ANOVA

Vinny Paris

December 2025

Motivation

We've done a test for one mean

Motivation

We've done a test for one mean

We've done a test for two means

Motivation

We've done a test for one mean

We've done a test for two means

Where do we go from here?

Motivation

We've done a test for one mean

We've done a test for two means

Where do we go from here?

Correct Answer: Regression Coefficient Testing (what we did...)

Motivation

We've done a test for one mean

We've done a test for two means

Where do we go from here?

Correct Answer: Regression Coefficient Testing (what we did...)

Second Best Answer: Tests for more than two means

What?

We have a t-test for one mean:

$$H_0 : \mu = 10 \tag{1}$$

We have a t-test for one mean:

$$H_0 : \mu_1 = \mu_2 \tag{2}$$

Now we want a test for

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 .... \tag{3}$$

Nomeclature

The easiest way for me to think of ANOVA is in the terms of a linear model with indicator variables so that is how I will discuss it with you guys...

A crop study was ran on the yeild of oats. There were three variety of oats (named Golden Rain, Marvelous, and Victory) that were randomly assigned to 7 fields with 4 replications at each field

- Factors:
- Levels:
- Treatments:

Replication is when a treatment is, independently, applied to two or more experimental units

Nomenclature

The easiest way for me to think of ANOVA is in the terms of a linear model with indicator variables so that is how I will discuss it with you guys...

A crop study was ran on the yeild of oats. There were three variety of oats (named Golden Rain, Marvelous, and Victory) that were randomly assigned to 7 fields with 4 replications at each field

- Factors: Variety of Oats
- Levels: Variety had three (Golden Rain, Marvelous, Victory)
- Treatments: same as the levels since there is 1 factor

Replication is when a treatment is, independently, applied to two or more experimental units

Model

The above *yields* a model, with Marvelous as the baseline, that can be wrote as...

$$\widehat{Yield} = \beta_0 + \beta_1 \mathbb{I}_{Golden\ Rain} + \beta_2 \mathbb{I}_{Victory}$$

Sure be nice if I could....

1. Check to see if there is evidence either β_1 or β_2 is 0
2. Check to see if there is a difference between β_1 and β_2
3. Check to see if variety actually matters or if it is random noise

But How?

But how?

Analysis **O**f **VA**riance

Big Picture

The total variance of the response (TSS) can be broken into what we can explain (SSM) and what we cannot explain (SSE)

We talked about that before

Big Picture

The total variance of the response (TSS) can be broken into what we can explain (SSM) and what we cannot explain (SSE)

We talked about that before

It turns out that SSM/SSE is a scaled F-distribution

- Named for Ronald Fisher
- Developed at Iowa State
- And independently at Indian Statistical Institute (Kolkata)

Big Picture

The total variance of the response (TSS) can be broken into what we can explain (SSM) and what we cannot explain (SSE)

We talked about that before

It turns out that SSM/SSE is a scaled F-distribution

It turns out further that SSM can itself be broken down into distinct parts associated with each explanatory variable

Hypothesis: Null

Before we go into the math let's discuss the null hypothesis to understand our starting point

$$H_0 : \beta_1 \ = \ \beta_2 \ = \ \beta_3 \ = \ \beta_4 .... - 0$$
$$H_A : \text{Otherwise}$$

Let's unpack H_0 first. What does it mean in words?

Hypothesis: Null

Before we go into the math let's discuss the hypothesis to understand our starting point

$$H_0 : \beta_1 \ = \ \beta_2 \ = \ \beta_3 \ = \ \beta_4 .... = 0$$
$$H_A : \text{Otherwise}$$

Let's unpack H_0 first. What does it mean in words?

The coefficients on the indicators are 0 so there is no linear relationship between explanatory variable and the response.

More plainly, it says the means of all the groups are the same

Hypothesis: Alt

Before we go into the math let's discuss the alt hypothesis to understand our starting point

$$H_0 : \beta_1 \;=\; \beta_2 \;=\; \beta_3 \;=\; \beta_4 .... = 0$$
$$H_A : \text{Otherwise}$$

Alright, so what does H_A mean?

