Activity 4: Analysis of Variance (ANOVA)

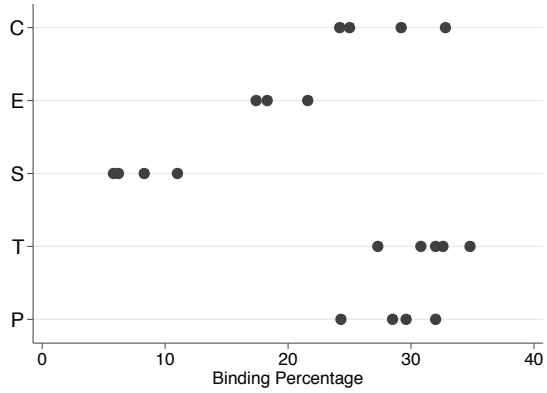
In a previous statistics course, you learned how to conduct an independent samples t-test to compare two population means. Today, we'll begin to study a method for comparing the means of three or more groups.

1) Last time, we learned that if we want to compare two variances, we _____ them and compare that value to _____.

Scenario: A drug's efficiency may be affected by the degree to which it binds to the proteins within blood plasma. The less bound a drug is, the more efficiently it can traverse cell membranes or diffuse.

Five drugs are administered to a total of 20 patients. The following table and graph represent the plasma protein binding percentages measured for each drug:

	Measurements					Mean	Std. Dev.	Sample
(P) Penicillin G	29.6	24.3	28.5	32.0		28.600	3.2177	$n_1 = 4$
(T) Tetracycline	27.3	32.6	30.8	34.8	32.0	31.500	2.7604	$n_2 = 5$
(S) Streptomycin	05.8	06.2	11.0	08.3		7.825	2.3838	$n_3 = 4$
(E) Erythromycin	21.6	17.4	18.3			19.100	2.2113	$n_4 = 3$
(C) Chloramphenicol	29.2	32.8	25.0	24.2		27.800	3.9900	$n_5 = 4$
				Total group:		$M = 23.585$	$s_t = 9.3889$	$N = 20$



2) What does M represent? Can we calculate it by taking the average of our 5 group means? Why?

3) What does s_t represent? How could we calculate it?

4) Based on the data and the graph, would you be willing to conclude the drugs have different average binding percentages? Why or why not?

5) We want to conduct a test to determine if at least one of the population means differs from the others. Write out appropriate null and alternate hypotheses.

6) If we want to compare the means of all 5 groups, we could use a series of independent samples t-tests. For example:
   Group 1 vs Group 2 Group 1 vs Group 3 Group 1 vs Group 4 Group 1 vs Group 5
   Group 2 vs Group 3 and so on...

   If we did this, how many independent samples t-tests would we need to make all possible pairwise comparisons? If we have **g** groups, how many t-tests would it take to compare all possible pairs of means?

7) What assumptions would we need to make in order to conduct an independent samples t-test? Do those assumptions look reasonable in this scenario?

8) In addition to being time-consuming, there's another reason why we might not want to conduct multiple t-tests – it inflates our overall α error rate (Type 1 error rate). What does α represent?

   α = _____

   Suppose we run 10 t-tests and set α=0.10 for each test. We might run these tests and think that we had a 10% chance of making this type of error. In fact, our overall α error rate increases substantially.

   Calculate the probability of making **at least one alpha error** if we were to conduct 10 independent samples t-tests.
   Hint: It may be easier to calculate the complement of that probability statement.

