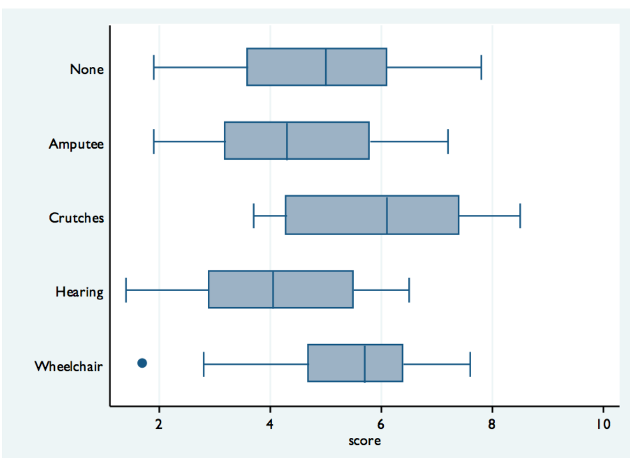
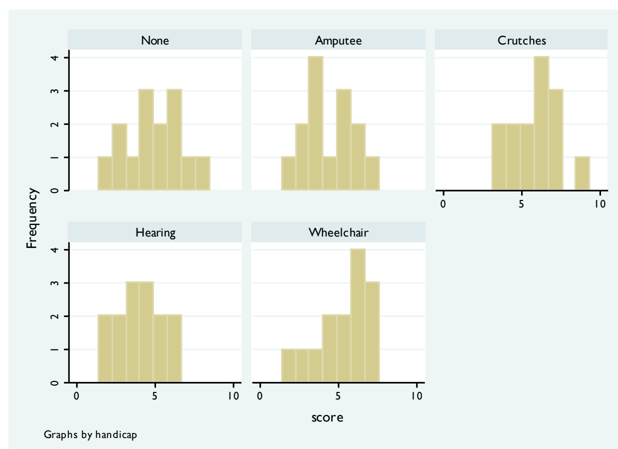
Unit 1 Assignment: Randomization Methods for Comparing 2+ Groups

In Activity #5, we'll learn about a 1990 paper that studied how physical disabilities affect perceptions of employment qualifications. For now, let's just take a look at the data that was obtained in the study. The following table displays job interview ratings for individuals with different disabilities:

	No Handicap	Amputee	Crutches	Hearing	Wheelchair
	1.90	1.90	3.70	1.40	1.70
	2.50	2.50	4.00	2.10	2.80
	3.00	2.60	4.30	2.40	3.50
	3.60	3.20	4.30	2.90	4.70
	4.10	3.60	5.10	3.40	4.80
	4.20	3.80	5.80	3.70	5.00
	4.90	4.00	6.00	3.90	5.30
	5.10	4.60	6.20	4.20	6.10
	5.40	5.30	6.30	4.30	6.10
	5.90	5.50	6.40	4.70	6.20
	6.10	5.80	7.40	5.50	6.40
	6.70	5.90	7.40	5.80	7.20
	7.40	6.10	7.50	5.90	7.40
	7.80	7.20	8.50	6.50	7.60
Mean	4.9000	4.4286	5.9124	4.0500	5.3429
StDev	1.7936	1.5857	1.4818	1.5325	1.7483





We'll conduct an ANOVA (and some post-hoc tests) on this data in Activity #5. For now, let's try a randomization approach to determine if the job interview ratings significantly differ among the 5 groups.

1) Write out the null and alternative hypotheses.

2) Suppose we wanted to conduct an ANOVA to test our hypothesis. What assumptions should we investigate prior to running an ANOVA?

3) Using Stata, I ran an ANOVA and obtained the following output:

```
Bartlett's test for equal variances:  chi2(4) =   0.7016  Prob>chi2 = 0.951

                        Analysis of Variance
    Source              SS         df       MS            F     Prob > F
------------------------------------------------------------------------
Between groups      30.5214294      4   7.63035734      2.86     0.0301
 Within groups      173.321429     65   2.66648353
------------------------------------------------------------------------
    Total           203.842859     69   2.95424433
```

What conclusion can you make from the "Bartlett's test for equal variances" line?

Explain what the SS and MS values represent, verify the degrees of freedom, and write a conclusion based on the ANOVA. What does the p-value represent? Calculate an effect size and interpret.

