If we want to compare three or more population means to see if they can be considered to be equal, we will use the mighty ANOVA!! (ANOVA ANOVA ANOVA)

Statistics
Class Notes
ANOVA: Comparing Three or More Means (Section 13.1)

Are babies born more frequently on one day of the week than other days? Do different teaching methods (online versus traditional versus hybrid versus student-centered) produce meaningfully different results on an end-of-semester exam? Do various types of soil/compost produce different mean yields?

Definition: ANOVA: Analysis of Variance (ANOVA) is an inferential method used to test the equality of three or more population means.

We will analyze three or more population means with the following null and alternative hypotheses.

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
H_1: At least one population mean is different from the others.

The reason we use ANOVA when confronted with three or more means instead of comparing the means two at a time with earlier methods is that doing multiple tests increases the probability that *at least* one test incorrectly rejects the null hypotheses (which is a type I error) often above our desired level of significance of α. (It also does *not* work to adjust the value of α for each test so that the overall probability of a type I error is where you want it, because this increases the chances of making a type II error (do *not* reject the null hypothesis when it is false).
So, yeah, don't do that.

Old, now-dead guy Ronald A. Fisher (1890 - 1962) gave us this method. What we do is *compare two estimates of the same population variance*, hence the name ANOVA! (ANOVA ANOVA ANOVA)

We are performing **one-way ANOVA** as there will be only one factor that distinguishes the populations, such as teaching method or type of soil/compost. **The data should come from a completely randomized design or random samples with a quantitative response variable.**

We could use ANOVA to test two population means but should *not*. The previous test for two means gives us more flexibility in the alternative hypothesis. Also, the ANOVA method assumes population variances are equal. That specifically is *not* assumed when we use the Welch's t-test as before for two population means. So, yeah, don't do that.

Requirements for One-way ANOVA Test:

1. There are k simple random samples from each of k populations, or a completely randomized experiment with k treatments.
2. The samples are independent of one another.
3. The populations are normally distributed.
4. The populations have the same variance, σ^2.

More on those Requirements:
The ANOVA method is robust. A small departure from normality is okay. If the population variances are *not* the same, that's okay too *as long as the sample sizes are the same*.

We will verify a population's normality with a normal probability plot as seen earlier.

When designing an experiment, try to roughly match sample sizes.

That Pesky Variance Requirement:
Of course, if we *had* population data to verify that the variances were equal, we would *not* be sampling and we would need absolutely none of this. We cannot truly verify such a thing in real life. So, we will use this metric. **If the largest sample standard deviation is *no more than twice* the smallest sample standard deviation, we will consider this requirement met.**

expl 1: A mathematics department is experimenting with four different delivery methods for content in their Algebra courses. One method is the traditional lecture (method I), the second is a hybrid format with half the class time online and the other half face-to-face (method II), the third is online (method III), and the fourth is a model from which students watch video lectures and do their work in a lab with an instructor available for assistance (method IV). To assess the effectiveness of the four methods, students in each approach are given a final exam with the results shown in the accompanying table.

Students were randomly assigned to each section. So, the first two requirements are satisfied. Let's check if the populations can be considered normal and if the equal variance requirement is met. To get started, fill out the following information.

Final Exam Scores

Method I	Method II	Method III	Method IV
79	83	85	82
77	56	60	92
86	83	74	85
70	79	71	59
85	62	67	80
75	41	76	69
79	59		
61	64		
64			
94			

Print Done

a.) The response variable is ____________________

and it is quantitative / qualitative (circle one).

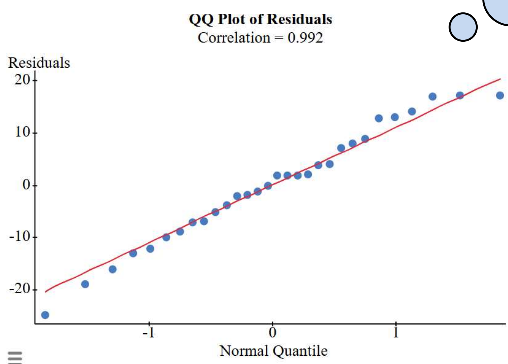
b.) The factor is _________________________ and

has _______ treatments.

c.) What are the null and alternative hypotheses?

expl 1 (continued):
d.) Here is a probability plot for the residuals (described more thoroughly in the book). Compare the correlation coefficient to the critical value for $n = 30$, the *combined* sample size, which is 0.960. Can we consider the populations to be normal? Explain.

