

Activity 3: Comparing multiple groups (Analysis of Variance, randomization-based methods, statistical modeling)

Scenario: If you've ever read a book with ambiguous language (like some statistics books), you know they can be difficult to comprehend. Read the following passage and explain what it describes.

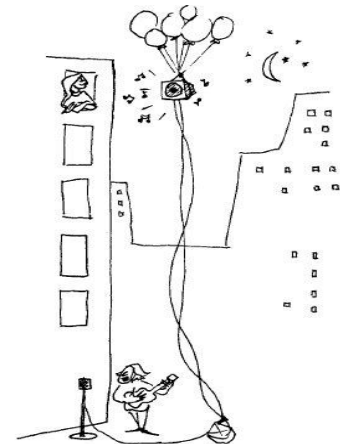
*If the balloons popped, the sound wouldn't be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong. (p. 719 of study)*

Would the picture to the right help you comprehend that passage?

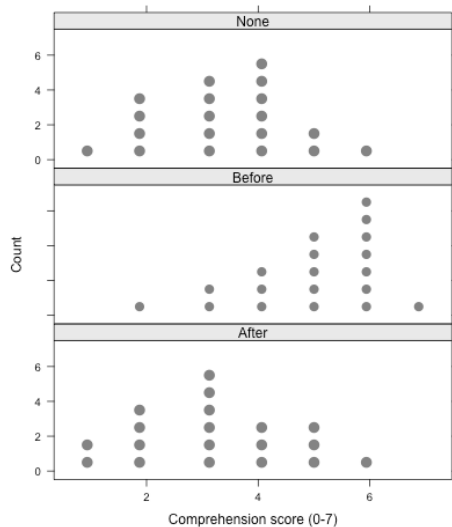
A 1972 study on memory encoding examined college students' comprehension of this ambiguous passage. Before reading the passage, 57 college students were randomly assigned to one of three groups:

- 19 students were shown the picture before hearing the passage
- 19 students were shown the picture after hearing the passage
- 19 students were not shown any picture before or after

The researchers were interested in determining if comprehension of an ambiguous prose passage is affected by viewing a picture.



Below is a visualization and summary statistics for the data in this study:



<u>Group</u>	<u>Count</u>	<u>Mean</u>	<u>Std. Dev</u>	<u>Median</u>
None	19	3.368	1.25656	3
Before	19	4.947	1.31122	5
After	19	3.211	1.39758	3
<b>Total</b>	<b>57</b>	<b>M = 3.842</b>	<b>s = 1.52115</b>	<b>4</b>

Data: <http://www.math.hope.edu/isi/data/chap9/Comprehension.txt>

Applet: <http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2>

Source of example: Introduction to Statistical Investigations – <http://math.hope.edu/isi/>

Actual Study: [http://memlab.yale.edu/sites/default/files/files/1972\\_Bransford\\_Johnson\\_JVLVB.pdf](http://memlab.yale.edu/sites/default/files/files/1972_Bransford_Johnson_JVLVB.pdf)

1. What do M=3.84 and s=1.52 represent? Does the data provide evidence that the picture improves comprehension?

2. For this sample of 57 students, it appears as though the **before** group had the highest comprehension. But we could have observed these results even if showing the picture did **not** improve comprehension. Perhaps the students with the highest comprehension abilities just happened to be randomly assigned to the **before** group.

Our key question is this: If the treatments (before, after, none) have no affect on comprehension, how unlikely is it to observe sample averages as different as what we observed in this study?

Let's begin by stating our null and alternative hypotheses. Under a null hypothesis, what would be true about the effect of the treatments on comprehension scores?

$H_0$ :

$H_1$ :

3. Let's rewrite these hypotheses as statistical models. A statistical model describes how we believe data in the study could have been generated. What factors influenced the comprehension score for individual  $i$  in treatment group  $j$ ?

Explain what each term in the following models represents:

Null model:  $Y_{ij} = \mu + e_i$

Alternate model:  $Y_{ij} = \mu + \alpha_j + e_i$

$$= \mu + (\mu_j - \mu) + (Y_{ij} - \mu_j)$$

$Y_{ij}$  represents:

$\mu$  represents:

$\alpha_j$  represents:

$e_i$  represents:

Demonstrate why our alternate model is a tautology:

What probability distribution do we assume generates the e terms in the models?

4. One way to test our hypotheses would be to compare the group means. In a previous statistics class, you most likely learned how to conduct an independent samples t-test to compare two group means.

Suppose we wanted to conduct multiple t-tests to compare all possible pairings of our 3 group means. To do this, we'd have to run 3 t-tests comparing: **none vs before**, **none vs. after**, and **before vs. after**. We'll learn, a little later, why we would want to avoid running multiple t-tests on a single set of data.

For now, remember that for an independent samples t-test, we calculate the following test statistic and compare it to a t-distribution with  $n_1 + n_2 - 2$  degrees of freedom:

$$t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{s_{\text{pooled}}^2}}$$

The numerator shows we subtract group means to compare them. Bigger values of the numerator indicate situations where the group means differ by a larger amount.

But how can we extend this to our scenario where we want to compare 3 group means?

