IST-772 Quantitative Reasoning in Data science

Week6/HW-6: ANOVA & Experimental Groups

Comparing Groups and Analyzing Experiments (Page 117-118: Problems 1-7)

Sharat Sripada (vssripad@syr.edu)

1. What are the dependent and independent variables in the InsectSprays dataset? Also, what are the total number of observations?

The InsectSprays dataset in R shows this:

```
> str(InsectSprays)
'data.frame':  72 obs. of  2 variables:
 $ count: num  10 7 20 14 14 12 10 23 17 20 ...
 $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Based on our learning in this module, categorical variables (factors/level) will be independent variables and the remaining will be dependent variables. Based on the output above:

- Independent variable: spray
- Dependent variable: count

From the same output we can see that the number of observations is 12.

2. After running the aov(), the Mean Sq for spray is 533.8 and Mean Sq for Residuals is 15.4. Which one of these the between-groups and variance, and which one of these is the within-groups variance?

The Mean Sq for spray is the between groups variance while the Mean Sq for Residuals is the within groups variance.

The Mean sq is obtained by first calculating the Sum of squares (SS) and then diving it by the Degrees of Freedom (DF)

SS between groups is calculated using formula:
sigma n (X-bar_j – G-bar) ^ 2

where X-bar is the corresponding group mean and G-bar is the Grand Mean.

SS within groups is calculated using formula:
sigma sigma (x_{ij} – X-bar_j) ^ 2

3. Calculate the F-ratio and say if you can reject null hypothesis. State why or why not

The F-ratio can be obtained by dividing the Mean squares.

F-ratio (5, 66) = 533.8/15.4 = 34.66

For an ANOVA result to be statistically significant the F-ratio should be substantially > 1.

However, to run a NULL hypothesis test we will obtain the significance level or P(>F) value based on the F-ratio and DF.

First, we will state the *NULL Hypothesis:*
All 6 groups were sampled from the same population

Using the calculator here https://www.socscistatistics.com/pvalues/fdistribution.aspx we obtain:
The *p*-value is < .00001

Since the p-value is << 0.05 we REJECT the NULL Hypothesis.

4. Calculate the DF between and within groups. Explain why the total is 1 less than the total number of observations in the dataset.

When obtaining the DF between groups, we subtract 1 from the total number of observations since we lose a DF when calculating the *Grand Mean* so, new df = 72 – 1 = 71.

Next, df between groups is obtained by subtracting 1 from the number of groups. This is because if we know the grand mean and means of all but one group, then that one group is no longer free to vary. So, df between groups = 6 – 1 = 5

Finally, the residual is 71 – 5 = 66

This is summarized as:

- DF between groups used along the independent variable Spray is 6 -1 = 5
- DF within groups used alongside the Residuals is 71 - 5 = 66

5. Run an ANOVA testing using the R aov() function. Discuss the results.

NULL Hypothesis: The 6 groups were sampled from the same population
Alternate Hypothesis: The 6 groups were sampled from different population

The output of aov and summary is captured below:

```
> insectOut <- aov(InsectSprays$count ~ InsectSprays$spray)
> summary(insectOut)
                   Df Sum Sq Mean Sq F value Pr(>F)
InsectSprays$spray  5   2669   533.8    34.7 <2e-16 ***
Residuals          66   1015    15.4
```

Given that Pr(>F) value is < 0.05 (alpha) we REJECT the null Hypothesis. This means, we REJECT the theory that the 6 groups were sampled from the same population.

6. Run the BayesFactor on the same dataset and state the strength of the result

Using the following R-code we can obtain the results for ANOVA based on Bayes theorem:

```
insect_df <- data.frame(InsectSprays$count, InsectSprays$spray)
insectBayesOut <- anovaBF(InsectSprays.count ~ InsectSprays.spray,
                          data=insect_df)
insectBayesOut
```

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

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

The result here is shows odds of 1.5:1. According to Kass and Raftery any ratio less than or equal to 3:1 is not worth mentioning.

7. Conduct a t-test on the same dataset. Interpret the results of the t-test

```
> grpc <- insect_df[insect_df$InsectSprays.spray == 'C', 1]
> grpf <- insect_df[insect_df$InsectSprays.spray == 'F', 1]
> t.test(grpc, grpf)

        Welch Two Sample t-test

data:  grpc and grpf
t = -7.7484, df = 13.201, p-value = 2.876e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.64308 -10.52358
sample estimates:
mean of x mean of y
 2.083333 16.666667
```

Based on the output, we can say that in 95% of the iterations, the true population mean difference lies in the interval (-18.64, -10.52).