Lecture 6 – Analysis of Variance

Analysis of variance is used to test the research hypothesis that at least one or more than two means are different from each other. This translates into testing the null hypothesis: $\mu_1 = \mu_2 = \mu_3$. If any of the means are different from any of the others, then this null hypothesis is not true. Another way of stating this null hypothesis is to say that there is no variance among the means of the different groups against the overall mean of all groups. In essence, it is an analysis of the two variances.

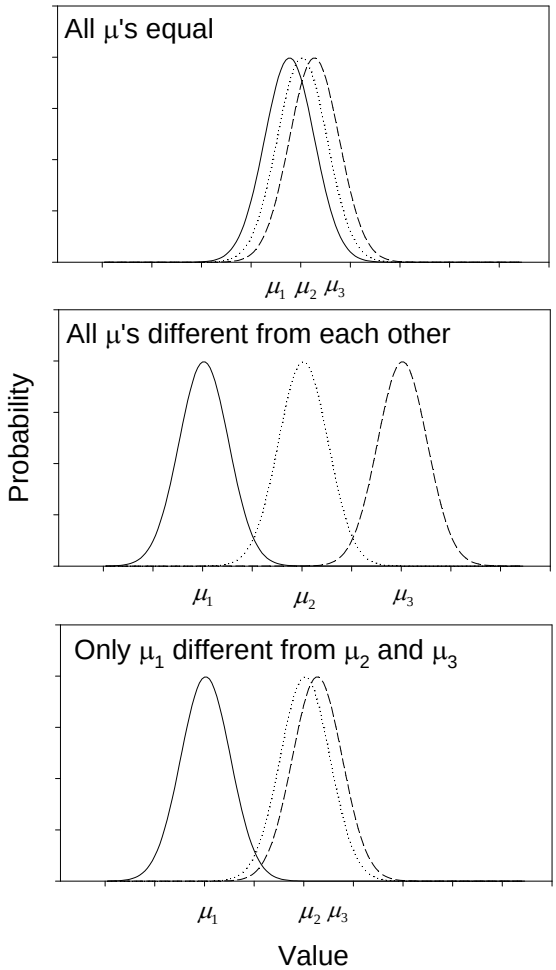

Figure 6.1. The above figure shows three population means (μ_1, μ_2, μ_3) and shows how either they are all different, all similar or at least one is different.

Our problem, as with other tests of inference, is that we don't know what the true means (parameters) are as we only have estimates from samples in the form of statistics. What we do know is that variance among sample means could be because of some effect together with sampling error.

ANOVA therefore tests whether more variance between the sample's means (from some sort of effect) or among the samples' means (from chance sampling error alone).

The result is a test to compare two variances – the between-groups variance and the within-group variance.