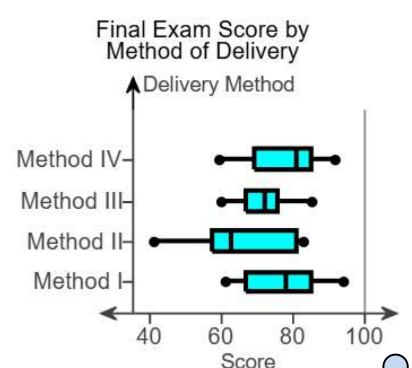
We can combine the data into one set to test for normality. Explained in the book but glossed over here, we analyze a probability plot of the residuals. If its correlation coefficient is greater than the critical value for the *total* sample size, we say the population is normal.



e.) The table here shows the standard deviations for each of the four samples. Check if we can assume the populations' variances are the same. Explain.

Method	**Standard deviation**
I	10.1105
II	14.8366
III	8.4715
IV	11.8898

f.) Here are the boxplots for the data. Does it look as though any mean is considerably different from the others?



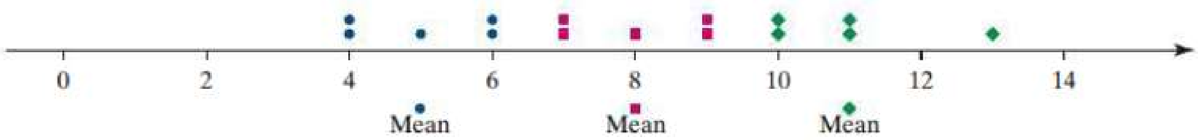
Recall the middle of the box is the median, *not* the mean.

Understanding the ANOVA Procedure:

Consider the data to the right (Table 2a) that represents three treatments of a factor under study. The sample means for each are given at the bottom of the table. Are the population means from whence they came equal?

To learn more, look at a dot plot of the data which is shown below. (The blue circles are x_1, the red squares are x_2, and the green diamonds are x_3.)

Table 2a		
x_1	x_2	x_3
4	7	10
5	8	10
6	9	11
6	7	11
4	9	13
$\bar{x}_1 = 5$	$\bar{x}_2 = 8$	$\bar{x}_3 = 11$



(a)

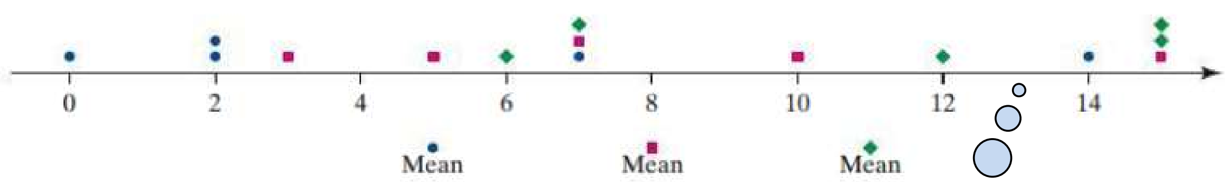
Notice how the data from each variable are clustered with its own values and *not* intermingled along the real number line.

We will speak of **within-sample variability** and **between-sample variability**. The variability of each sample individually is **within-sample variability** and the variability among sample *means* is **between-sample variability**. For the data above, the between-sample variability is much higher than the within-sample variability.

Contrast that with this second set of data. The sample means for each are given at the bottom of the table; notice they match the previous data. Are the population means from whence *they* came equal?

Let's look at a dot plot of this data below. (The blue circles are y_1, the red squares are y_2, and the green diamonds are y_3.)

Table 2b		
y_1	y_2	y_3
14	10	6
2	3	7
2	15	12
7	7	15
0	5	15
$\bar{y}_1 = 5$	$\bar{y}_2 = 8$	$\bar{y}_3 = 11$



(b)

Here, the data sets are *not* clustered.
We say the within-sample
variability is high.

When testing a null hypothesis, as always, we assume it to be true. In other words, we assume that the samples come from the same normal population with mean μ and variance σ^2.

