STAT 100B Lab 4

Wesley Chang

Summer 2020 Session B

Setup for Lab

```r
# load InsectSprays data into workspace
data(InsectSprays)
```

Lab Exercises

Exercise 1

- *What are the variables in this data set?*
- *How many cases are in the sample?*

```r
# number of variables
names(InsectSprays)
```

```
## [1] "count" "spray"
```

```r
# number of cases
dim(InsectSprays)
```

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

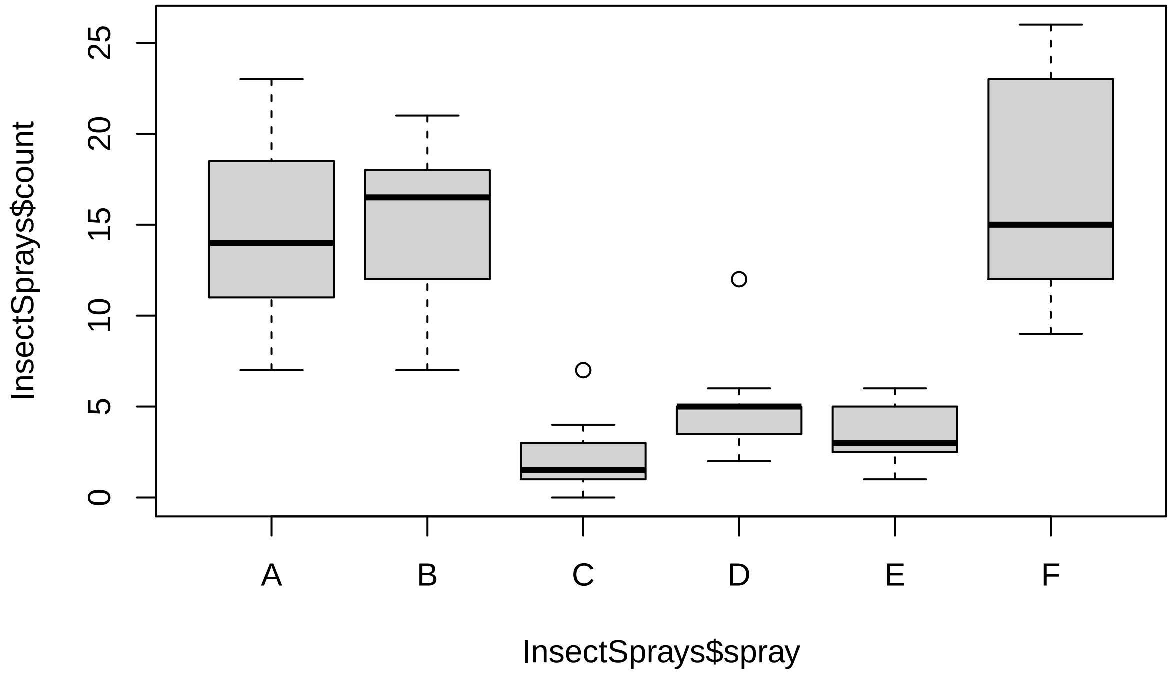
Answer

There are two variables in this data set, "count" and "sprays", and 72 cases.

Exercise 2

Based on the boxplot, does it suggest the `spray` *used can affect the* `count`*?*