We could add up all the pairwise differences: **Potential test statistic** =  $(\bar{X}_1 - \bar{X}_2) + (\bar{X}_1 - \bar{X}_3) + (\bar{X}_2 - \bar{X}_3)$

or we could find the average pairwise difference: **Potential test statistic** =  $\frac{(\bar{X}_1 - \bar{X}_2) + (\bar{X}_1 - \bar{X}_3) + (\bar{X}_2 - \bar{X}_3)}{3}$

Ideally, we'd choose a test statistic that gets bigger as the means differ by a larger amount (and becomes zero when the means are exactly the same). What's wrong with both of these potential test statistics?

To fix this problem, we could consider using the **MAD (mean absolute difference)**:  $\frac{|\bar{X}_1 - \bar{X}_2| + |\bar{X}_1 - \bar{X}_3| + |\bar{X}_2 - \bar{X}_3|}{3}$

Calculate and interpret the MAD for the data in our study.

Group	Count	Mean	Std. Dev
None	19	3.368	1.25656
Before	19	4.947	1.31122
After	19	3.211	1.39758
Total	57	M = 3.842	s = 1.52115

Under our null hypothesis, what would we expect for the value of the MAD?

5. Under our null hypothesis, the treatments had no effect on comprehension. Let's assume this is true.

As you can see in the table below, the first subject in this study was randomly assigned to the **none** group and ended with a comprehension score of **4**.

Suppose we go back in time and, once again, randomly assign the 57 subjects to the 3 groups. Because the assignment is random, we'd expect some of the group assignments to change. Suppose, for example, the first subject was now randomly assigned to the **before** group. How would this change that subject's comprehension?

Subject	Actual Data	Randomization #1	Randomization #2	Randomization #3
1	None 4	Before ____	After 4	None 4
2	Before 7	Before ____	None 7	None 7
3	After 6	After ____	Before 6	None 6
.	... ..	... ..	... ..	... ..
.	... ..	... ..	... ..	... ..
.	... ..	... ..	... ..	... ..
57	After 5	Before ____	None 5	After 5
Means:	None: 3.368 Before: 4.947 After: 3.211 <b>MAD: 1.158</b>	None: 3.632 Before: 4.158 After: 3.737 <b>MAD: 0.351</b>	None: 4.632 Before: 3.158 After: 3.737 <b>MAD: 0.982</b>	None: 5.211 Before: 2.948 After: 3.368 <b>MAD: 1.509</b>

The table shows 4 different values of MAD we could get by randomly assigning the 57 subjects to 3 groups. If the groups have no effect on comprehension, each of these MAD values are equally likely to occur.

Recall our key question: If the treatments (before, after, none) have no affect on comprehension, how unlikely is it to observe sample averages as different as what we observed in this study?

To answer this question, we simply need to estimate how unlikely it was to get a MAD value of 1.158 or greater.

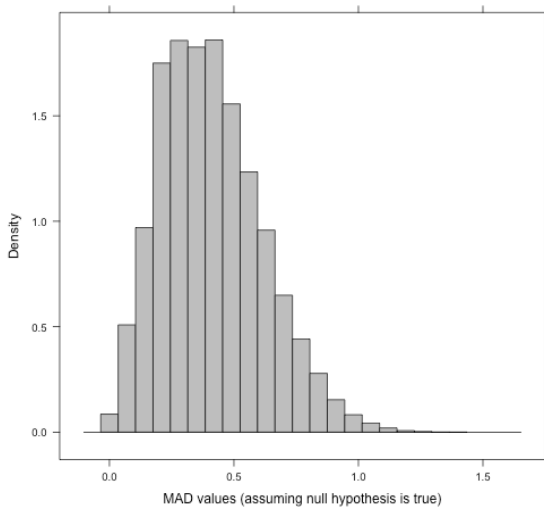
One way to estimate this likelihood would be to list out all possible randomizations from this study and count the proportion that are greater than 1.158. Unfortunately, that's difficult to do since there are **lots** of ways to randomly assign 57 subjects to 3 groups:

$$\frac{\binom{57}{19} \binom{38}{19} \binom{19}{19}}{3!} = 3,752,394,405,341,098,927,410,000 = 3.75 \text{ septillion randomizations}$$

We could, instead, have a computer simulate lots and lots of these randomizations and calculate MAD for each one. While we won't get all 3.75 septillion possible MAD values, we'll get a nice representative sample.

The top of the next page shows the distribution of MAD values we get from 100,000 randomizations of this experiment. Remember that these MAD values are only possible if we assume the treatments (none, before, and after) have **no effect on comprehension**.

mean = 0.394476140350877 ; std. dev = 0.2042262867596



This graph shows many of the possible differences in means (MAD values) we could get if the treatments did not affect comprehension.

The value we got from our actual study was  $MAD = 1.158$ .

The computer tells me that of the 100,000 randomizations conducted, only 69 yielded MAD values of 1.158 or greater.

Therefore, our p-value is 0.00069.

Interpret this p-value and draw a conclusion from this study.

A p-value represents: \_\_\_\_\_

Conclusion(s): \_\_\_\_\_

Which group, if any, yields the highest average comprehension score: \_\_\_\_\_

6. Look at the sampling distribution for the MAD statistic. Notice that this distribution is not normal and is not centered at zero. Should it be?

