7. Analysis of Variance (ANOVA)

In this case we have **independent** samples from k populations. We wish to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 \ldots = \mu_k,$$

where μ_i is the population mean from which the i-th sample is taken, against the alternative

H_A: at least one of the population means is different from the remaining population means.

Assumptions of ANOVA

1. The observations in each sample are from a normal distribution.
2. The population variances are equal (denoted as σ^2).
3. The samples are independent. [e.g. we **cannot** have three samples of weights taken for one group a) before a diet, b) a month into a diet, c) at the end of the diet].

7.1 Notation and intuition

1. $\overline{X}$ denotes the mean of all the observations.
2. $\overline{X}_j$ denotes the mean of the j-th sample.
3. s_j^2 denotes the sample variance for the j-th sample.
4. n_j is the number of observations in the j-th sample.
5. $X_{i,j}$ denotes the i-th observation in the j-th sample.
6. $n = \sum_{j=1}^{k} n_j$ is the total number of observations.
   Note: summing over j, we sum over all the groups.

Mean of all observations

We have

$$\overline{X} = \frac{\sum_{j=1}^{k} n_j \overline{X}_j}{n}$$

Note that the sample means are weighted according to the number of observations in that sample.

In particular, if there are the same number of observations in each sample then, the mean of all the observations is simply the mean of the sample means.

Total sum of squares (TSS)

The total sum of squares (TSS) is

$$TSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{i,j} - \overline{X})^2$$

Note: Summing over both i and j, we sum over all the observations. Here we sum the squares of the deviations from the mean of all observations.

This is a measure of the variation in the observations as a whole.

The number of degrees of freedom is $n - 1$ (n=total number of observations).

Sum of squares between groups (SSB)

The sum of squares between groups (SSB) is

$$SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2$$

(note that the sum is taken over the groups).

The number of degrees of freedom associated with this sum is $k - 1$ (k is the number of groups).

This is a measure of the variation between the means of the sample. The more the $\overline{X}_j$ differ, the larger this sum will be.

Sum of squares within groups (SSW)

The sum of the squares within groups (SSW) is given by

$$SSW = \sum_{j=1}^{k} (n_j - 1)s_j^2$$

This is measure of the level of variation within groups.
The number of degrees of freedom is $n - k$ (the total number of observations minus the number of groups).
It can be shown that

$$TSS = SSB + SSW$$

Mean sums of squares

The mean square of variation between groups (MSB) is the sum of squares between groups divided the number of degrees of freedom.

$$MSB = \frac{SSB}{k-1}$$

The mean square of variation within groups (MSW) is the sum of squares within groups divided the number of degrees of freedom.

$$MSW = \frac{SSW}{n-k}$$

The F test

We use the F test, in which the test statistic is

$$F = \frac{MSB}{MSW}$$

Under the null hypothesis this test statistic has an $F_{k-1,n-k}$ distribution.

The realisation of this statistic (the value obtained using the data) will be denoted by f.

Intuition of the F test

If the population means are all equal, the mean variation between groups will tend to be small in comparison to the mean variation within groups.

If these population means vary, the mean variation between groups will tend to be large in comparison to the mean variation within groups.

High values of f indicate that the null hypothesis is false.

Critical values for the F test

We reject H_0 at a significance level of $100\alpha\%$ if $f > F_{k-1,n-k,\alpha}$, where

$$P(F > F_{k-1,n-k,\alpha}) = \alpha.$$

The critical values for the F distribution are given in Table 9 of the coursebook.

Relation between the F test and the test for a difference between population means for 2 independent samples

It should be noted that this test is a generalisation of the test for the difference between 2 population means for independent samples.

When ANOVA is used for $k = 2$, the realisation of the test statistic f is related to t (the realisation in the 2-sample test) as follows

$$f = t^2$$

The conclusion made from both tests is the same.

Example 7.1

The heights of a sample of Irish, American and Japanese males are given below:

Americans: 177, 183, 192, 168
Irish: 165, 174, 183
Japanese 162, 165, 171, 174

Test the null hypothesis that Americans, Irish and Japanese men have the same average height.

Solution to Example 7.1-Hypotheses

Hence, we have

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

where μ_i represents the mean height of the population from which the i-th sample is taken.

H_A : at least one of the population means differs from the other.

Solution to Example 7.1-Calculation of sample means

First we calculate the means and variances of the samples.