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



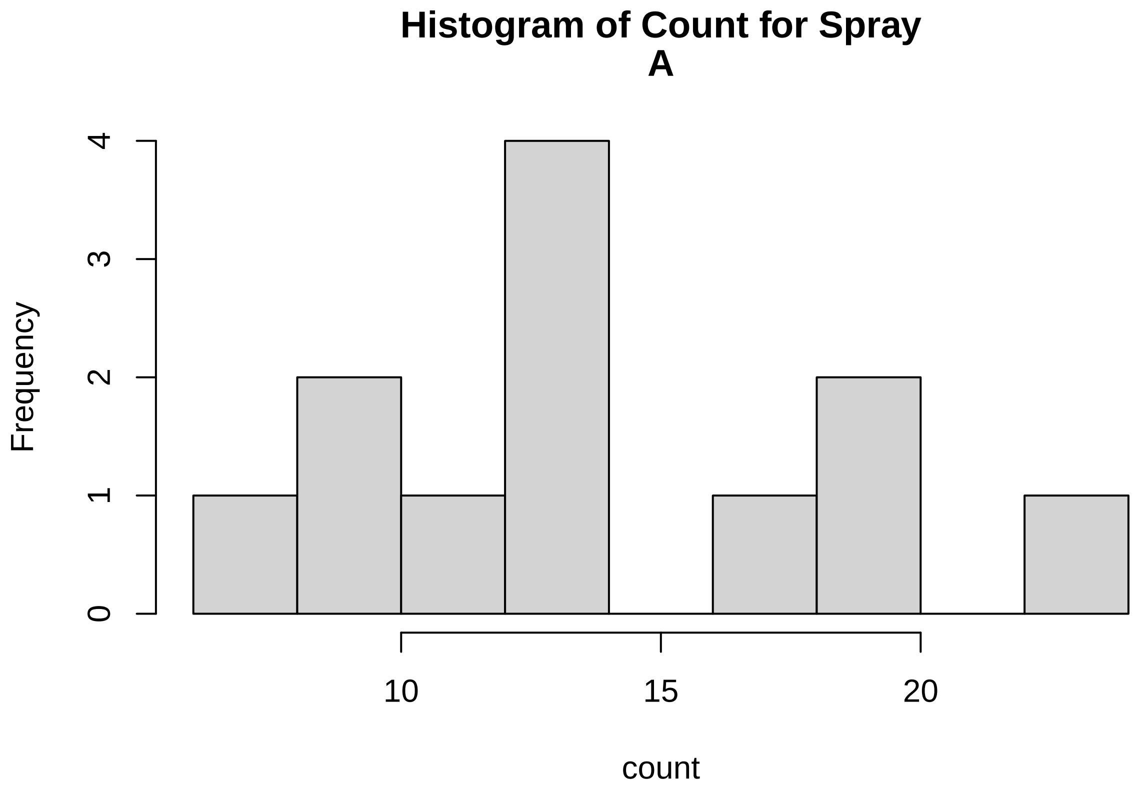
Answer

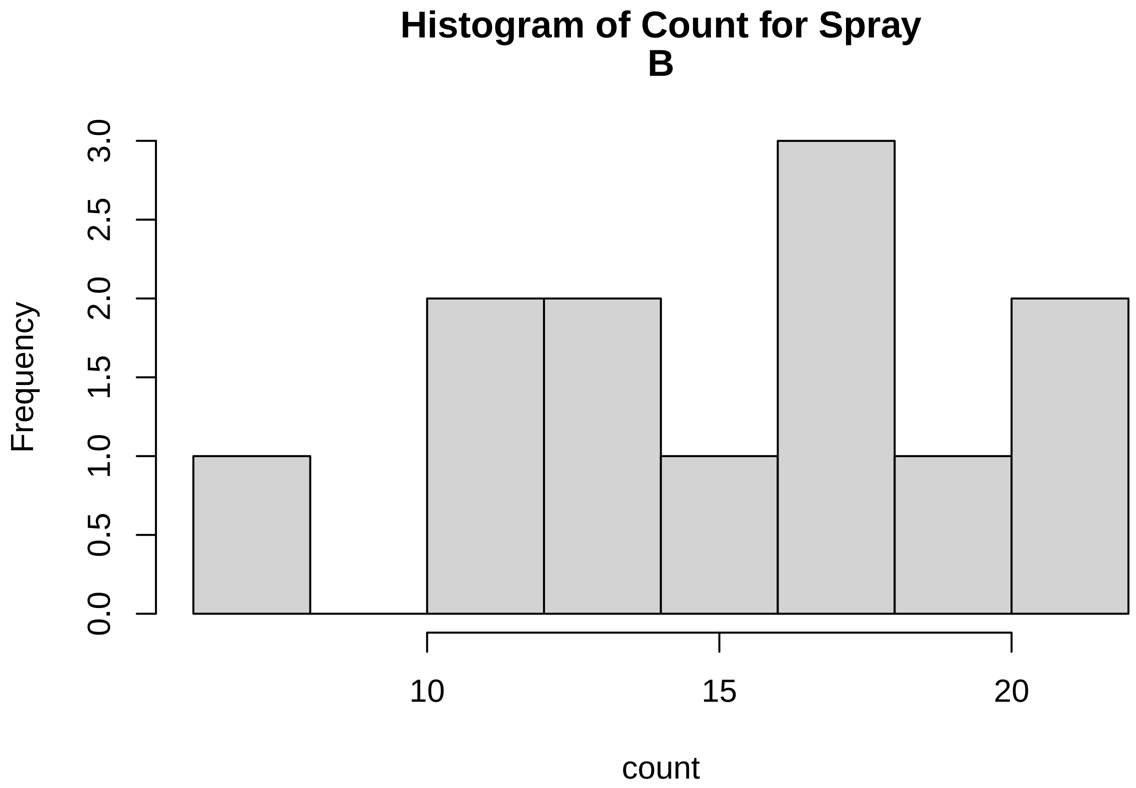
Yes, it does appear like the `spray` used can affect the count.

