Analysis of Variance (ANOVA)

First Name __________ Last Name ___________SID (last 4 digits only)_____

Insect sprays

The `InsectSprays` data in `R` gives the counts of insects in agricultural experimental units treated with six different insectisides. We can use this data to examine the relative effectiveness of the difference insectisides. (If the insectiside is effective, we would expect to see relatively few insects on the corresponding experimental unit.)

Exploratory analysis

Load the `InsectSprays` data set into our workspace.

```
data(InsectSprays)
?InsectSprays
```

We have observations on 2 different variables, one categorical and one numerical.

Exercise 1 What are the variables in this data set? (1pt) ___________________ (Hint: you can use the "names" function)
How many cases are in the sample? (1pt) ____________________________ (Hint: you can use the "dim" function)

As a first step in the analysis, we should consider summaries of the data.

```
summary(InsectSprays)
dim(InsectSprays)
hist(InsectSprays$count)
```

This gives information for the number of experimental units assigned to each spray and the spread of the number of insects across all units. It does not tell us anything about how these variables might interact. We can consider this interaction using side-by-side boxplots.

```
boxplot(InsectSprays$count ~ InsectSprays$spray)
```

Exercise 2 Based on the boxplot, does it suggest the `spray` used can affect the `count` ? (1pt) _________ (Yes/No)

ANOVA

To examine whether the conditions for ANOVA are satisfied, we can start by examining plots of the `count` for each `spray` group.

```
hist(InsectSprays$count[InsectSprays$spray == "A"], breaks=6,
     main ="Histogram of Count for Spray A", xlab="count")
```

Exercise 3 Repeat for sprays B - F.

Do the groups appear to be normally distributed for spray B? (1pt)______

Do the groups appear to be normally distributed for spray C? (1pt)______

Do the groups appear to be normally distributed for spray D? (1pt)______ Do the groups appear to be normally distributed for spray E? (1pt)______

Do the groups appear to be normally distributed for spray F? (1pt)______

We can also check for normality using a Shapiro-Wilk test. The null hypothesis for the Shapiro-Wilk test is H0: the population is normally distributed. This means that if we *reject*, then the data is *not* normally distributed. If we *fail to reject*, then we say that it is reasonable to assume normality.

```
shapiro.test(InsectSprays$count[InsectSprays$spray == "F"])
```

Exercise 4 Repeat for sprays B - F.

Is it reasonable to assume normality in sray B? (1pt) _______

Is it reasonable to assume normality in sray D? (1pt) _______

Is it reasonable to assume normality in sray D? (1pt) _______

Is it reasonable to assume normality in sray E? (1pt) _______

Is it reasonable to assume normality in sray F? (1pt) _______

Exercise 5 Find the standard deviation of `count` for each `spray` . You can use the following code to find standar deviation of counts in srapy A.

```
filter <- InsectSprays$spray == 'A'
sd(InsectSprays$count[filter])
```

Stardard devation for counts in spray (6pts)

sd_A = _________ sd_B =____________ sd_C = _____________

sd_D = _________ sd_E =____________ sd_F = _____________

Do you think it is reasonable to assume that the six groups have equal variances? (1pt)________(Yes/No)
You can also find the standard deviation by spray using the “aggregate” function

```
av <- aggregate(. ~spray, InsectSprays, function(x) sd = sd(x))
av
```

Exercise 6 Write the hypotheses for the ANOVA for these data (2pts)

We will assume that the assumptions are satisfied and will run the ANOVA. We do this using the function `aov` and the `summary` function. The first, `aov`, calculates the ANOVA. The latter `summary` formats the result into what ANOVA that we are familiar with.

```
summary(aov(InsectSprays$count ~ InsectSprays$spray))
```

Exercise 7 What can you conclude based on this ANOVA table? Test at the 5% level of significance.(2pts)

Pairwise Comparison

Finally, we want to examine where the differences are. We will use the function `pairwise.t.test` to examine all possible comparisons between the groups. Because we need to control for Type I error inflation, we will correct using the Bonferonni correction.

```
pairwise.t.test(x=InsectSprays$count, g=InsectSprays$spray, p.adj="bonf")
```

This outputs a matrix with the p-value for comparing each `spray` group, where the null hypothesis is that the means for the two groups are equal. Since probabilities are always between 0 and 1, if the Bonferonni correction makes a probability greater than 1, `R` will report 1. For example, for sprays A and B, the p-value for the test that mean(A) = mean(B) is 1.

Exercise 8 Which sprays are significantly different from one another? Among the following pairs, mark those are significantly different from each other. (3pts)

A and B ______
A and C ______
A and D ______ A and E ______ A and F _______ B and F _______ C and D _______
C and E _______

On your own

Use the following `R` code to read in a subset of the `chickwts` dataset in `R`. This data is for an experiment designed to measure the relative effectiveness of various feed supplements on chicken weights. Assume the different feeds are randomly assigned to the chickens. We are removing two of the feeds, horsebean and meatmeal, to simplify the problem slightly. To learn more about the full data, use `?chickwts`.

```
chickwts <- chickwts[which(chickwts$feed != "horsebean" & chickwts$feed != "meatmeal"),]
```

1. What are the hypotheses for the ANOVA corresponding to these data? (2pts) The null hypothesis is: The alternative hypothesis is:

2. Check the ANOVA condition for these data. (9pts)

   Independency: Normality: Equal Variances:

3. Assume the assumptions are satisfied. Conduct the ANOVA for these data. Include the ANOVA table in your report and make sure to report your results with a plain language explanation in the context of the study. (3pts)

   The ANOVA table: Conclusion:

4. Run a post hoc test to determine where the differences exist. If there are no differences, use your post hoc test to confirm. Report the all pairs where difference exist. (2pts)____