The mean heights are

$$\overline{X}_1 = \frac{177 + 183 + 192 + 168}{4} = 180$$
$$\overline{X}_2 = \frac{165 + 174 + 183}{3} = 174$$
$$\overline{X}_3 = \frac{162 + 165 + 171 + 174}{4} = 168$$

Solution to Example 7.1-Calculation of the mean of all observations

The total number of observations is $\sum_{j=1}^{k} n_j = 4 + 3 + 4 = 11$

$$\begin{aligned}\overline{X} &= \frac{\sum_{j=1}^{k} n_j \overline{X}_j}{n} \\ &= \frac{4 \times 180 + 3 \times 174 + 4 \times 168}{11} = 174\end{aligned}$$

Solution to Example 7.1-Calculation of the sample variances

The sample variances are

$$s_1^2 = \frac{(X_{1,1} - \overline{X}_1)^2 + (X_{2,1} - \overline{X}_2)^2 + \ldots + (X_{n_1,1} - \overline{X}_2)^2}{n_1 - 1}$$

$$= \frac{(177 - 180)^2 + (183 - 180)^2 + (192 - 180)^2 + (168 - 180)^2}{3} = 102$$

$$s_2^2 = \frac{(165 - 174)^2 + (174 - 174)^2 + (183 - 174)^2}{2} = 81$$

$$s_3^2 = \frac{(162 - 168)^2 + (165 - 168)^2 + (171 - 168)^2 + (174 - 168)^2}{3} = 30$$

Solution to Example 7.1-Calculation of the sum of squares between groups

Now we calculate SSB (the measure of the variation between groups).

$$SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2$$
$$= 4 \times (180 - 174)^2 + 3 \times (174 - 174)^2 + 4 \times (168 - 174)^2 = 288.$$

Solution to Example 7.1-Calculation of the sum of squares within groups

Now we calculate SSW (the measure of variation within groups)

$$\begin{aligned} SSW &= \sum_{j=1}^{k} (n_j - 1) s_j^2 \\ &= (4-1) \times 102 + (3-1) \times 81 + (4-1) \times 30 = 558. \end{aligned}$$

Solution to Example 7.1-Calculation of the mean squares

The mean square between groups (MSB) is

$$MSB = \frac{SSB}{k-1} = \frac{288}{3-1} = 144.$$

(k is the number of groups)

The mean square variation within groups is

$$MSW = \frac{SSW}{n-k} = \frac{558}{11-3} = 69.75$$

(n is the total number of observations)

Solution to Example 7.1-Calculation of the realisation of the test statistic

The test statistic is

$$F = \frac{MSB}{MSW}$$

The realisation of this test statistic is

$$f = \frac{144}{69.75} \approx 2.06$$

Solution to Example 7.1-Critical value for the test statistic

The critical value is $F_{k-1,n-k,\alpha}$, where α is the significance level.

At the 5% level

$$F_{k-1,n-k,\alpha} = F_{2,8,0.05}$$

Solution to Example 7.1-Reading the critical value-Table 9 in coursebook

1. The first index corresponds to the column and the second index corresponds to the row.
2. Each cell in the table contains 4 critical values. The first corresponds to $\alpha = 0.05$. The second, which is in brackets, corresponds to $\alpha = 0.025$. The remaining two correspond to $\alpha = 0.01$ and $\alpha = 0.001$, respectively.

Thus, $F_{2,8,0.05} = 4.46$.

Solution to Example 7.1-Conclusion

Since $f = 2.06 < F_{2,8,0.05} = 4.46$.

We do not reject H_0 at a significance level of 5%. We have no evidence that the mean height of these populations vary.

Note: If ANOVA indicates that there is a difference between these population means, then we should use 'post-hoc' procedures to check how they differ.

This is only considered in the laboratory classes.

The ANOVA table

The results of analysis of variance are often presented in the form of a table.

Source	D.F.	S.S.	M.S.	F
Between	$df_B = k - 1$	SSB	$MSB = \frac{SSB}{df_B}$	$F = \frac{MSB}{MSE}$
Within	$df_W = n - k$	SSW	$MSW = \frac{SSW}{df_W}$	
Total	$df_T = n - 1$	TSS		

Note that $df_T = df_W + df_B$.

The ANOVA table for Example 7.1

Source	D.F.	S.S.	M. S.	F
Between	$k-1=2$	288	$\frac{288}{2}=144$	$f=\frac{144}{69.75}=2.06$
Within	$n-k=8$	558	$\frac{558}{8}=69.75$	
Total	10	846		