7. If you want to replicate this example (and you do not have R installed), you can use the following websites:

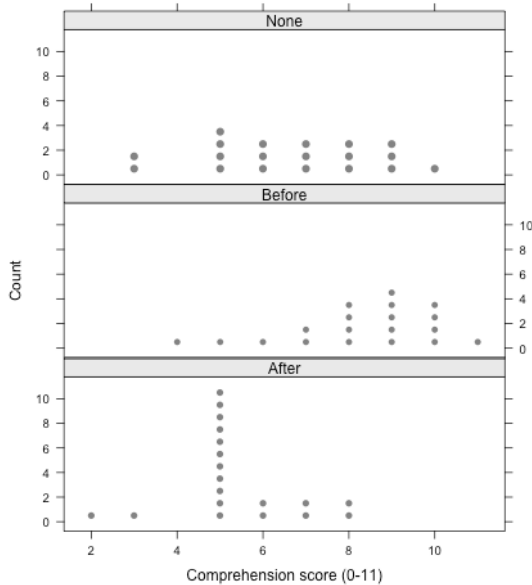
Data are available at: <http://www.math.hope.edu/isi/data/chap9/Comprehension.txt>

Randomization applet available at: <http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2>

You'll have an opportunity to practice using this MAD-randomization-based approach in an assignment.

For now, let's repeat this randomization-based method using a different test statistic.

Scenario: The students in this comprehension study were retested a few hours later to see how well they could recall the ambiguous passage. The scores from this retest are displayed below:



Group	Count	Mean	Std. Dev
None	19	6.632	2.00584
Before	19	8.263	1.82093
After	19	5.369	1.46099
Total	57	M = 6.754	s = 2.11526

**MAD = 1.930**

Data: <http://www.math.hope.edu/isi/data/chap9/Recall.txt>

Applet: <http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2>

Source of example: Introduction to Statistical Investigations – <http://math.hope.edu/isi/>

Actual Study: [http://memlab.yale.edu/sites/default/files/files/1972\\_Bransford\\_JVLVB.pdf](http://memlab.yale.edu/sites/default/files/files/1972_Bransford_JVLVB.pdf)

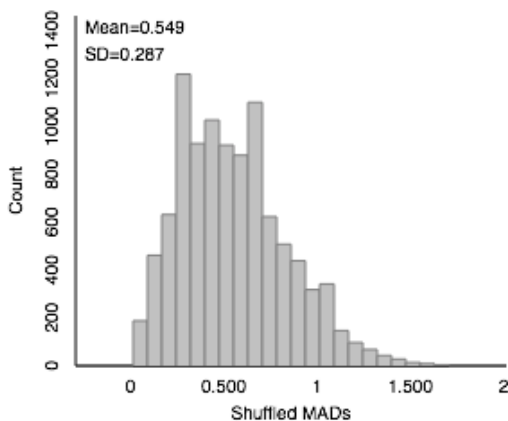
8. Let's begin with the same competing hypotheses and models:

Null hypothesis: All three of the long-run average recall scores for students in each condition are the same.

Alternative hypothesis: At least one of the long-run average recall scores is different from the others

Under the null hypothesis, we can pretend to go back in time and run the study repeatedly (each time with a different set of students randomly assigned to each group). For each of those randomizations, we can calculate the MAD. Below, I've pasted this distribution of MAD values along with the p-value.

Total Shuffles = 10000



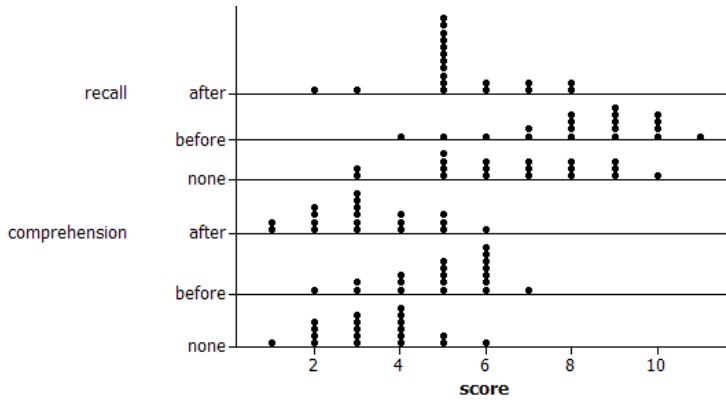
Notice that this distribution of MAD values is, once again, skewed to the right.

The p-value is estimated to be less than 0.0001. What does that mean?

The MAD value for this (recall) dataset is 1.930. The MAD value for our previous (comprehension) dataset was 1.158. Can we directly compare these values and conclude this recall dataset represents a scenario with greater differences among means?