   P(making at least one α error in 10 tests) = _____

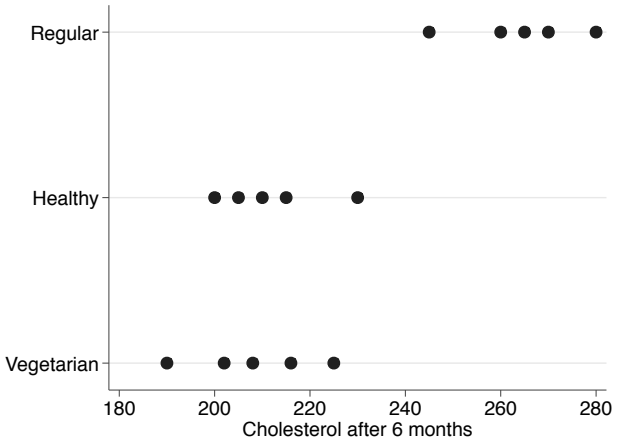
   See if you can generalize this formula. What formula could we use to calculate the overall α error rate if we compared all possible pairs of means from **g** groups?

9) Suppose we want an overall α of 0.10 when we conduct 10 t-tests. Do you see anything we could do to control the overall α? Is there any downside to doing this?

We need to develop a procedure that will allow us to compare 2 or more population means without inflating our overall α-error rate. As we'll learn, this procedure is called analysis of variance (ANOVA). Despite its name, the goal of ANOVA is not to compare variances; the goal is to compare means.

10) To simplify things, let's work with a new dataset. 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
Vegetarian	190	202	208	216	225	208.2	13.349	$n_1 = 5$
Healthy	200	205	210	215	220	212.0	11.511	$n_2 = 5$
Regular	245	260	265	270	280	264.0	12.942	$n_3 = 5$
					Total:	$M = 228.07$	$s_t = 28.8257$	$N = 15$



11) 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, to conduct an ANOVA, we'll assume independence, normality, and equal variances. 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.

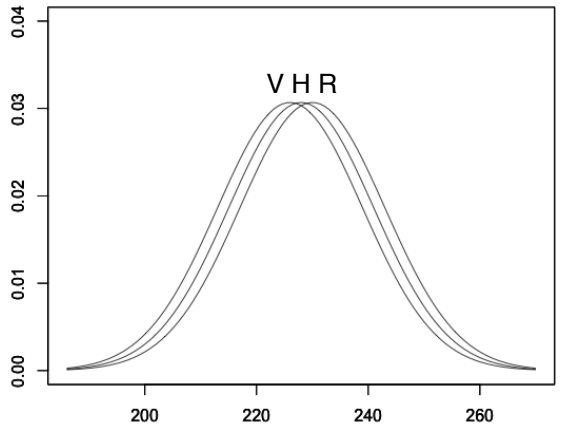
Cholesterol

True null hypothesis

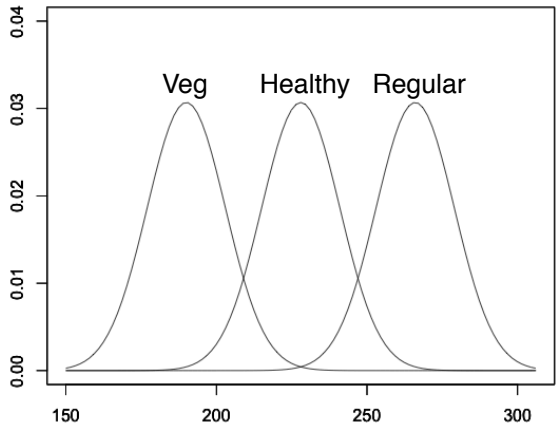
Cholesterol

False null hypothesis

True null hypothesis



False null hypothesis



14) Before we go on, I want you to think about three individuals in this study. The first person, randomly assigned to the vegetarian group, wound up with a cholesterol level of 190. Another person in the vegetarian group ended with a cholesterol level of 225. A third person, assigned to the regular diet group, ended with a cholesterol level of 280.

