STAT 313/513 Assignment 1
Due Monday, 4/5

Melons, p. 20
Read the problem description, but answer the following questions rather than those listed in the text.

1. What are the null and alternative hypotheses relevant to this problem?

   \[ H_0: \text{The mean yield is the same for all four varieties of melon} \]
   \[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]

   \[ H_A: \text{The mean yield is not the same for all four varieties of melon} \]
   \[ H_A: \text{the } \mu_i's \text{ are not all equal} \]

2. Using R, create a plot similar to the one in Figure 1.15 using the stripchart command. The data is downloadable from the data page on our website. Display your code and plot (suitable for publication) as the solution to this problem.

   ```r
   > with(melons, stripchart(Yield ~ Variety, vertical = TRUE,
   + pch = 16, xlab = 'Melon Variety', ylab = 'Yield',
   + xlim = c(0.5, 4.5)))
   ```

   Note: I’d like to specify the units of \textit{Yield}, but alas, the book provides no information about the units.

3. Using R, compute the mean, standard deviation, standard error of the mean, and sample size of the yields for each of the four melon varieties. Display the code used to produce the summaries and a formatted table (in word) that presents your results (suitable for publication).
> rbind(Mean = yield.m, SD = yield.sd, SE = yield.se, n = yield.n)
A B C D
Mean 20.490000 37.403333 20.462500 29.896667
SD 4.694422 3.950497 4.764692 2.2298939
SE 1.916490 1.612784 2.382346 0.9103504
n 6.000000 6.000000 4.000000 6.000000

This is decent, but it would be better to format this into a proper table:

<table>
<thead>
<tr>
<th>Yield</th>
<th>Melon Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Mean</td>
<td>20.490</td>
</tr>
<tr>
<td>SD</td>
<td>4.694</td>
</tr>
<tr>
<td>SE</td>
<td>1.916</td>
</tr>
<tr>
<td>n</td>
<td>6</td>
</tr>
</tbody>
</table>

4. Use R to conduct an ANOVA that addresses your hypotheses in part 1. Display your code, the ANOVA output, and your conclusions regarding these hypotheses being sure to cite relevant quantities in the ANOVA output to defend your answer.

> ml = aov(Yield ~ Variety, data = melons)
> summary(ml)

    Df Sum Sq Mean Sq  F value    Pr(>F)
Variety     3 1115.28 371.760 23.7980 1.735e-06 ***
Residuals 18  281.19  15.622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

There is very strong evidence ($F_{3,18} = 23.80, P < 0.0001$) that the mean yield is not the same for all four varieties.

If you want to take this problem further (not required) you could also run the following:

> TukeyHSD(ml, ordered = TRUE)

Tukey multiple comparisons of means
95% family-wise confidence level
factor levels have been ordered

Fit: aov(formula = Yield ~ Variety, data = melons)

$Variety$

<table>
<thead>
<tr>
<th></th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-C</td>
<td>0.027500</td>
<td>-7.183145</td>
<td>7.238145</td>
<td>0.9999995</td>
</tr>
<tr>
<td>D-C</td>
<td>9.434167</td>
<td>2.223522</td>
<td>16.644812</td>
<td>0.0081484</td>
</tr>
<tr>
<td>B-C</td>
<td>16.940833</td>
<td>9.730188</td>
<td>24.151478</td>
<td>0.0000172</td>
</tr>
<tr>
<td>D-A</td>
<td>9.406667</td>
<td>2.957270</td>
<td>15.856064</td>
<td>0.0032628</td>
</tr>
<tr>
<td>B-A</td>
<td>16.913333</td>
<td>10.463936</td>
<td>23.362730</td>
<td>0.0000040</td>
</tr>
<tr>
<td>B-D</td>
<td>7.506667</td>
<td>1.057270</td>
<td>13.956064</td>
<td>0.0193814</td>
</tr>
</tbody>
</table>

We see that all of the varieties except for A and C are significantly different from one another as all Tukey P-values are < 0.05 (or equivalently, all intervals do not contain zero, except for the A-C interval.
Dioecious trees, p. 21
Read the problem description, but answer the following questions rather than those listed in the text.

1. Using R, create a meaningful plot that is useful to compare the number of flowers produced by male and female trees. Display your code and plot (suitable for publication) as the solution to this problem.

Either of the following plots are okay. I prefer the simplicity of the boxplot because there is so much data.

> with(trees, stripchart(Flowers ~ Sex, vertical = TRUE,
+   pch = 16, method = 'jitter',
+   xlab = 'Sex', ylab = 'Number of Flowers',
+   xlim = c(.5,2.5)))
2. Is there significant evidence that male and female trees produce different numbers of flowers on average? Use R to conduct both a two-sample t-test and a Wilcoxon-Mann-Whitney test. Do these two analyses agree? Which do you think is more appropriate and why?

```r
> with(trees, boxplot(Flowers ~ Sex,
+ xlab = 'Sex', ylab = 'Number of Flowers', boxwex = .3))

> t.test(Flowers ~ Sex, data = trees)

Welch Two Sample t-test

data:  Flowers by Sex
t = 1.2386, df = 43.64, p-value = 0.2221
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -75.08973  314.42306
sample estimates:
mean in group Female  mean in group Male
  431.3667            311.7000

From the two-sample t-test there is no significant evidence that the mean number of flowers differs among the sexes ($t_{44} = -1.24, P = 0.2221$).

> wilcox.test(Flowers ~ Sex, data = trees)

Wilcoxon rank sum test with continuity correction

data:  Flowers by Sex
W = 298, p-value = 0.9763
alternative hypothesis: true location shift is not equal to 0

From the Wilcoxon-Mann-Whitney test we also obtain similar results. There is no significant evidence that the mean number of flowers differs among the sexes ($W = 298, P = 0.9763$).
The conclusions of both tests are the same, so we don’t need to worry too much about which test is more appropriate. Because of the reasonably large sample sizes, we can trust the t-test unless the data for each group is exceptionally non-normal. The normal probability plots below indicate that the female distribution is somewhat skewed. A histogram for each sex would also reveal this.

```r
> with(subset(trees, subset = Sex == 'Male'), qqnorm(Flowers, main = 'Male'))
> with(subset(trees, subset = Sex == 'Female'), qqnorm(Flowers, main = 'Female'))
```