Count Samples     
 Count = 0/10000 (0.0000)

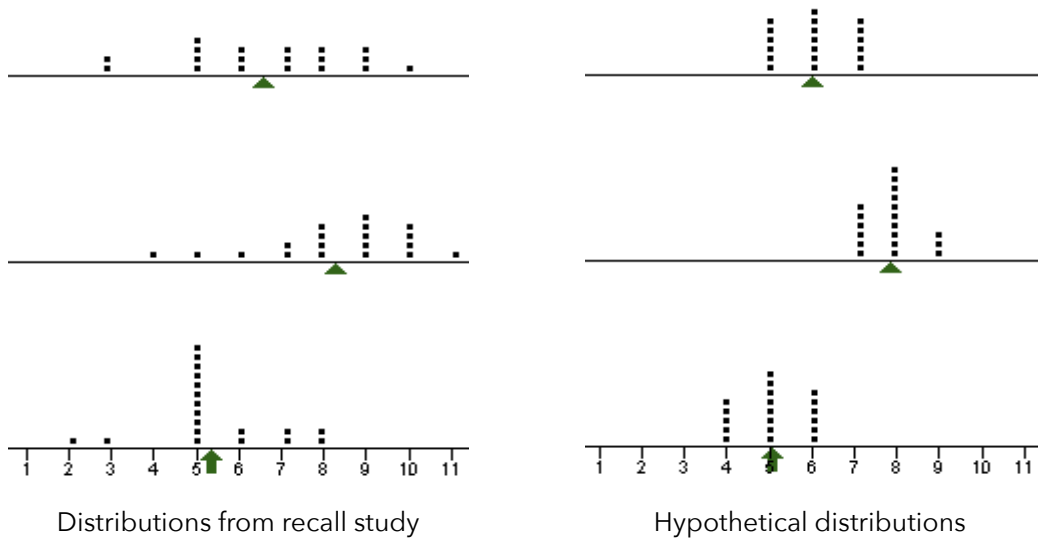
9. Take a look at the results for both the comprehension and recall datasets.



Which dataset (recall or comprehension) provides the stronger evidence against the null hypothesis? Why?

10. Maybe the MAD isn't the best measure of the differences among group means. To demonstrate why, take a look at the following distributions. The distributions on the left are the actual distributions from the recall study (with the triangles pointing at the group means).

The distributions on the right were constructed to have similar group means but smaller variances.



Which set of distributions provides stronger evidence that at least one group mean differs?

11. Instead of using the MAD statistic as our measure of group mean differences, perhaps we should create a statistic that standardizes those differences by comparing the variability between group means to the variability within each group. That's what we're going to do.

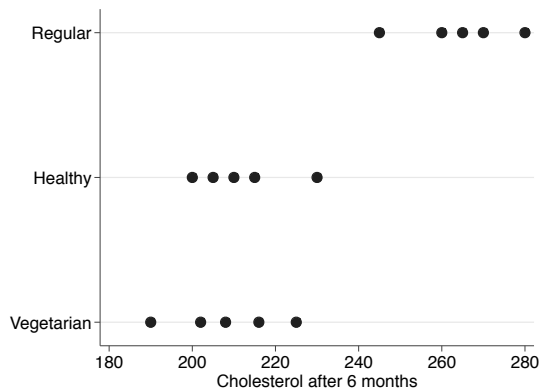
Assuming we figure out how to calculate the variance between the group means and the variance within the groups, how will we compare those two variances?

Our goal is to calculate an F-statistic that represents the ratio of the variance between the groups to the variance within groups. Because we'll be analyzing variances, we call this process ANOVA (analysis of variance). Despite its name, the goal of ANOVA is not to compare variances; the goal is to compare means.

To see how we could do this, let's take a look at a simple (fictitious) dataset:

Scenario: Suppose we're interested in the effects of diet on cholesterol. To study this effect, we randomly select 15 individuals and randomly assign them to one of three diets: vegetarian, healthy, and regular. After six months of this diet, we then measure the cholesterol level of each individual.

	Measurements					Mean	Std. Dev.	Sample
<b>Vegetarian</b>	<b>190</b>	<b>202</b>	<b>208</b>	<b>216</b>	<b>225</b>	<b>208.2</b>	<b>13.349</b>	<b>n<sub>1</sub> = 5</b>
<b>Healthy</b>	<b>200</b>	<b>205</b>	<b>210</b>	<b>215</b>	<b>220</b>	<b>212.0</b>	<b>11.511</b>	<b>n<sub>2</sub> = 5</b>
<b>Regular</b>	<b>245</b>	<b>260</b>	<b>265</b>	<b>270</b>	<b>280</b>	<b>264.0</b>	<b>12.942</b>	<b>n<sub>3</sub> = 5</b>
<b>Total:</b>						<b>M = 228.07</b>	<b>s<sub>t</sub> = 28.8257</b>	<b>N = 15</b>



Write out a null hypothesis for this study. Based on this data, would you be willing to conclude that diet impacts cholesterol? Why or why not?

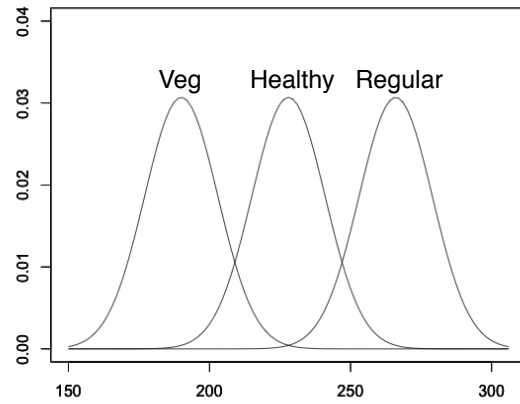
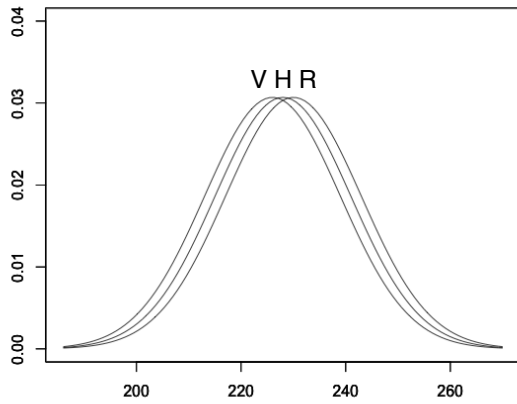
12. As we'll see, we need to assume independence, normality, and equal variances to conduct an ANOVA. Based on this data and study, are you comfortable making these assumptions? How can we determine if the assumptions are satisfied?