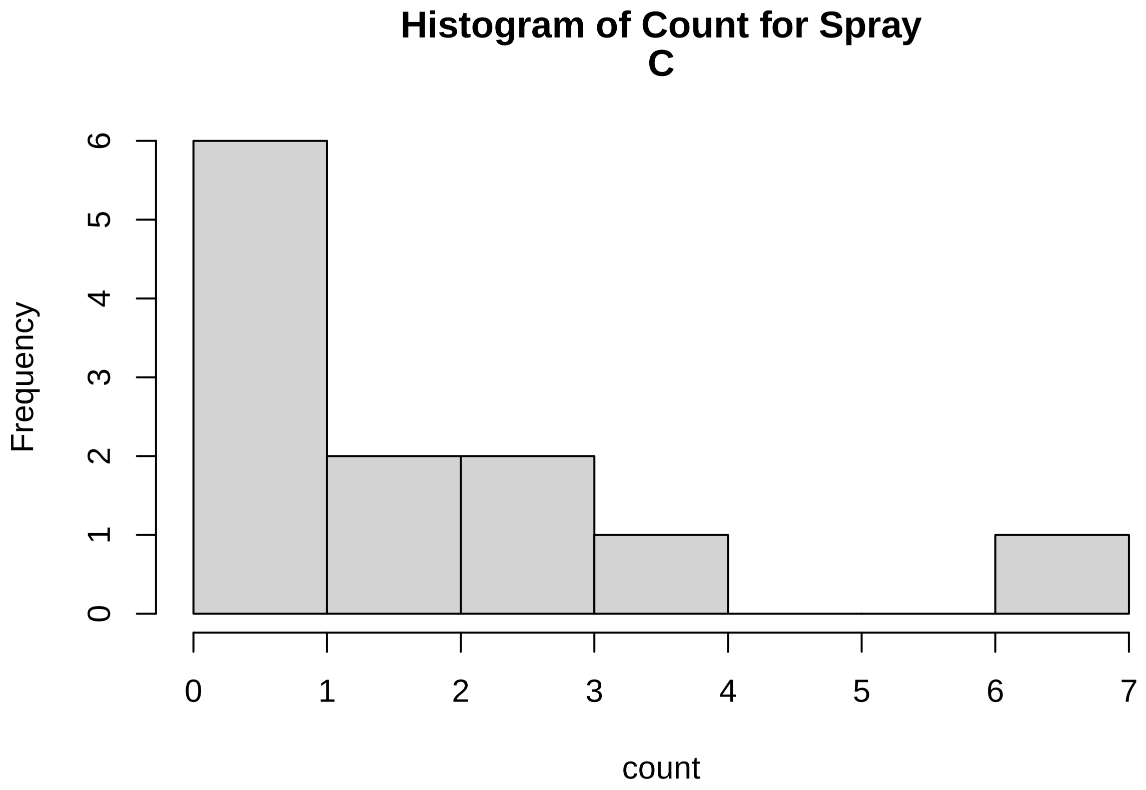
Exercise 3

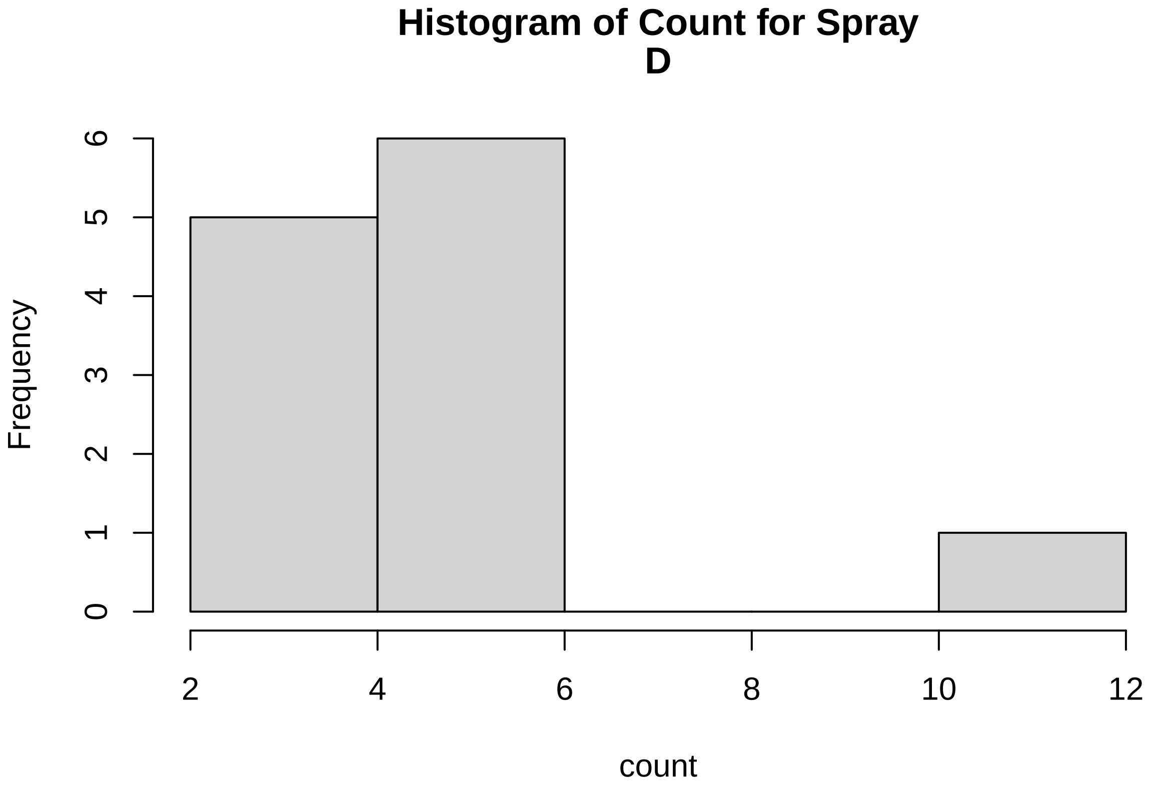
Generate plots of `count` *for each* `spray` *group, A-F*

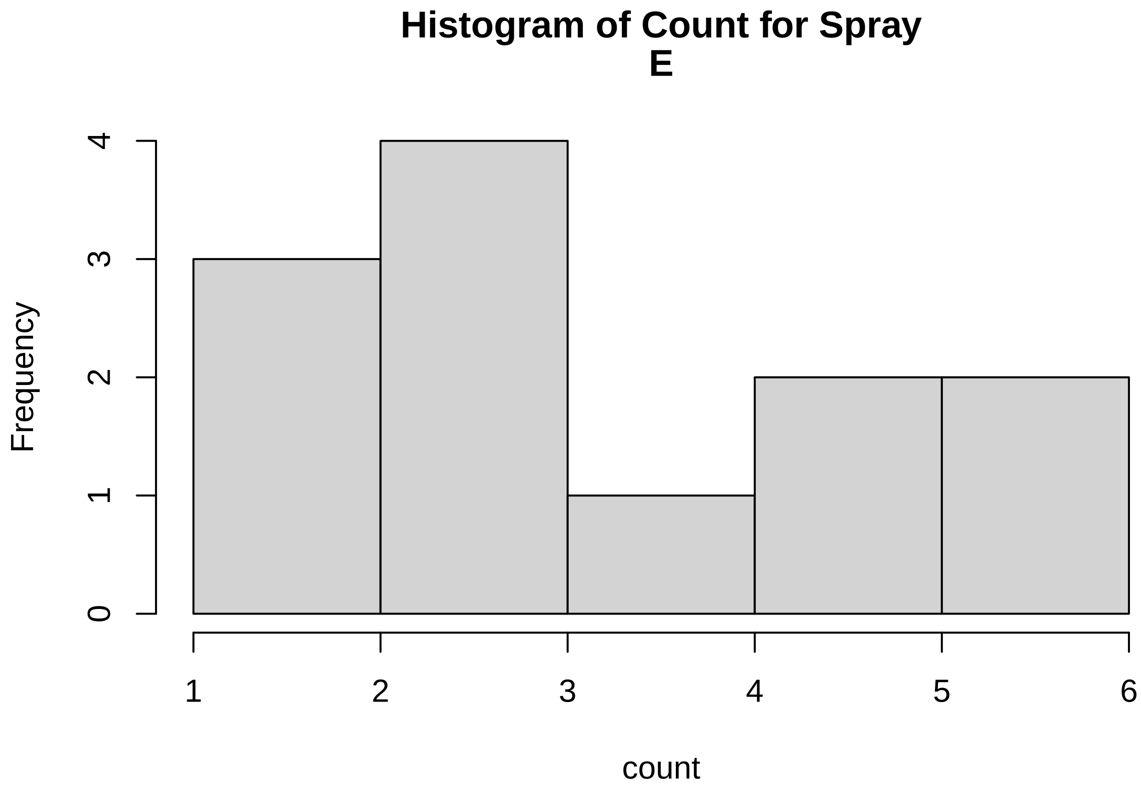
```
# using for loop to generate histogram of counts
for(i in levels(InsectSprays$spray)){
    hist(InsectSprays$count[InsectSprays$spray == i],
    breaks=6, main =c("Histogram of Count for Spray", i), xlab="count")
  }
```

