Analysis of Variance:

At its simplest, analysis of variance is a technique for dealing with more than two samples of data.

It works by (indirectly) comparing variances, which is why it's called **AN**alysis **O**f **VA**riance.

But this is a gross simplification of things. ANOVA is generally the main subject in "experimental design" type courses.

The reasons for this is that ANOVA can deal with so many different "types" of experiments.

Common terms that you may have heard that deal with ANOVA include such things as:

Nested design	Replications	Blocking
Latin squares	Random effects vs. fixed effects	
Two way ANOVA	Tukey's procedure	Contrasts

These are only a few terms associated with ANOVA. We won't discuss all of these now, and we may not discuss some of these at all.

The point is that ANOVA is a complicated topic, and there are whole semester courses offered on just ANOVA (including one here at GMU).

We'll start simple and consider ANOVA as an extension of the two sample t-test to three or more samples.

One way ANOVA:

(The presentation in your text is very similar to that in the text for 214/312)

First - why can't we just use *t*-tests for everything?

Because the probability of making a type I error would explode:

Suppose we wanted to compare three samples, A, B, and C

So we would need to test A vs. B, A vs. C, and B vs. C

That's three *t*-tests, not one.

Now suppose we pick $\alpha = 0.05$

What is α? $\alpha = \Pr\{\text{reject } H_0\}$ if H_0 is true.

So $1-\alpha$ = Pr{failing to reject H_0} if H_0 is true

and if $\alpha = 0.05$ then $1 - \alpha = 0.95$.

But now I'm doing three tests. So I need:

(Pr{failing to reject H_0} if H_0 is true)$^3 = 0.95^3 = 0.86$

This implies that α becomes 1-0.86 = 0.14

This is much higher than 0.05.

In other words, the probability of making a type I error increases with the number of tests you actually carry out.

See table 10.1, p. 190 in the text for a complete breakdown.

So we need to use ANOVA because it gives us "one" test instead of three to compare three groups (or *k* groups in general).

So what are our hypotheses?

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = ... = \mu_k$

H_1: *not all μ's are the same*

Let's follow the example in the text (10.1), where we are comparing four different kinds of feed and the effect on the weight of pigs raised on these diets:

	Feed 1	Feed 2	Feed 3	Feed 4
	60.8	68.7	69.6	61.9
	67.0	67.7	77.1	64.2
	65.0	75.0	75.2	63.1
	68.6	73.3	71.5	66.7
	61.7	71.8		60.3
i	1	2	3	4
n_i	5	5	4	5
Sum = $\sum_{j=1}^{n_i} y_{ij}$	323.1	356.5	293.4	316.2
$\bar{y}$	64.62	71.30	73.35	63.24

And the overall mean = $\bar{y} = \dfrac{323.1 + 356.5 + 293.4 + 316.2}{19} = 67.8526$

Comment on notation:

Note that "j" refers to the "j^{th}" observation of group "i". For example, $y_{31} = 69.6$.

Also note:

N = total sample size k = number of groups
$\bar{y}_i$ = mean of group i $\bar{y}$ = overall mean

(if you had ANOVA in 312, $N = n^*$ and $\bar{y} = \bar{\bar{y}}$)

The way that ANOVA works is that it compares variances. In particular, it compares the variance within each group (e.g.,. the variance for group 1, the variance for group 2, ... , the variance for group k) with the variance between groups (e.g., between groups 1, 2, ... , k).

In other words, we need to get an "average" variance with groups, and compare this to the variance of the means for the different groups.

To do this, we (of course) need to calculate Sums of Squares.

Let's figure out the total sum of squares first.

Suppose all our data were in a single column, just calculate that sum of squares for that column:

$$SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

You're taking each observation, subtracting the grand mean, and then squaring this. It's simply a Sum of Squares for everything.

We could get our usual variance from this, but it's actually not used in ANOVA, so we won't.

Here's the total sum of squares for the pig data:

$(60.8 - 67.8526)^2 + (67.0 - 67.8526)^2 +$
$\quad ... + (60.3 - 67.8526)^2 =$ **479.6874**

Now let's deal with the sum of squares for the means between our groups:

$$SS_{groups} = SS_{between} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$ (Your text uses "groups" instead of "between".)

Here we take each mean, subtract off the grand mean, and square that.

What about the n_i? It's a weighting factor. Or if you really want the details, remember that:

$$s_y^2 = \frac{s^2}{n} \quad \textit{which implies} \quad s^2 = ns_y^2$$

So if we want a variance to compare to another variance, we have to multiply by n since otherwise we'd have a variance of $\bar{y}$.