4) Suppose we were concerned about the normality assumption necessary to conduct an ANOVA. When we have this concern, we can conduct a *nonparametric* test of our hypotheses. Nonparametric methods do not rely on assumptions that data are drawn from a given probability distribution (like a normal distribution).

Recall (from MATH 300) that we can replace our observed data with ranks -- the lowest value is ranked 1, the next lowest value is ranked 2, and so on. If we conduct an ANOVA on these ranks, then we're actually conducting a nonparametric test called the *Kruskal-Wallis One-Way ANOVA by Ranks* (Kruskal-Wallis, for short).

The Kruskal-Wallis test (which is fairly easy to calculate, although we won't go into any details here) tests the equality of population medians among groups. While it does not assume normality, it does assume each group has identically-shaped distributions.

Using Stata, I conducted a Kruskal-Wallis test and obtained the following output. What conclusions can you draw?

```
  +---------------------------+
  | handicap | Obs | Rank Sum |
  |----------+-----+----------|
  |     None |  14 |   491.50 |
  |  Amputee |  14 |   406.00 |
  | Crutches |  14 |   660.50 |
  |  Hearing |  14 |   353.00 |
  | Wheelcha |  14 |   574.00 |
  +---------------------------+

chi-squared =    10.642 with 4 d.f.
probability =     0.0309
```

5) Let's conduct one more analysis on this data. We're still interested in determining if job interview scores differ significantly among disability groups. This time, instead of conducting an ANOVA or ANOVA based on ranks, let's use randomization methods.

Recall that randomization methods require us to:

(1) Pool all the data into one big pool/group
(2) Randomly assign observations to groups (assuming the groups have no impact on the observations)
(3) Calculate a test statistic
(4) Repeat steps 1-3 many, many times and record the test statistic each time
(5) Determine the likelihood of the observed data based on all these test statistics

Remember that when we were comparing two groups, this process was easy. As a simple example, suppose we wanted to compare two groups:

Group A	Group B
9	4
10	11
17	12
Sum = 36	Sum = 27
Average = 12	Average = 9

From this sample, it appears as though Group A scored higher than Group B. If we used randomization methods to compare these groups, we would:

(1) Pool all the data into one big pool/group:
(2) Randomly assign observations to groups ("X" = score was assigned to Group A)
(3) Calculate a test statistic; (4) repeat

Scores:	4	9	10	11	12	17	SUM	SUM>=36
Trial 1	X	X	X				23	
Trial 2	X	X		X			24	
Trial 3	X	X			X		25	
Trial 4	X	X				X	30	
Trial 5	X		X	X			25	
Trial 6	X		X		X		26	
Trial 7	X		X			X	31	
Trial 8	X			X	X		27	
Trial 9	X			X		X	32	
Trial 10	X				X	X	33	
Trial 11		X	X	X			30	
Trial 12		X	X		X		31	
Trial 13		X	X			X	36	X
Trial 14		X		X	X		32	
Trial 15		X		X		X	37	X
Trial 16		X			X	X	38	X
Trial 17			X	X	X		33	
Trial 18			X	X		X	38	X
Trial 19			X		X	X	39	X
Trial 20				X	X	X	40	X

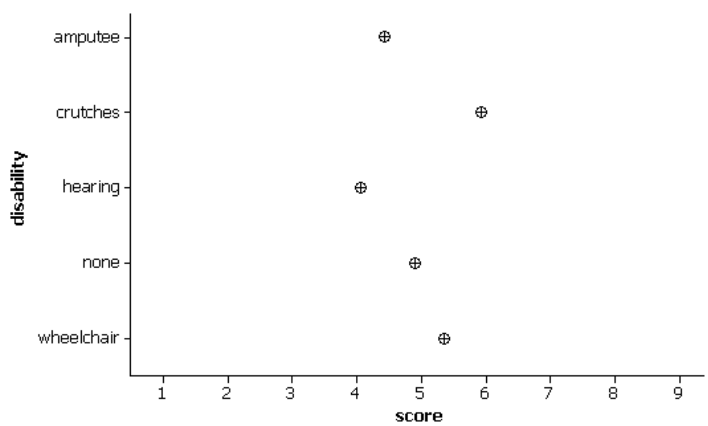
From this table, we would conclude the following:

In our actual data, we observed that Group A summed to 36. If the Groups had no impact on the scores, the likelihood of observing such a high sum would be 6/20 = 0.30. Since this likelihood is reasonably big, we cannot reject the hypothesis that our results happened by chance.

