References:

- Chapter 6 of Saville and Wood, *Statistical Methods: A Geometric Approach*.
- Appendix A of Saville and Wood, *Statistical Methods: A Geometric Approach*.

Questions about Two Populations

How much does fertilizer affect yield? Last time we encoded the two treatment groups via dummy variables:

Treatment	Value	x_1	x_2
1	$y_{11} = 29.80$	1	0
1	$y_{12} = 28.57$	1	0
1	$y_{13} = 29.97$	1	0
2	$y_{21} = 30.33$	0	1
2	$y_{22} = 31.27$	0	1
2	$y_{23} = 30.37$	0	1

We also wrote our model in a few ways, like:

$$y = \begin{cases} \mu_1 & \text{if } y \text{ is from treatment } 1 \\ \mu_2 & \text{if } y \text{ is from treatment } 2 \end{cases}$$

Or via dummy variables. Either way, μ_i is estimated as the sample mean of each group:

$$\hat{\mu}_i = \bar{y}_{i\bullet} = \frac{y_{i1} + y_{i2} + y_{i3}}{3}$$

Fitted model:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{bmatrix}$$

The space of all possible observation is 6-dimensional space so $\mathrm{df}_{\mathrm{Total}} = 6$.

The model space is a 2-dimensional subspace with free parameters μ_1 and μ_2 so $\mathrm{df}_{\mathrm{Model}} = 2$.

This leaves the error space with dimension $\mathrm{df}_{\mathrm{Error}} = 6 - 2 = 4$.

Again, the model vector $(\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2)^T$ is chosen to be the closest point in the model space to the observation vector $(y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23})^T$, so the error vector is orthogonal to the model vector.

Fitted model (with actual values):

$$\underbrace{\begin{bmatrix} 29.80 \\ 28.57 \\ 29.97 \\ 30.33 \\ 31.27 \\ 30.37 \end{bmatrix}}_{\substack{\text{Observation} \\ \text{vector}}} = \underbrace{\begin{bmatrix} 29.447 \\ 29.447 \\ 29.447 \\ 30.657 \\ 30.657 \\ 30.657 \end{bmatrix}}_{\substack{\text{Fitted model} \\ \text{vector}}} + \underbrace{\begin{bmatrix} .353 \\ -.877 \\ .523 \\ -.327 \\ .613 \\ -.287 \end{bmatrix}}_{\substack{\text{Fitted error} \\ \text{vector}}}$$

Hypothesis testing (preliminary):
Two orthogonal directions that span our model space $(\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2)^T$ are:

- the vector $(1, 1, 1, 0, 0, 0)^T$ which tests the hypothesis $H_0 : \mu_1 = 0$, and
- the vector $(0, 0, 0, 1, 1, 1)^T$ which tests the hypothesis $H_0 : \mu_2 = 0$.

Hypothesis testing:
More often, we are interested in the hypothesis $H_0 : \mu_1 = \mu_2$ (comparison of treatments) or equivalently $\mu_1 - \mu_2 = 0$ which corresponds to the vector $(1, 1, 1, -1, -1, -1)^T$.

An orthogonal vector to this which still lies in our model space is $(1, 1, 1, 1, 1, 1)^T$ which corresponds to the hypothesis $H_0 : \mu = 0$. Here μ is the **grand mean** or **overall mean**. It is estimated by the mean $\bar{y}_{\bullet,\bullet}$ taken over all observations.

ANOVA table:
A more refined orthogonal decomposition of our observation vector is therefore:

$$y = \bar{y}_{..} + (\bar{y}_{i\cdot} - \bar{y}_{..}) + (y - \bar{y}_{i\cdot})$$

Or with actual values:

$$\underbrace{\begin{bmatrix} 29.80 \\ 28.57 \\ 29.97 \\ 30.33 \\ 31.27 \\ 30.37 \end{bmatrix}}_{\substack{\text{Observation} \\ \text{vector}}} = \underbrace{\underbrace{\begin{bmatrix} 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \\ 30.052 \end{bmatrix}}_{\substack{\text{Overall mean} \\ \text{vector}}} + \underbrace{\begin{bmatrix} -.605 \\ -.605 \\ -.605 \\ .605 \\ .605 \\ .605 \end{bmatrix}}_{\substack{\text{Treatment} \\ \text{vector}}}}_{\text{Model}} + \underbrace{\begin{bmatrix} .353 \\ -.877 \\ .523 \\ -.327 \\ .613 \\ -.287 \end{bmatrix}}_{\substack{\text{Fitted error} \\ \text{vector}}}$$