The *F*-distribution (no, that does *not* stand for what you think):
We will use the F-distribution (named for Ronald Fisher from before). It was covered in a section we skipped, but here it is in quick form (courtesy of the book).

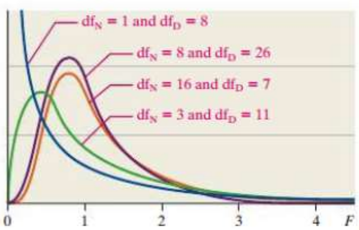
Fisher's *F*-distribution

If $\sigma_1^2 = \sigma_2^2$ and s_1^2 and s_2^2 are sample variances from independent simple random samples of size n_1 and n_2, respectively, drawn from normal populations, then

$$F = \frac{s_1^2}{s_2^2}$$

follows the F-distribution with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator.

Figure 15
F-distributions



Characteristics of the *F*-distribution

1. The F-distribution is skewed right.
2. The shape of the F-distribution depends on the degrees of freedom in the numerator and denominator. See Figure 15. This is similar to the χ^2-distribution and Student's t-distribution, whose shapes depend on their degrees of freedom.
3. The total area under the curve is 1.
4. The values of F are always greater than or equal to zero.

Here is the formula we use to find the F-test statistic.

ANOVA *F*-Test Statistic

The analysis of variance ***F*-test statistic** is given by

$$F_0 = \frac{\text{between-sample variability}}{\text{within-sample variability}}$$

We'll reject H_0 if this is too large.

But how do we find these variabilities? To make a long story a bit shorter, the **between-sample variability** compares each sample mean with the overall mean. The **within-sample variability** is a weighted average of the samples' variances.

They are both estimates of the population variance σ^2. We will find them and compare them in the ratio that is the F-test statistic.

We call the between-sample variability the **mean square due to treatment (MST)**. It is calculated $MST = \dfrac{\sum_{i=1}^{k} n_i\left(\bar{x}_i - \bar{x}\right)^2}{k-1}$. Here, k is the number of treatments.

We call the within-sample variability the **mean square due to error (MSE)**. It is calculated

$MSE = \dfrac{\sum_{i=1}^{k}\left(n_i - 1\right)s_i^2}{n-k}$. Here, n is the total sample size (sum of all n_i).

Again, both the MST and MSE are considered estimates of the population variance σ^2. The MSE is an unbiased estimator of σ^2 *whether the null hypothesis* of equal means *is true or not*. However, the MST is *only* an unbiased estimator of σ^2 *if the null hypothesis is true*.

Hence, *if the null hypothesis is true*, then the ratio $\dfrac{MST}{MSE}$ (our F-test statistic) should be close to one. *If the null hypothesis is false*, then at least one of the sample means would be far away from the overall mean causing the MST to overestimate the value of σ^2. This would result in a large F-test statistic.

Returning to the data on page 4, the overall mean ($\bar{x}$) is 8 for both sets 2*a* and 2*b*. The MST for both sets is the same at 45. Do you see why these would match?

However, the MSE for data set 2*a* (which we see each sample clustered on the dot plot) was 1.1667. Dividing, we get an F-test statistic of 38.57.

In contrast, the MSE for data set 2*b* (which we see each sample spread out on the dot plot) was 24.1667. Dividing, we get an F-test statistic of 1.86.

As we will see, we would reject H_0 for the data in set 2*a* because F_0 is too large. There is evidence that the samples do *not* come from populations with the same mean.

Computing the F-test Statistic by Hand:
We will *not* do this as a rule. However, one homework problem will require it. Instructions are given in the book. It involves finding all of the parts of MST and MSE and pluggin' and chuggin'. For the most part, we will leave that work to our future overlords, the computers.

ANOVA (Classical Method):
As before, we would compute the F-test statistic and compare it to a critical value gotten from technology or a look-up table (Table IX in book). However, we will *not* be practicing this method.

ANOVA (*P*-value Method):
As before, technology will provide us with a P-value. We will compare this to the level of significance, α. If the P-value is less than α, then we reject the null hypothesis.

ANOVA Tables:
Technology will provide us a handy table with all we need.