13. Suppose all the assumptions are satisfied. Sketch the population distributions of cholesterol levels for each diet assuming the null hypothesis is false and again assuming the null hypothesis is reasonable?

\_\_\_\_\_ Cholesterol  
True null hypothesis

\_\_\_\_\_ Cholesterol  
False null hypothesis





14. Before we go on, I want you to think about three individuals in this study:

- A person randomly assigned to the vegetarian group who ended with a cholesterol level of 190.
- Another person in the vegetarian group who ended with a cholesterol level of 225
- A third person, assigned to the regular diet group, who ended with a cholesterol level of 280.

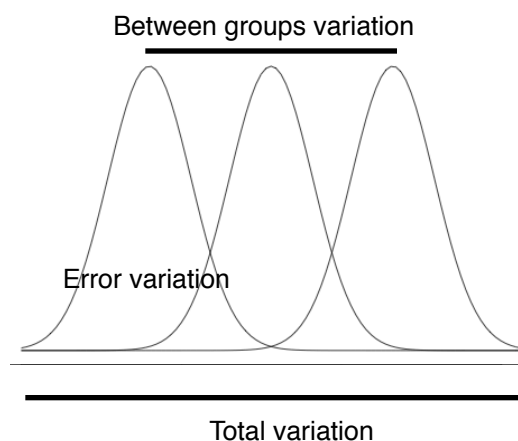
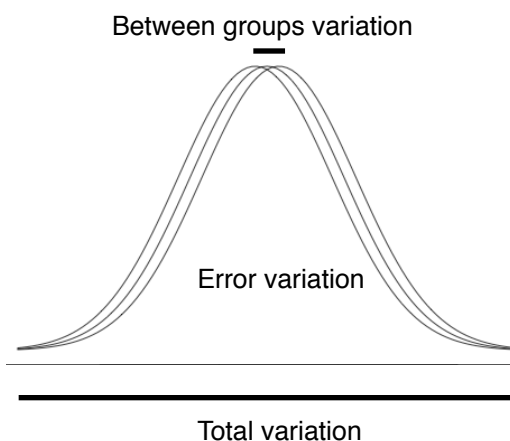
List several potential reasons why these individuals ended the study with different cholesterol levels.

15. In an ANOVA, we're going to look at the total variation in our data. In a sense, this is the overall difference in cholesterol levels among all the individuals in the study.

We're then going to divide that total variation into two components. The first component, called a **between-groups** or treatment component, will represent the variation that is due to the treatments (or groups). In this study, it would represent the variation in cholesterol that is due to diet.

The second component, called **within-groups** or error, will represent the variation that is not due to the treatment. In other words, it would represent the variation we have within each group. In this study, it would represent the variation in cholesterol levels among individuals in the same diet.

We can visualize these sources of variation:



16. If we were able to calculate the variance **between** groups and the variance **within** groups, we could take their ratio and compare that to an F-distribution.

Suppose we did this. Suppose we estimated both variances and then calculated:  $\frac{\text{variance between groups}}{\text{variance within groups}}$

If that ratio turns out to be a relatively large number, what does that say about our null hypothesis? .

If that ratio turns out to be a relatively small number, what does that say about our null hypothesis? .

17. Let's see if we can figure out formulas to estimate these variances. To get us started, let's look at the formula for the unbiased estimate of the population variance:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1} = \frac{\text{sum of squared deviations from the mean}}{\text{degrees of freedom}} = \frac{SS}{df} = \text{"Mean Square"}$$

What do SS and df represent? Why can we think of a variance as a mean square? What does that mean?

18. Let's calculate the total variation in our data. Calculate the total sum of squares (SST), explain what it represents, and determine its degrees of freedom. Then, calculate MST and explain what it represents. Finally, given what MST represents, can you think of another formula we could use to calculate SST?

$$MS_{\text{Total}} = \frac{SS_{\text{Total}}}{df_{\text{Total}}} = \frac{\sum ( \quad - \quad )^2}{( \quad - \quad )} = \frac{\quad}{( \quad - \quad )}$$

19. We know the total variation in our data is approximately  $SS_T = 11633$ . As was stated earlier, we're going to partition that variation into two components. The first component we'll calculate represents the average variation **within** each group (or the variation that is **not** due to our treatments).

Derive formulas for  $SS_E$ ,  $df_E$ , and  $MS_E$ . Explain what they represent. Given this explanation, can you think of another formula we could use to calculate  $SS_E$ ? Finally, calculate  $MS_E$  for our data

$$MS_E = \frac{SS_E}{df_E} = \frac{\sum ( \quad - \quad )^2}{( \quad - \quad )} = \frac{\quad}{( \quad - \quad )}$$

20. We can also find how much of the total variation in our data is due to the treatments. To do this, we estimate the variation among the group means.