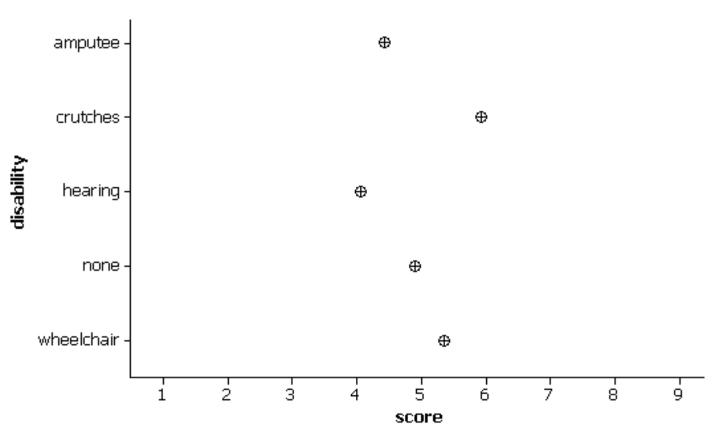
This example is easy, since we only had 2 groups to compare. When we compare 2 groups, we can simply find the difference between the group means. How can we compare 5 groups simultaneously? What test statistic will give us a measure of the overall dispersion of the 5 group means? Before we attempt to answer these questions, let's examine the data again:

	No Handicap	Amputee	Crutches	Hearing	Wheelchair
	1.90	1.90	3.70	1.40	1.70
	2.50	2.50	4.00	2.10	2.80
	3.00	2.60	4.30	2.40	3.50
	3.60	3.20	4.30	2.90	4.70
	4.10	3.60	5.10	3.40	4.80
	4.20	3.80	5.80	3.70	5.00
	4.90	4.00	6.00	3.90	5.30
	5.10	4.60	6.20	4.20	6.10
	5.40	5.30	6.30	4.30	6.10
	5.90	5.50	6.40	4.70	6.20
	6.10	5.80	7.40	5.50	6.40
	6.70	5.90	7.40	5.80	7.20
	7.40	6.10	7.50	5.90	7.40
	7.80	7.20	8.50	6.50	7.60
Mean	4.9000	4.4286	5.9124	4.0500	5.3429
StDev	1.7936	1.5857	1.4818	1.5325	1.7483

The following axes show the group means. Sketch boxplots around these means that would convince you that the 5 distributions differed significantly.



Now sketch boxplots that would convince you the 5 distributions were coming from the same overall population:



6) When comparing more than 2 groups, we need a measure of the differences (variability) *between* the group means that also considers the variability *within* each group.

If the variability *between* the group means is significantly larger than the variability *within* the groups, then the boxplots will not overlap much and we will have evidence that the groups differ significantly. If, on the other hand, the within-group variability is as large as, or larger than, the between-group variability, the boxplots will overlap and we will **not** have evidence to support the conclusion that the groups differ significantly.

So, as we discovered in Activity #4, the F-statistic provides a ratio of the between-group variability to the within-group variability:

$$F = \frac{\text{between-groups variability}}{\text{within-group variability}} = \frac{SS_A / df_A}{SS_E / df_E} = \frac{\sum_{a=1}^{a}\left(\bar{X}_a - M\right)^2 / (a-1)}{\sum_{a=1}^{a}\left(n_a - 1\right)s_a^2 / (N-a)}$$

We already know that for our observed data, F = 2.86. This means that the between-groups variability is 2.86 times larger than the within-groups variability. Does this value provide convincing evidence against our null hypothesis that the group means are equal ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$)? Normally, we'd look in our F-table and make a decision. In this example, we're going to use randomization methods.

7) What type of value for our F-statistic (e.g., large, small, less than zero, greater than 1) would be considered evidence against the null hypothesis?

8) If the null hypothesis is true, then the numerator and denominator of our F-statistic are both measuring "variability in the data" and we would expect the F-statistic to be around 1.0. If the group means are far apart (in comparison to the variation within each group), then the F-statistic would increase. Thus, large values of our F-statistic provide evidence against the null hypothesis. Our p-value, then, would be the probability of obtaining an F-statistic (assuming the group means are equal) at least as large as the F-statistic we obtained from our data.