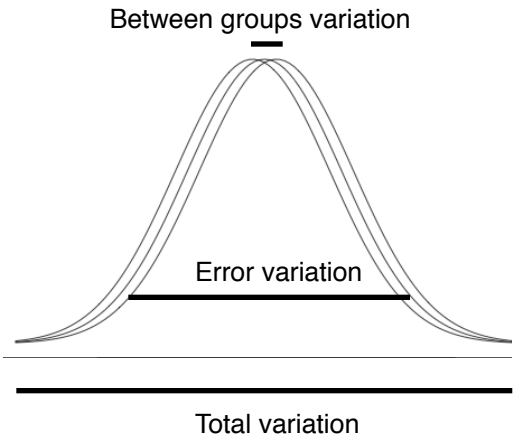
List several potential reasons why these individuals ended the study with different cholesterol levels.

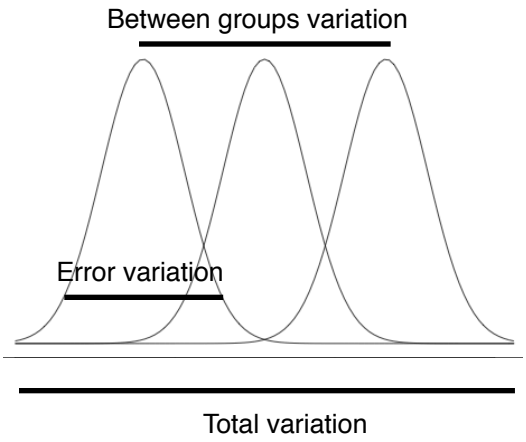
15) In an ANOVA, we're going to look at the total variation in our data. In a sense, this is the overall difference in cholesterol levels among all the individuals in the study.

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 compare them by taking their ratio.

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

If we calculate that ratio to be a relatively large number, what does that say about our null hypothesis?

Likewise, 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 (SS_T), explain what it represents, and determine its degrees of freedom. Then, calculate MS_T and explain what it represents. Finally, given what MS_T represents, can you think of another formula we could use to calculate SS_T?

$$MS_T = \frac{SS_T}{df_t} = \frac{\sum(\quad - \quad)^2}{(\quad - \quad)} = \frac{\quad}{(\quad - \quad)}$$

19) We now 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.

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

We can create an ANOVA summary table to summarize the formulas and calculations we've done thus far.

Source	Sum of Squares	df	Mean Square	MS Ratio
Treatment (Among groups)	$\mathrm{SS}_A = \sum n_a \left(\bar{X}_a - M\right)^2$	$a-1$	$MS_A = \dfrac{\mathrm{SS}_A}{df_A}$	$MSR = \dfrac{MS_A}{MS_E}$
Error (Within groups)	$\mathrm{SS}_E = \sum \left(X_i - \bar{X}_a\right)^2$ $\mathrm{SS}_E = \sum \left(n_a - 1\right) s_a^2$	$N-a$	$MS_E = \dfrac{\mathrm{SS}_E}{df_E}$	
Total	$\mathrm{SS}_T = \sum \left(X_i - M\right)^2$ $\mathrm{SS}_T = \sum \left(N-1\right) s_{total}^2$ $\mathrm{SS}_T = \mathrm{SS}_A + \mathrm{SS}_E$	$N-1$	$MS_T = \dfrac{\mathrm{SS}_T}{df_T} = \mathrm{SS}_T = s_{total}^2$	

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

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

$$SS_T = \sum\sum \left(x_i - M\right)^2$$

$$= \sum\sum \left[\left(x_i - \bar{X}\right) + \left(\bar{X} - M\right)\right]^2 = \sum\sum \left[\left(\bar{X} - M\right) + \left(x_i - \bar{X}\right)\right]^2$$

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

since:

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

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

$$SS_A + SS_E$$

22) 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:

$$s^s_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_1^2 + (n_3-1)s_1^2}{(n_1-1)+(n_2-1)+(n_3-1)} =$$

$$= \frac{(n_1-1)\left(\frac{\sum\left(x_1-\overline{X}_1\right)^2}{(n_1-1)}\right) + (n_2-1)\left(\frac{\sum\left(x_2-\overline{X}_2\right)^2}{(n_2-1)}\right) + (n_3-1)\left(\frac{\sum\left(x_3-\overline{X}_{31}\right)^2}{(n_3-1)}\right)}{(n_1-1)+(n_2-1)+(n_3-1)}$$

$$= \frac{\sum\left(x_i-\overline{X}_{a_1}\right)^2 + \sum\left(x_i-\overline{X}_{a_2}\right)^2 + \sum\left(x_i-\overline{X}_{a_3}\right)^2}{(n_1-1)+(n_2-1)+(n_3-1)} = \frac{\sum\left(x_i-\overline{X}_a\right)^2}{N-a} = MS_E$$

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

Consider, once again, the two possible outcomes of our study:

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 _____.

23) Now let's turn our attention back to MS_A. When would MS_A be larger – when the null hypothesis is true or false? Why?

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

24) 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.

	True null hypothesis	False null hypothesis
Expected value of MS_A	_____	_____
Expected value of MS_E	_____	_____

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).

25) 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?

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

27) Let's complete our cholesterol study. The following ANOVA summary table displays our calculations. Calculate the MSR and find the critical value from the F-distribution (using α=0.05). Use your calculator to estimate the p-value.

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

Calculator p-value: DISTR --> FCDF(left bound, right bound, df numerator, df denominator)

28) Suppose we define something called "eta-squared" to be: $\eta^2 = \frac{SS_A}{SS_T} = \frac{9720}{11632.8} \approx 0.84.$ Interpret this value. What does the 84% represent?

29) What conclusions can we make from this study? Can we conclude vegetarians have lower cholesterol levels than individuals on regular diets?

30) Let's take the opportunity to run this ANOVA on a computer. Using a statistical application (R, Stata, SPSS) or an applet, see if you can verify the calculations in this example.

31) Before we move on, I want to briefly introduce ANOVA through the lens of a statistical model. For our cholesterol study, we're actually testing two different models:

A null model: $X_{ij} = \mu + \varepsilon_i$ (each individual's cholesterol level is due to a common factor and random error)

An alternate model: $X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ (cholesterol is due to a common factor, diet, and random error)

We can write out the alternate model more fully as: $X_{ij} = \mu + \left(\mu_j - \mu\right) + \left(X_{ij} - \mu_j\right)$

How do we know this equation holds?

Notice our null and alternate models differ by a single term. With this, we can rewrite our null hypothesis as: $H_0: \alpha_j = 0$

32) Complete an ANOVA for our drug binding study. Write out a null hypothesis, check the necessary assumptions, and create an ANOVA summary table. Calculate eta-squared and write out any conclusions you can make.

	Measurements					Mean	Std. Dev.	Sample
(P) Penicillin G	29.6	24.3	28.5	32.0		28.600	3.2177	$n_1 = 4$
(T) Tetracycline	27.3	32.6	30.8	34.8	32.0	31.500	2.7604	$n_2 = 5$
(S) Streptomycin	05.8	06.2	11.0	08.3		7.825	2.3838	$n_3 = 4$
(E) Erythromycin	21.6	17.4	18.3			19.100	2.2113	$n_4 = 3$
(C) Chloramphenicol	29.2	32.8	25.0	24.2		27.800	3.9900	$n_5 = 4$
				Total group:		M = 23.585	$s_t = 9.3889$	N = 20

Output from Stata:

```
                    Number of obs =      20     R-squared     =  0.9187
                    Root MSE      =  3.0125     Adj R-squared =  0.8971

      Source |  Partial SS    df       MS           F     Prob > F
-------------+----------------------------------------------------
       Model |  1538.75794     4  384.689486       42.39     0.0000
             |
        var1 |  1538.75794     4  384.689486       42.39     0.0000
             |
    Residual |    136.1275    15  9.07516666
-------------+----------------------------------------------------
       Total |  1674.88544    19  88.1518654
```

33) Let's finish this activity by comparing ANOVA to the independent samples t-test. The test statistic for a t-test can be calculated by:

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

Squaring this value and doing some algebraic manipulations...

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