Hypothesis: Alt

Before we go into the math let's discuss the alt hypothesis to understand our starting point

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = \beta_3 = \beta_4 .... = 0 \\
H_A &: \text{Otherwise}
\end{aligned}
$$

Alright, so what does H_A mean?

Not everything is equal to 0

Hypothesis: Alt Warning

There is a very dangerous subtlety here

$$H_0 : \beta_1 \ = \ \beta_2 \ = \ \beta_3 \ = \ \beta_4 .... = 0$$
$$H_A \text{ :Otherwise}$$

H_A does NOT ONLY say $\beta_i \ \neq \ 0$ for some β

ie H_A does not say at least two group means are different

Hypothesis: Alt Warning

There is a very dangerous subtlety here

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 .... = 0$$
$$H_A : \text{Otherwise}$$

H_A does NOT ONLY say $\beta_i \neq 0$ for some i (ie H_A does not say at least two group means are different)

The F-test also simultaneously tests things like...

$$\frac{\beta_1}{2} + \frac{\beta_2}{4} + \frac{\beta_3}{4} = 0$$

Ie H_A includes if any *linear combination* (weighted averages) of coefficients is not 0

Hypothesis: Alt

Why the distinction?

There exists times when the ANOVA will "reject" H_0 but no t-test for any given parameter will return strong evidence

It's not a paradox nor a problem, it's that the hypothesis for the F-test is complicated. I won't go into the math.

Hypothesis: Try it

We are interested to see if using different varieties matters or not. Using the three varieties we have Marvelous as our baseline and both Golden Rain and Victory have indicator variables in our model is...

$$\widehat{Yield} = \beta_0 + \beta_1 \mathbb{I}_{Golden\ Rain} + \beta_2 \mathbb{I}_{Victory}$$

H_0:
H_A:

Hypothesis: Try it

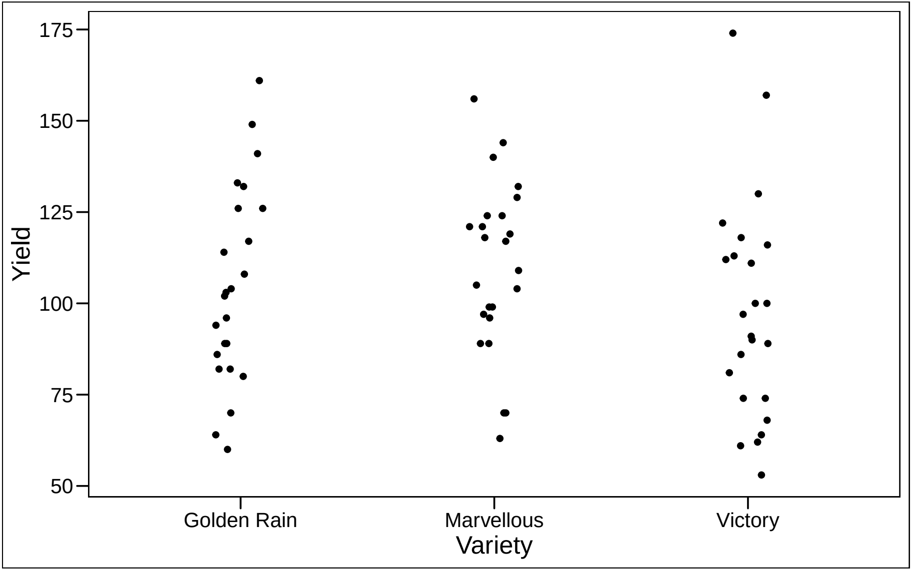
We are interested to see if using different varieties matters or not. Using the three varieties we have Marvelous as our baseline and both Golden Rain and Victory have indicator variables in our model is...

$$\widehat{Yield} = \beta_0 + \beta_1 \mathbb{I}_{Golden\ Rain} + \beta_2 \mathbb{I}_{Victory}$$

H_0: $\beta_1 = \beta_2 = 0$
H_A: Otherwise

So we are claiming the coefficients are 0 and that the effect of variety is irrelevant

Visualize Your Data



Assumptions

Remember how I said it's easiest if you think about it as a linear model?

