Matthew L. Pergolski
IST 772
Dr. Block
11/16/2021

Homework 6

Beginning Statement

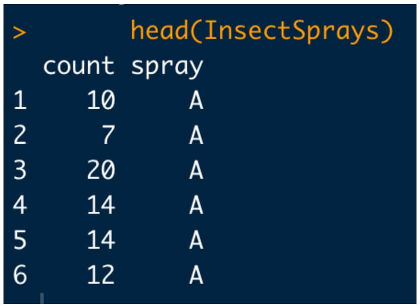
"I produced the material below with no assistance [direct quote from IST 772 class syllabus]." Note: Homework questions from the book may have been copied/pasted into the document for both the student and viewer's convenience.

The homework for week 6 is exercises 1-7 on pages 117 and 118.

Homework Question 1

Question:
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?

Answer/Student Response:
The head function was ran on InsectSprays and the following was observbed:



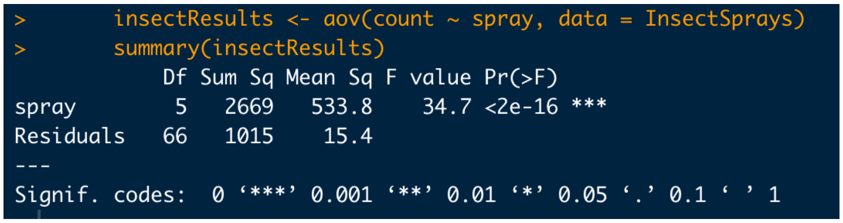
```
>        head(InsectSprays)
  count spray
1    10     A
2     7     A
3    20     A
4    14     A
5    14     A
6    12     A
```

The dependent variable would be the amount of insects killed and the independent variable(s) would be the different types, or categories, of spray; total number of observations is 72.

Homework Question 2

Question:
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.

Answer/Student Response:
The following was generated and observed:



```
>       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
```

The 533.8 value corresponds to the 'Between Groups' variance while the 15.4 value corresponds to the 'Within Groups' variance. The ratio of Between Groups divided by Within Groups variance gives us our F-value -- which, in this case, is 34.7 (533.8 / 15.4).

Homework Question 3

Question:
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?

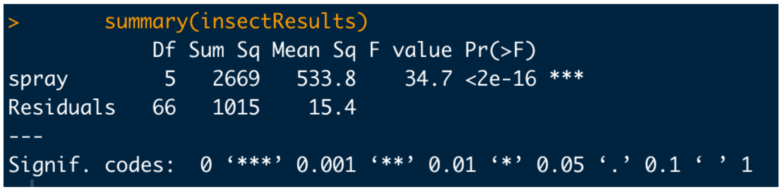
Answer/Student Response:
F-ratio calculated: 533.8 / 15.4 = 34.7.

Based on the conventional rules of the F-value, it would generally equal ~1 if there were no real, significant differences between the groups; however, if we spot outliers, or values that are far greater than 1, it gives us evidence to suggest that a significant difference is real, and thus we'd be inclined to reject the null hypothesis, provided we could confirm that our p-value is less than the specified alpha value (which may be 0.05 in many traditional academic cases).

Homework Question 4

Question:
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.

Answer/Student Response:
To aid in this analysis, the following was again generated and observed:



```
>       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
```

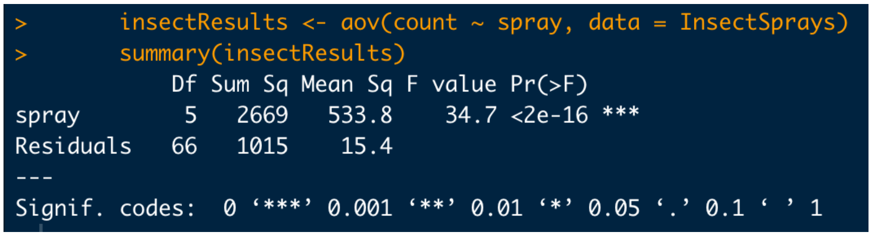
Within our summary(insectResults) command, we can spot the 'DF' section that indicates our degrees of freedom. If we add up the residuals value of 66 and spray value of 5, we get one less than the total number of observations within the group.

As we perform calculations on a data set, we are essentially using up information that leaves us with less moving forward. This reduction in information/data increases our uncertainty, and as a result we need to subtract in order to have an accurate approach to finding the most unbiased statistic/end-value. An example would be calculating the mean when attempting to find the variance -- this intermediate step of calculating the mean would require us to subtract one (1).

Homework Question 5

Question:
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.