To estimate this p-value, we will repeatedly assign observations to the 5 groups at random. Then, for each repetition, we will calculate the F-statistic. Explained another way...

(1) Take all 70 observations from our data and randomly assign 14 of them to each of 5 groups.
(2) Calculate the F-statistic
(3) Repeat steps 1-2 many times and record the F-statistic each time
(4) Determine the likelihood of our actual F-statistic of 2.86.

9) For extra credit, go ahead and run through every possible randomization of this data by hand (every possible way of assigning these 70 observations into 5 groups with 14 observations each). If you were to try this, you would find there are more than **2,378,829,280,000,000,000,000,000,000,000,000,000,000,000,000** different randomizations of this data.

Even for a computer, this would take a long time. Rather than trying to account for every possible randomization, let's have a computer run 10,000 randomizations and see what we get.

In Stata, the code to run 10,000 randomizations of an ANOVA is:

```
permute score F=e(F), reps(10000) : anova score handicap
```

This code tells Stata to compute the F statistic for 10,000 replications of an ANOVA where "score" is our outcome and "handicap" is our grouping variable.

The output obtained is:

```
Monte Carlo permutation results                    Number of obs   =         70

      command:  anova score handicap
            F:  e(F)
  permute var:  score

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
           F |   2.861581     322   10000  0.0322  0.0018  .0288269    .035849
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```

Let's try to interpret this output.

- On the top-left, we can see Stata provided Monte Carlo permutation results. *Monte Carlo* methods use simulations based on random sampling (what we call randomization methods). Recall that *permutations* refer to rearrangements of objects in different orders.
- On the top-right, we see Stata used all 70 of our observations
- In the table, "T" represents "test statistic." So our observed test statistic -- T(obs) -- is 2.861581. This is our observed F-statistic of 2.86 that we calculated earlier.
- The n of 10,000 represents the number of F-statistics that were calculated from our randomizations
- As the note on the bottom attempts to explain, "c" represents the number of randomizations that were greater or equal to our observed test statistic. In other words, 322 of our 10,000 randomizations resulted in F-statistics greater or equal to 2.86.
- Our p-value is p = c/n = 322 / 10000 = 0.0322. This is similar to the p-values we calculated in our original ANOVA and in the Kruskal-Wallis analysis.
- SE(p) is a standard error of our p-value. How did Stata calculate this value? We'll see a bit later.
- The last two columns represent the 95% confidence interval for our p-value. Since the confidence interval is (0.029, 0.036), we can reject our null hypothesis at a 0.05 level of significance.

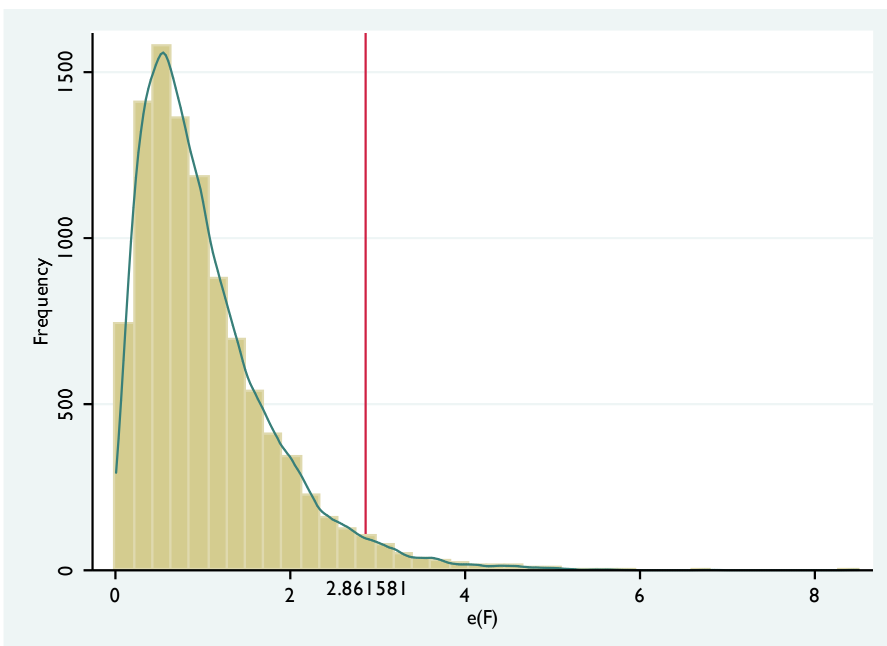
10) Based on the Stata output, briefly write any conclusions you can make from this study.

11) I had Stata record the values of the F-statistic for all 10,000 randomizations. A summary and histogram of these F-statistics are displayed below:

```
                              e(F)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0752249       .0074058
 5%     .1763627       .0082038
10%      .258638       .0107121       Obs                10000
25%     .4813985       .0112823       Sum of Wgt.        10000

50%      .840827                      Mean            1.033114
                        Largest       Std. Dev.       .7766495
75%     1.374217       5.613302
90%     2.045394       5.800845       Variance        .6031845
95%     2.550499       6.793564       Skewness        1.661402
99%     3.696429       8.501678       Kurtosis        7.401396
```

Let's try to interpret this output.

- The average value of the 10,000 F-statistics is 1.0331. The median is 0.8408. The standard deviation among the F-statistics is 0.7766.
- 10% of the F-statistics were less than 0.2586; 75% of the F-statistics were less than 1.3742.
- 5% of the F-statistics were greater than 2.5505; 1% of the F-statistics were greater than 3.6964
- The smallest F-statistics calculated were 0.0074 and 0.0082; the largest were 8.5017 and 6.7936



The red line in the histogram references our observed F-statistic of 2.861581. Remember that this graph represents F-statistics we could get if the groups had no impact on the values (in other words, if disability type had no impact on job interview scores).

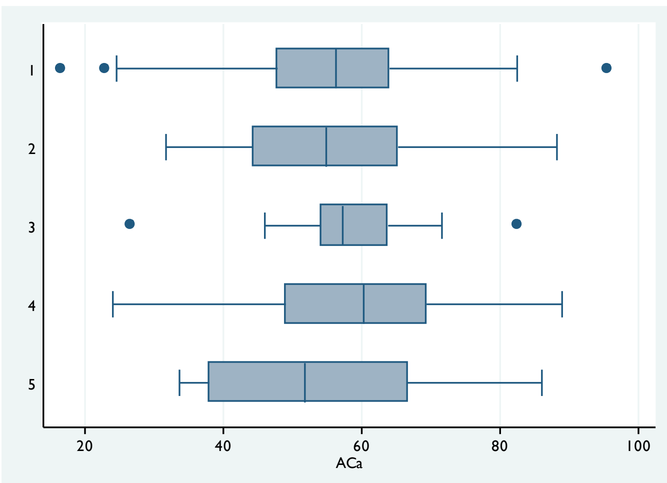
Out of 10,000 randomizations, only 322 provided F-statistics greater than our observed F-statistic of 2.86. What can you conclude from this?

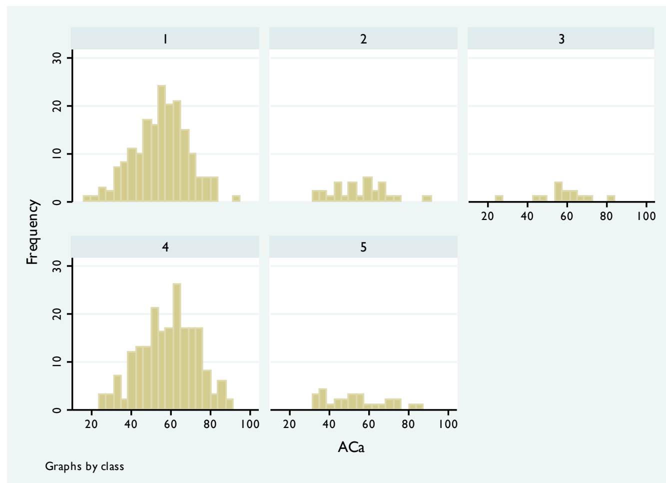
12) Let's look at another example. In 2009, SAU students completed the NSSE (National Survey of Student Engagement). Based on responses to several items, the NSSE assigns each student a "Academic Challenge" score. This score represents each student's perception of how academically challenged he or she is.

The following table summarizes the results for freshmen, sophomores, juniors, seniors, and other students:

	Freshmen (group 1)	**Sophomores (group 2)**	**Juniors (group 3)**	**Seniors (group 4)**	**Other (group 5)**
n	182	29	14	203	27
Mean	55.496	54.582	57.890	59.126	53.787
Std. Dev	13.474	13.194	13.130	13.893	15.234

13) Based on the following boxplots and histograms, do you think we will find significant differences among the group means?





14) Are you concerned about any of the assumptions necessary to run an ANOVA? Explain.

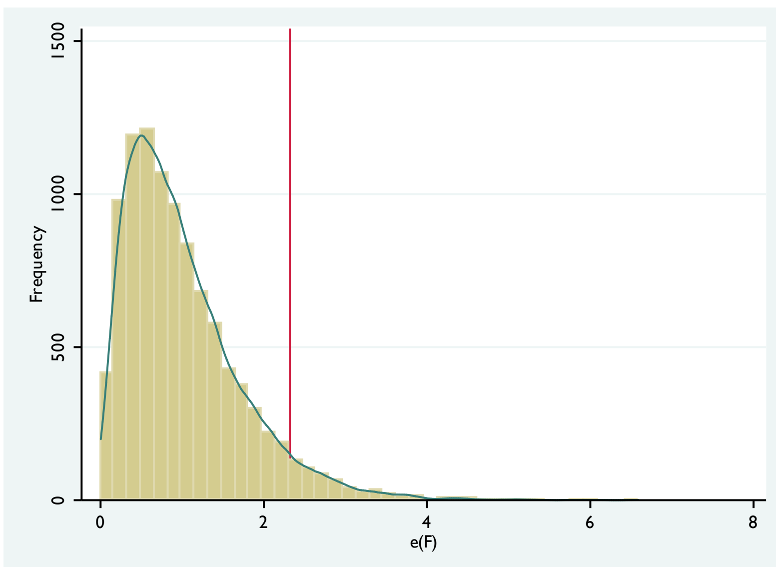
15) What conclusions can you make from the following ANOVA output?