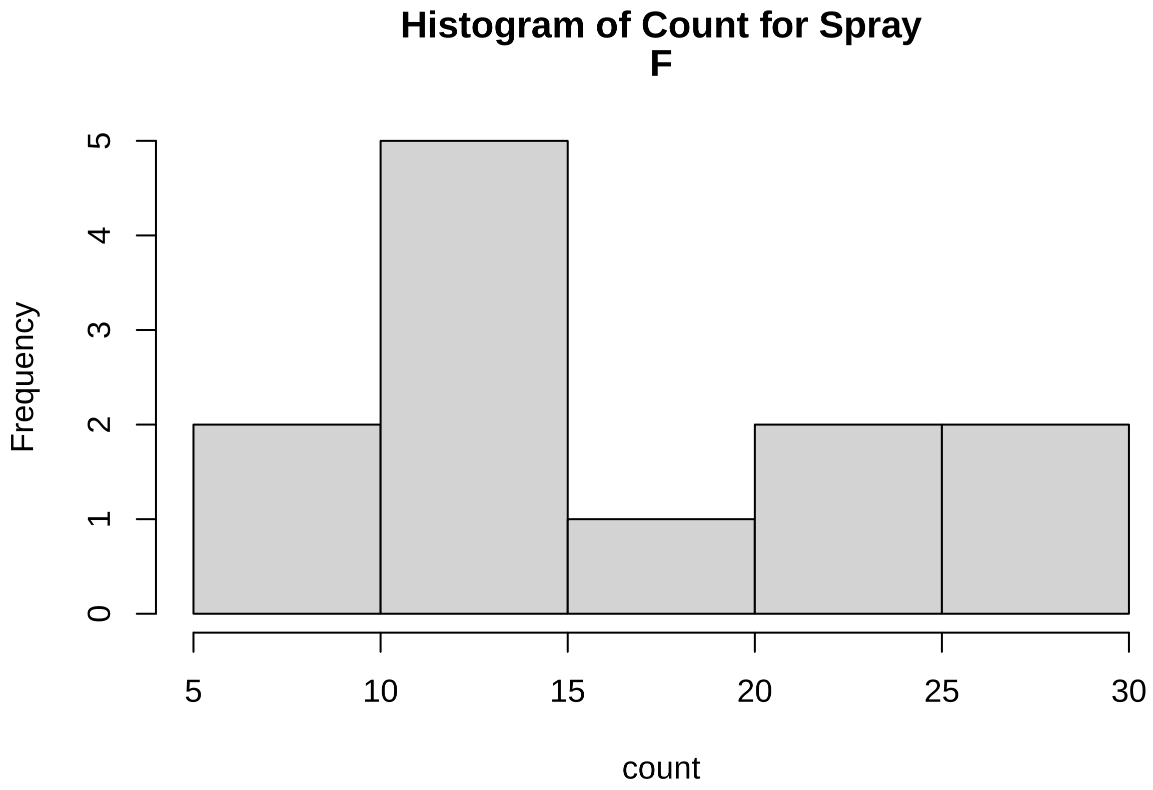

Histogram of Count for Spray
A
Frequency
0
1
2
3
4
10
15
20
count


Histogram of Count for Spray B
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
10
15
20
count


Histogram of Count for Spray
C
Frequency
0
1
2
3
4
5
6
0
1
2
3
4
5
6
7
count


Histogram of Count for Spray
D
Frequency
0
1
2
3
4
5
6
2
4
6
8
10
12
count


Histogram of Count for Spray E
Frequency
0
1
2
3
4
1
2
3
4
5
6
count



Answer

The groups appear to only be normally distributed for groups A and B. The rest appear skewed.

