```

```
## [1] 72  2
```

```
summary(InsectSprays)
```

```
##      count       spray
##  Min.   : 0.00   A:12
##  1st Qu.: 3.00   B:12
##  Median : 7.00   C:12
##  Mean   : 9.50   D:12
##  3rd Qu.:14.25   E:12
##  Max.   :26.00   F:12
```

Answer: The *count* variable is the dependent variable (outcome) and the independent variable is the *spray* variable. In other words, after some analysis, we can predict the number of insects killed by the spray using the *spray* type/label. There are **72** observations in this dataset (12 observations for each of the *sprays*).

---

2. After running the aov() procedure on the InsectSprays data set, the "Mean Sq" for spray is 533.8 and the "Mean Sq" for Residuals is 15.4. 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.

```
aov(count ~ spray, data=InsectSprays)
```

```
## Call:
##    aov(formula = count ~ spray, data = InsectSprays)
##
## Terms:
##                     spray Residuals
## Sum of Squares  2668.833  1015.167
## Deg. of Freedom        5        66
##
## Residual standard error: 3.921902
## Estimated effects may be unbalanced
```

```
sprayAOV <- aov(count ~ spray, data=InsectSprays)
summary(sprayAOV)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: The p-value on the F-tests is **0.00000000000000022** and has surpassed the traditional 0.05 level of alpha and is therefore statistically significant (this can also be seen by the 3 stars next to the value in the summary output). We must reject the null hypothesis based on the analysis. The *spray* Mean Sq value of **533.8** is the *between-groups variance* and the *Residuals* value of **15.4** is the *within-groups variance*. The *spray* is the predictor where as the *Residuals* can be thought of as the errors.

---

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?

```
between_group_variance <- 533.8
within_groups_variance <- 15.4

f_ratio <- between_group_variance/within_groups_variance

f_ratio
```

```
## [1] 34.66234
```

Answer: The F-ratio in this case is **34.66234**. Any F-ratio that is substantially larger than **1.0** is considered possible evidence that at least one of the groups is from a population with a different mean. **34.66234** is quite a bit larger than **1**, we would reject the null.

---

4. Continuing with the InsectSprays example, there are six groups where each one has n = 12 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.

```
# Get the number of unique observations
unique_length <- length(unique(InsectSprays$spray))

# Get total number of observations
observations <- length(InsectSprays$count)

# Subtract 1 from the length to get the between-groups degrees of freedom
spray_df <- unique_length - 1

# Subtract the spray degrees of freedom from the total number of observations minus 1 to get the Residu

residual_df <- (observations - 1) - spray_df

print(spray_df)
```

```
## [1] 5
```

```
print(residual_df)
```

```
## [1] 66
```

Answer: There are **6** unique categories in the dataset (*A,B,C,D,E,F*). You calculate the *degrees of freedom* by subtracting **1** from the number of categories, therefore there are **5 degrees of freedom** for the *spray*. The *residual degrees of freedom* is calculated by subtracting **1** from total number of observations (in this case, there are **72** total observations) and then subtracting the *between-groups* degrees of freedom: 72 observations - 1 = 71. 71 - 5 between-groups degrees of freedom = **66 Residual degrees of freedom**.

