IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: September 22, 2020
Chapter Number: # 6
Title of Chapter: Comparing Groups and Analyzing Experiments

ANOVA

Analysis of variance (ANOVA): examine combinations of different factors.

- Core concept of ANOVA is that it uses two different variance estimates to assess whether group means differ.
- The technique partitions the overall variance among all of the observations into **between-groups** (the variability among means) and **within-groups variance** (the variability of scores with respect to their group means) to evaluate whether the samples might have come from the same underlying population.
- By pooling together all of the data we have collected from all of the different groups, we can create a reasonable estimate of the underlying population variance, under the assumption that all of the data from all of the groups was drawn from the same population.
- If one or more of the groups have been sampled from a population with a very different mean, however, the variance among the sample means will tend to exceed what we would expect if the samples all came from the same population.

F-ratio: numerator of the ratio is the between-groups variance while the denominator of the ratio is the within-groups variance.

- Under the assumption (which is also the null hypothesis) that we are sampling all of the groups from the same population, *most of our F-ratios should be very close to 1.0*. Namely, if we draw multiple groups from the same population, the scaled between-groups variance and the within-groups variance will generally be about equal to each other.
- *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.*
- For an **ANOVA** result to **be statistically significant, *F* must substantially exceed one.** For ***F* to be significant** according to the logic of the null hypothesis significance test, **the value of Pr(>F) must be less than the alpha level chosen by the experimenter before conducting the test.**

Categorical vs Metrix:

- **Categorical variable**: variable indicates membership in two or more discrete categories.
- **Metric:** means that the variable ascertains an ordered measurement of some type, such as ratings on a 10-point scale, weight, or number of bytes of memory.

Bayes factor: it is nothing more or less than an odds ratio that results from the comparison of one hypothesis to another

- Every Bayes factor represents a comparison between two statistical models, such as an alternative hypothesis versus a null hypothesis. However, it is possible that both models that we are comparing may be poor choices.

- The strength of evidence you may need depends upon your research situation.

Formulas for ANOVA

Total Sum-of-Squares: $SS_{\text{total}} = \sum \left(x - \bar{G}\right)^2$ or $SS_{\text{total}} = SS_{\text{within}} + SS_{\text{between}}$

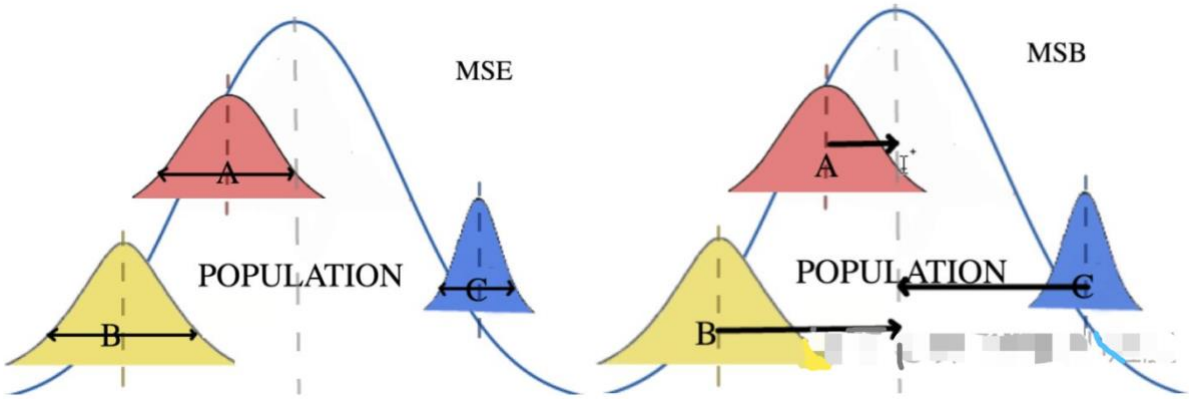
The total sum-of-squares is the sum, across all scores (x), of *the squared difference between each score and the* ***grand mean*** (which is the mean of all of the scores in the whole data set). This formula might seem very familiar because it is at the heart of calculating the regular old variance and the standard deviation.

Within-Groups Sum-of-Squares: $SS_{\text{within}} = \sum \sum \left(x_{ij} - \bar{X}_j\right)^2$

The within-groups sum-of-squares consists of the *sum of the squared deviations of each score from its respective group mean, with the results from all the groups summed together*.

