Agenda for Week 4, Hour 3

Another one-way ANOVA example

Introduction to two-way ANOVA

Today's dataset: farms.csv

Correction from Tuesday (Week 4 Hour 1,2):

When testing for equal variance, there are two different tests:

Testing if the variances of two groups are equal is the ***Levene Test***.

Testing if the variances of three or more groups are equal is the ***Bartlett Test***.

The filled notes for 4-1 and 4-2 include the correction.

Another one-way ANOVA example.

Consider the dataset ‘farms.csv’ . In this dataset, we have the yields of 40 equally sized plots of land.

We also know the type of fertilizer (skotz, nature touch, greeno, none) and the topography of the plot (Flat or Sloped).

Our goal is to find if the average yield is different for different kinds of fertilizer.

In terms of a hypothesis test, we wish to test:

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, vs

Ha: At least one pair of means is not equal,

Where μ_1 is the average yield you would always get from the first fertilizer (i.e. the parameter).

We solve this with a ***one-way*** analysis of variance (ANOVA).

It's called a one-way ANOVA because there is only one ***grouping variable*** we are investigating.

The other variable, land type, isn't being used at the moment, so it doesn't matter that we have it.

Recall that ANOVA is used to see if the MEANS of groups are the same, but it requires that the standard deviations of groups be roughly the same.

We can look at the descriptive statistics to check this informally:

```
by(farms$yield, farms$fertilizer, sd)
by(farms$yield, farms$fertilizer, mean)
```

Fertilizer	N	Mean	SD
Skotz	10	66	10.12
Nature Touch	10	72.2	7.22
Greeno	10	81.7	11.78
none	10	60.1	9.52

Is any standard deviation at least ***three times as large*** as any other? No.

So even if the assumption of equal variance is violated, it's not a major violation.

For the sake of practice (and correct my mistake from Tuesday), we can do a formal hypothesis test.

```
bartlett.test( response, grouping_variable)
```

```
> bartlett.test(farms$yield, farms$fertilizer)

        Bartlett test of homogeneity of variances

data:  farms$yield and farms$fertilizer
Bartlett's K-squared = 2.0111, df = 3, p-value = 0.5701
```

P > 0.10, so there is no evidence that the variances are unequal.

If there WAS major inequality in the group variances, we have a couple of options.

First, we could apply a ***transformation*** such as the logarithmic transform. (the one used in the GDP vs Life dataset in Assignment 1.)

This would work best if the groups with the largest standard deviation were also the ones with the largest means.

Many common transformations also require that the response values are all positive.

Sometimes a transformation either doesn't help, or doesn't make sense for the data at hand.

In those cases, we could resort to using another ANOVA-style test called the ***Kruskal-Wallis*** test.

There's less we can do with a Kruskal-Wallis test compared to ANOVA, so it's better not to use it unless we have to.

Since none of the groups have a dominant amount of variation over any other, we can move on to doing a one-way ANOVA.

```
> mod = lm(yield ~ fertilizer, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)
fertilizer  3 2557.4  852.47  8.8835 0.0001538 ***
Residuals  36 3454.6   95.96
```

```
> mod = lm(yield ~ fertilizer, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)
fertilizer  3 2557.4  852.47  8.8835 0.0001538 ***
Residuals  36 3454.6   95.96
```

What do we know from this?

The null hypothesis is that the means are equal, and the p-value is very small (0.00015), there is very strong evidence against it. **We confidently can say the means are different.**

```
> mod = lm(yield ~ fertilizer, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)
fertilizer  3 2557.4  852.47  8.8835 0.0001538 ***
Residuals  36 3454.6   95.96
```

Df = 36, which makes sense because there are 40 observations and 4 parameters of interest.

The mean squared error (MSE) is 95.96, which could be useful later.

Comparing three or more means? Ano vat to do!

We may also want to know which of the fertilizers is related to a statistically significantly higher yield. That calls for a ***Tukey's HSD*** analysis.

```
> mod = aov(yield ~ fertilizer, data=farms)
> TukeyHSD(mod)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ fertilizer, data = farms)

$fertilizer
                     diff       lwr        upr     p adj
nature touch-skotz    6.2  -5.59874 17.9987396 0.4983377
greeno-skotz         15.7   3.90126 27.4987396 0.0052640
none-skotz           -5.9 -17.69874  5.8987396 0.5400718
greeno-nature touch   9.5  -2.29874 21.2987396 0.1515366
none-nature touch   -12.1 -23.89874 -0.3012604 0.0426463
none-greeno         -21.6 -33.39874 -9.8012604 0.0001057
```

Using a family-wise / experiment-wise alpha of 0.05, we can see that there is a significant difference between…

Group means of "skotz" and "greeno" ($p = 0.005$)

Group means "nature touch" and no fertilizer ($p = 0.042$)

Group means "greeno" and no fertilizer ($p < 0.001$)

And that there are no other significance differences.

