Lecture 19: ANOVA Part I

Chapter 5.5

Previously... Conditions/Assumption for Using t Dist'n

The key situation to use the t distribution is when you have a small sample.

- Independence of observations: To ensure this, either
    - collect a simple random sample that is less than 10% of the population
    - or if it was an experiment or random process check that each observation was independent
- Observations come from a nearly normal distribution: This second condition is difficult to verify with small data sets:
    - take a look at a plot of the data for obvious departures from the normal model
    - consider whether any previous experiences alert us that the data may not be nearly normal

Previously... Confidence Intervals and *t*-Test

We have the same two methods for inference as in Chapter 4, but:

1. Confidence intervals: Now we use t^*_{df} instead of z^*

$$[\overline{x} - t^*_{df} SE,\ \overline{x} + t^*_{df} SE] = \left[\overline{x} - t^*_{df} \times \frac{s}{\sqrt{n}},\ \overline{x} + t^*_{df} \times \frac{s}{\sqrt{n}}\right]$$

2. Hypothesis testing: Now we use the t-test

$$t = \frac{\overline{x} - \text{null value}}{SE}$$

   using the t-table on page 410 (instead of z-table) for $df = n - 1$

New Topic: Analysis of Variance (ANOVA)

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

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

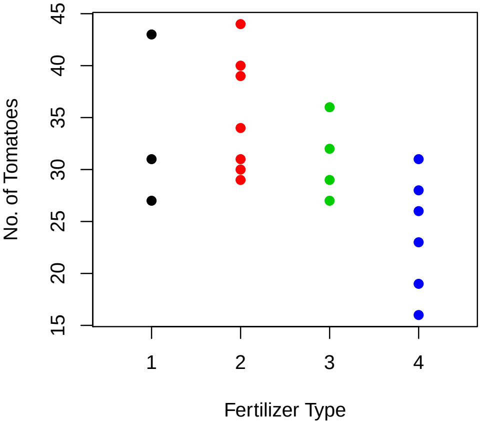
- assign n_i plants to each of the $k = 4$ fertilizers as follows:

n_1	n_2	n_3	n_4	total n
3	7	4	6	20

- we evaluate the number of tomatoes on each plant

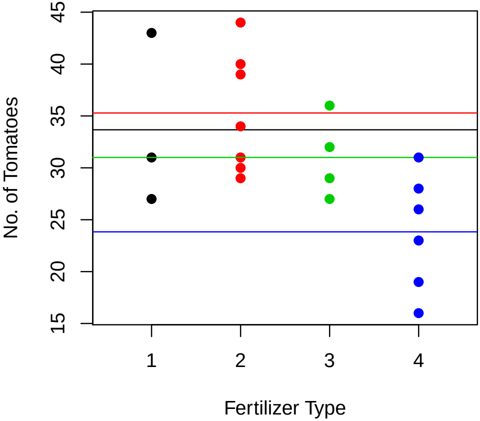
Tomato Fertilizer

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



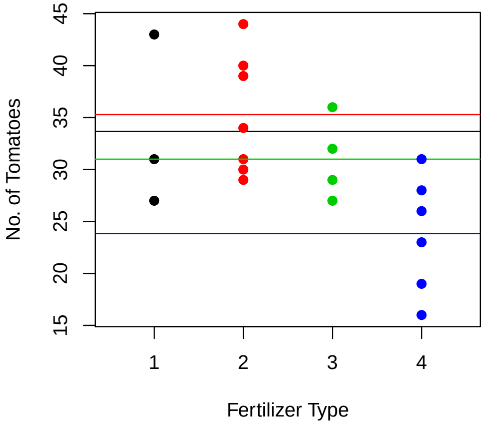
Tomato Fertilizer

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



Tomato Fertilizer

We observe the following data, 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:

$$\mu_1 \text{ and } \mu_2 \text{ and } \dots \text{ and } \mu_k$$

We could do $\binom{k}{2}$ individual two-sample tests. Ex. for groups 1 & 2:

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

i.e. no difference in means

Analysis of Variance

Or, rather than conducting all $\binom{k}{2}$ tests, we do a single overall test via Analysis of Variance ANOVA:

The hypothesis test is:

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

How ANOVA Tests Work

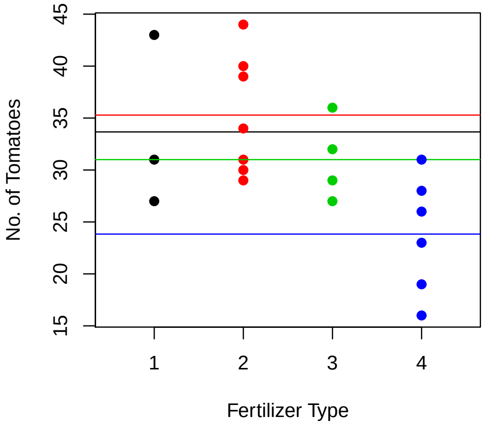
ANOVA asks: where is the overall variability of the data originating from?

