ANOV A, an introduction

A) Warnings & comment.

- we won’t learn too many details here, though it’ll sound like a lot. Once we
  moved past t-tests, there are often entire semester classes devoted to topics like
  this.

- If you find yourself in a situation where you need to do a more complicated type
  of ANOV A, talk to a statistician!!!! Hopefully you’ll know enough so that you
  can get your ideas across.

B) So what is ANOV A?

- ANOV A stands for ANalysis Of VAriance. Though it only does this indirectly.

1) Back to the t-test (aren’t you tired of this yet???).

- suppose you have three, four, or more groups you want to compare
  instead of two. For example (using example 11.1 on p. 453[ 11.15, p. 490]
  [11.7.1, p. 449] (note that the 3rd and 4th editions do this a little
differently, but it’s the same example), you subject soybeans to four
different treatments:

  - low light, control (A)
  - low light, stress (B)
  - moderate light, control (C)
  - moderate light, stress (D)

- one approach a lot of people take is to do 6 t-tests as follows:

  A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, C vs. D

- what is the problem with this???

  - what happens to the probability of making a type I error?

  - it explodes!

- for a single test, $\alpha$ is the probability of making a type I error.

- but now if you’re doing two tests, the probability of making a type I error
  goes up. For six tests, it went way up.

  {- incidentally, a similar problem can result if you take a bunch of
    measurements on two things (e.g, height, weight, length, diameter


in males and females), and then do a t-test on each of these measurements (t-test using height, t-test using weight, etc.). This is a totally different situation from ANOVA; we'll talk about it just a little during the very last lecture.)

- this is the WRONG way to do this kind of thing.

- Examine the probability this way - if you buy a lottery ticket what is the probability that you win? - what happens when you buy two? Table 11.2 on page 455 [11.2, p. 465] lists the overall probabilities of making a type I error.

2) So what do you need to do? You need to do a single test that tests all of your means at once. ANOVA will do this. For our example above, your null hypothesis becomes:

\[ H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \]

- but what about our alternative hypothesis? You obviously can’t use a one-sided alternative. Your alternative becomes:

\[ H_1: \text{not all } \mu's \text{ are the same} \]

or

\[ H_1: \text{at least one } \mu \text{ is different} \]

(these are actually the same, just different ways of phrasing)

- incidentally, this is wrong (why??):

\[ H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \]

3) The rest of the ANOVA then follows:

- determine \( \alpha \)

- verify assumptions (more on this later)

- calculate your test statistic (\( F^* \))

- compare \( F^* \) to the tabulated \( F \), if \( F^* \geq F_{\text{table}} \), then you conclude that at least one mean is significantly different.

C) So here are some details (see section 11.2, starting on p. 457 [11.2, p. 467] lists the overall probabilities of making a type I error).
1) First, we need to deal with notation. We now have more than two groups of things we’re trying to compare, so our notation becomes more complicated. We need to keep track as to which group an observation belongs to.

\[ y_{ij} = \text{observation } j \text{ in group } i \]

- \( i \) is now the "group" that our observation belongs to, not the observation number. Our observation number is now indicated by "\( j \)".

- \( i \) goes from 1 to \( k \), where \( k = \# \text{ of groups} \) (your text uses \( I \) (so \( I = k \)).

- also:

\[ n_i = \text{number of observations in group } i \]

\[ \bar{y}_i = \text{the mean for group } i \]

So, for example, the sample mean for group \( i \) becomes:

\[
\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \bar{y}_i = \frac{(y_{i1} + y_{i2} + \ldots + y_{in})}{n_i}
\]

Why do we need all this?? Well, stay tuned.

2) Here are a couple of quantities that we will need as we go on:

- What is the grand mean? The overall mean of everything (we lump all our observations into one big group, regardless of where they come from, and take the average):

\[
\bar{y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{n^*}
\]

- before we go on, let’s make sure we understand what this says (also note \( y \) has two bars to indicate the “grand mean”).

- first, sum all the observations in each group.

- then sum all these totals (i.e., add up the group totals).

- so what about \( n^* = n, \text{ in the fourth edition} \)?
$n^*$ is just the sum of all the $n$’s (the $n$ for the first group, the $n$ for the second group, etc.)


3) Now we get even more complicated. What we want is a way to describe the variation within each group, or better, the “average” variation within each group. If you have just one group, this is easy. But instead, we now have $k$ groups, so we need to average this somehow. Before we look at variation, we’ll need to look at Sums of Squares.