Between Groups Sum-of-Squares: $SS_{\text{between}} = \sum n \left(\bar{X}_j - \bar{G}\right)^2$

The *squared deviation between each group mean and the grand mean and adds the results together*. Note that after calculating each squared deviation, the result is multiplied by the number of observations in that particular group.



More Information about Degrees of Freedom

Sample variance (aka. unbiased estimator): $s^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{(n-1)}$

The sample variance is a fraction where the *numerator contains the sum of all squared deviations from the sample mean and the denominator is 1 less than the sample size.* The use of *(n–1) naturally makes the variance slightly larger* than it would be if you used *n* in the denominator, suggesting that if we were to (incorrectly) use *n* instead, it might lead to an underestimation of population variance.

Giving Some Thought to Priors

- This insight has led some statisticians toward the idea that **uninformative pri- ors** may suffice except when working with small samples of data.
- **Noncommittal priors** signify that they are perhaps more meaningful than uninformative priors, but without making a commitment to specific values.

Exercise Review

1.
The dependent variable is the count (number of insects killed by the spray) whist the independent variable is the brand or type of spray. The total number of observations is 72.

2.
```
> insect <- aov(count ~ spray, data=InsectSprays)
> summary(insect)
            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
```
533.8 is for between-groups variance whist 15.4 is for within-groups variance.

3.
The F-ratio is 34.7 (which is absolutely higher than 1.0), indicating that at least one of the groups is from a population with a different mean. What's more, the F value is significant since the Pr(>F) value is significantly lower than 0.001. Therefore, it is reasonable to reject the null hypothesis.

4.
Df for between-group is 5 whist the df for within-group is 66. The total observation is 72. Intuitively, the df should be 72. However, the statisticians call the sample variance (with denominator n–1) an unbiased estimator**.** Therefore, using n-1 is been perceived as being more properly than using n since using (n–1) in the denominator of the variance calculation corrects for the uncertainty raised by sampling error. Then we only have df = 71.
Since we still consider calculating the between-group- and inside-group-variance, we need to spear 5 df to represent the between-group variance.
With total *df* = 71 and between-groups borrowing 5 *df* from the total, the remaining degrees of freedom are allocated to the within-groups variance (*df* = 66). Together, between-groups *df* and within-groups *df* always add up to total *df* (in this case 5 + 66 = 71).

5.
```
> insect <- aov(count ~ spray, data=InsectSprays)
> summary(insect)
            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
```

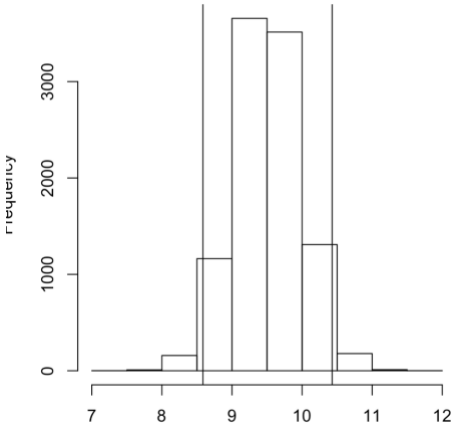
IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Null: All six groups were sampled from the same population such that any variation among means was attributable to sampling error.
Alternative: all six groups were not sampled from the same population such that any variation among means was not only attributable to sampling error.
Conclusion: reject the null hypothesis.

6.
Code:

```
result <- anovaBF(count ~ spray, data=InsectSprays)
mcmcOut2 <- posterior(result,iterations=10000)

hist(mcmcOut2[,"mu"])
abline(v=quantile(mcmcOut2[,"mu"],c(0.025)), col='black')
abline(v=quantile(mcmcOut2[,"mu"],c(0.975)), col='black')
quantile(mcmcOut2[,"mu"],c(0.025)) # 8.585475
quantile(mcmcOut2[,"mu"],c(0.975)) # 10.42916
```




```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

           Mean     SD Naive SE Time-series SE
mu        9.504 0.4728 0.004728       0.004728
spray-A   4.828 1.0473 0.010473       0.010871
spray-B   5.644 1.0379 0.010379       0.010614
spray-C  -7.144 1.0542 0.010542       0.010837
spray-D  -4.431 1.0395 0.010395       0.010640
spray-E  -5.785 1.0524 0.010524       0.010786
spray-F   6.889 1.0499 0.010499       0.010768
sig2     16.094 2.8930 0.028930       0.034570
g_spray   3.431 3.3321 0.033321       0.034873
```

The boundaries of the 95% HDI is 8.585475 to 10.42916. Here the HDI is not belonged to any groups. It is the HDI for the grand mu. Accordingly, most of the means are deviated from the grand mu (which is 9.504). Therefore, there are groups which deviate meaningfully from the population, indicating that the group means are credibly different from the grand mean and that these six groups were not drawn from the same population.

