18 - ANOVA

Today we redo ANOVA lesson with more hands-on practice and then proceed further:

1. First we review one-population ANOVA. This subsumes Lesson 11's one-population t-test in introductory statistics class. Our reference is Chapter 5 of *Geometric Approach*.
2. Next we review two-population ANOVA. This subsumes Lesson 12a's two-population t-test. Our reference is Chapter 6 of *Geometric Approach*.
3. Finally we treat the case of ANOVA for several populations with the new concept of "contrasts". Our reference is Chapter 8 of *Geometric Approach*.

1. ANOVA: One Population

When analyzing a single population, we model the data by a single number, the mean:

$$y = \mu$$

Given a sample of four observations (8, 10, 12, 14), we can decompose this data vector into a mean vector (the overall mean is 11) and an error vector:

$$\begin{matrix} \text{sample} \\ \begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} \end{matrix} = \begin{matrix} \text{mean} \\ \begin{bmatrix} 11 \\ 11 \\ 11 \\ 11 \end{bmatrix} \end{matrix} + \begin{matrix} \text{error} \\ \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix} \end{matrix}$$

Computing the squared lengths of each vector gives the Sum of Squares decomposition:

$$\begin{matrix} \text{Total (SST)} \\ \begin{bmatrix} 8^2 \\ + \\ 10^2 \\ + \\ 12^2 \\ + \\ 14^2 \end{bmatrix} \\ 504 \end{matrix} = \begin{matrix} \text{Model (SSM)} \\ \begin{bmatrix} 11^2 \\ + \\ 11^2 \\ + \\ 11^2 \\ + \\ 11^2 \end{bmatrix} \\ 484 \end{matrix} + \begin{matrix} \text{Error (SSE)} \\ \begin{bmatrix} (-3)^2 \\ + \\ (-1)^2 \\ + \\ 1^2 \\ + \\ 3^2 \end{bmatrix} \\ 20 \end{matrix}$$

Note that this particular SSE is just the variance of our data vector.

SAS Implementation

In SAS, the ANOVA table results by calling `PROC REG` or `PROC GLM`.

```
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;
```

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

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	4
Number of Observations Used	4

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			

Root MSE	2.58199	R-Square	0.0000
Dependent Mean	11.00000	Adj R-Sq	0.0000
Coeff Var	23.47263		

Parameter Estimates

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

The GLM Procedure

Dependent Variable: y

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			

R-Square	Coeff Var	Root MSE	y Mean
0.000000	23.47263	2.581989	11.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Intercept	1	484.0000000	484.0000000	72.60	0.0034

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Intercept	1	484.0000000	484.0000000	72.60	0.0034

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

By default, procedures like `PROC REG` and `PROC GLM` test the null hypothesis $H_0 : \mu = 0$. If you want to test against a specific non-zero value like $H_0 : \mu = 10$, you must either transform the response variable, say, from y to $y - 10$, or use `PROC TTEST`:

```
proc ttest data=sample h0=10;
    var y;
run;
```

2. ANOVA: Two Populations

Suppose our population above actually come from two subpopulations: 8 and 10 are the number of shoes owned by two men, while 12 and 13 were the number of shoes owned by two women. We can model two populations with two means:

$$y = \begin{cases} \mu_1 & \text{if } y \text{ is from M} \\ \mu_2 & \text{if } y \text{ is from F} \end{cases}$$

In this case, we can be more nuanced with how we decompose our data vector.

First we model the data vector by their subgroup means:

$$\begin{matrix} \text{sample} \\ \begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} \end{matrix} = \begin{matrix} \text{subgroup means} \\ \begin{bmatrix} 9 \\ 9 \\ 13 \\ 13 \end{bmatrix} \end{matrix} + \begin{matrix} \text{error} \\ \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \end{matrix}$$

Second, we further average the subgroup means into **grand mean**. We call the remaining part the **treatment vector**:

$$\begin{matrix} \text{subgroup means} \\ \begin{bmatrix} 9 \\ 9 \\ 13 \\ 13 \end{bmatrix} \end{matrix} = \begin{matrix} \text{grand mean} \\ \begin{bmatrix} 11 \\ 11 \\ 11 \\ 11 \end{bmatrix} \end{matrix} + \begin{matrix} \text{treatment} \\ \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix} \end{matrix}$$

Substituting this back into the previous step gives the decomposition:

$$\begin{matrix} \text{sample} \\ \begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} \end{matrix} = \begin{matrix} \text{grand mean} \\ \begin{bmatrix} 11 \\ 11 \\ 11 \\ 11 \end{bmatrix} \end{matrix} + \begin{matrix} \text{treatment} \\ \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix} \end{matrix} + \begin{matrix} \text{error} \\ \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \end{matrix}$$

In two population questions, we try to detect whether $\mu_1 = \mu_2$ or not. We don't care about how μ_1 and μ_2 compare to the grand mean μ.