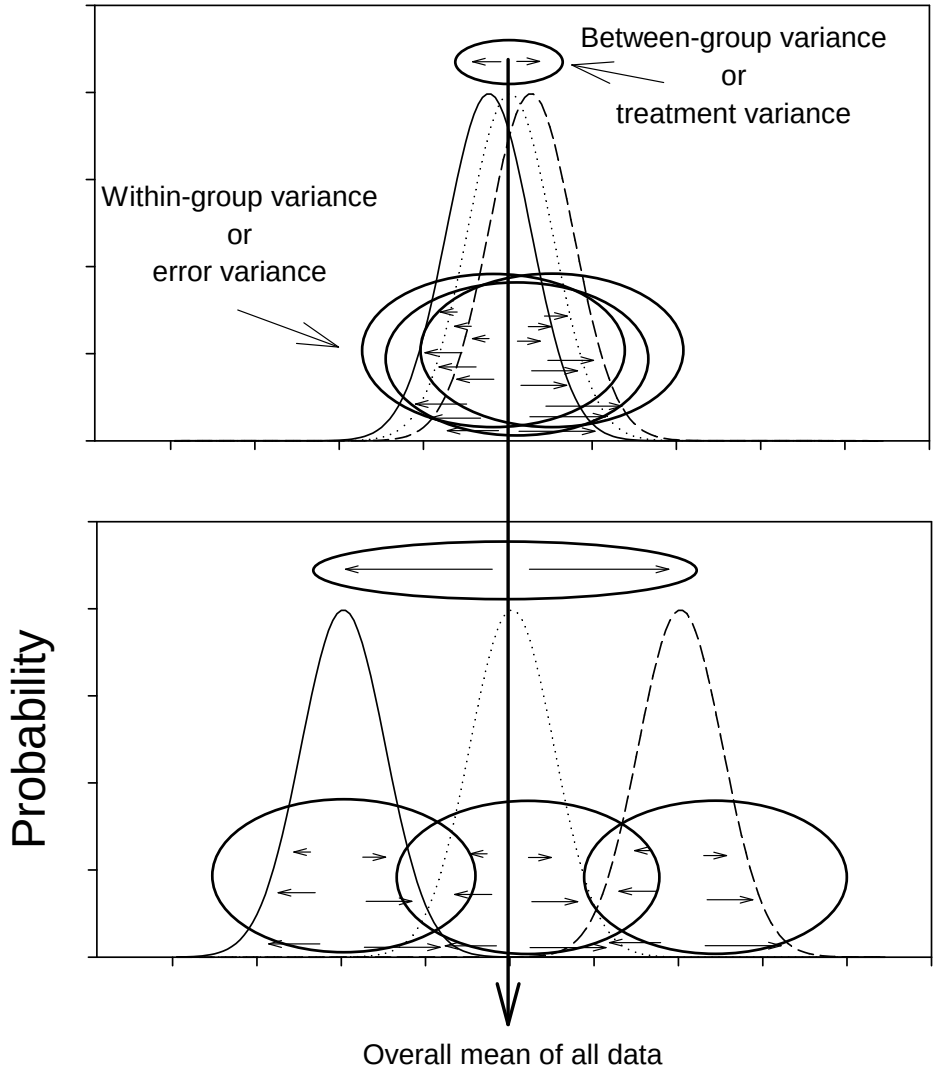

Figure 6.2. The above figure shows two scenarios where the populations either have, or have not, similar means. For the sake of argument, all three samples have the same variance and only differ any their means. Therefore, in this scenario, the within group variances are the same for both the upper and lower panels. In ANOVA, the test statistic compares the between-group and within-group variances. If the between-group variances are small then the averages are similar to each other. Alternatively, if the variance is high then the means are possibly different from each other.

Calculating variances

Variances, or in ANOVA-speak Mean Squared Errors, are calculated through calculating the sums-of-squares and then dividing by the correct degrees of freedom.

Two variances are to be calculated:

- Treatments : treatment averages around grand mean (this called the between-group variance).
- Error : individual data from all treatments around grand mean (also called the within-group variance).

This course will use treatment and error to separate out the ascribed causes of the variance, together with linking the test to an experimental protocol.

Variances are generally calculated from the sum-of-squared differences divided by the requisite degrees of freedom, such that.

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Sums-of-squares

The sums-of-squares calculated from the each of the *k* group means, $\bar{y}_t$, and the one overall mean, $\bar{y}_\bullet$, where *t* refers to the t^{th} treatment and $\bullet$ the for all the treatments.

$$SS_{TREAT} = \sum_{t=1}^{k}\sum_{i=1}^{n}(\bar{y}_t - \bar{y}_\bullet)^2 = \sum_{t=1}^{k} n_t(\bar{y}_t - \bar{y}_\bullet)^2$$