Yeah just use MLR assumptions and you are golden

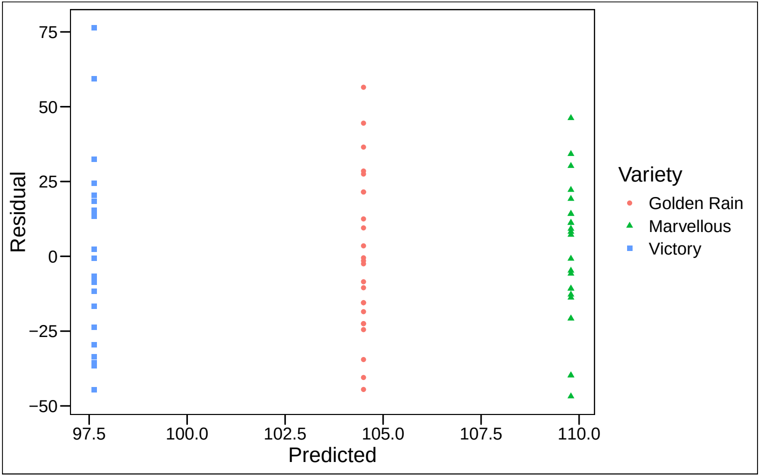
- Random
- Residuals from your model are independent and identically distributed
  - Big one here is looking for heteroskedasticity
  - ANOVA's are apocryphally said to be sensitive to changes in variance
- Population (of residuals) is normal or n is large

Assumptions

Random: Treatments were randomly assigned so yes

IID: Yes

Normal: Yes



Test Statistic: Sum of Squares

$$\text{Total Sum of Squares} = \Sigma_{i=1}^{n} \left(y_i - \bar{y}\right)$$

$$\text{Sum of Squares of the Errors} = \Sigma_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

$$\text{Sum of Squares of the Model} = \Sigma_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

$$\text{TSS} = \text{SSE} + \text{SSM}$$

Turns out the ratio of SSM/SSE (times some stuff) has a sampling distribution.....

ANOVA

Let's make a table to keep this all straight....

Source	Sum of Squares
Model	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
Errors	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$

What's the stuff?

Before we can introduce the sampling distribution we need to talk about degrees of freedom for ANOVA's. This requires us to keep track of multiple (degrees of freedom)s.

1. Explanatory Variables:
   - It's the number of beta's the variable uses (2 for our crop study)
   - In particular the degrees of freedom for the categorical explanatory variable = number of levels - 1 = k - 1
     - We had three varieties (levels) so k = 3
     - df = 2 in our crop study example
   - Numeric explanatory variables have 1 df (only have to estimate one parameter)
2. df_{TSS}: n - 1 = sample size - 1
3. df_{SSE}: df_{TSS} - df_{SSM}

ANOVA

Let's make a table to keep this all straight with our explanatory variable being a categorical variable with k levels....

Source	Formula	df
Model	$\Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k -1
Errors	$\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2$	$\text{df}_{TSS} - df_{SSM}$ =n-k
Total	$\Sigma_{i=1}^{n} (y_i - \bar{y})^2$	n - 1

Mean Squares

We want to compare the amount of "noise" there is by average number of degrees of freedom...

- Idea being something with a lot of parameters will naturally do better
- So we average the Sum of Squares by their degrees of freedom to get the mean squares

Let's make a table to keep this all straight with our explanatory variable being a categorical variable with k levels....

Source	SS	df	MS
Model	$\Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k -1	$\frac{\Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{k-1}$
Error	$\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n - k$	$\frac{\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k}$
Total	$\Sigma_{i=1}^{n} (y_i - \bar{y})^2$	n - 1	

ANOVA: F-Test Statistic

We can now calculate our test statistic! It's...

$$\frac{MS_{Model}}{MS_{Errors}}$$

which has a sampling distribution of an F-distribution with df_{Model} and df_{Errors} which means p-values

ANOVA

Source	SS	df	MS	F	p-value
Model	$\Sigma_{i=1}^{n}\ (\hat{y}_i - \bar{y})^2$	k -1	$\frac{\Sigma_{i=1}^{n}\ (\hat{y}_i - \bar{y})^2}{k-1}$	$\frac{MS_{Model}}{MS_{Errors}}$	(from R)
Error	$\Sigma_{i=1}^{n}\ (y_i - \hat{y}_i)^2$	$n - k$	$\frac{\Sigma_{i=1}^{n}\ (y_i - \hat{y}_i)^2}{n-k}$		
Total	$\Sigma_{i=1}^{n}\ (y_i - \bar{y})^2$	n - 1			