```
Bartlett's test for equal variances:  chi2(4) =   0.9106  Prob>chi2 = 0.923

                        Analysis of Variance
    Source              SS         df       MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1752.35621      4   438.089052      2.32     0.0562
 Within groups      84996.4808    450   188.881068
------------------------------------------------------------------------
    Total            86748.837    454   191.076734
```

16) What conclusions can you make from the following output from 10,000 randomizations?

Stata code: permute aca F=e(F), reps(10000) nodots saving(/Users/Brad/Desktop/nssereps): anova aca class

```
Monte Carlo permutation results                   Number of obs   =        480

      command:  anova aca class
            F:  e(F)
  permute var:  aca

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
           F |   2.319391     555   10000  0.0555  0.0023  .0510926   .0601685
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```



Assigned: Absolute Mean Differences

A psychologist was interested in whether different TV shows lead to a more positive outlook on life. 25 subjects were split into 4 groups and then taken to a room to view a program. One group watched *The Muppet Show*, another watched *Futurama*, and another watched the *BBC News*. The fourth group did not watch any tv program. After the program a blood sample was taken and serotonin levels measured. A higher serotonin level represents more happiness. The following data were obtained:

The Muppet Show	***Futurama***	***BBC News***	**No Program**
11	4	4	7
7	8	3	7
8	6	2	5
14	11	2	4
11	9	3	3
10	8	6	4
5			4
			4
n = 7 Mean = 9.42857 Std. Dev = 2.99205	n = 6 Mean = 7.66667 Std. Dev = 2.42212	n = 6 Mean = 3.33333 Std. Dev = 1.50555	n = 8 Mean = 4.75000 Std. Dev = 1.48805

N = 27
Grand Mean = 6.296296
Total Standard Deviation = 3.172172

17) We know we need to assume homogeneous variances in order to conduct an ANOVA. Do the data in this study satisfy this assumption? Support your answer with an appropriate test.