---

5. Use R or R-Studio to run the aov() command on the InsectSprays 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 insectResults. Run the summary() command on insectResults 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.

```
insectResults <- aov(count ~ spray, data=InsectSprays)
summary(insectResults)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: The *null hypothesis* is that there is no difference in the means of the populations. The *alternative hypothesis* is that there is a difference in the means of populations. The p-value is less than 0.05, therefore we must reject the null hypothesis based on the analysis

- **Df**: degrees of freedom, which is how many elements are free to vary after subtracting 1 from the number of observations.

- **Sum Sq**: A measure of variability or deviation from the mean
- **Mean Sq**: The same thing as variance. This is calculated by dividing the Sum Sq by Df.

- **F value**: This is the F ratio, which is calculated by dividing the Mean Sq of the between-group by the Mean Sq of the within-group
- **Pr(>F)**: This is the p-value of the test and helps determine statistical significance of the test.

---

6. Load the BayesFactor package and run the anovaBF() command on the InsectSprays 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. Interpret the results briefly in your own words, including an interpretation of the BayesFactor produced by the grouping variable. 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?

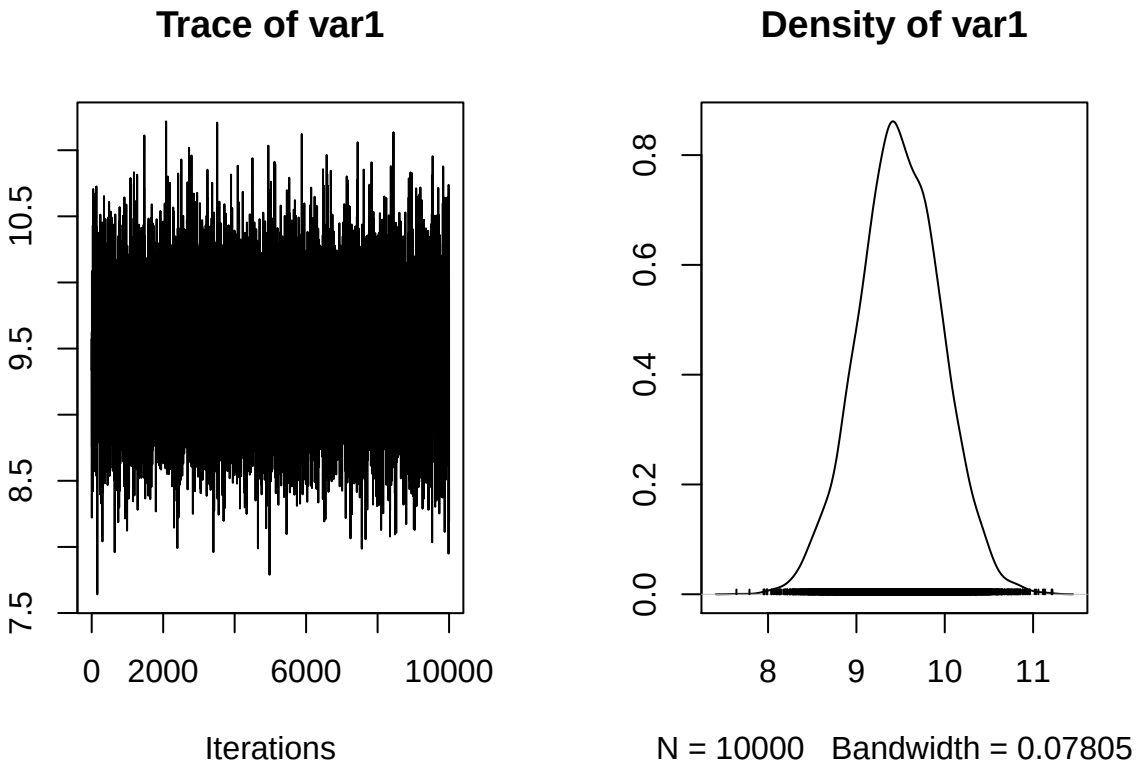
```
library(BayesFactor)
```

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

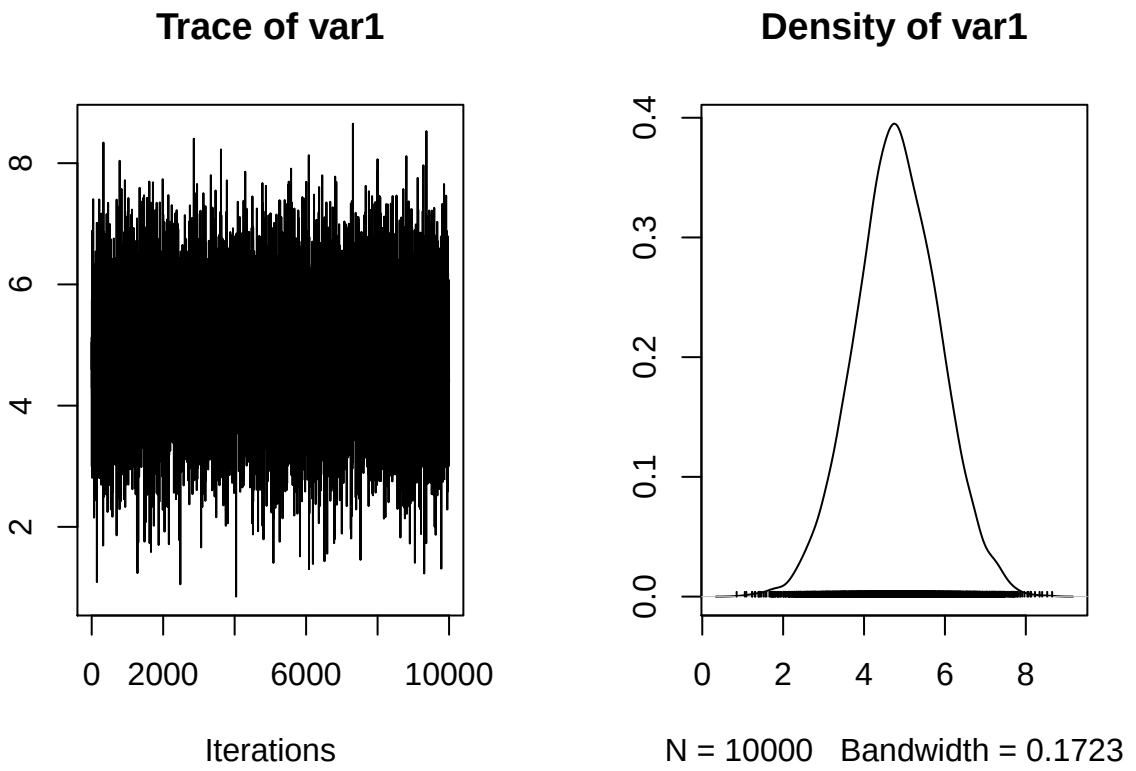
```
## Loading required package: Matrix
```

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

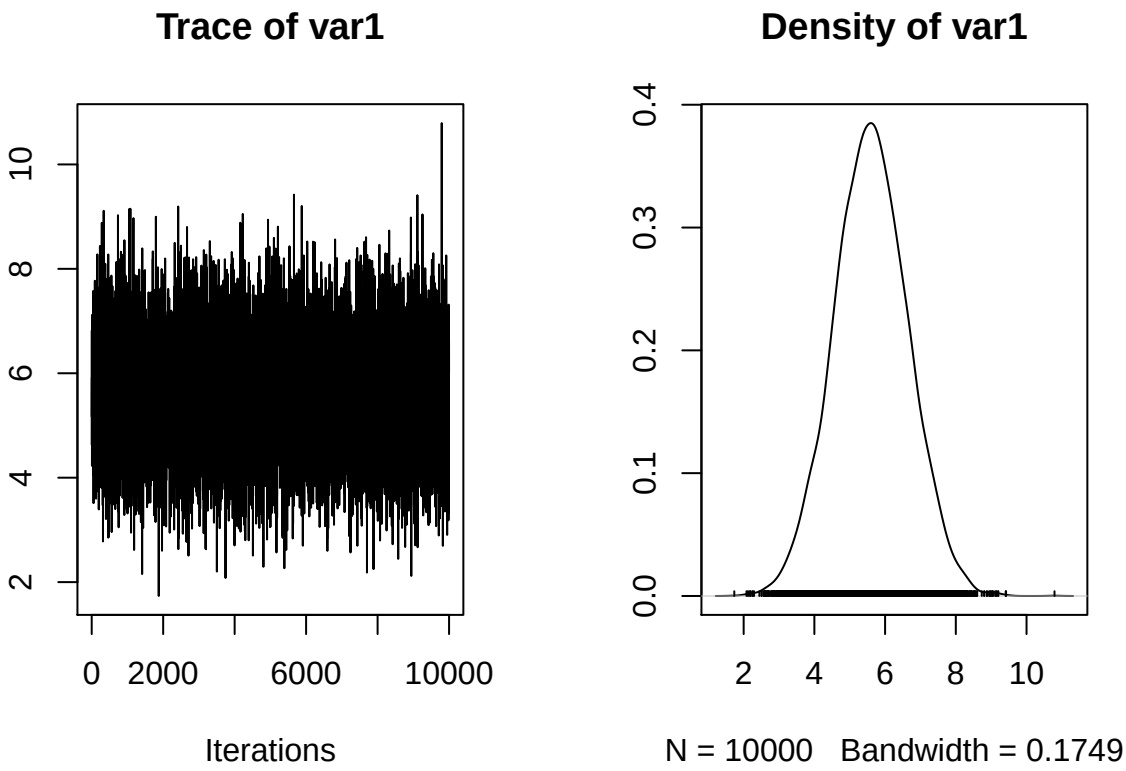
```
spray_bayes_out <- anovaBF(count ~ spray, data=InsectSprays)
mcmcOut <- posterior(spray_bayes_out,iterations = 10000)
plot(mcmcOut[,'mu'])
```



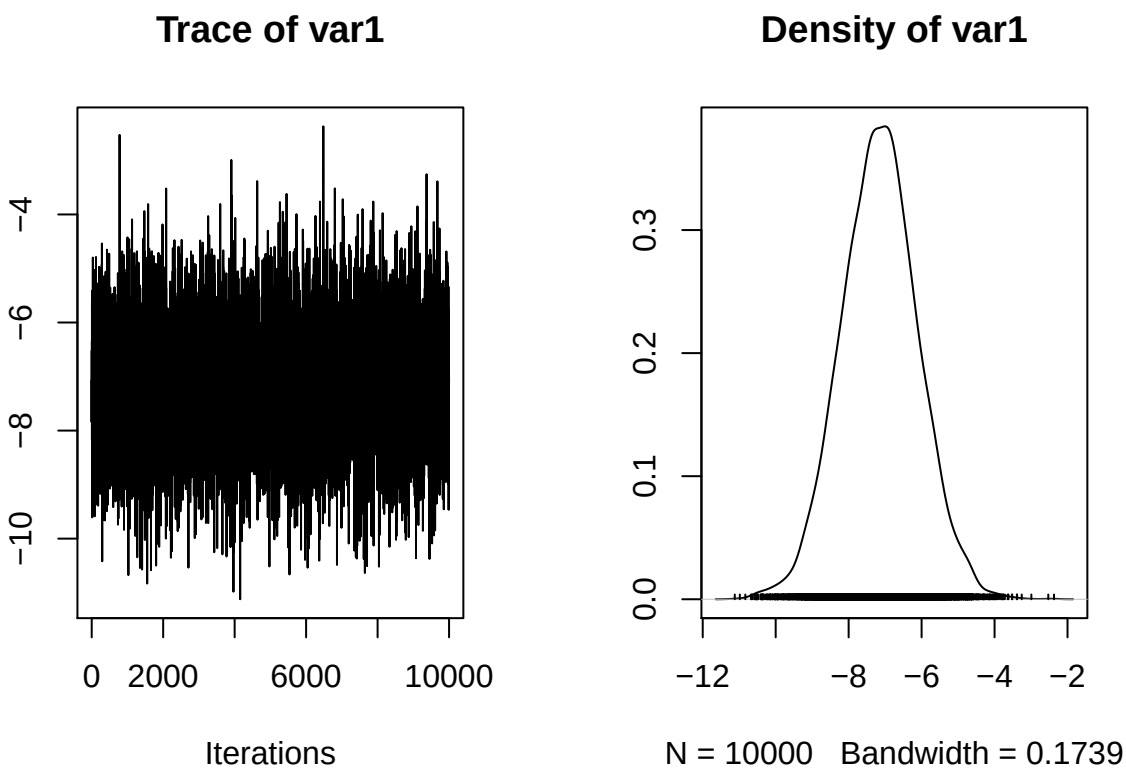