(note the absence of a subscript *i* in the equation – therefore the differences are a constant and added n_t times for each treatment. For example, if treatments A, B and C had 10, 15 and 12 samples, then n_A = 10, n_B =15, and n_C =12, respectively)

$$SS_{ERROR} = \sum_{t=1}^{k}\sum_{i=1}^{n}(y_{ti} - \bar{y}_t)^2$$

The sum-of-squares are additive such that $SS_{TOTAL} = SS_{TREAT} + SS_{ERROR}$ therefore the TOTAL sum-of-squares can be partitioned as follows

$$SS_{TOTAL} = \sum_{t=1}^{k}\sum_{i=1}^{n}(y_{ti} - \bar{y}_\bullet)^2 = \sum_{t=1}^{k}\sum_{i=1}^{n}[(y_{ti} - \bar{y}_t) + (\bar{y}_t - \bar{y}_\bullet)]^2$$

$$SS_{TOTAL} = \sum_{t=1}^{k}\sum_{i=1}^{n}(y_{ti} - \bar{y}_t)^2 + \sum_{t=1}^{k} n_t(\bar{y}_t - \bar{y}_\bullet)^2 = SS_{ERROR} + SS_{TREAT}$$

These are also illustrated in the Figure 6.3.

The solution to how the sum-of-squares is partitioned is deceptively simple, if one does some transformations such that $a = y_{ti}$, $c = \bar{y}_t$ and $b = \bar{y}_\bullet$, and dropping the summations.

Then let us show that $(a-b)^2 = [(a-c)+(c-b)]^2$ (by throwing in a "dummy" term *c*)

For the LHS we have

$$(a-b)^2 = a^2 - 2ab + b^2$$

For the RHS we have

$$\begin{aligned}[(a-c)+(c-b)]^2 &= (a-c)^2 + 2(a-c)(c-b) + (c-b)^2 \\ &= a^2 - 2ac + c^2 + 2ac - 2ab - 2c^2 + 2cb + c^2 - 2cb + b^2\end{aligned}$$

and cancel out the terms, such that

$$[(a-c)+(c-b)]^2 = a^2 - 2ab + b^2$$

Therefore, $(a-b)^2 = [(a-c)+(c-b)]^2$.