Instructions for StatCrunch:
(There are instructions for the TI calculators in the book.)
1. Enter the data for each sample or treatment in a separate column. Label the columns. As an alternative, you can put all values in one column and then use a second column for indicator variables for each sample.

2. Select **Stat** > **ANOVA** > **One Way**.

3. If you used multiple columns for each sample, choose to **Compare: Selected columns**. If you used the alternative way to enter data, you select **Values in a single column**. In either case, tell it which columns contain the data. (Under **Options**, you will see a **Tukey** test. We will investigate that in the next section.) Under **Graphs**, select **QQ Plot of residuals with corr.**. Click **Compute!**. (Screaming "Compute!" is optional at this time.)

If we reject the null, we have evidence that the means differ but which one(s)? We will use Tukey's test in the next section.

expl 2: Let’s investigate the methods used to teach math class in example 1.

Recall: A mathematics department is experimenting with four different delivery methods for content in their Algebra courses. One method is the traditional lecture (method I), the second is a hybrid format with half the class time online and the other half face-to-face (method II), the third is online (method III), and the fourth is a model from which students watch video lectures and do their work in a lab with an instructor available for assistance (method IV). To assess the effectiveness of the four methods, students in each approach are given a final exam with the results shown in the accompanying table.

Students were randomly assigned to each section.

Here are the final exam scores for students in the four different sections. Their means are, respectively from Method I to IV, 77, 65.9, 72.2, and 77.8. Can we conclude these samples come from populations with the same mean? In other words, does the teaching method make *no* difference on exam scores?

We will analyze the ANOVA table given in StatCrunch. The output is below.

Final Exam Scores

Method I	Method II	Method III	Method IV
79	83	85	82
77	56	60	92
86	83	74	85
70	79	71	59
85	62	67	80
75	41	76	69
79	59		
61	64		
64			
94			

Print Done

Options (1 of 2)

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
Method I	10	77	10.110501	3.197221
Method II	8	65.875	14.83661	5.2455338
Method III	6	72.166667	8.4715209	3.4584839
Method IV	6	77.833333	11.889772	4.8539789

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	3	708.825	236.275	1.7419757	0.1831
Error	26	3526.5417	135.63622		
Total	29	4235.3667			

< >

Locate the *P*-value and make your conclusion at the 5% level. Write a full sentence explaining the conclusion.

expl 3: I "fixed" some of the lower scores for Method I students. My adulterated data is to the right here. Let's analyze the ANOVA table now to see if we can say that the population means are likely different.

Below is the StatCrunch output. Locate the *P*-value and make your conclusion at the 5% level.

Method I	Method II	Method III	Method IV
79	83	85	82
77	56	60	92
86	83	74	85
75	79	71	59
85	62	67	80
75	41	76	69
79	59		
75	64		
80			
94			

Options (1 of 2)

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
Method I	10	80.5	6.1508807	1.9450793
Method II	8	65.875	14.83661	5.2455338
Method III	6	72.166667	8.4715209	3.4584839
Method IV	6	77.833333	11.889772	4.8539789

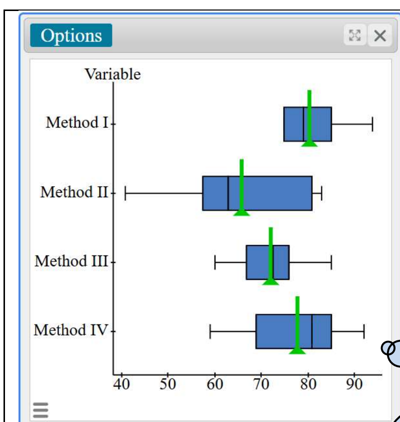
ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	3	1054.1583	351.38611	3.1000712	0.044
Error	26	2947.0417	113.34776		
Total	29	4001.2			

Compare the means as given in the top table. Do you see why the null hypothesis was rejected?

More Evidence: Boxplots:

We will use side-by-side boxplots. Below are ones I drew in StatCrunch.



StatCrunch Instructions:

1. Select **Graph** > **Boxplot**.
2. **Select columns** (all of them). Under **Other options**, you can tell it to **Draw boxes horizontally**. Under **Markers**, you can tell it to show the **Means**. They show up as green vertical segments.
3. Leave all other options alone and click **Compute!**.

How do the boxplots confirm our conclusion?