Derive formulas for  $SS_A$ ,  $df_A$ , and  $MS_A$ . Explain what they represent. Then, calculate  $MS_E$  for our data.

$$MS_A = \frac{SS_A}{df_A} = \frac{\sum ( \quad - \quad )^2}{( \quad - \quad )} =$$

21. We can summarize all our calculations in an ANOVA summary table (shown on the next page).

<u>Source</u>	<u>Sum of Squares</u>	<u>df</u>	<u>Mean Square</u>	<u>MS Ratio</u>
Treatment (Among groups)	$SS_A = \sum n_a (\bar{X}_a - M)^2$	$a - 1$	$MS_A = \frac{SS_A}{df_A}$	$MSR = \frac{MS_A}{MS_E}$
Error (Within groups)	$SS_E = \sum (X_i - \bar{X}_a)^2$ $SS_E = \sum (n_a - 1) s_a^2$	$N - a$	$MS_E = \frac{SS_E}{df_E}$	
Total	$SS_T = \sum (X_i - M)^2$ $SS_T = \sum (N - 1) s_{total}^2$ $SS_T = SS_A + SS_E$	$N - 1$	$MS_T = \frac{SS_T}{df_T} = SS_T = s_{total}^2$	

<u>Source</u>	<u>Sum of Squares</u>	<u>df</u>	<u>Mean Square</u>	<u>MS Ratio</u>
<b>Treatment</b>	<b>9720</b>	<b>2</b>	<b>4860</b>	
<b>Error</b>	<b>1912.8</b>	<b>12</b>	<b>159.4</b>	
<b>Total</b>	<b>11632.8</b>	<b>14</b>	<b>830.9</b>	

22. Let's prove that we have partitioned the variation:  $SS_T = SS_A + SS_E$ . Explain what is happening at each step.

$$\begin{aligned}
 SS_T &= \sum \sum (x_i - M)^2 \\
 &= \sum \sum [(x_i - \bar{X}) + (\bar{X} - M)]^2 = \sum \sum [(\bar{X} - M) + (x_i - \bar{X})]^2 \\
 &= \sum \sum (\bar{X} - M)^2 + \sum \sum (x_i - \bar{X})^2 + 2 \sum \sum (\bar{X} - M)(x_i - \bar{X})
 \end{aligned}$$

since:

$$\sum (x_i - \bar{X}) = 0$$

$$= \sum n_a (\bar{X} - M)^2 + \sum (x_i - \bar{X})^2$$

$$SS_A + SS_E$$

23. Let's examine  $MS_E$  in a little more detail. Since *mean squares* is another term for *variance*,  $MS_E$  represents error variance. It represents the variance within a distribution (or the variance of a single group).

How can  $MS_E$  represent the variance of a single group of observations if we have 3 (or more) groups in our study? Remember, to conduct an ANOVA, we assume our groups have equal population variances. With this assumption, we can think of  $MS_E$  as the average variance of our groups.

This concept of an average variance should be familiar. When you learned how to conduct an independent samples t-test, you learned about the *pooled standard deviation* (or pooled standard error).

If we take the formula we used to calculate a pooled standard deviation and extend it to a situation with 3 groups, we would have:

$$\begin{aligned}
 S_{pooled}^s &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)} = \\
 &= \frac{(n_1 - 1) \left( \frac{\sum (x_1 - \bar{X}_1)^2}{(n_1 - 1)} \right) + (n_2 - 1) \left( \frac{\sum (x_2 - \bar{X}_2)^2}{(n_2 - 1)} \right) + (n_3 - 1) \left( \frac{\sum (x_3 - \bar{X}_{31})^2}{(n_3 - 1)} \right)}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)} \\
 &= \frac{\sum (x_i - \bar{X}_{a_1})^2 + \sum (x_i - \bar{X}_{a_2})^2 + \sum (x_i - \bar{X}_{a_3})^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)} = \frac{\sum (x_i - \bar{X}_a)^2}{N - a} = MS_E
 \end{aligned}$$

This demonstrates, again, that  $MS_E$  represents the average variance within our groups.

Consider, once again, the two possible outcomes of our study. Fill-in-the-blanks:

A) If the null hypothesis is true, the expected value of MSE would be \_\_\_\_\_.

B) If the null hypothesis is false, the expected value of MSE would be \_\_\_\_\_.

24. Now let's turn our attention back to  $MS_A$ . When would  $MS_A$  be larger - when the null hypothesis is true or false?

$$MS_A = \frac{\sum n_a (\bar{X}_a - M)^2}{a - 1}$$

25. Suppose the null hypothesis were true. Suppose the population means of our treatments were equal. In that case, the sample means we calculated from our data differed because of random error. Under this true null hypothesis, what would be the expected values of  $MS_A$  and  $MS_E$ ?

What would be the expected values of  $MS_A$  and  $MS_E$  under a false null hypothesis (where the population means of our treatments differed)? Fill in the table.