The degrees of freedom are shown in the following ANOVA table:

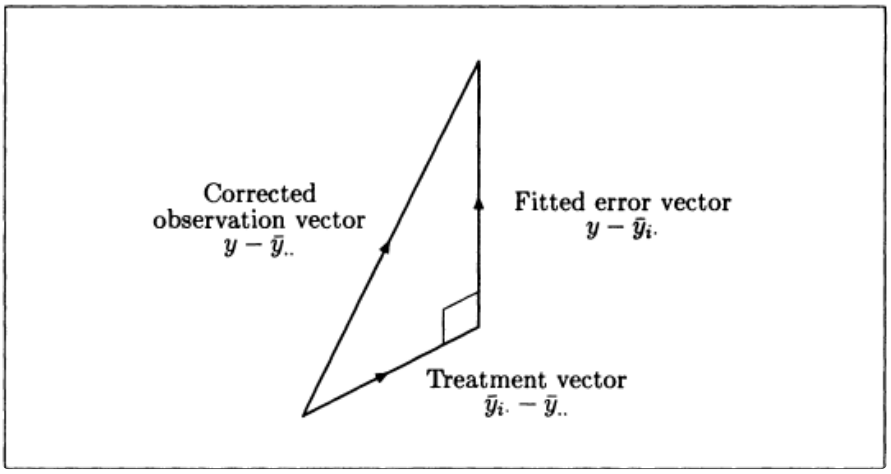
Source of Variation	df	SS	MS	F
Overall mean ($H_0 : \mu = 0$)	1	5418.6160	5418.6160	
Treatments ($H_0 : \mu_1 = \mu_2$)	1	2.1962	2.1962	5.07 (ns)
Error	4	1.7324	0.4331	
Total	6	5422.5445		

Table 6.2: ANOVA table for our two treatment bulkometer example.

Simplified ANOVA table:
Since we mostly care about the comparison $\mu_1 - \mu_2$ and not the relative values of μ_1 and μ_2 to the grand mean μ, we often simplify the model by subtracting over the overall mean to the left side:

$$\begin{matrix} & & \overset{\substack{\text{variability} \\ \text{within} \\ \text{groups}}}{} & \overset{\substack{\text{variability} \\ \text{between} \\ \text{groups}}}{} \\ y - \bar{y}_{..} & = & (\bar{y}_{i\cdot} - \bar{y}_{..}) & + & (y - \bar{y}_{i\cdot}) \end{matrix}$$

$$\begin{bmatrix} -.252 \\ -1.482 \\ -.082 \\ .278 \\ 1.218 \\ .318 \end{bmatrix} = \begin{bmatrix} -.605 \\ -.605 \\ -.605 \\ .605 \\ .605 \\ .605 \end{bmatrix} + \begin{bmatrix} .353 \\ -.877 \\ .523 \\ -.327 \\ .613 \\ -.287 \end{bmatrix}$$



This simplified form is what is more commonly shown in ANOVA outputs:

Source of Variation	df	SS	MS	F
Treatments ($H_0 : \mu_1 = \mu_2$)	1	2.1962	2.1962	5.07 (ns)
Error	4	1.7324	.4331	
Total (corrected)	5	3.9286		

Table 6.3: The more usual form of the ANOVA table for the two treatment bulkometer example.

General case:
In general, suppose you have two groups of data:

- Group 1 with n_1-many values $y_{11}, y_{12}, \ldots, y_{1n_1}$;
- Group 2 with n_2-many values $y_{21}, y_{22}, \ldots, y_{2n_2}$;

whose means μ_1 and μ_2 you wish to compare. The general (simplified) ANOVA table is as follows:

Source of Variation	**df**	**SS**	**MS**	**F**
Treatments	1	$\|\bar{y}_{i\cdot} - \bar{y}_{..}\|^2$	$\|\bar{y}_{i\cdot} - \bar{y}_{..}\|^2$	$\dfrac{\|\bar{y}_{i\cdot} - \bar{y}_{..}\|^2}{\|y - \bar{y}_{i\cdot}\|^2/(n_1+n_2-2)}$
Error	$n_1 + n_2 - 2$	$\|y - \bar{y}_{i\cdot}\|^2$	$\dfrac{\|y - \bar{y}_{i\cdot}\|^2}{n_1+n_2-2}$	
Total (corrected)	$n_1 + n_2 - 1$	$\|y - \bar{y}_{..}\|^2$		

Assumptions:
This model assumes each group has equal variance. If this assumption fails, then use a more general form of ANOVA called Welch's ANOVA (like Welch's t-test).