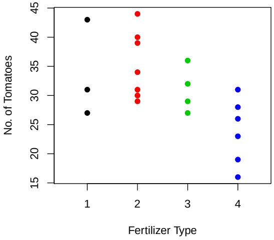
Lecture 19: ANOVA Part I

Chapter 5.5

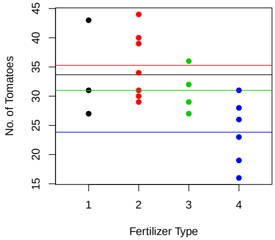
Analysis of Variance (ANOVA)

A farmer has the choice of four tomato fertilizers and wants to compare their performance in terms of crop yield.



Analysis of Variance (ANOVA)

A farmer has the choice of four tomato fertilizers and wants to compare their performance in terms of crop yield.



Analysis of Variance (ANOVA)

We have $k = 4$ groups AKA levels of a factor: the 4 types of fertilizer.

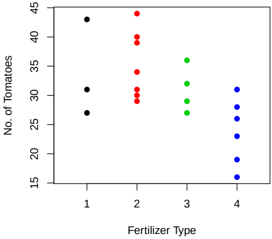
- n_i plants assigned to each of the $k = 4$ fertilizers:

n_1	n_2	n_3	n_4	total n
3	7	4	6	20

- Count the number of tomatoes on each plant

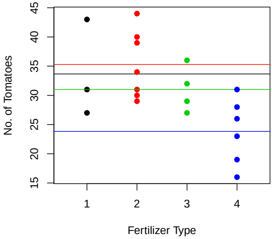
Tomato Fertilizer

We observe the following, where each point is one tomato plant.



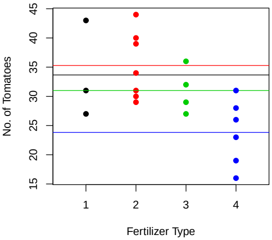
Tomato Fertilizer

We observe the following, where each point is one tomato plant.
Plot the sample mean of each level.



Tomato Fertilizer

We observe the following, where each point is one tomato plant. Plot the sample mean of each level. Question: are the mean tomato yields different?



Analysis of Variance

Say we have k groups and want to compare the k means:

$$\mu_1, \mu_2, \ldots, \mu_k$$

We could do $\binom{k}{2}$ individual two-sample tests.

Ex. for groups 1 & 2:

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 \\ \text{vs. } H_a : &\quad \mu_1 \neq \mu_2 \end{aligned}$$

Analysis of Variance

Or we do a single overall test via Analysis of Variance ANOVA:

The hypothesis test is:

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 = \ldots = \mu_k \\ \text{vs. } H_a : &\quad \text{at least one of the } \mu_i\text{'s are different} \end{aligned}$$

Analysis of Variance

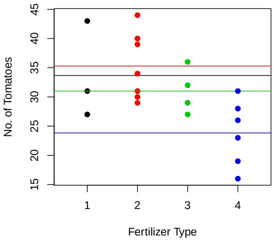
ANOVA asks: where is the overall variability of the observations originate from?

The test statistic used to compute a p-value is now the F-statistic:

$$F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$$

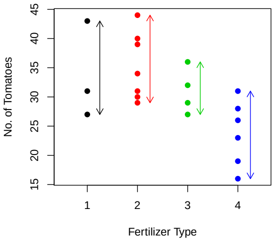
Tomato Fertilizer Example

Numerator: the between-group variation refers to the variability between the levels (the 4 horizontal lines):



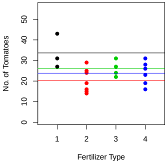
Tomato Fertilizer Example

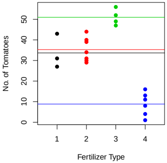
Denominator: the within-group variation refers to the variability within each level (the 4 vertical arrows):



Tomato Fertilizer Example

Now compare the following two plots. Which has "more different" means?





Tomato Fertilizer Example

- They have the same within-group variability. Call this value W
- The right plot has higher between group variability b/c the 4 means are more different. Call these values B_{left} and B_{right} with $B_{left} < B_{right}$
- Recall $F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$
- Since $\frac{B_{left}}{W} < \frac{B_{right}}{W}$, thus $F_{left} < F_{right}$ The right plot as a larger F-statistic

F Distributions

Assuming H_0 is true (that $\mu_1 = \mu_2 = \ldots = \mu_k$), the F-statistic

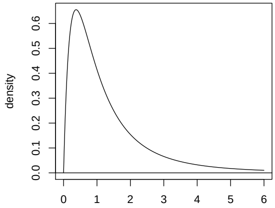
$$F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$$

follows the F distribution with $df_1 = k - 1$ and $df_2 = n - k$ degrees of freedom where

- n = total number of observations
- k = number of groups

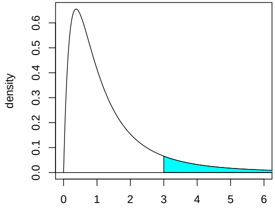
F Distributions

For $df_1 = 4$ and $df_2 = 6$, the F distribution looks like:



F Distributions

p-values are computed where "more extreme" means larger. Say the $F = 3$, the p-value is the area to the right of 3 and is computed in R: `pf(3,df1=4,df2=6,lower.tail=FALSE)`



Conducting An F-Test

The results are typically summarized in an ANOVA table:

Source of Variation	df	SS	MS	F	p-value
Between groups	$k-1$	$SSTr$	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$	p
Within groups	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

Conditions

1. The observations have to be independent. 10% rule.
2. Trade off of n and normality of observations within each group.
3. Each of the groups has constant variance $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$. Check via:
   - boxplots
   - comparing the sample standard deviations $s_1, \ldots, s_k$