18) Output from an ANOVA is displayed below. It might be a good idea to verify these calculations (especially the easy ones, such as degrees-of-freedom and SST) to ensure you can use the formulas. Briefly write out any conclusions you can make from this ANOVA.

```
                        Analysis of Variance
    Source              SS         df       MS            F     Prob > F
------------------------------------------------------------------------
Between groups      151.748677      3   50.5828924     10.59      0.0001
 Within groups      109.880952     23   4.77743271
------------------------------------------------------------------------
    Total            261.62963     26   10.0626781

Bartlett's test for equal variances:  chi2(3) =   4.0184  Prob>chi2 = 0.259
```

19) We're going to try a randomization-based approach with this data, but we're going to do one thing differently. In the previous examples, we took 10,000 randomizations and calculated an F statistic for each. This time, let's calculate a different test statistic. Let's calculate the sum of all the absolute mean differences. Go ahead and calculate this value for our sample:

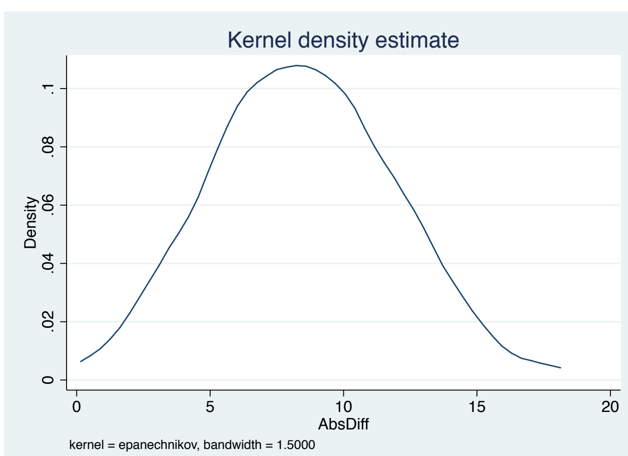
$$\left|\bar{X}_{\text{Muppet}} - \bar{X}_{\text{Futurama}}\right| + \left|\bar{X}_{\text{Muppet}} - \bar{X}_{\text{BBC}}\right| + \left|\bar{X}_{\text{Muppet}} - \bar{X}_{\text{None}}\right| + \left|\bar{X}_{\text{Futurama}} - \bar{X}_{\text{BBC}}\right| + \left|\bar{X}_{\text{Futurama}} - \bar{X}_{\text{None}}\right| + \left|\bar{X}_{\text{BBC}} - \bar{X}_{\text{None}}\right| = \_\_\_\_\_\_\_\_\_\_$$

20) Below, I've pasted the Stata syntax used to run this randomization test. I've included some comments in case you want to try to follow along. Don't worry too much about the syntax; I just pasted it here to remind myself how I did this.

```
tabulate Show, generate(dum)              -- This creates dummy variables for the 4 tv shows (dum1, dum2, dum3, dum4)

regress Happiness dum2 dum3 dum4          -- This runs a regression I need to generate coefficients

generate show1=_b[_cons]                  -- This generates a variable show1 that is equal to the mean of The Muppet Show
generate show2=_b[_cons]+_b[dum2]         -- This generates a variable show2 that is equal to the mean of Futurama
generate show3=_b[_cons]+_b[dum3]         -- This generates a variable show3 that is equal to the mean of BBC News
generate show4=_b[_cons]+_b[dum4]         -- This generates a variable show4 that is equal to the mean of No Program

generate absdiff=abs(show1- show2)+abs(show1- show3)+abs(show1- show4)+abs(show2-
         show3)+abs(show2- show4)+abs(show3- show4)
                                          -- This calculates the sum of the absolute mean differences

permute Show absdiff=abs((_b[_cons]- (_b[_cons]+_b[dum2]))+abs(_b[_cons]- (_b[_cons]
+_b[dum3]))+abs(_b[_cons]- (_b[_cons]+_b[dum4]))+abs((_b[_cons]+_b[dum2])- (_b[_cons]
+_b[dum3]))+abs((_b[_cons]+_b[dum2])- (_b[_cons]+_b[dum4]))+abs((_b[_cons]+_b[dum3])-
(_b[_cons]+_b[dum4]))),reps(5000) nodots saving(/Users/Brad/Desktop/anovaabsdiff.dta):
regress Happiness dum2 dum3 dum4
                                          -- This calculates the sum of the absolute mean differences
```

```
Output
Monte Carlo permutation results                    Number of obs   =         27

------------------------------------------------------------------------------
T            |     T(obs)       c       n   p=c/n   SE(p) [95% Conf. Interval]
-------------+----------------------------------------------------------------
     absdiff |   21.20238    0001    5000  0.0002  0.0000    0.00     .0009375
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{|T| >= |T(obs)|}
```

A graph of the absolute mean differences calculated for each of the 5000 randomizations is displayed below:



Based on this output and the graph, briefly state any conclusions you can make from this study. What is the approximate p-value?