Exercise 4

Test for normality using the Shapiro-Wilk test for groups A-F

```
# shapiro.test
# using for loop
for(i in levels(InsectSprays$spray)){
  print(paste0("Spray ", i))
  print(shapiro.test(InsectSprays$count[InsectSprays$spray == i]))
  }
```

```
## [1] "Spray A"
##
##  Shapiro-Wilk normality test
##
## data:  InsectSprays$count[InsectSprays$spray == i]
## W = 0.95757, p-value = 0.7487
##
## [1] "Spray B"
##
##  Shapiro-Wilk normality test
##
```

```
## data:  InsectSprays$count[InsectSprays$spray == i]
## W = 0.95031, p-value = 0.6415
##
## [1] "Spray C"
##
##  Shapiro-Wilk normality test
##
## data:  InsectSprays$count[InsectSprays$spray == i]
## W = 0.85907, p-value = 0.04759
##
## [1] "Spray D"
##
##  Shapiro-Wilk normality test
##
## data:  InsectSprays$count[InsectSprays$spray == i]
## W = 0.75063, p-value = 0.002713
##
## [1] "Spray E"
##
##  Shapiro-Wilk normality test
##
## data:  InsectSprays$count[InsectSprays$spray == i]
## W = 0.92128, p-value = 0.2967
##
## [1] "Spray F"
##
##  Shapiro-Wilk normality test
##
## data:  InsectSprays$count[InsectSprays$spray == i]
## W = 0.88475, p-value = 0.1009
```

Answer

The Shapiro-Wilks test checks for normality with a null hypothesis of normality. Therefore, when the results of the test generate a p-value less than or equal to 0.05, the hypothesis of normality is rejected. For the Sprays A-F, we reject normality in Sprays C and D. Sprays A, B, E, and F are considered normal according to this test.

Exercise 5

Find the standard deviation of `count` *for each* `spray` *A to F.*

Answer

```r
## Method 1
# using filter() function
# for loop

for(i in levels(InsectSprays$spray)){
  print(paste0("Spray ", i))
  filter <- InsectSprays$spray == i
  print(sd(InsectSprays$count[filter]))
}
```

```
## [1] "Spray A"
## [1] 4.719399
## [1] "Spray B"
## [1] 4.271115
## [1] "Spray C"
## [1] 1.975225
## [1] "Spray D"
## [1] 2.503028
## [1] "Spray E"
## [1] 1.732051
## [1] "Spray F"
## [1] 6.213378
```

```
## Method 2
# using aggregate() function
av <- aggregate(. ~spray, InsectSprays, function(x) sd = sd(x))
av
```

```
##   spray    count
## 1     A 4.719399
## 2     B 4.271115
## 3     C 1.975225
## 4     D 2.503028
## 5     E 1.732051
## 6     F 6.213378
```

I do not think it is reasonable to assume that the six groups have equal variances.

Exercise 6

Write the hypotheses for the ANOVA for these data

The null hypothesis for the ANOVA of these data is that the means of the effect of each insecticide on the count of insects are equal. The alternate is that at least one or more of these means are not equal.

Exercise 7

What can you conclude based on this ANOVA table? Test at the 5% level of significance.

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

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## InsectSprays$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 At the 5% level of signficance with alpha = 0.05, we reject the null hypothesis based on the small p-value `<2e=16`, and do not conclude that *all* of the means are equal.

Exercise 8

Which sprays are significantly different from one another? Among the following pairs, mark those that are significantly different from each other.

```r
# use a pairwise t test to examine all possible comparisons between the groups
# to control for Type 1 error inflation, we use the Bonferonni correction

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

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  InsectSprays$count and InsectSprays$spray
##
##   A       B       C       D       E
## B 1       -       -       -       -
## C 1.1e-09 1.3e-10 -       -       -
## D 1.5e-06 1.8e-07 1       -       -
## E 4.1e-08 4.9e-09 1       1       -
## F 1       1       4.2e-12 6.1e-09 1.6e-10
##
## P value adjustment method: bonferroni
```

Answer

Significantly Different *(p-values are less than or equal to 0.05)*: A and C, A and D, A and E, B and C, B and D, B and E, C and F, D and F, E and F.

Not Significantly Different *(p-values are greater than 0.05)*: A and B, A and F, B and F. C and D, C and E, D and E.

Significantly Different	Not Significantly Different
A and C	A and B
A and D	A and F
A and E	B and F
B and C	C and D
B and D	C and E
B and E	D and E
C and F	—
D and F	—
E and F	—
p-value $\leq$ *0.05*	*p-value* $>$ *0.05*

On Your Own

```r
# setup data
chickwts <- chickwts[which(chickwts$feed != "horsebean" & chickwts$feed != "meatmeal"),]
chickwts$feed <- factor(chickwts$feed)
```

Problem 1

What are the hypothesis for the ANOVA corresponding to these data?

Answer

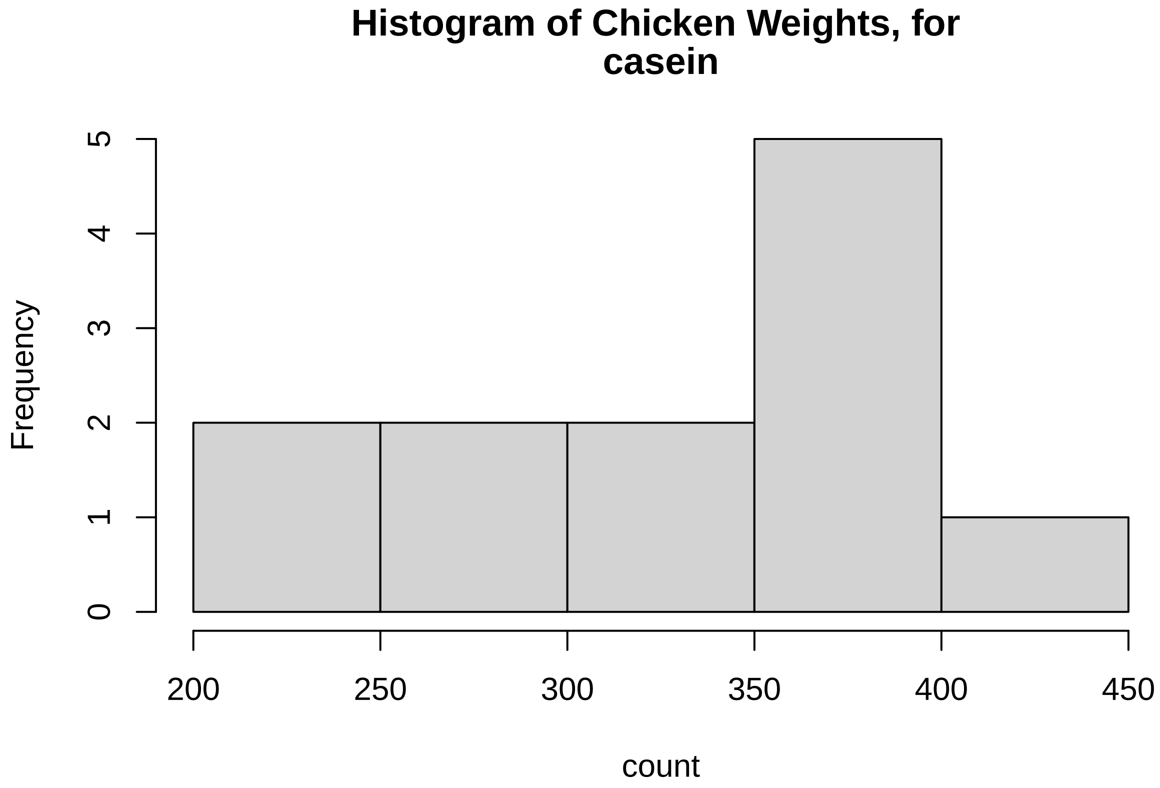
The null hypothesis is that the means of chicken weights across feed types are equal. The alternate hypothesis is that one or more of these means are not equal.