Answer/Student Response:
Again, we run the following code and observe:



```
>       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
```

The null hypothesis in this situation suggests that there are is no significant difference between the groups/sprays being compared. On the opposite end, our alternative hypothesis suggests that there is indeed a significant difference between the various types of sprays within the data set. Based on the results of the frequentest aov() R command, we see that the F-value is significantly larger than one (1). If the variance from the differing groups were about the same, we would expect an F-value of about 1.

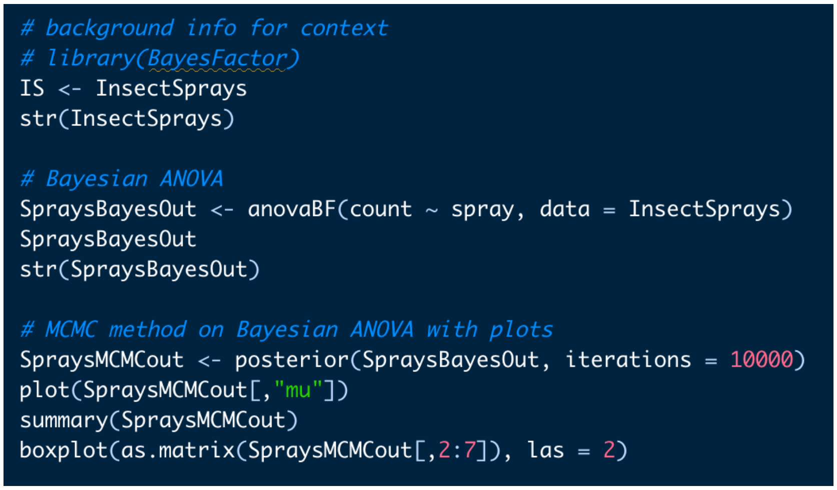
This result along with the fact that our P value is significantly under 0.05, we can indicate that we reject the null-hypothesis and conclude there may be evidence to suggest a significant difference exists between our independent variables.

Homework Question 6

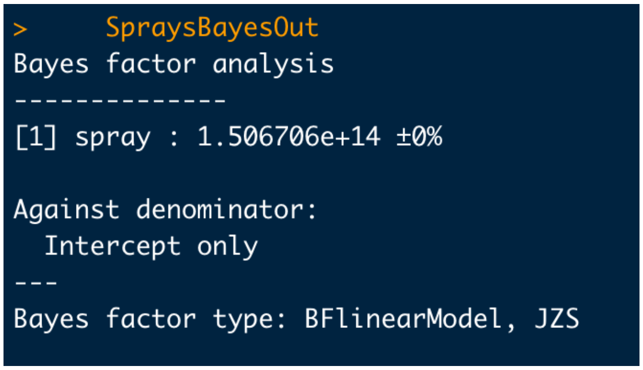
Question:
6. Load the BayesFactor package and run the anovaBF() command on the InsectSprays data set. You will have to specify the model correctly using the "~" character to separate the dependent variable from the independent variable. Produce posterior distributions with the posterior() command and display the resulting HDs. 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?

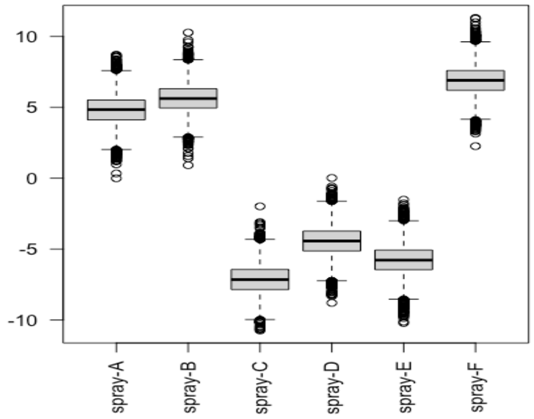
Answer/Student Response:
The following was run and observed:

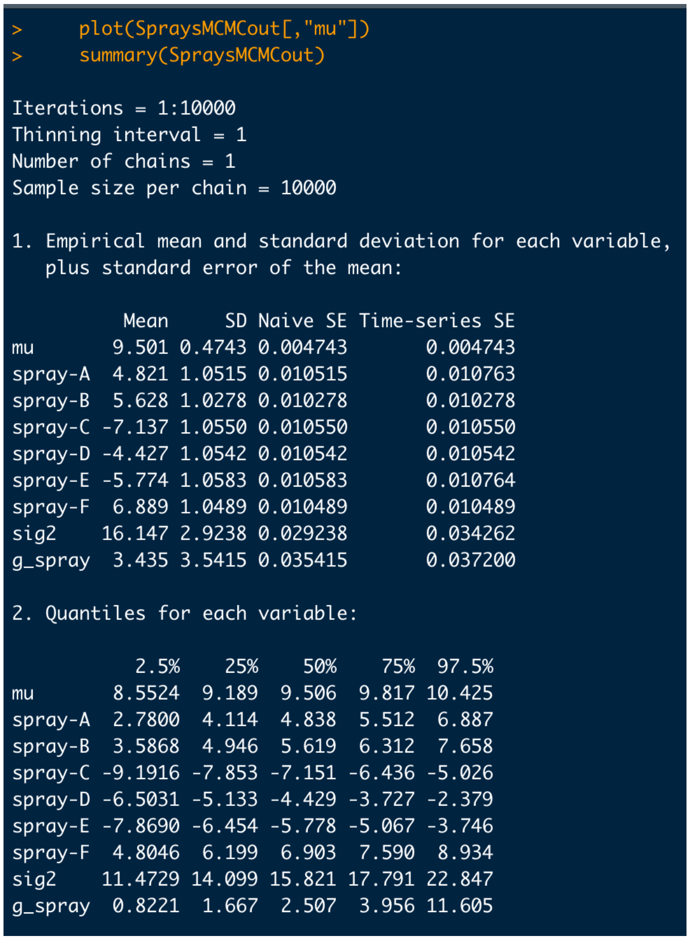


```
# background info for context
# library(BayesFactor)
IS <- InsectSprays
str(InsectSprays)

# Bayesian ANOVA
SpraysBayesOut <- anovaBF(count ~ spray, data = InsectSprays)
SpraysBayesOut
str(SpraysBayesOut)

# MCMC method on Bayesian ANOVA with plots
SpraysMCMCout <- posterior(SpraysBayesOut, iterations = 10000)
plot(SpraysMCMCout[,"mu"])
summary(SpraysMCMCout)
boxplot(as.matrix(SpraysMCMCout[,2:7]), las = 2)
```



```
>     SpraysBayesOut
Bayes factor analysis
--------------
[1] spray : 1.506706e+14 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```





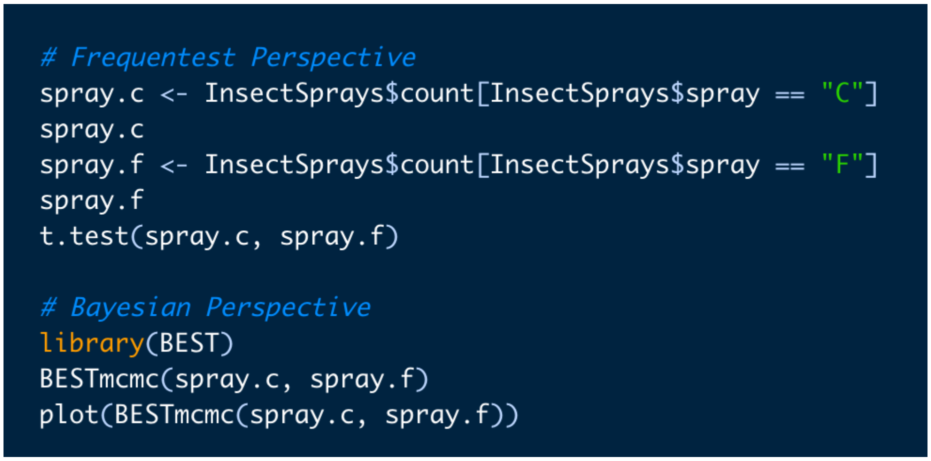
```
>     plot(SpraysMCMCout[,"mu"])
>     summary(SpraysMCMCout)

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       9.501 0.4743 0.004743       0.004743
spray-A  4.821 1.0515 0.010515       0.010763
spray-B  5.628 1.0278 0.010278       0.010278
spray-C -7.137 1.0550 0.010550       0.010550
spray-D -4.427 1.0542 0.010542       0.010542
spray-E -5.774 1.0583 0.010583       0.010764
spray-F  6.889 1.0489 0.010489       0.010489
sig2    16.147 2.9238 0.029238       0.034262
g_spray  3.435 3.5415 0.035415       0.037200