```
plot(mcmcOut[,'spray-A'])
```



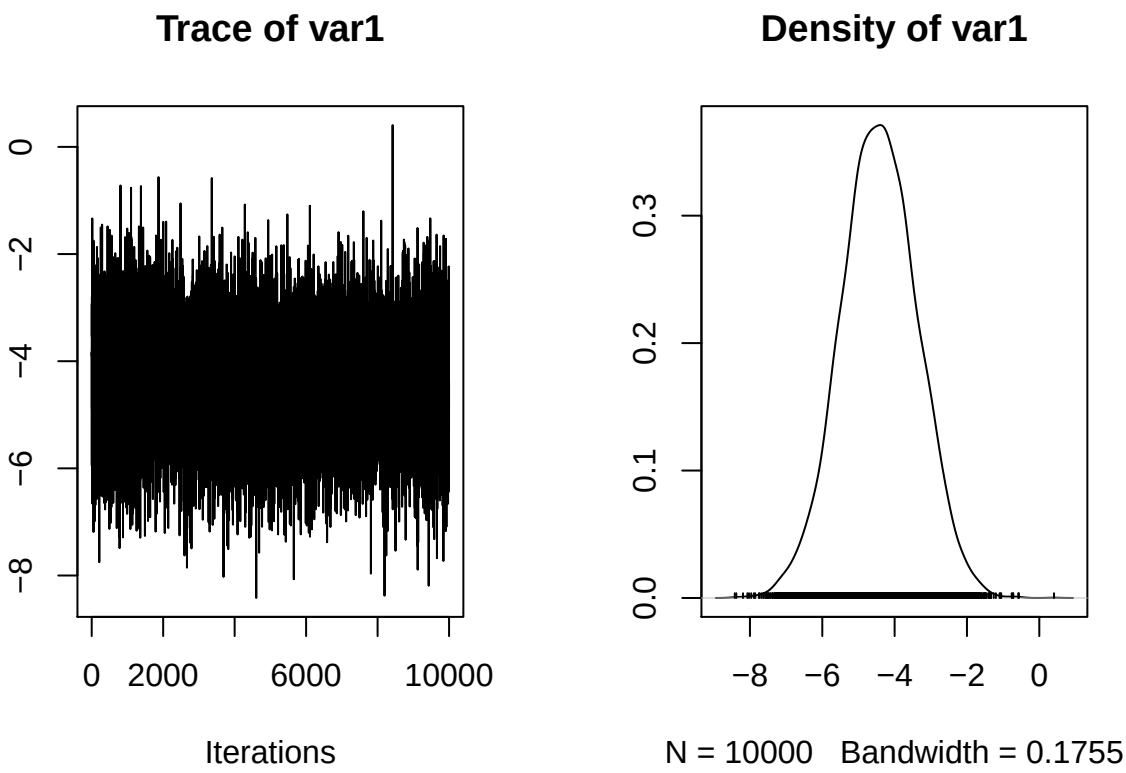
```
plot(mcmcOut[,'spray-B'])
```



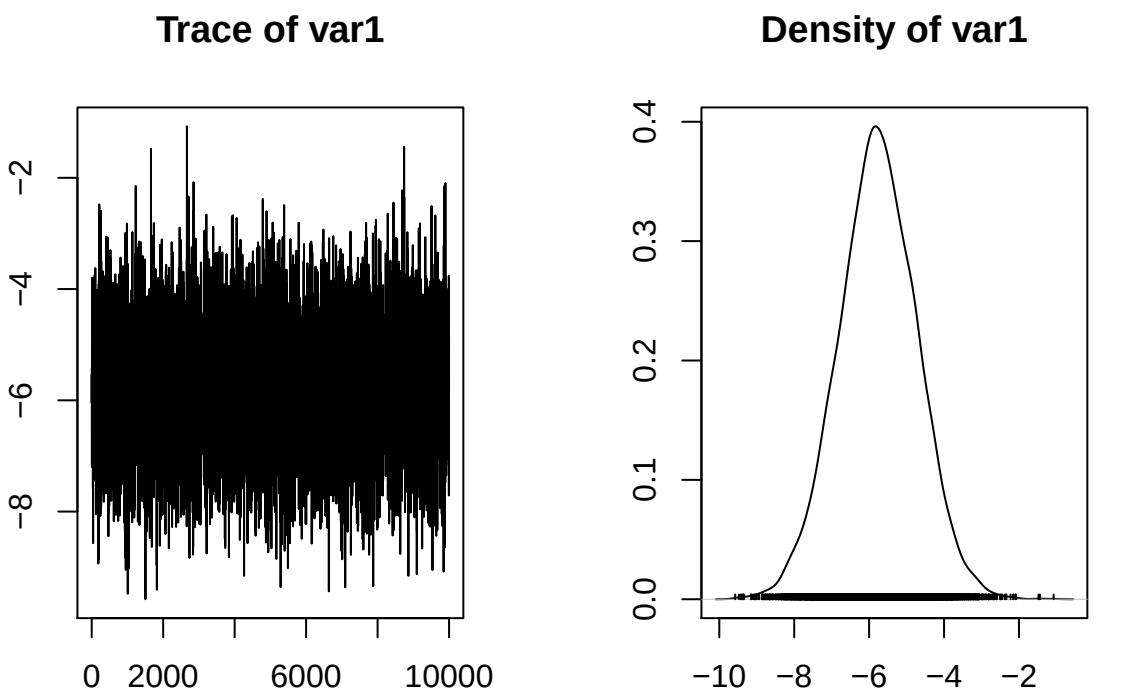