- The Sum of Squares for within groups is termed the “within group Sum of Squares”, and is given as follows:

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

- This says the following:

  - first, get the sum of squares for each group (that’s what the $j$ is doing if you ignore the $i$)

  - then add up these totals for each group (that’s what the $i$ is for, it adds up the $SS$ for each group).

- {4th edition note: the formula presented is a bit different; but see the footnote on p. 422, where the above formula is given.}

- Let’s keep going with the lamb weight example, p. 460 [p. 469, but see note just above].

- Now, just like we take our $SS$ and divide by $df$ (usually $n-1$ for a single sample) to get a variance, we now take this $SS_{within}$ and divide by $df$ to get something called the Mean Square$_{within}$ (abbreviated $MS_{within}$). But we need to know what $df$ is. It’s as follows:

$$df_{within} = n^* - k$$

simply, the grand total number of observations minus the number of groups we have.

so we have:
this describes, sort of, the “average” variance for our groups [but also, incidentally, estimates the overall σ squared for our population].

- your text goes into a few more details on p. 229 and 230 [224-225] \{420 - 422\}. Without getting that detailed, here are the basics:

- $MS_{within}$ is identical to our $s^2_{pooled}$ if we’re assuming equal variances and only have two groups [this is really obvious in the 4th edition].

- this measures ONLY the variation within each group. For example, if you take one group and add 7 to each number, you will not change the value for $MS_{within}$. (Remember, if you add a constant to a single sample, you’re not changing the variance).

- feel free to look at the example in the text.

4) But now we also want to measure the variation between our groups. As before, we start with the Sum of Squares:

- Here’s how we do it

$$SS_{between} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

- Now, remember that just as above, we want to convert this value into a sort of “variance”. As usual, we do this, by dividing by $df$. In this case, we have:

$$df_{between} = k - 1$$

- or, in words, our $df$ is the number of groups minus 1.

And so we get:

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

- this shouldn’t be a big surprise. You basically wind up calculating a Sum of Squares for the means of the groups and multiplying by $n_i$ (if you want
to know about \( n_i \), see the note below “for the curious”). [Incidentally, \( MS_{\text{between}} \) ALSO estimates the overall \( \sigma^2 \) squared for our population. We now have two estimates for the same thing!]

- 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, \( MS_{\text{between}} \), would be almost the same as our \( MS_{\text{within}} \).
  
  - if \( H_0 \) is false, then \( MS_{\text{between}} \) should be nice and big, much bigger than \( MS_{\text{within}} \) since the means should be “more different” than the individual observations in each group.

the following is only for the curious:

- what about the \( n_i \) in \( SS_{\text{between}} \)? Where did that come from?

- As it turns out, if \( H_0 \) is true, both \( MS_{\text{within}} \) and \( MS_{\text{between}} \) are “designed” so that they estimate \( \sigma^2 \) for the whole population.

- so remember the following?

\[
s_y^2 = \frac{s^2}{n}
\]

which implies

\[
s^2 = n s_y^2
\]

- well, since we’re estimating \( \sigma^2 \), we need to multiply our \( SS_{\text{between}} \) by \( n \) to get back to an estimate for the overall variance, since our \( SS_{\text{between}} \) is based on \( \bar{y} \), not \( y \) (we “want” the quantity on the left in the bottom equation above).

- so let’s continue with the lamb weight gain example, now on p. 463 [p. 472] [422 - 423].

5) We’re almost done with all this, but we need a few more things.

- We also want a measure of the total variability in our data. This is easy. If we just lump everything together (all our groups are now one group) and calculate the Sum of Squares, we have what we want. Using our new notation this is:
\[ SS_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \]

- again, you’re basically taking each observation, subtracting the grand mean, and then squaring this quantity. A Sum of Squares for everything.

- we also want a df for everything. If we had just one sample, this would be \( n-1 \). Same here. It’s just \( n^* - 1 \) (i.e., the total number of observations we have minus 1).

- but even though we want \( SS_{total} \) and \( df_{total} \), we generally don’t bother calculating an \( MS_{total} \). It’s not needed, but would be real easy to calculate.

Before we come to our grand conclusions, let’s finish with the lamb example, p. 464 [p. 473] [p. 425].

6) Now to finish up this part. A fundamental relationship in ANOVA is the following:

\[ SS_{total} = SS_{within} + SS_{between} \]

\[ df_{total} = df_{within} + df_{between} \]

This is what makes ANOVA work. But to see this, we’ll have to wait until next time.

7) We’ve finished up with most of the math we need to worry about, now we just have to figure out how to use all this stuff to get a test statistic and how to compare this to some kind of tabulated value.