We have divided the total variaiblity in the model into two and we can now test their ratio

ANOVA Example

Source	SS	df	MS	F	p-value
Model	1786	3 -1 = 2	893.18	1.228	0.2993
Error	50200	72-3 = 69	727.53		
Total	51986	72-1 = 71			

```
> my_oats <- lm(yield ~ Variety, data = Oats)
> Anova(my_oats, type = 3)
Anova Table (Type III tests)

Response: yield
            Sum Sq Df  F value Pr(>F)
(Intercept) 262086  1 360.2407 <2e-16 ***
Variety       1786  2   1.2277 0.2993
Residuals    50200 69
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Conclusion

Same as always:

There is little to no evidence to suggest that the variety has an effect on the yield

(I generally write it in a broad form like the above because the interpretation of H_A is so messy with "...any linear combination of β's...")

ANOVA: Example 2

We have four experimental units per treatment in each block (with 6 blocks)

Problems with our old model?

ANOVA: Example with 2 Explanatory Var.s

We have four experimental units per treatment in each field (with 6 fields)

Problems with our old model?

The fields act as *blocks* which are exogenous, physically real, and uninteresting variables that (can) affect our response variable

Our current model...

- has a lurking variable (field) so violates the IID assumption
- Goes against the Kempthorne Principle
  - This principle isn't named but I'm making it a thing

ANOVA: Example with 2 Explanatory Var.s

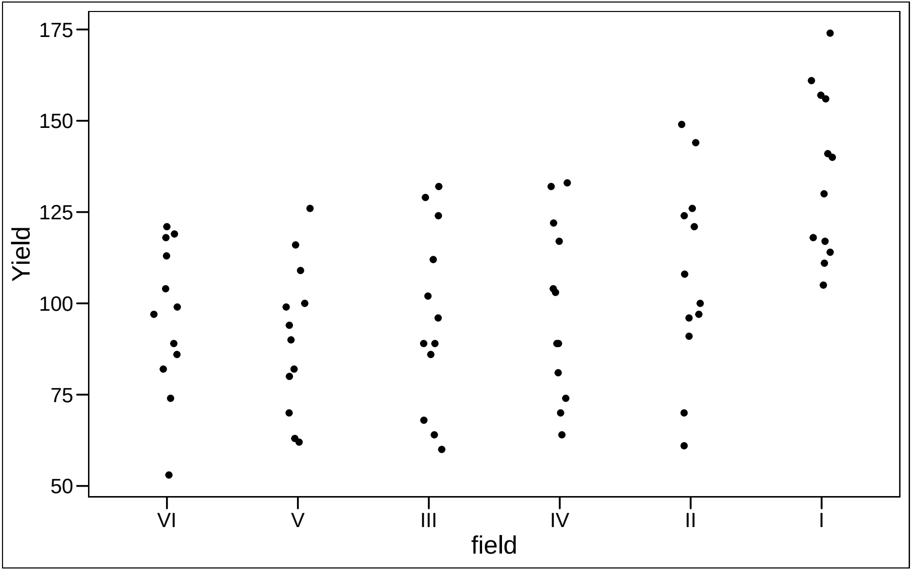
We are interested to see if using different varieties matters or not. Using the three varieties we have Marvelous as our baseline and both Golden Rain and Victory have indicator variables in our model is...

$$\widehat{Yield} = \beta_0 + \beta_1 \mathbb{I}_{Golden\ Rain} + \beta_2 \mathbb{I}_{Victory} + \beta_3 \mathbb{I}_{Field1} + \beta_4 \mathbb{I}_{Field2} \ldots + \beta_8 \mathbb{I}_{Field6}$$

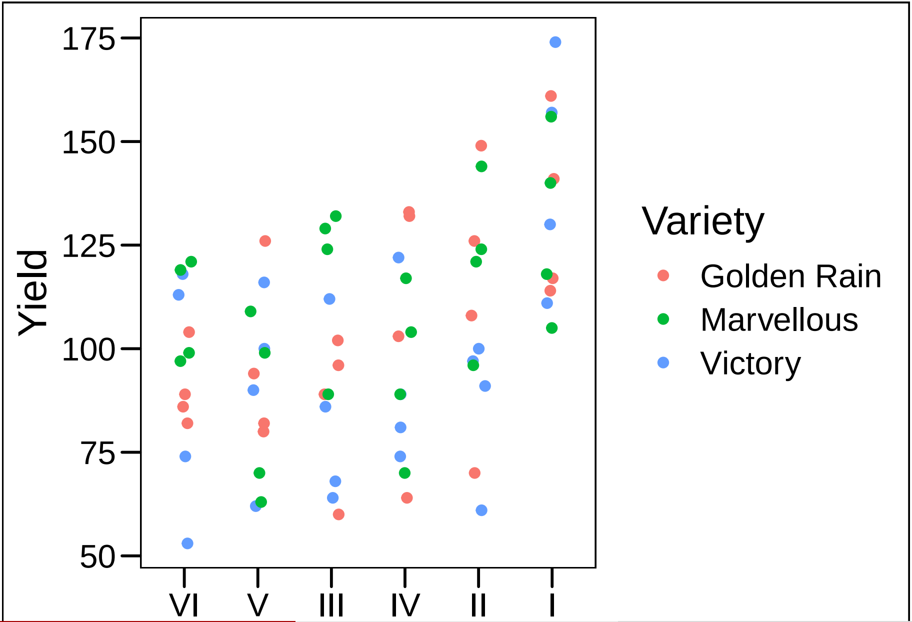
H_0: $\beta_1 = \beta_2 = 0$
H_A: Otherwise

So we are claiming the coefficients for variety are 0 and that the effect of variety is irrelevant after controlling for the effects of the fields!

Visualize Your Data

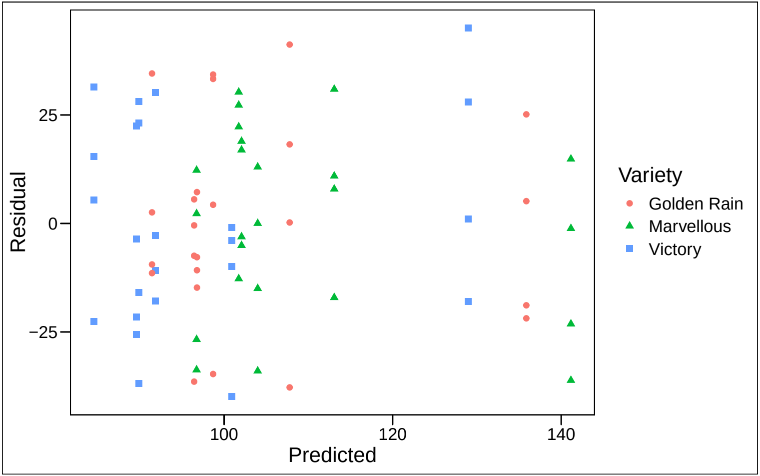


Visualize Your Data



Assumptions

Random: Yes, IID: Yes? Normal: Yes?



ANOVA: 2 Explanatory Vars

It turns out, as mentioned earlier, that SSM can be broken into parts, each associated with an explanatory variable.

Source	SS	df	MS	F	p-value
Variety	1786	3 -1 = 2	893.18	1.665	0.1972
Field	15875	6-1 = 5	3175	5.920	.000147
Error	34324	72 - 8 = 64	536.3125		
Total	51986	72-1 = 71			