Using our pig data we get:

$$5(64.62 - 67.8526)^2 + 5(71.30 - 67.8526)^2 + 4(73.35 - 67.8526)^2 + 5(63.24 - 67.8526)^2 = \mathbf{338.9372}$$

Finally, let's get a sum of squares for within groups:

$$SS_{within} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Now we're getting the sum of squares for each group, then adding the sum of squares of the next group, and so on.

But what we're measuring is actually the sum of squares within groups.

Again, using our pig data:

$$(60.8 - 64.62)^2 + ... + (68.7 - 71.30)^2 + ... + (69.6 - 73.35)^2 + ... + (61.9 - 63.24)^2 + ... = \mathbf{140.7500}$$

Now we have all our Sum of Squares. Next we need to convert these into something that resembles a variance:

Normally we take our sum of squares and divide by n-1. This quantity (n-1) is also called the degrees of freedom.

So for our sums of squares above, we need to know:

$df_{between} = k - 1$

$df_{within} = N - k$

$df_{total} = N - 1$

This will then give us our variances, which are called “Mean Squares” in ANOVA:

$$MS_{between} = \frac{SS_{between}}{k-1}$$

$$MS_{within} = \frac{SS_{within}}{n^* - k}$$

(and we don't bother with MS_{total} since it's not used.)

So we need to get $MS_{between}$ and MS_{within} for our pigs:

$$MS_{between} = \frac{338.9372}{4-1} = 112.9791$$

and

$$MS_{withn} = \frac{140.7500}{19-4} = 9.3833$$

And we're almost done with our calculations. But before we go on, something to think about:

If H_0 is true, then you would expect that there would not be a lot of difference in the $\bar{y}$'s. In fact, our variance, $MS_{between}$, would be almost the same as our MS_{within}.

If H_0 is false, then $MS_{between}$ should be nice and big, much bigger than MS_{within}, since the means should be "more different" than the individual observations in each group.

(Think about what would happen to the above calculations if you added 40 to each of the weights for feed 4:

MS_{within} would basically not change, but $MS_{between}$ would explode.)

What we wind up doing is comparing $MS_{between}$ with MS_{within}.

Finally, we take all of the above stuff that we calculated and arrange it into an ANOVA table:

ANOVA table for Pig data:

Source	SS	df	MS	F	prob.
groups = between	338.9374	3	113.9791	12.04	0.00029
within = error = residuals	140.7500	15	9.3833		
TOTAL	479.6874	18			

So what does it all mean?

Introducing the F distribution:

Generally speaking, if one divides one variance by another, one gets an *F* distribution. In other words, s_1^2/s_2^2 is distributed as *F*. Since our *MS*'s above are variances, we shouldn't be surprised to see that if we calculate $MS_{between}/MS_{within}$, we get an *F*-distribution.

Just like the *t*-distribution, there are many different *F*-distributions. The *F* distribution depends on two parameters, the "numerator degrees of freedom" and the "denominator degrees of freedom".

Instead of just one value for *df*, we now need to worry about two.

But at least it doesn't involve complicated formulas.

- the numerator degrees of freedom is simply the *df* from the $df_{between}$ row, and the denominator is the *df* from the df_{within} row.

To look up an *F*-value, go into the table B.4, p. 680 in your text and look up the appropriate value using the numerator *df* at the top, the denominator *df* from the side, and then the appropriate column for whatever alpha you want. Some *F*-tables are set up differently, but you should be able to figure them out.

Then you proceed as always. Compare your F^* to the tabulated *F*, and if $F^* \geq F_{table}$, you reject and conclude - ?? - what do you conclude??

At least one of the means is not the same as the others.

So for our pigs, we go into the *F* table with 3 and 15 *d.f.*, and $\alpha = 0.05$, and we get $F_{table} = 4.15$

Since our $F^* = 12.04 \geq F_{table} = 4.15$, we reject our H_0 and conclude that at least one feed type is different (causes a different weight in our pigs).

What about that last column? The one labeled "prob"? Most computer packages will simply give you an F^* value, and then give you the associated probability in the last column. This makes it easy (remember, all you need to do is compare α to this *p*-value, and if the *p*-value $\leq \alpha$, you reject).

Doing ANOVA in R:

You need to put your data in the right format (using different columns works, but get's increasingly cumbersome in R (or any other statistical packages for that matter)).

So for our pig data you need to do:

60.8	feed1
67.0	feed1
.	.
.	.
.	.
68.7	feed2
67.7	feed2
.	.
etc.	

In other words, put all your measurements in the first column and an "identifier" (e.g., feed1, feed2, feed3, feeed4) in the second column.

Make sure your identifier variable is actually considered a factor in R (or you can get weird results). This is particularly important if you used numbers (e.g., 1,2,3, etc.) as your identifier. If you called your identifier variable "diet", you would do:

```
diet <- factor(diet)
```

Now you can just do:

```
pigs <- aov(weight~diet)
summary(pigs)
```

And here is the result, which should be self explanatory:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
diet         3  338.9  112.98   12.04 0.000283 ***
Residuals   15  140.8    9.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Some comments about ANOVA:

If you only have two groups, what should you use? ANOVA works just fine for two groups:

As it turns out, ANOVA for two groups is identical (in terms of power/decision) to a *t*-test with pooled variances. So it doesn't make any difference which you use if you *can assume equal variances*.

You can do a one sided ANOVA with just two groups, but it's a bit silly. Just do a *t*-test instead

Our table does give you one sided values, but you should probably just ignore them.

This raises the question: what are the assumptions of ANOVA?

They're identical to the pooled variance *t*-test

- equal variances

- data in each group is normally distributed

- as usual, data is random (though we'll learn lots of ways to deal with partially non-random data).

What happens if you violate the assumptions? As usual, this depends, although ANOVA is fairly robust (robust = it can deal with violating the assumptions).

- if sample size is large, don't worry about the normal assumption too much.

- if the *n*'s (i.e. sample sizes) are similar, don't worry about the variance assumption too much (remember -the formulas for the *t** are identical if $n_1 = n_2$ - a similar thing happens here).

- but if you have seriously non-normal data, and *n*'s that are very different, then you probably ought to worry. Talk to a statistician!!

- there is a non-parametric test similar to the Mann-Whitney *U*-test (the Kruskall-Wallace test) that's pretty good and we will learn it

For two groups, it will give the same results as the Mann-Whitney U-test.

Multiple comparisons procedures (emphasizing Tukey's procedure):

Often, when you're done with the initial ANOVA, you want to find out which means are different from each other.

ANOVA only tells you that at least one of the means is different. But you don't know which one.

Tukey's is one of a whole group of what are called "post-hoc" tests. Sometimes also called "all pairwise comparisons".

There are many others:

SNK

Fisher's least significant difference

Dunnet's

Scheffe

Bonferroni

REGWQ

We won't go into any of these (except, briefly, the last two). Tukey's is a pretty good test compared to most of these (except the last).

In all cases, the results are presented the same way (by you! - you need to arrange and present the results):

1) Arrange the means from smallest to larges. For example:

B E A C D

(B would be the smallest mean, D, the largest)

2) Draw lines over (or under) the means that are not significantly different:

_____ _____
B E A C D

This would show that B and E are not significantly different, and C and D are not significantly different. Everything else is significantly different.

_____
_____ _____
B E A C D

This would show that B and E are not significantly different, E and A are not significantly different and C and D are not significantly different. However, what about B and A? They *are* significantly different!

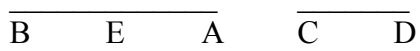
Remember, if something is N.S., this does *not* mean they're the same.

We can not “prove” the null hypothesis!

So: we have enough evidence to show B and A are different, but not enough evidence to show B and E or E and A are different. It does, actually, work.

Fortunately, this doesn't happen too often (but it does happen!)

Incidentally, here's what the picture would look like if B and A had been the same (note the difference):



3) You should always include this diagram after the result of any significant ANOVA.

Multiple comparisons are an add on to ANOVA. You really shouldn't do these unless you've done an ANOVA first.

You will find a difference of opinion on this sometimes (particularly with Tukey's procedure)

However, we'll assume you did the ANOVA first (you probably really should).

NEVER do any multiple comparison procedure if the ANOVA is not significant.

The details of Tukey's procedure:

Tukey's procedure is a way to figure out which means are different from each other (out of all pairwise comparisons) while preserving the overall level of α.

Tukey's is very similar to doing a series of t-tests (kind of what was discouraged right at the outset here).

However, it uses a different distribution (q) instead of the t-distribution.

It also uses a different way of computing the denominator.

Tukey's procedure can be used in two different ways. Both are given in your text (they're identical, just two different ways of looking at things:

Do individual tests on each pairwise comparison

Look at the confidence intervals between all possible pairs

If the CI includes 0, then “fail to reject” or we don't see a difference.

We'll stick with the second approach (it's what R does, and in this case it's easier to do than to print out a whole bunch of different tests).

Your text shows Tukey's in chapter 11.