	<u>True null hypothesis</u>	<u>False null hypothesis</u>
<u>Expected value of <math>MS_A</math></u>	_____	_____
<u>Expected value of <math>MS_E</math></u>	_____	_____

Remember the goal of ANOVA is to test whether the means of 2+ groups are equal. If  $H_0$  is true, the treatments have no impact and both  $MS_A$  and  $MS_E$  provide unbiased estimates of the variance within a group. If  $H_0$  is false, then  $MS_A$  becomes larger.

This is the key to understanding ANOVA. We compare two estimates of variance:  $MS_A$  and  $MS_E$ . If  $MS_A$  is significantly larger than  $MS_E$ , we conclude that the null hypothesis is false (and that at least one treatment mean differs from the others). If  $MS_A$  and  $MS_E$  are similar, we conclude that the null hypothesis is true (and that the treatment means do not significantly differ).

26. How do we compare our two estimated variances:  $MS_A$  and  $MS_E$ ? What sampling distribution does this value come from? How many degrees of freedom does our test statistic have?

27. If the null hypothesis is true, what value would we expect for our mean square ratio?

28. Let's complete this cholesterol study. The following ANOVA summary table displays our calculations. Calculate the MSR and find the critical value from the F-distribution (using  $\alpha=0.05$ ). Use a computer to estimate the p-value.

<b>Source</b>	<b>Sum of Squares</b>	<b>df</b>	<b>Mean Square</b>	<b>MS Ratio</b>
<b>Treatment</b>	<b>9720</b>	<b>2</b>	<b>4860</b>	
<b>Error</b>	<b>1912.8</b>	<b>12</b>	<b>159.4</b>	
<b>Total</b>	<b>11632.8</b>	<b>14</b>	<b>830.9</b>	

Applet to calculate probabilities under the F-distribution: [http://lock5stat.com/statkey/theoretical\\_distribution/theoretical\\_distribution.html#F](http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#F)  
 Calculator method: DISTR --> FCDF(left bound, right bound, df numerator, df denominator)

29. What does that p-value tell us? Does it tell us anything about the magnitude of the differences among the group means?

30. Suppose I define something called "eta-squared" to be  $\eta^2 = \frac{SS_A}{SS_T} = \frac{9720}{11632.8} \approx 0.836$ . Interpret this value.

31. Assuming this study was real, what conclusions could we make? Can we conclude vegetarians have lower cholesterol levels than individuals on regular diets

32. Use a computer to conduct an ANOVA using the comprehension data (page 1) and the recall data (question #8).

If you don't have access to a statistical application (like R, Stata, or SPSS), you can use one of these websites:

ANOVA calculator from summary data: <http://danielsoper.com/statcalc3/calc.aspx?id=43>

ANOVA calculator from raw data: <http://vassarstats.net/anova1u.html>

If you use R, the code to run an ANOVA (for the comprehension data) would be:

```
## Run an ANOVA modeling comprehension as a function of condition (treatments)
mod1 <- aov(Comprehension ~ Condition, data=ambiguous)

## Summarize the model in ANOVA format
anova(mod1)
```

In Stata, the code would simply be: `anova comprehension condition`

The output for both programs is displayed below (for the comprehension study). Calculate eta-squared and write out any conclusions you can make.

#### Analysis of Variance Table

Response: Comprehension

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Condition	2	35.053	17.5263	10.012	0.0002002
Residuals	54	94.526	1.7505		

Number of obs =	57	R-squared =	0.2705
Root MSE =	1.32306	Adj R-squared =	0.2435

R output above  
Stata output to the right

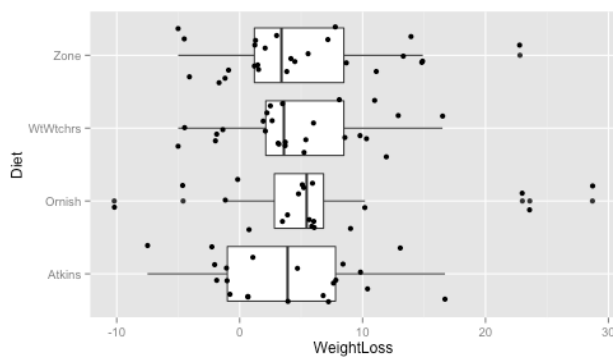
Source	Partial SS	df	MS	F	Prob > F
condition	35.0526316	2	17.5263158	10.01	0.0002
Residual	94.5263158	54	1.75048733		
Total	129.578947	56	2.31390977		

33. Do the results from that ANOVA agree with the results from the MAD-randomization-based method we conducted earlier? Why aren't the results exactly the same?

Scenario: If you want to lose weight, do some diets work better than others? Is a low-carb diet better than a low-fat diet? To investigate this, researchers randomly assigned subjects to one of three diets:

- Atkins (very low-carb)
- Zone (40:30:30 ratio of carbs:protein:fat)
- Ornish (low-fat)
- Weight Watchers

The 93 subjects who volunteered for the study were educated on their assigned diet and were observed periodically as they stayed on the diet for one year. At the end of the year, the researchers calculated the changes in weight for each subject.