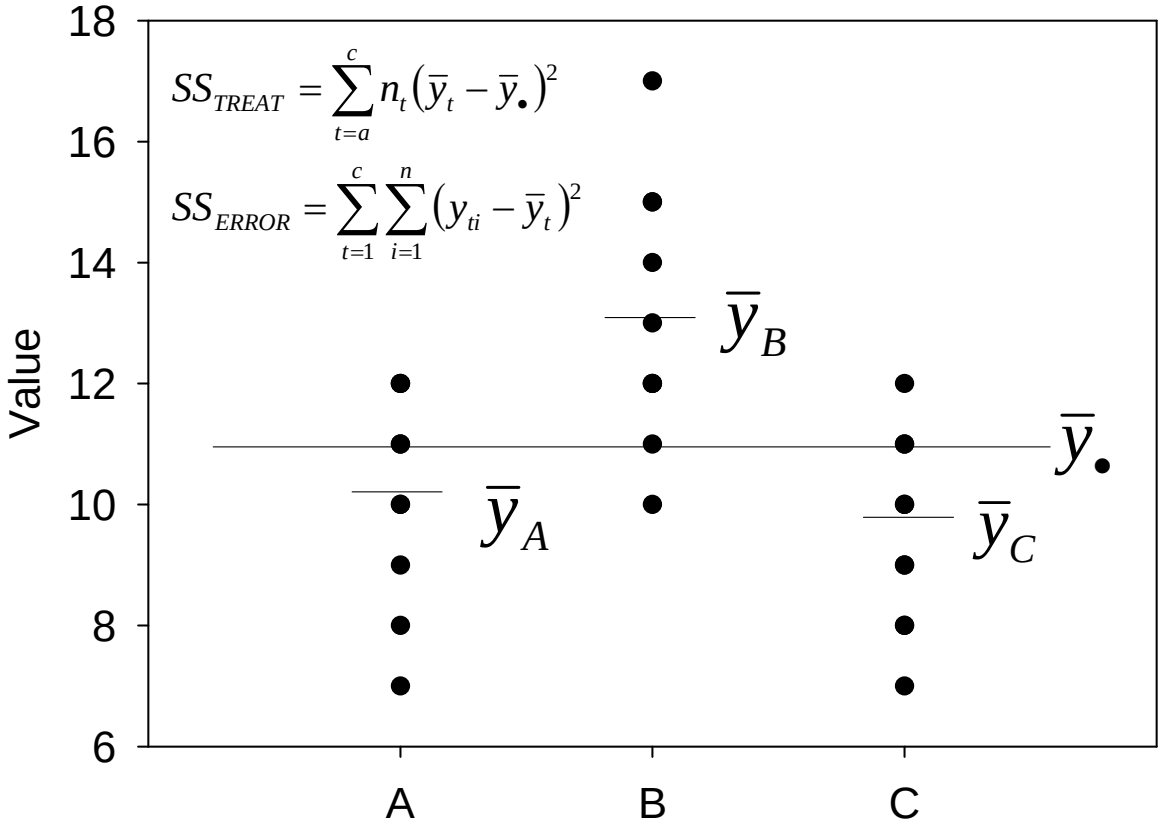

Figure 6.2. The partitioning of the sums-of-squares into its TREATMENT and ERROR components

Degrees of freedom

- In the treatment variance there will be one less degree of freedom than the *k* treatments considered – because we have one known grand mean
- In the error variance we have the total number of data points from *k* sample less the *k* averages.

Therefore $df_{TREAT} = k - 1$ and $df_{ERROR} = \left(\sum_{t=1}^{k} n_t \right) - k$.

Comparing variances

The hypothesis that both variances are equal is simply via an F-test between variances such that

$$F_{k-1,\sum n-k} = \frac{s^2_{TREAT}}{s^2_{ERROR}}$$

where

$$s^2_{TREAT} = \frac{SS_{TREAT}}{df_{TREAT}} \text{ and } s^2_{ERROR} = \frac{SS_{ERROR}}{df_{ERROR}}$$

The results from an ANOVA are usually displayed in table form.

	SS	Df	VAR	F	P
TREATMENT	SS_{TREAT}	k-1	$\frac{SS_{TREAT}}{df_{TREAT}}$	$\frac{VAR_{TREAT}}{VAR_{ERROR}}$	
ERROR	SS_{ERROR}	$\sum_k n_k - k$	$\frac{SS_{ERROR}}{df_{ERROR}}$		
TOTAL	$SS_{TREAT} + SS_{ERROR}$	n-1			

Anova's and t-tests

An ANOVA with $k = 2$ is mathematically equivalent to a two-sample *t*-test. i.e. $T^2 = F$.

Assumptions about ANOVA

- Random samples
- Residuals are Normal distributed (many textbook will state the "Data" – but as we will see in the linear regression section it is the residuals)
- Equal variances for all populations., i.e., the variances are homoscedastic.

Example 1:

Here are samples from three different treatments:

Sample A: 170, 146,120,112,132
Sample B: 224, 196,163,231,195
Sample C: 155, 153, 104,143, 198

Test the hypothesis that at least one mean is different

Step 1: Hypothesis: $H_0 : \mu_A = \mu_B = \mu_C$
$H_a : \mu_A \neq \mu_B \neq \mu_C$

Step 2: Test statistic: Do an ANOVA and calculate *F*

Step 3: Rejection region: Reject H_0 if $F > F_{(0.05,2,12)} = 3.89$

Step 4: Calculating the test statistic

To calculate the ANOVA we calculate four means,
$\bar{y}_A = 136.0$, $\bar{y}_B = 201.8$, $\bar{y}_C = 150.6$ and the grand mean of all 15 data points
$\bar{y}_\bullet = 162.8$.

We know that the treatment variance will have k-1 or 2 degrees of freedom and the error variance $\sum_k n_k - k$ or 15-3=12 degrees of freedom.

Now to calculate the different sums-of-squares.

$$SS_{TREAT} = \sum_{t=a}^{c} n_t \left(\bar{y}_t - \bar{y}_\bullet\right)^2 = 5(136.0 - 162.8)^2 + 5(201.8 - 162.8)^2 + 5(150.6 - 162.8)^2 = 11940.4$$

$$SS_{ERROR} = \sum_{t=a}^{c} \sum_{i=1}^{n} \left(y_{ti} - \bar{y}_t\right)^2 =$$

$$= (170 - 136.0)^2 + (146 - 136.0)^2 + (120 - 136.0)^2 + (112 - 136.0)^2 + (132 - 136.0)^2$$
$$+ (224 - 201.8)^2 + (196 - 201.8)^2 + (163 - 201.8)^2 + (231 - 201.8)^2 + (195 - 201.8)^2$$
$$+ (155 - 150.6)^2 + (153 - 150.6)^2 + (104 - 150.6)^2 + (143 - 150.6)^2 + (198 - 150.6)^2$$
$$= 9536.0$$

The variances, or mean squared errors, are

$$s^2_{TREAT} = \frac{11940.4}{2} = 5970.2$$

and $s^2_{ERROR} = \frac{9536.0}{12} = 794.67$

such that $F = \frac{5970.2}{794.67} = 7.51$

Step 5: As this value is greater than 3.89 we reject the null hypothesis and conclude that at least one mean is different.

The results can be summarised in a table as

	SS	df	MSE	F
TREATMENT	11940.4	2	5970.2	7.51
ERROR	9536.0	12	794.66	
TOTAL	21476			

Example 1:

Here are samples from three different treatments:

Treatment A: 111.3, 88.9, 88.6, 114.8, 111.8
Treatment B: 88.2, 97.1, 85.2, 93.7, 90.2, 97.2, 94.0
Treatment C: 71.3, 65.4, 68, 78.2, 61.7, 75.7, 71.4, 76.5, 63.4

Test the hypothesis that at least one mean is different.

(Note that this ANOVA will be unbalanced as the sample sizes are different. Do not collect your data like this.)

Step 1: Hypothesis: $H_0 : \mu_A = \mu_B = \mu_C$
$H_a : \mu_A \neq \mu_B \neq \mu_C$

Step 2: Test statistic: Do an ANOVA and calculate *F*

Step 3: Rejection region: Reject H_0 if $F > F_{(0.05,2,18)} = 3.55$

Step 4: Calculating the test statistic

To calculate the ANOVA we calculate four means,
$\bar{y}_A = 103.08, \bar{y}_B = 92.23, \bar{y}_C = 70.18$ and the grand mean of all 15 data points
$\bar{y}_\bullet = 85.36$.

We know that the treatment variance will have *k*-1 or 2 degrees of freedom and the error variance $\sum_{t=A}^{C} n_t - k$ or 21-3=18 degrees of freedom.

Now to calculate the different sums-of-squares.

$$SS_{TREAT} = \sum_{t=a}^{c} n_t (\bar{y}_t - \bar{y}_\bullet)^2 = 5(103.08 - 85.36)^2 + 7(92.23 - 85.36)^2 + 9(70.18 - 85.36)^2 = 3974.73$$

$$SS_{ERROR} = \sum_{t=a}^{c} \sum_{i=1}^{n} (y_{ti} - \bar{y}_t)^2 = 1098.16$$

The variances, or mean squared errors, are

$$s^2_{TREAT} = \frac{3974.73}{2} = 1987.37$$

and $s^2_{ERROR} = \frac{1098.16}{18} = 61.01$

such that $F = \frac{1987.37}{61.01} = 32.57$

Step 5: As this value is greater than the critical *F*-statistic, we reject the null hypothesis and conclude that at least one mean is different.

The results can be summarised in a table as

	SS	df	MSE	F
TREATMENT	3974.73	2	1987.37	32.57
ERROR	1098.16	18	61.01	
TOTAL	5072.89			