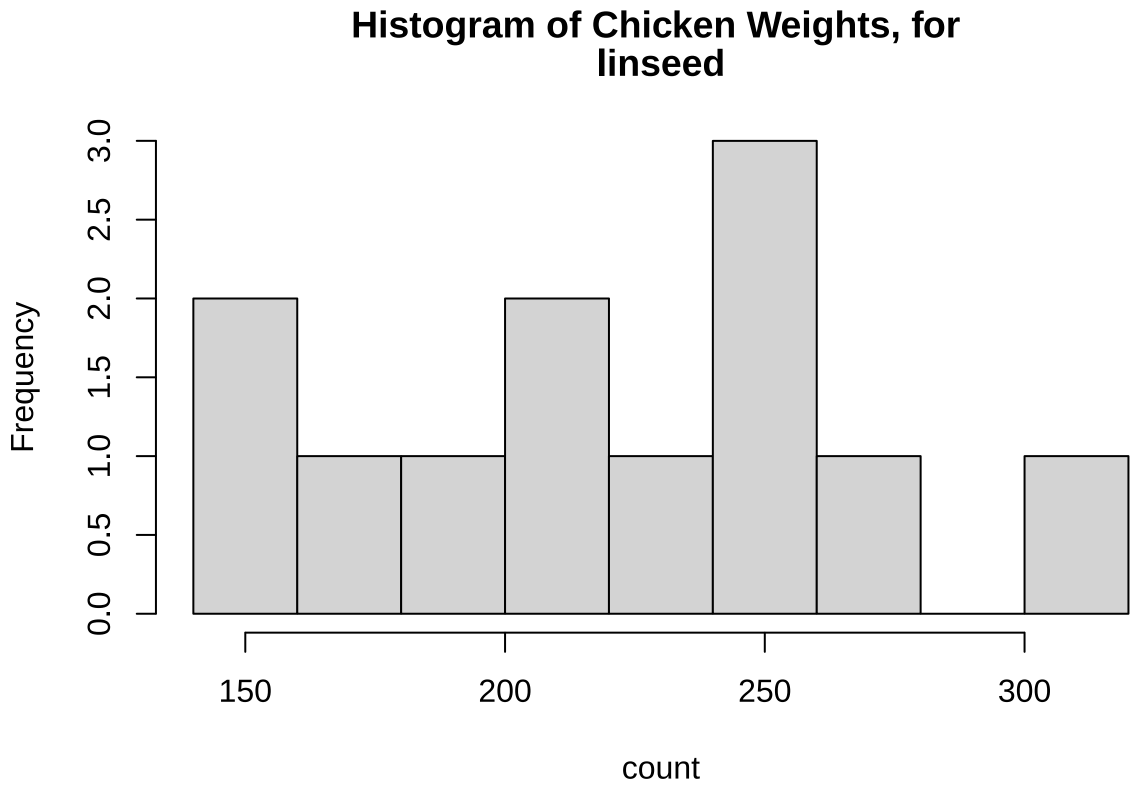
Problem 2

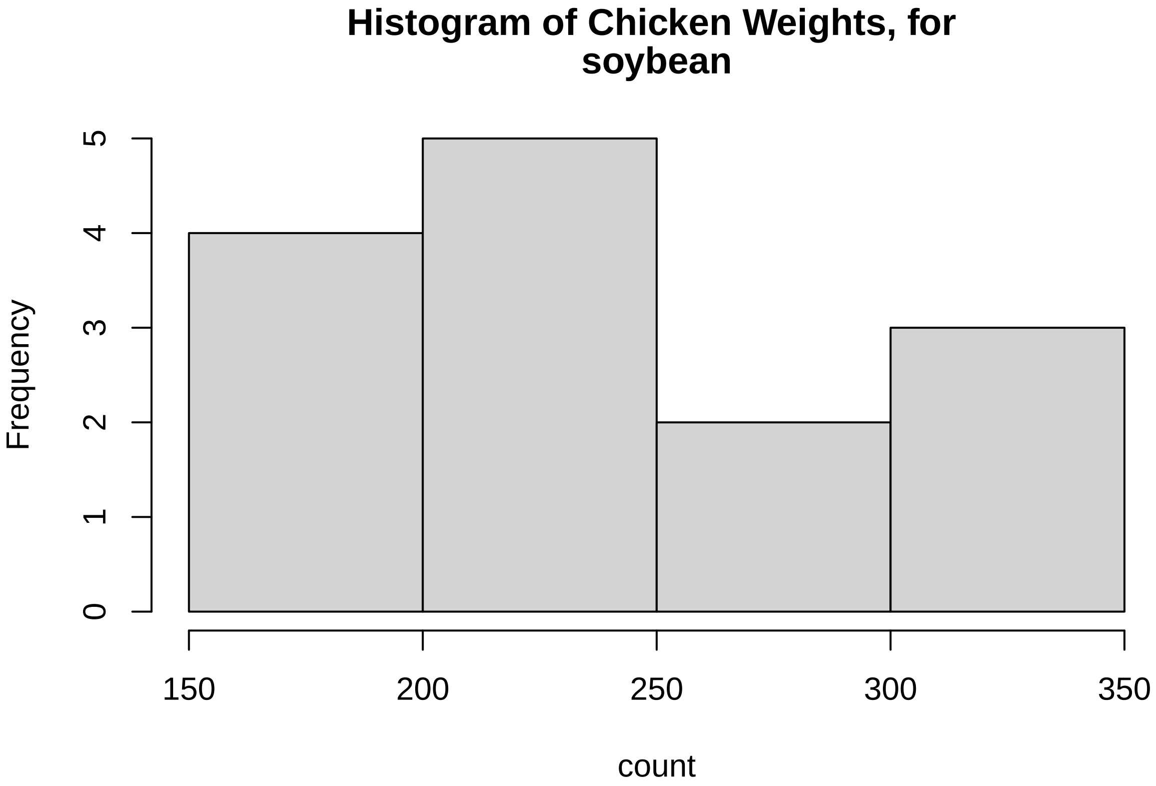
Check ANOVA conditions of Independency, Normality, and Equal Variances.

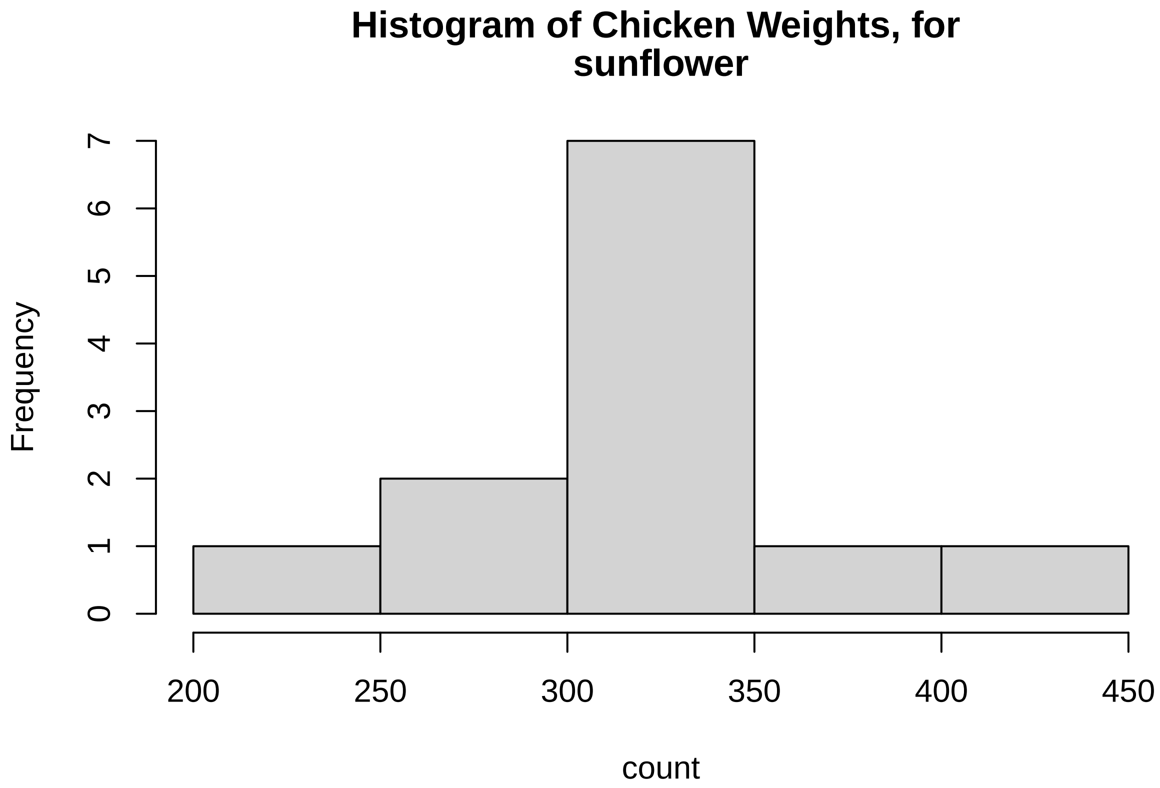
```
## checking for normality
# Two methods

# 1: compare histograms
for(i in levels(chickwts$feed)){
    hist(chickwts$weight[chickwts$feed == i],
    breaks=6, main =c("Histogram of Chicken Weights, for ", i), xlab="count")
}
```




Histogram of Chicken Weights, for linseed
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
150
200
250
300
count


Histogram of Chicken Weights, for
soybean
Frequency
0
1
2
3
4
5
150
200
250
300
350
count