Conversely, we find no evidence of a difference in the following pairs:

Nature Touch and Skotz,

Nature Touch and Greeno,

No fertilizer and Skotz.

Let’s look a couple of these in combination, is there a paradox here?

(Nature Touch) is similar to (Skotz), and

(No fertilizer) is similar to (Skotz), but

(No fertilizer) is NOT similar to (Nature Touch).

There is no paradox.

When a p-value against two means being equal is large, it does NOT imply that there is no difference.

It ONLY means that there is ***not enough evidence*** to detect a difference.

A better way to think of the Tukey output would be:

Between (Nature Touch) and (Skotz), there is a small difference,

Between (No fertilizer) and (Skotz), there is a small difference, and

Between (No fertilizer) and (Nature Touch) , there is a LARGE difference.

In other words, the large difference between (Nature Touch) and (No fertilizer) is made of small differences added together.

Fertilizer	N	Mean	SD
Greeno	10	81.7	11.78
Nature Touch	10	72.2	7.22
Skotz	10	66	10.12
none	10	60.1	9.52

To summarize the groups into clusters:

Start with the group with the highest mean: Greeno

(Greeno) is not significantly different from (Nature Touch), so they get put in a cluster together.

However, (Greeno) is significantly higher than the other two means, so there are only these first two in ‘Cluster a’.

There are no significant differences between group means that share a cluster letter.

Fertilizer	N	Mean	SD	Cluster
Greeno	10	81.7	11.78	a
Nature Touch	10	72.2	7.22	a
Skotz	10	66	10.12	
none	10	60.1	9.52	

Next, move on to the 2nd highest-mean group, that's Nature Touch. Compare it to the two means that are lower than it.

(Nature touch) is not significantly different than (Greeno).

It IS significantly different than (no fertilizer). (Nature Touch) and (Greeno) get a new cluster that we will call 'Cluster b'.

Fertilizer	N	Mean	SD	Cluster
Greeno	10	81.7	11.78	a
Nature Touch	10	72.2	7.22	ab
Skotz	10	66	10.12	b
none	10	60.1	9.52	

Move to the 3rd highest-mean group, Skotz.

(Skotz) is not significantly different from no fertilizer, so both of those groups to into 'Cluster c'.

Fertilizer	N	Mean	SD	Cluster
Greeno	10	81.7	11.78	a
Nature Touch	10	72.2	7.22	ab
Skotz	10	66	10.12	bc
none	10	60.1	9.52	c

To repeat: There are no significant differences between group means that share a cluster letter.

Take a second, and get ready for two-way ANOVA.

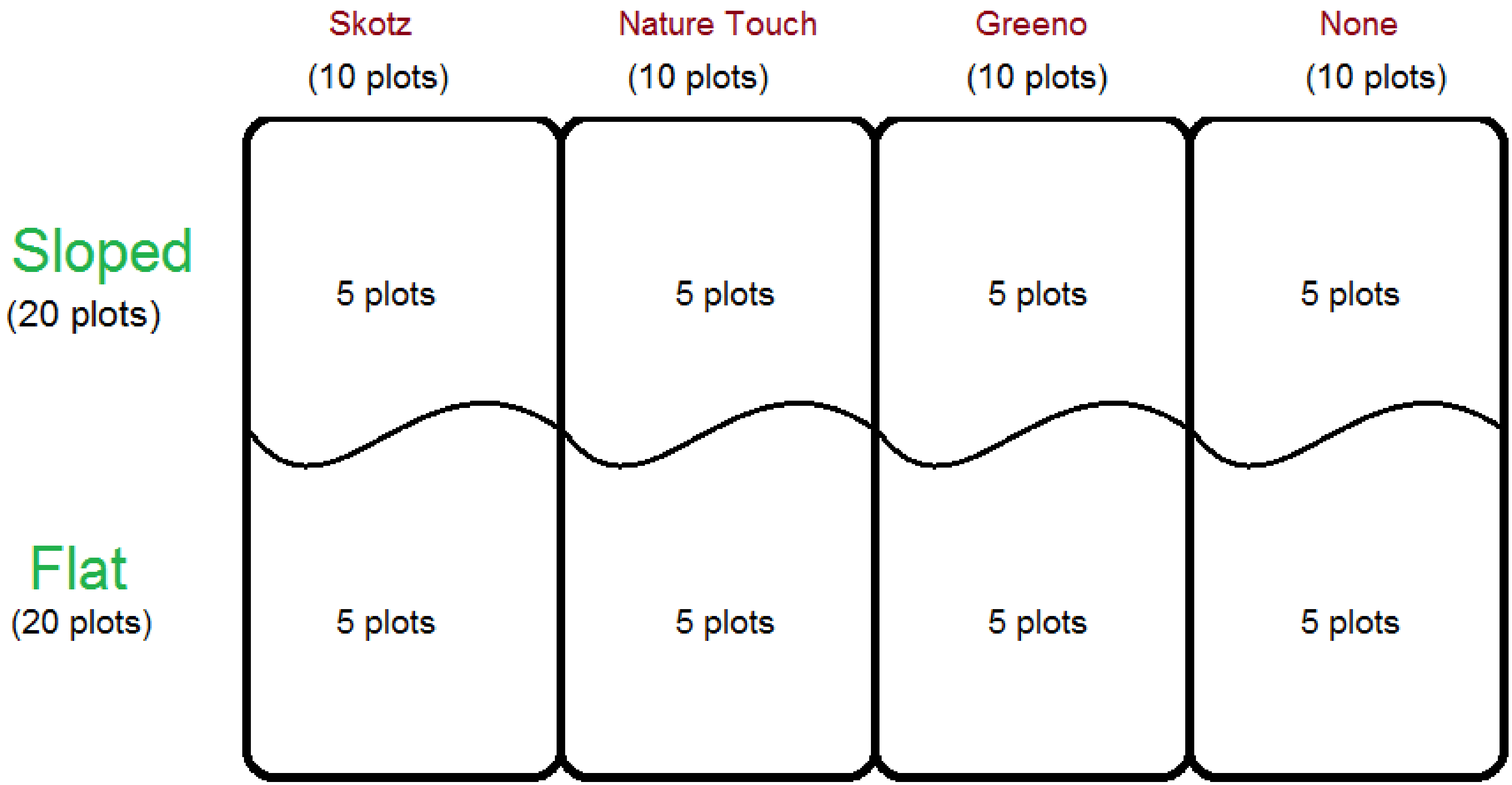
ANOVA isn't just a test, it's a method of breaking down the sources of variation in a set of responses. It can be taken a lot farther than being used as an improved t-test.