Result of the Bayes Factor

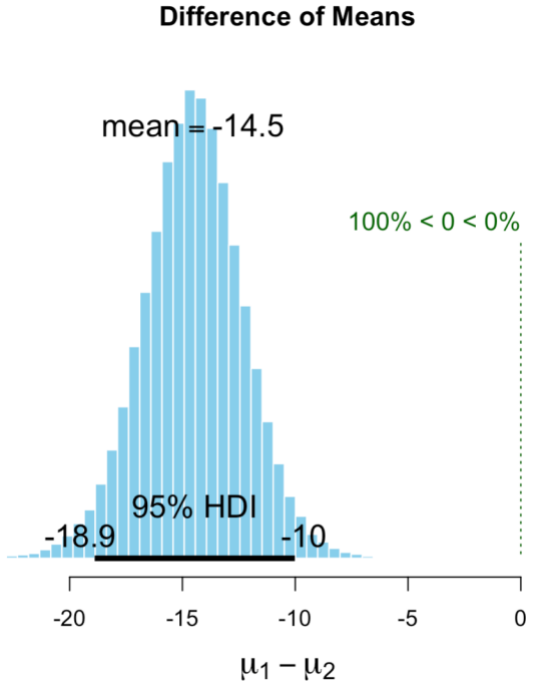
```
> result # to get the Bayes Factor
Bayes factor analysis
--------------
[1] spray : 1.506706e+14 ±0%

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

This analysis shows odds of 150,670,600,000,000:1 in favor of the alternative hypothesis. 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.

7.

```
plot(BESTmcmc(InsectSprays[InsectSprays$spray=='C',1],
              InsectSprays[InsectSprays$spray=='F',1]))
```



The interpretation is that 95% of the likely values of the population mean difference lie in the bell-shaped area between -10 and -18.9, indicating that the number of insects killed by Spray C is lower than that of insects killed by Spray F. Spray F has a better effect.

R Code Fragment and Explanation

Code: precipOut <- aov(precipAmount ~ precipGrp, data=precipDF)
Explanation: Test precipAmount as the dependent variable and make it a function of the independent variable(s) that follow the "~" character (in this case just the grouping variable precip-Grp).

Code: summary(precipOut)

Result:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
precipGrp	2	90	45.11	0.247	0.782
Residuals	57	10404	182.53		

- f—degrees of freedom: A statistical quantity indicating how many elements of a set are free to vary once an initial set of statistics has been calculated; from a data set of 60 we lose one degree of freedom for calculating the grand mean; among the three group means only two can vary freely; this leaves 57 degrees of freedom within groups (aka residuals);
- Sum Sq—sum of squares: A raw initial calculation of variability; the first line is the "between-groups" sum of squares discussed above; the second line is the "within-groups" sum of squares;

- Mean Sq—mean squares, the sum of squares divided by the degrees of freedom, aka variance: the first line is the "between-groups" variance; the second line is the "within-groups" variance;
- **F-value**—the *F*-ratio: quite literally a ratio of the mean squares from the first line (between groups) and the mean squares of the second line (within groups), that is, 43.47 divided by 180.19;
- **Pr(>F)—**the probability of a larger *F*-ratio: when examining a random distribution of F-ratios for the degrees of freedom appearing in this table, this is the probability of finding an *F*-value at least this high (in this case at least 0.247, which is a really small *F*-value). The F distribution only has a positive tail, so for us to reject the null hypothesis, we must look for extreme values of *F* that appear in the tail of the distribution.
- Residuals—this line accounts for all of the within-groups variability. A residual is what is left over when all of the systematic variance is removed—in this case everything that is left over after "precipGrp" is taken into account.

Code for Bayes Factors:

```
chicksBayesOut <- anovaBF(weight ~ feed, data=chickwts)
mcmcOut2 <- posterior(chicksBayesOut,iterations=10000)
boxplot(as.matrix(mcmcOut2[,2:7]))
summary(mcmcOut2)
chicksBayesOut # to get the Bayes Factor
```

Your Question:

1. When is the good time to use aov() when is good for anovaBF()? Is there a difference between using these two functions?