So the third thing we do is to subtract the grand mean vector from both sides:

$$\begin{bmatrix} 8 \\ 10 \\ 12 \\ 14 \end{bmatrix} - \begin{bmatrix} 11 \\ 11 \\ 11 \\ 11 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \implies \begin{bmatrix} -3 \\ -1 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$$

This simplified decomposition is what most statistical software use as the starting point. The squared lengths of these vectors are the Corrected Sum of Squares:

$$\begin{matrix} \text{Corrected Total} \\ \begin{bmatrix} (-3)^2 \\ + \\ (-1)^2 \\ + \\ 1^2 \\ + \\ 3^2 \end{bmatrix} \\ = 20 \end{matrix} = \begin{matrix} \text{Treatment} \\ \begin{bmatrix} (-2)^2 \\ + \\ (-2)^2 \\ + \\ 2^2 \\ + \\ 2^2 \end{bmatrix} \\ = 16 \end{matrix} + \begin{matrix} \text{Error} \\ \begin{bmatrix} (-1)^2 \\ + \\ 1^2 \\ + \\ (-1)^2 \\ + \\ 1^2 \end{bmatrix} \\ = 4 \end{matrix}$$

SAS Implementation

In SAS, the ANOVA table results by calling `PROC REG` or `PROC GLM`. `PROC REG` cannot accept character strings and so we must use our dummy-variable encoding of (F, M) with it.

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

```
proc glm data=sample;
    class x;
    model y = x;
run;
```

Truncated output shared by both `PROC` :

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			

Introduction to Contrasts

To test hypotheses like $H_0 : \mu_1 = \mu_2$ in SAS, we use "contrasts". Informally, a contrast is an equational comparison of group means. Formally, a vector is a **contrast vector** if its coefficients sum to zero: this ensures the vector is strictly orthogonal to the all-ones vector (which corresponds to the grand mean).

For instance, $H_0 : \mu_1 = \mu_2$ can be rearranged into $\mu_1 - \mu_2 = 0$, the coefficients are $+1, -1$. So the null hypothesis $H_0 : \mu_1 = \mu_2$ corresponds to the contrast vector $c = (1, -1)$.

If you run this in SAS, you get an extra output stating the result of this hypothesis test:

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

SAS Output:

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

3. ANOVA: Several Populations

When you have k different groups, you can formulate up to $k - 1$ mutually orthogonal contrasts. When the sample sizes of each group are equal, then these orthogonal contrasts represent statistically independent questions about the data.

Consider an experiment comparing the yield of two legumes (Peas, Lupins) and one non-legume (Mustard). Since $k = 3$, we can ask $k - 1 = 2$ independent questions:

Question 1: Do the legumes (peas, lupins) differ from the nonlegume (mustard)?

Answer: We formulate this as the hypothesis $H_0 : \dfrac{\mu_1 + \mu_3}{2} = \mu_2$.

Rearranging the equation gives $\mu_1 - 2\mu_2 + \mu_3 = 0$. The coefficient or contrast vector is therefore $c_1 = (1, -2, 1)$.

Question 2: Do the lupins differ from the peas?

Answer: We formulate this as the hypothesis $H_0 : \mu_1 = \mu_3$.

Rearranging the equation gives $\mu_1 - \mu_3 = 0$ or the contrast $c_2 = (1, 0, -1)$.

Orthogonality Check:

The dot product of c_1 and c_2 is $(1)(1) + (-2)(0) + (1)(-1) = 0$ so c_1 and c_2 are orthogonal.

SAS Execution & Interpretation

SAS `CLASS` statements sort categorical levels alphabetically. For Lupins, Mustard, and Peas, the internal ordering

is 1, 2, 3. Your contrast coefficients in code must match this exact alphabetical order:

```
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;

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;
```

SAS output:

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

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Treatment	2	0.91714286	0.45857143	4.59	0.0922

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Treatment	2	0.91714286	0.45857143	4.59	0.0922

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

Based on this output, if we assume normality and equal variance of the data is checked, then the p-value for Legumes vs. Non-legume (0.0397) shows a statistically significant difference at the $\alpha = 0.05$ level. On the other hand, the p-value for Lupins vs. Peas (0.5265) shows no convincing evidence of a difference.

Note that if we had equal sample sizes in each population or group, the contrast sum of squares of a complete set of orthogonal contrasts would add to Sum of Squares Model by the Pythagorean Theorem.

4. Constructing a Complete Set of Orthogonal Contrasts

To systematically generate all $k-1$ orthogonal contrasts for a given experiment, use a recursive splitting strategy: continuously divide the set of treatments into two distinct subgroups until groups can no longer be split.

Example. Suppose the following are the $k = 5$ Herbicide treatments in our dataset:

1. Control (No herbicide)
2. Systemic Herbicides A (Herbicide)
3. Systemic Herbicides B (Herbicide)

4. Contact Herbicides X (Herbicide)
5. Contact Herbicides Y (Herbicide)

Here the construction of a complete set of orthogonal contrasts will have $k - 1 = 4$ elements. We can structure the splits as follows:

Comparison / Hypothesis	Contrast Vector
Control vs. All Herbicides	$\mathbf{c}_1 = (-4, 1, 1, 1, 1)$
Systemic vs. Contact	$\mathbf{c}_2 = (0, 1, 1, -1, -1)$
Systemic A vs. Systemic B	$\mathbf{c}_3 = (0, 1, -1, 0, 0)$
Contact X vs. Contact Y	$\mathbf{c}_4 = (0, 0, 0, 1, -1)$

We can visualize the splits also as tree diagram:

