ANOVA: One population

Model one population by one mean $y = \mu$. ANOVA by hand:

$$\begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} = \begin{bmatrix} 11 \\ 11 \\ 11 \\ 11 \end{bmatrix} + \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix} \rightarrow \begin{bmatrix} 8^2 \\ + \\ 10^2 \\ + \\ 12^2 \\ + \\ 14^2 \end{bmatrix} = \begin{bmatrix} 11^2 \\ + \\ 11^2 \\ + \\ 11^2 \\ + \\ 11^2 \end{bmatrix} + \begin{bmatrix} (-3)^2 \\ + \\ (-1)^2 \\ + \\ 1^2 \\ + \\ 3^2 \end{bmatrix}$$

sample: df = 4 (4 samples) | mean: df = 1 (1 parameter) | error: df = 4 - 1 = 3

Sum of Squares Total (SST) = 504 | Sum of Squares Model (SSM) = 484 | Sum of Squares Error (SSE) = 20

ANOVA by SAS:

Input data.

```
data sample;
    input x $ dummy y;
    datalines;
M  0 8
M  0 10
F  1 12
F  1 14
;
run;
```

```
proc reg data=sample;
    model y = ;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	3	20.00000	6.66667		
Corrected Total	3	20.00000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.00000	1.29099	8.52	0.0034

```
proc glm data=sample;
    model y = / intercept;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	484.0000000	484.0000000	72.60	0.0034
Error	3	20.0000000	6.6666667		
Uncorrected Total	4	504.0000000			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.00000000	1.29099445	8.52	0.0034

```
/* Test H0: mu = 10 */
proc ttest data=sample h0=10;
    var y;
run;
```

DF	t Value	Pr > |t|
3	0.77	0.4950

PROC REG and GLM intercept tests $\mu = 0$. So transform y to y-10 or use PROC TTEST to test $\mu = 10$.

ANOVA: Two populations

Model two population by two means $y = \mu_1$, $y = \mu_2$. Get ANOVA decomposition:

1. Break into subgroups:
2. Further break into grand mean:
3. Subtract grand mean vector:

$$\begin{matrix} \text{Pop 1} \\ \\ \text{Pop 2} \end{matrix} \begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} = \begin{bmatrix} 9 \\ 9 \\ 13 \\ 13 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} = \begin{bmatrix} 11 \\ 11 \\ 11 \\ 11 \end{bmatrix} + \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$$

sample: df = 4 (4 samples) | subgroup means: df = 2 (2 parameters) | error: df = 3 - 1 = 2

grand mean: df = 1 | treatment: df = 1 | error: df = 2

corrected: df = 4-1 | treatment: df = 1 | error: df = 2

4. Squared lengths:

$$\begin{bmatrix} (-3)^2 \\ + \\ (-1)^2 \\ + \\ 1^2 \\ + \\ 3^2 \end{bmatrix} = \begin{bmatrix} (-2)^2 \\ + \\ (-2)^2 \\ + \\ 2^2 \\ + \\ 2^2 \end{bmatrix} + \begin{bmatrix} (-1)^2 \\ + \\ 1^2 \\ + \\ (-1)^2 \\ + \\ 1^2 \end{bmatrix}$$

Corrected SST = 504 | Sum of Squares Treatment = 16 | SSE = 4

```
proc reg data=sample;
    model y = dummy;
run;
quit;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16.00000	16.00000	8.00	0.1056
Error	2	4.00000	2.00000		
Corrected Total	3	20.00000			

Root MSE	1.41421	R-Square	0.8000
Dependent Mean	11.00000	Adj R-Sq	0.7000