```
plot(mcmcOut[,'spray-C'])
```



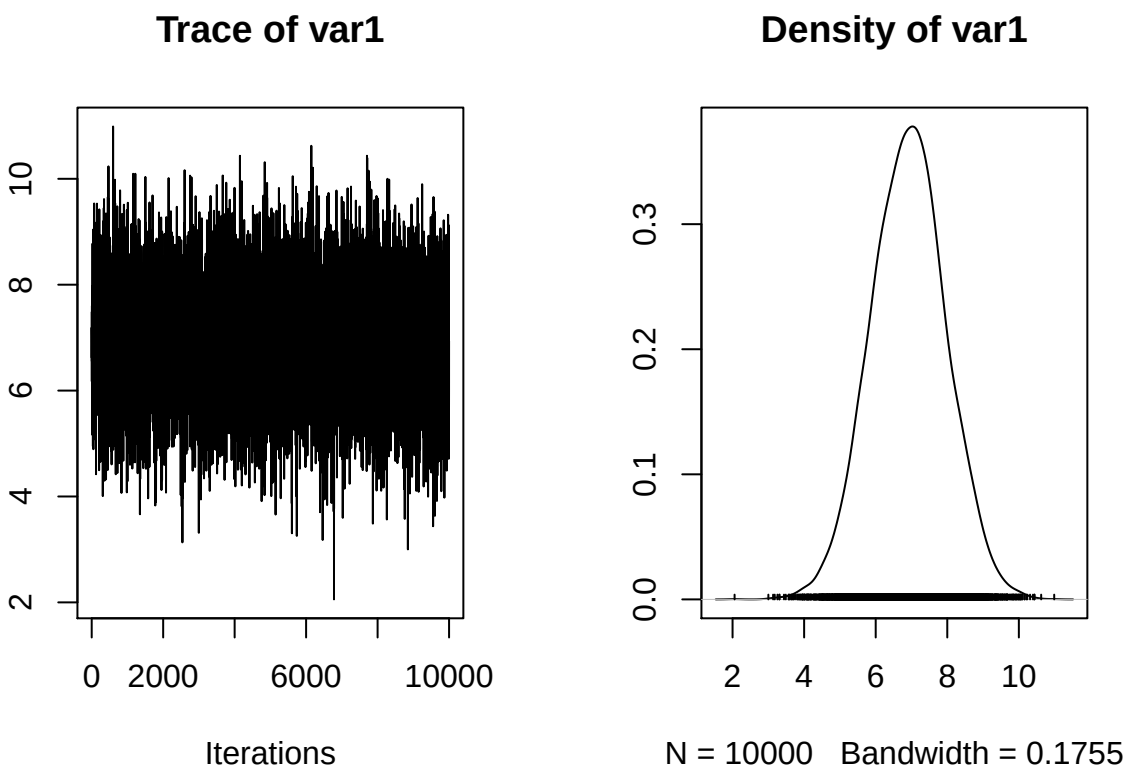
```
plot(mcmcOut[,'spray-D'])
```