```
> my_oats_2 <- lm(yield ~ Variety + field, data = Oats)
> Anova(my_oats_2, type = 3)
Anova Table (Type III tests)

Response: yield
            Sum Sq Df  F value    Pr(>F)
(Intercept) 262086  1 488.6772 < 2.2e-16 ***
Variety       1786  2   1.6654 0.1972060
field        15875  5   5.9201 0.0001472 ***
Residuals    34324 64
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Note on language

I'm blurring the (meaningless?) lines between some things with different names....

- One-way ANOVA: Has 1 categorical explanatory variable
  - Oldest form
  - Type 1 = Type 2 = Type 3 since there isn't multicolinearity
- Two-way ANOVA: Has 2 (or more) categorical explanatory variables
  - Generalizes and is more useful than one-way ANOVA
- ANCOVA: Has numeric explanatory variables
  - ANalysis of COVAriance
  - Distinction is meaningless to me

(There is also MANOVA if your response is a matrix and TANOVA if reponse is a (hyper-) cube)

What's up with "type"?

Don't have time to get into it but the idea is, due to multicolinear explanatory variables "explaining" some of the same information ANOVA Types 1, 2, and 3 dictate how the overlap is dealt with

- Type 1 works sequentially and should be avoided
  - How you write your lm() function will change your results!

- Type 3 removes all variability associated with the other variables before testing the variable of interest
  - Most conservative
  - Sum of Squares adds up to TSS (Type 1 and 2 don't!!!)
  - Probably the one you want

- Type II is in-between 1 and 3
  - Higher order interactions of the variable are removed from the model before the lower order variable is tested
  - Variables and interactions not associated with the variable being tested are left in the model

ANOVA Type 1

Using all three explanatory variables (nitrogen wasn't brought up) with all interactions we get...

```
> anova(moddy)
Analysis of Variance Table

Response: yield
                    Df  Sum Sq Mean Sq  F value    Pr(>F)    
field                5 15875.3  3175.1  18.4578 4.652e-09 ***
Variety              2  1786.4   893.2   5.1924  0.010440 *  
nitro                1 19536.4 19536.4 113.5727 1.104e-12 ***
field:Variety       10  6013.3   601.3   3.4958  0.002679 ** 
field:nitro          5   655.5   131.1   0.7621  0.583084    
Variety:nitro        2   168.3    84.2   0.4893  0.617048    
field:Variety:nitro 10  1758.2   175.8   1.0221  0.444825    
Residuals           36  6192.6   172.0                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA Type 2

Using all three explanatory variables (nitrogen wasn't brought up) with all interactions we get...

```
> Anova(moddy, type = 2)
Anova Table (Type II tests)

Response: yield
                     Sum Sq Df  F value    Pr(>F)    
field               15875.3  5  18.4578 4.652e-09 ***
Variety              1786.4  2   5.1924  0.010440 *  
nitro               19536.4  1 113.5727 1.104e-12 ***
field:Variety        6013.3 10   3.4958  0.002679 ** 
field:nitro           655.5  5   0.7621  0.583084    
Variety:nitro         168.3  2   0.4893  0.617048    
field:Variety:nitro  1758.2 10   1.0221  0.444825    
Residuals            6192.6 36                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

ANOVA Type 3

Using all three explanatory variables (nitrogen wasn't brought up) with all interactions we get...

```
Anova Table (Type III tests)

Response: yield
                     Sum Sq Df  F value    Pr(>F)
(Intercept)           57494  1 334.2339 < 2.2e-16 ***
field                  2036  5   2.3667   0.05898 .
Variety                1256  2   3.6498   0.03604 *
nitro                  6810  1  39.5900 2.834e-07 ***
field:Variety          1805 10   1.0493   0.42453
field:nitro            1630  5   1.8955   0.11940
Variety:nitro           168  2   0.4893   0.61705
field:Variety:nitro    1758 10   1.0221   0.44482
Residuals              6193 36
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Where does ANOVA fit in?

First we have to talk about the downside of the ANOVA:

Any guesses on what's annoying with it?

Where does ANOVA fit in?

First we have to talk about the downside of the ANOVA:

Any guesses on what's annoying with it?

If our p-value is small we have evidence that something is different from something.....and that's not super useful by itself

Where does ANOVA fit in?

Often, but not always, ANOVA is a fist pass

- It let's you know which parameters look important
- And saves on the multiple comparisons problem instead of a lot of t-tests

Once the sig. variables have been identified then..

- t-tests for specific parameters
- confidence intervals around our means

Next Time

We will work through some examples of ANOVA's

Extend ANOVA to multiple explanatory variables and continuous variables as well

Briefly touch on expected counts in tables maybe