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

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

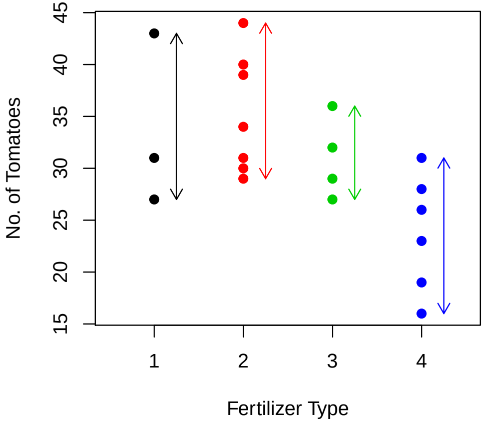
Tomato Fertilizer Example

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



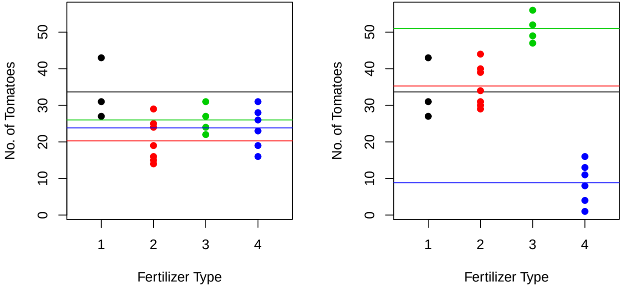
Tomato Fertilizer Example

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



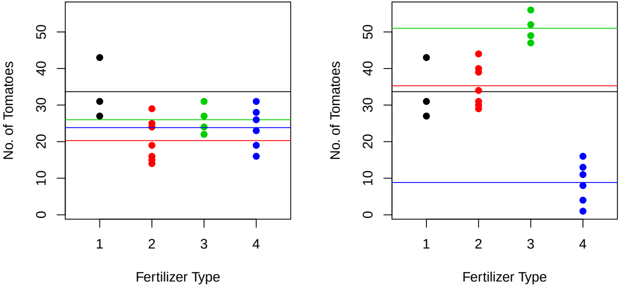
Tomato Fertilizer Example

Now compare the following two plots:



- 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 $B_{left} < B_{right}$

Tomato Fertilizer Example



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

$$\begin{array}{lccc} \text{Since} & \dfrac{B_{left}}{W} & < & \dfrac{B_{right}}{W} \\ \text{thus} & F_{left} & < & F_{right} \end{array}$$

F Distributions

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

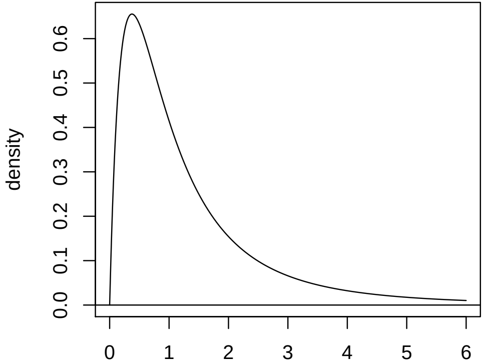
$$F = \frac{\text{measure of between-group variability}}{\text{measure of within-group variability}}$$

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

- n is the total number of observations
- k is the number of groups

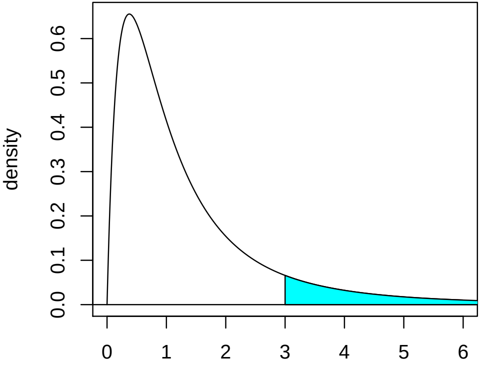
F Distributions

Much like the t distribution with degrees of freedom df, the parameters for the F distribution are $df_1 = k - 1$ and $df_2 = n - k$. Example with $df_1 = 4$ and $df_2 = 6$:



F Distributions

p-values are computed as before where “more extreme” means larger. You can compute these in R using `pf(F,df1,df2)`. Say the *F*-statistic is equal to 3, the *p*-value is the area to the right of 3.



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. If the sample sizes are small within each group, normality of the data is important. If not small, we can be lax about this.
3. Each of the groups has constant variance $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$. We can check this with boxplots and by comparing the sample standard deviations $s_1, \ldots, s_k$