```
plot(mcmcOut[,'spray-E'])
```



```
plot(mcmcOut[,'spray-F'])
```

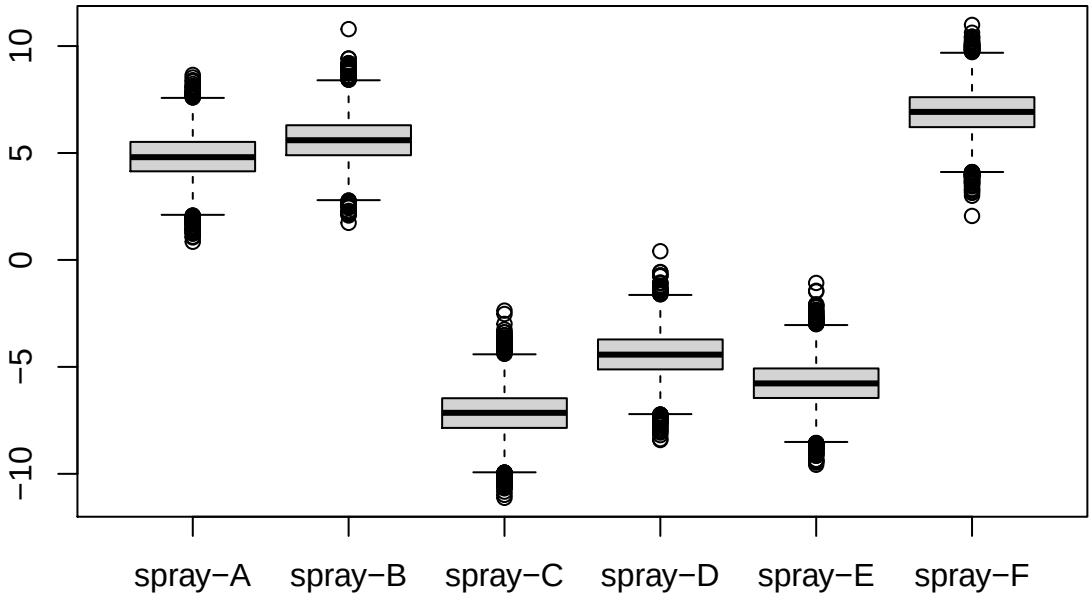



```
summary(mcmcOut)
```

```
##
## 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        9.490 0.4648 0.004648        0.00481
## spray-A   4.818 1.0360 0.010360        0.01036
## spray-B   5.604 1.0409 0.010409        0.01057
## spray-C  -7.152 1.0495 0.010495        0.01083
## spray-D  -4.421 1.0472 0.010472        0.01047
## spray-E  -5.767 1.0397 0.010397        0.01065
## spray-F   6.918 1.0491 0.010491        0.01080
## sig2     16.083 2.9055 0.029055        0.03448
## g_spray   3.462 3.4724 0.034724        0.03671
##
## 2. Quantiles for each variable:
##
##             2.5%    25%    50%    75%  97.5%
## mu        8.5691  9.183  9.483  9.805 10.404
```

```
## spray-A  2.7619  4.144  4.805  5.518  6.843
## spray-B  3.5715  4.896  5.597  6.298  7.646
## spray-C -9.1905 -7.853 -7.147 -6.466 -5.062
## spray-D -6.4976 -5.117 -4.427 -3.718 -2.393
## spray-E -7.8361 -6.454 -5.771 -5.073 -3.756
## spray-F  4.8748  6.210  6.921  7.610  8.972
## sig2    11.3873 14.023 15.761 17.746 22.783
## g_spray  0.8271  1.677  2.552  4.023 12.272
```

```
boxplot(as.matrix(mcmcOut[,2:7]))
```




```
spray_bayes_out
```

```
## Bayes factor analysis
## --------------
## [1] spray : 1.506706e+14 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