Group	Count	Mean	Std. Dev
Atkins	21	3.919	6.0452
Ornish	20	6.560	9.2912
WtWch	26	4.592	5.3927
Zone	26	4.885	6.9155
<b>Total</b>	<b>93</b>	<b>M = 4.945</b>	<b>s = 6.8931</b>

Data: <http://www.math.hope.edu/isi/data/chap9/Diets2.txt>

Source: Journal of the American Medical Association ( Dansinger, Griffith, Gleason et al., 2005)

34. Suppose we conducted independent samples t-tests to compare all pairs of group means. How many t-tests would we need to run? If we used  $\alpha = 0.05$  for each test, what would be the probability that we would make at least one alpha error in all these tests?

35. Can you think of any way to control the overall alpha error rate? What would be the downside of doing this?



36. State your null and alternative hypotheses for this study.

37. First, I compared the group means using randomization-based methods with the MAD test statistic.

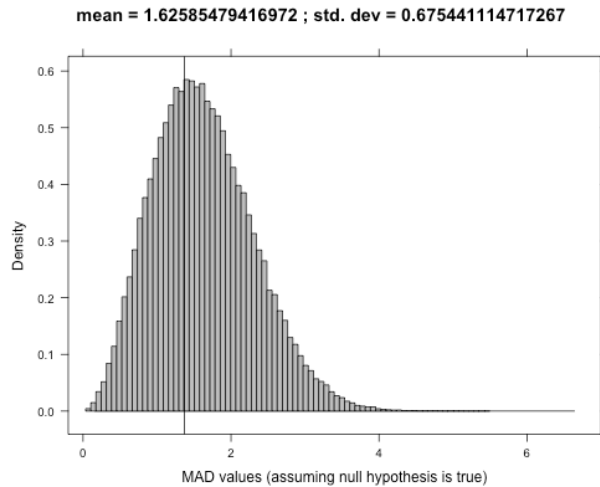
I copied the data from: <http://www.math.hope.edu/isi/data/chap9/Diets2.txt>

and used this applet: <http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2>

The results are pasted below. What conclusions can you draw? What assumptions did you make in this analysis?

**Output:**

**Observed MAD = 1.369**  
**p-value = 0.618**



38. I then had the computer conduct an ANOVA. Interpret the output and calculate an effect size. What assumptions did you make in this analysis?

39. I then had the computer conduct an ANOVA. Interpret the output and calculate an effect size. What assumptions did you make in this analysis?

**Analysis of Variance Table**

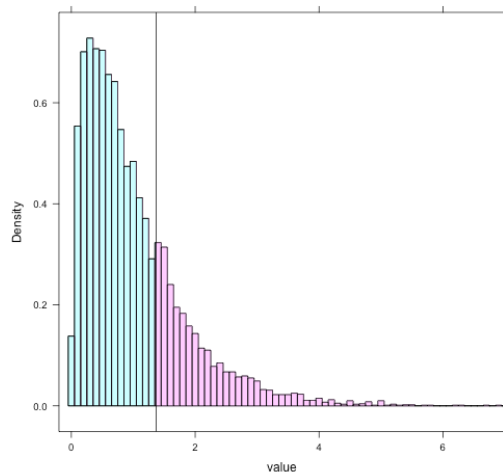
**Response: WeightLoss**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	77.6	25.866	0.5361	0.6587
Residuals	89	4293.7	48.244		

40. Finally, I conducted one more analysis with this data. I combined the randomization-based methods with the ANOVA. To do this, I:

- Calculated a value of the F-statistic for my observed data ( $F = 0.54$ )
- Randomly shuffle the group assignments (diets) to each observation 10,000 times
- Calculate the F-statistic for each of those 100,000 randomizations
- Found the proportion of those 100,000 randomizations that yielded an F-statistic  $>$  my observed F-statistic

The output is displayed below. Compare this output to the two previous analyses. Do they all produce similar p-values? What are the assumptions underlying each analysis method? Identify advantages and disadvantages of each method.



Observed  $F = 0.54$

10,000 randomizations yielded the distribution to the left

$P(F > 0.54) = 0.657$

I conducted this analysis using R, but you can replicate it by:

copying this data: <http://www.math.hope.edu/isi/data/chap9/Diets2.txt>

using this applet: [http://lock5stat.com/statkey/advanced\\_1\\_quant\\_1\\_cat/advanced\\_1\\_quant\\_1\\_cat.html](http://lock5stat.com/statkey/advanced_1_quant_1_cat/advanced_1_quant_1_cat.html)

41. As we learned last time, the F-distribution is a good approximation to the null distribution of the F-statistic if:

- The population distributions are approximately normal
- The variability in the population distributions are equal

We'll learn more about these assumptions next time. For now, explain how you could determine if these assumptions are valid for this diet study.

42. Let's (finally) finish this activity by looking at our independent samples t-test. Follow along and explain the implications of the following:

$$t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{s_{\text{pooled}}^2}}$$

Squaring this value and doing some algebraic manipulations...

$$t_{n_1+n_2-2}^2 = \frac{(\bar{X}_1 - \bar{X}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{\text{pooled}}^2} = \frac{(n_1+n_2)(\bar{X}_1 - \bar{X}_2)^2}{s_{\text{pooled}}^2} = \frac{n_1(\bar{X}_1 - \bar{X}_2)^2 + n_2(\bar{X}_1 - \bar{X}_2)^2}{s_{\text{pooled}}^2} = \frac{\sum n_a (\bar{X}_a - M)^2 / 1}{MS_E} = F$$