```
proc glm data=sample;
    class x;
    model y = x;
    contrast "M vs F" x 1 -1;
run;
quit;
```

(optional here)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16.00000000	16.00000000	8.00	0.1056
Error	2	4.00000000	2.00000000		
Corrected Total	3	20.00000000			

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
M vs F	1	16.00000000	16.00000000	8.00	0.1056

* PROC REG and GLM's F-value test $H_0 : \mu_1 = \mu_2$.

  We do not compare to the grand mean so we removed it in 3.

* (Informal) A contrast is a comparison of group means (even groups of groups).
  - Ex. the comparison $\mu_1 = \mu_2$ rewrites to $\mu_1 - \mu_2 = 0$ whose coefficients are +1, -1.
  - Testing "Group 1 = Group 2" requires a contrast's coefficients to sum to zero.

* (Formal) A vector is a **contrast vector** if its coefficients sum to zero.

  (I.e. a contrast vector is orthogonal to the all-ones vector representing grand mean.)

ANOVA: Several populations

if all groups have same sample size

With k groups, you can use up to k-1 orthogonal contrasts (= k-1 statistically independent questions) comparing various group means.

Q1: Do the legumes (peas, lupins) differ from the nonlegume?

$$H_0: \frac{\mu_1+\mu_3}{2} = \mu_2 \quad \Rightarrow \quad \mu_1 - 2\mu_2 + \mu_3 = 0 \quad \Rightarrow \quad c_1 = (1,-2,1)$$

Q2: Do the lupins differ from the peas?

$$H_0: \mu_1 = \mu_3 \quad \Rightarrow \quad \mu_1 - \mu_3 = 0 \quad \Rightarrow \quad c_2 = (1,0,-1)$$

Aside: Are these contrasts orthogonal?

Answer. Take dot products: $c_1^T c_2 = [1,-2,1]\begin{bmatrix}1\\0\\-1\end{bmatrix} = 0$ so yes.

```
data CropYields;
    input Treatment $ Yield;
    datalines;
Lupins 3.4
Lupins 3.8
Lupins 3.3
Mustard 4.4
Mustard 4.0
Peas 3.0
Peas 3.6
;
run;
```

μ_1 μ_2 μ_3

Note: SAS sorts CLASS levels alphabetically: Lupins, Mustard, Peas. Order contrast coefficients likewise.

```
proc glm data=CropYields;
    class Treatment;
    model Yield = Treatment;
    contrast 'Legumes vs Non-legume' Treatment 1 -2 1;
    contrast 'Lupins vs Peas' Treatment 1 0 -1;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.91714286	0.45857143	4.59	0.0922
Error	4	0.40000000	0.10000000		
Corrected Total	6	1.31714286			

R-Square	Coeff Var	Root MSE	Yield Mean
0.696312	8.680762	0.316228	3.642857

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Legumes vs Non-legume	1	0.90352941	0.90352941	9.04	0.0397
Lupins vs Peas	1	0.04800000	0.04800000	0.48	0.5265

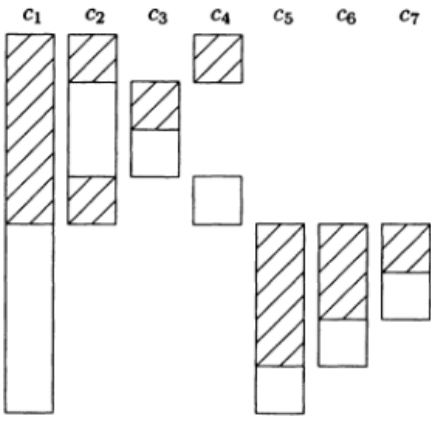
Interpret results:
(First check each pop. is normal with approx. same variance).

There is convincing evidence (at the 5% significance level) that avg legume yield and avg non-legume yield are different.

There is not convincing evidence (at α=5%) that avg lupin yield and avg pea yield are different.

General way to get complete set of k-1 orthogonal class comparisons:
Split and compare the set of treatments into two classes. Iterate.

Treatment		c_1	c_2	c_3	c_4	c_5	c_6	c_7
Brigadier mangels	μ_1	1	1	0	1	0	0	0
York globe mangels	μ_2	1	−1	1	0	0	0	0
Orange globe mangels	μ_3	1	−1	−1	0	0	0	0
Red intermediate mangels	μ_4	1	1	0	−1	0	0	0
Mono rosa fodder beet	μ_5	−1	0	0	0	1	1	1
Mono blanc fodder beet	μ_6	−1	0	0	0	1	1	−1
Mono bomba fodder beet	μ_7	−1	0	0	0	1	−2	0
Yellow daeno fodder beet	μ_8	−1	0	0	0	−3	0	0



Ex. Construct a complete set of orthogonal contrasts:

		c_1	c_2	c_3	c_4
1. Control	(No herbicide)	−4	0	0	0
2. Systemic herbicide A	(Herbicide)	1	1	1	0
3. Systemic herbicide B	(Herbicide)	1	1	−1	0
4. Contact herbicide X	(Herbicide)	1	−1	0	1
5. Contact herbicide Y	(Herbicide)	1	−1	0	−1