Answer: The *null hypothesis* is that the spray has no effect on the count. The *alternative hypothesis* is that the spray has an effect on the count. The odds are approximately **1.506706:1**. error!). According to the rules of thumb provided by Kass and Raftery (1995), any odds ratio in excess of 150:1 is considered very Strong evidence. Therefore we must reject the null hypothesis

---

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 InsectSprays data, conduct a t-test to compare groups C and F (preferably a Bayesian t-test). Interpret the results of this t-test.

```
library(BEST)
```

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

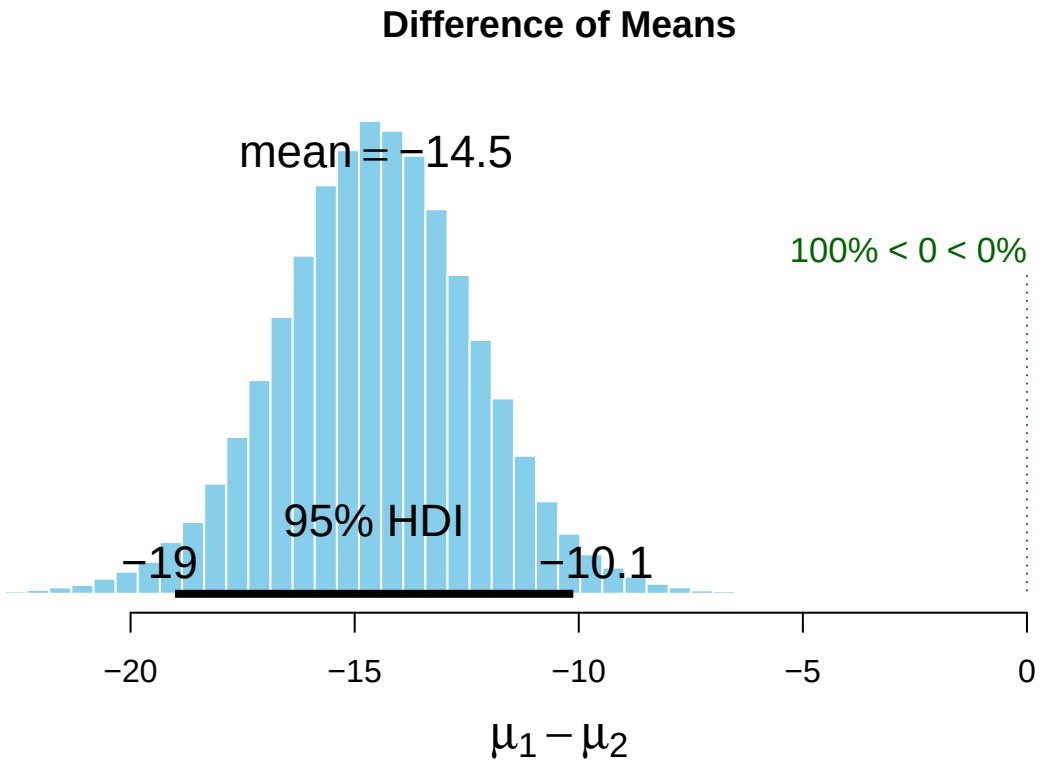
```
# Slice the data
CandF <- InsectSprays[InsectSprays$spray=="C" | InsectSprays$spray=="F",]

sprayC <- InsectSprays[InsectSprays$spray=="C",1]
sprayF <- InsectSprays[InsectSprays$spray=="F",1]

obs <-data.frame("C" = sprayC, ".F" = sprayF)
plot(BESTmcmc(InsectSprays[InsectSprays$spray=="C",1],
              InsectSprays[InsectSprays$spray=="F",1]))
```

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

```
## done.
```




```
sprayBFOut <- anovaBF(count ~ spray, data=CandF)
summary(sprayBFOut)
```

```
## Bayes factor analysis
## --------------
## [1] spray : 90005.78 ±0%
##
## Against denominator:
```

```
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

Answer: The entire distribution is below **0**, as seen in the $\mathbf{100\% < 0 < 0\%}$. There is also a 95% chance of the mean differences being between **-18.9** and **-10**. In addition, the most likely mean difference is **-14.5** It appears that *spray-F* performes better than *spray-C*.