2. Quantiles for each variable:

           2.5%    25%    50%    75%  97.5%
mu       8.5524  9.189  9.506  9.817 10.425
spray-A  2.7800  4.114  4.838  5.512  6.887
spray-B  3.5868  4.946  5.619  6.312  7.658
spray-C -9.1916 -7.853 -7.151 -6.436 -5.026
spray-D -6.5031 -5.133 -4.429 -3.727 -2.379
spray-E -7.8690 -6.454 -5.778 -5.067 -3.746
spray-F  4.8046  6.199  6.903  7.590  8.934
sig2    11.4729 14.099 15.821 17.791 22.847
g_spray  0.8221  1.667  2.507  3.956 11.605
```

Our null hypothesis suggests that there is no significant difference in the effectiveness of the different sprays (i.e., independent variables) within our data set. The alternative hypothesis indicates the opposite (i.e., there is a significant difference between the variables). When conducting the Bayesian ANOVA test, we see that the spray variable has a 1.506706e+14 ±0% to 1 ratio supporting the alternative hypothesis, which is exceptional odds when validating against Kass and Raftery's rule of thumb (i.e., odds ratios of more than 150:1 are very strong evidences for the favored hypothesis) when visualizing the HDIs in boxplot form, we see a difference in the measured effectiveness of the various sprays; Spray C seems to be the worst performer while Spray F appears to be the best.

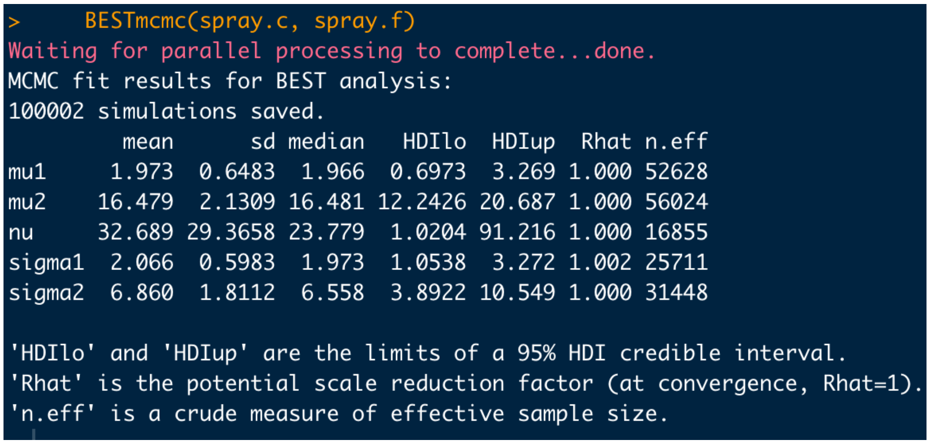
Homework Question 7

Question:
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 -test). Interpret the results of this t-test.

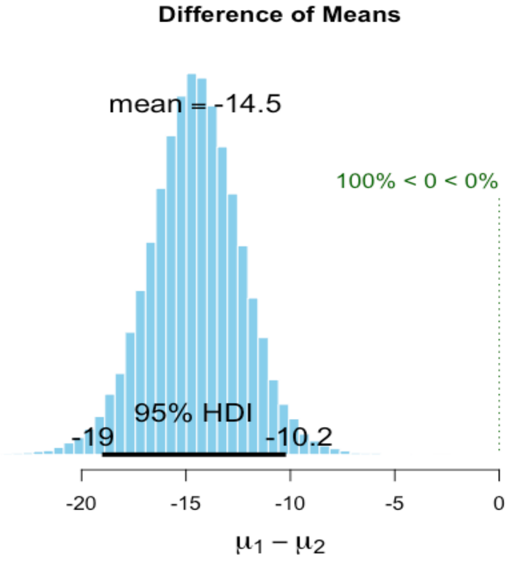
Answer/Student Response:
The following was generated and observed:



```
# Frequentest Perspective
spray.c <- InsectSprays$count[InsectSprays$spray == "C"]
spray.c
spray.f <- InsectSprays$count[InsectSprays$spray == "F"]
spray.f
t.test(spray.c, spray.f)

# Bayesian Perspective
library(BEST)
BESTmcmc(spray.c, spray.f)
plot(BESTmcmc(spray.c, spray.f))
```



```
>     BESTmcmc(spray.c, spray.f)
Waiting for parallel processing to complete...done.
MCMC fit results for BEST analysis:
100002 simulations saved.
         mean      sd median   HDIlo  HDIup  Rhat n.eff
mu1     1.973  0.6483  1.966  0.6973  3.269 1.000 52628
mu2    16.479  2.1309 16.481 12.2426 20.687 1.000 56024
nu     32.689 29.3658 23.779  1.0204 91.216 1.000 16855
sigma1  2.066  0.5983  1.973  1.0538  3.272 1.002 25711
sigma2  6.860  1.8112  6.558  3.8922 10.549 1.000 31448

'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.
```



Within the t.test, from a Bayesian perspective (i.e., BESTmcmc()), we can see that 100% of the data is showing on the negative section of the graph, with an HDI spanning from -18.9 to -10.1. This indicates that we can expect, with 95% confidence, that the difference between spray C and F is within this range, with the most likely difference being the point-estimate, or -14.5. This would suggest that spray C has a disadvantage (i.e., negative performance) of 14.5 units compared to spray F; if we also compare this to the frequentest perspective, we observe a p-value significantly less than the conventional alpha value of 0.05 -- so we can reject the null hypothesis, which assumed there was no difference between the two variables.