For example, we can use ANOVA to examine the effect of 2 or more grouping variables at the same time.

An ANOVA that looks at two grouping variables together is called a two-way ANOVA.

Returning to the farms.csv dataset. There were four levels of fertilizer treatment, but there were also two levels of “land type” treatment that we haven’t considered yet.

In this particular dataset, not only are there 20 sloped plots and 20 flat plots, but there are 5 plots of every combination of land type and fertilizer.



This is sort of setup, where there are an equal number of observations in every category (or combination of categories) is called a ***balanced design***.

Balanced designs are useful because some estimates are only as good as the smallest group, but they are not necessary for ANOVA procedures.

Balanced treatment categories take a lot of practical setup, and for something complex like a human, or large like a plot of land, balanced designs are a luxury.

We already know that fertilizer matters for yield, but what about land type?

To test both, we can make a linear model that includes both variables, and perform a two-way ANOVA on that.

mod = lm(yield ~ fertilizer + land, data=farms)

anova(mod)

```
> mod = lm(yield ~ fertilizer + land, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)    
fertilizer  3 2557.4  852.47  17.985 3.11e-07 ***
land        1 1795.6 1795.60  37.882 4.83e-07 ***
Residuals  35 1659.0   47.40                      
```

What can we tell from this?

```
> mod = lm(yield ~ fertilizer + land, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)    
fertilizer  3 2557.4  852.47  17.985 3.11e-07 ***
land        1 1795.6 1795.60  37.882 4.83e-07 ***
Residuals  35 1659.0   47.40
```

That both fertilizer AND land type matter. They both have their own distinct, ***additive*** effect on the yield.

```
> mod = lm(yield ~ fertilizer + land, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)    
fertilizer  3 2557.4  852.47  17.985 3.11e-07 ***
land        1 1795.6 1795.60  37.882 4.83e-07 ***
Residuals  35 1659.0   47.40                      
```

What may be surprising is that the p-value for the fertilizer effect is smaller in the two-way ANOVA.

This indicates that, accounting for land type makes the evidence that fertilizer type matters even stronger.

This strengthening of evidence is a big advantage a two-way ANOVA has over using two separate, but simpler one-way ANOVAs.

By including both variables, a 'filtering' effect happens. The ANOVA procedure finds the variance explained by land type, but only out of the variance left unexplained by fertilizer type.

```
> mod = lm(yield ~ fertilizer, data=farms)
> anova(mod)
Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value    Pr(>F)
fertilizer  3 2557.4  852.47  8.8835 0.0001538 ***
Residuals  36 3454.6   95.96
```

Also notice that in the one-way ANOVA, the Mean Squared ***Error (MSE)*** is 95.96.

In the two-way ANOVA, it is 47.40.

Significance tests use the MSE to find the standard error.

A smaller standard error is better because it allows us to find differences between groups more easily, therefore a smaller MSE is better too.

However, the MSE will only get smaller if the new variable (e.g. land type) actually has a significant effect on the response.

In other words, models only improve when meaningful information is included.

In this case, considering both fertilizer type and land type together produces a better model of crop yield that either one separately.