So here is how it works:

1) Calculate:

$$(\bar{y}_1 - \bar{y}_2) \pm (q_{\alpha,\nu,k})(SE)$$

where q = q from table B.5, p. 717 with:

ν = error *d.f.* = $N - k$
k = number of groups
α = α, as usual

and:

$$SE = \sqrt{\frac{s^2}{n}}$$

where:

n = sample size of each group (assuming all n's are equal)
s^2 = error mean square from the ANOVA table.

or, if the n's are not all equal:

$$SE = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_b} + \frac{1}{n_a}\right)}$$

with the same definitions as above (n_b and n_a ought to be obvious).

2) Note that the SE only changes if the sample sizes are different (otherwise you get to use the same SE every time (which does make things a bit easier)

3) If the CI includes 0, that implies no significant difference.

4) Your test suggests starting with the biggest difference of means, then proceeding from there.

If a difference is N.S., then you shouldn't do any more comparisons of means within this difference.

For example, suppose we have

B E A C D

If you test B vs. A, and it's N.S., you should NOT test B vs. E or E vs. A (because B and A are N.S.)

This is irrelevant if all the *n*'s are the same or very close. But it can occasionally give you contradictory results (e.g., B vs. A is N.S. and B vs. E is significant) if the *n*'s are very different.

5) R prints all the results in any case; but if you wind up with something weird, you should check out point (4).

Let's do just part of an example (it's too tedious to do everything, and R does all this for you). Let's look at the pig feed example again, and compare means 3 vs. 4:

$$73.35 - 63.24 \pm q_{\alpha,\nu,k}\sqrt{\frac{9.38}{2}\left(\frac{1}{5}+\frac{1}{4}\right)} = 10.11 \pm 4.076(1.453) = 10.11 \pm 5.92 = (4.19, 16.03)$$

Then you would do all the others (3 vs. 1, 3 vs. 2, 2 vs. 4, 2 vs. 1, 1 vs. 4). Yes it's tedious, and much easier with R.

Doing Tukey's in R:

This is fairly straightforward. Once you've done your ANOVA, just do:

```
TukeyHSD(pigs)
```

Assuming you did "pigs <- aov(weight~diet)"; in other words, make sure you do your ANOVA first and assign it to a variable ("pigs" in this example).

The results are then:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ diet)

$diet
               diff        lwr       upr     p adj
feed2-feed1    6.68   1.096263 12.263737 0.0168421
feed3-feed1    8.73   2.807553 14.652447 0.0034914
feed4-feed1   -1.38  -6.963737  4.203737 0.8906642
feed3-feed2    2.05  -3.872447  7.972447 0.7530266
feed4-feed2   -8.06 -13.643737 -2.476263 0.0041505
feed4-feed3  -10.11 -16.032447 -4.187553 0.0009497
```

And then note that feeds 4 and 1 are shown to be N.S., and feeds 3 and 2 are shown to be N.S. (their CI's include 0)

Incidentally, notice that 4 vs. 3 (at the bottom) gives us the same results as the manual calculation above (except (-), because R did 4 vs. 3, and I did 3 vs. 4. But this is irrelevant!

Tukey's defaults to 95% CI's. If you want to change that, do:

```
TukeyHSD(pigs, conf.level = 0.99)
```

(or whatever you want to use).

Finally, we want to take all the means and arrange them into our little graph (abbreviating "feed" with "f":

f4 f1 f2 f3

And we are finally done with Tukey's.

Final comments:

Remember - you should do a pairwise comparison *after* you do an ANOVA.

Never do a pairwise comparison if the ANOVA is N.S.

Two other procedures worth considering:

1) Bonferroni:

This is really (really) easy to do, and works under any circumstances (not just ANOVA):

a) Figure out how many comparisons you're making. Let's call this "*c*".

b) Take α and divide it by "*c*":

$$New\,\alpha = \frac{Old\,\alpha}{c}$$

c) proceed with your test using this new value of α:

- if you want to do *t*-tests, do however many you want, just use this new value for α.

- if you want to do a bunch of correlations, again, just use the new value of α.

- whatever it is you want to do, just use the new value of α.

d) the problem:

- Bonferroni has very low (= terrible) power. You really shouldn't use it unless you don't have another choice.

2) REGWQ

This test is actually names after 4 different people, and the letters are the first letters of their last names (look it up!) (Q stands for “q” distribution or test).

It is actually quite good, and not as conservative as Tukey's (it has a bit more power).

If you really get into doing a lot of pairwise comparisons, you should check it out, as it can do better than Tukey's.

It's not as well established as the others, but it is in R (you'll have to install an add-on package (look for the “mutoss” package)).