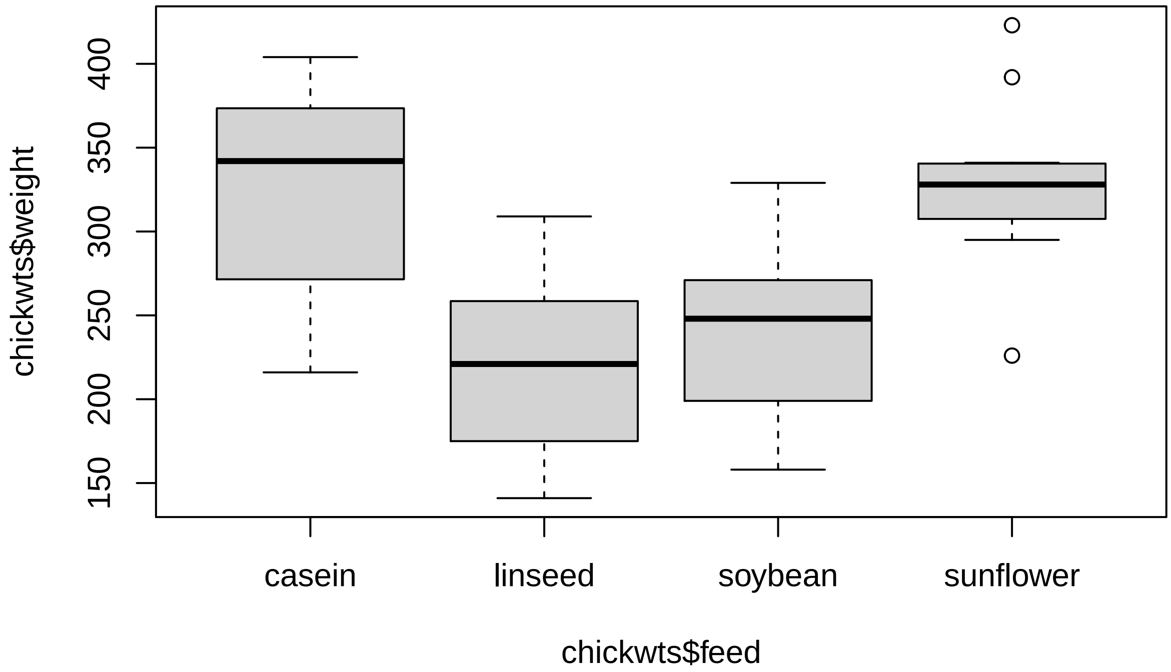
```r
# 2: shapiro test
for(i in levels(chickwts$feed)){
  print(paste0("Normality test for ", i))
  print(shapiro.test(chickwts$weight[chickwts$feed == i]))
}
```

```
## [1] "Normality test for casein"
##
##  Shapiro-Wilk normality test
##
## data:  chickwts$weight[chickwts$feed == i]
## W = 0.91663, p-value = 0.2592
##
## [1] "Normality test for linseed"
##
##  Shapiro-Wilk normality test
##
## data:  chickwts$weight[chickwts$feed == i]
## W = 0.96931, p-value = 0.9035
##
## [1] "Normality test for soybean"
##
##  Shapiro-Wilk normality test
##
## data:  chickwts$weight[chickwts$feed == i]
## W = 0.9464, p-value = 0.5064
```

```
##
## [1] "Normality test for sunflower"
##
##  Shapiro-Wilk normality test
##
## data:  chickwts$weight[chickwts$feed == i]
## W = 0.92809, p-value = 0.3603
```

```
## checking for equality of variances
# two ways

# 1: compare boxplots
boxplot(chickwts$weight ~ chickwts$feed)
```



```
# 2: compare variances

## method i
for(i in levels(chickwts$feed)){
  print(paste0("Variance of ", i))
  filter <- chickwts$feed == i
  print(var(chickwts$weight[filter]))
}
```

```
## [1] "Variance of casein"
## [1] 4151.72
## [1] "Variance of linseed"
## [1] 2728.568
## [1] "Variance of soybean"
## [1] 2929.956
## [1] "Variance of sunflower"
## [1] 2384.992
```

```
## method ii
av <- aggregate(. ~feed, chickwts, function(x) var = var(x))
av
```

```
##        feed   weight
## 1    casein 4151.720
## 2   linseed 2728.568
## 3   soybean 2929.956
## 4 sunflower 2384.992
```

Answer

- **Independency**: Satisfied by random assignment of feed supplements.
- **Normality**: Satisfied, based on the results of the Shapiro-Wilks test (results of all weights for each feed have a p-value greater than 0.05, do not reject null hypothesis).
- **Equality of Variances**: By directly comparing the variances, I would not think that it is reasonable to assume that there is equality of variances.

Problem 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.

```
# anova using aov() and summary()
summary(aov(chickwts$weight ~ chickwts$feed))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## chickwts$feed  3 112125   37375   12.28  5e-06 ***
## Residuals     46 140008    3044                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Answer

Based on the results of the ANOVA with a p-value `5e-06`, we reject the null hypothesis at the 0.05 level of significance. We conclude that one or more of the means are not equal. In other words, at least one of the mean weights of the chicken grouped by feed type are not equal.

Problem 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 all pairs where a difference exists

```r
# run post hoc test, using a pairwise t test
# we know that there is a difference in at least one of the means
# now we want to know which one(s) are different
# use the bonferonni correction

pairwise.t.test(x=chickwts$weight, g=chickwts$feed, p.adj="bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  chickwts$weight and chickwts$feed
##
##           casein  linseed soybean
## linseed   0.00017 -       -
## soybean   0.00533 1.00000 -
## sunflower 1.00000 7.6e-05 0.00254
##
## P value adjustment method: bonferroni
```

Answer

Significantly Different *(p-values are less than or equal to 0.05)*: casein and linseed, casein and soybean, linseed and sunflower, soybean and sunflower

Not Significantly Different *(p-values are greater than 0.05)*: casein and sunflower, linseed and soybean

Significantly Different	Not Significantly Different
casein and linseed	casein and sunflower
casein and soybean	linseed and soybean
linseed and sunflower	—
soybean and sunflower	—
$p\text{-value} \leq 0.05$	$p\text{-value} > 0.05$