The Analysis of Variance

20.1 INTRODUCTION

Earlier, we compared two population means by using a two-sample -test. However, we are often required to compare more than two opulation means simultaneously. We might be tempted to apply the wo-sample t-test to all possible pairwise comparisons of means. For xample, if we wish to compare 4 population means, there will be $\binom{4}{2} = 6$ eparate pairs and to test the null hypothesis that all four population neans are equal, would require six two-sample t-tests. Similarly, to test he null hypothesis that 10 population means are equal, we would need $\binom{10}{2} = 45$ separate two-sample t-tests. This sort of running multiple two-sample t-tests for comparing means has two disadvantages. *First*, the procedure is tedious and time consuming, and *secondly*, the overall level of significance greatly increases as the number of t-tests increases. Thus a series of two-sample t-tests is not an appropriate procedure to test the equality of several means simultaneously.

Evidently, we require a procedure for carrying out a test on several means simultaneously. One such procedure is the *analysis of variance*, introduced by Sir R.A. Fisher (1890–1962) in 1923. The *analysis of variance* (abbreviated as *ANOVA*) is a technique that partitions the total variation–a term distinct from variance and measured by the sum of squares of deviations from the mean–into its component parts, each of which is associated with a different source of variation. These component parts of variance are then analysed (hence the name, analysis of variance) in such a manner that certain hypotheses can be tested. This technique is based on the facts that (i) the more the sample means differ the larger the variance becomes, and (ii) the separate components provide independent and unbiased estimates of the common population variance. The anlaysis of variance procedure therefore compares different estimates of variance by using F-distribution to deterr whether the population means are equal. The analysis of variance been shown the most powerful and useful technique whenever statistical data can be categorised in groups.

When each observation is classified into one sample or anot according to a single criterion, we have a *one-way classification* while classification of each observation on the basis of two criteria classification, is called a *two-way classification*. In a similar way, a *mu way classification* is defined. We discuss the analysis of variar procedures for the first two classifications only as a **one-way analy of variance** and a **two-way analysis of variance** respectively.

20.2 ONE-WAY ANALYSIS OF VARIANCE

The *one-way analysis of variance* is also called the *one-variable-o classification analysis of variance*. The data are classified into k classe or groups, etc. on the basis of a single criterion. The technical term for a class, a group or a variable of classification is *treatment*.

Suppose we have k samples of equal size r (the case of unequal sample sizes will be discussed later), selected randomly and independently, one from each of k normal populations with means $\mu_1, \mu_2, ..., \mu_k$ and common variance σ^2; and we wish to test the nu hypothesis that all the k-population means are equal, *i.e.*

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

against the alternative hypothesis

$$H_1 : \text{Not all means are equal.}$$

Let X_{ij} denote the ith observation of the jth sample (or treatment). Then the data can be arranged as in table below:

Observation	Samples (or Treatments)						Total
	1	2	...	j	...	k	
1	X_{11}	X_{12}		X_{1j}		X_{1k}	
2	X_{21}	X_{22}		X_{2j}		X_{2k}	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
i	X_{i1}	X_{i2}		X_{ij}		X_{ik}	
.	.	.		.		.	
.	.	.		.		.	
.	.	.		.		.	
r	X_{r1}	X_{r2}		X_{rj}		X_{rk}	
Totals	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.k}$	$T_{..}$
Means	$\bar{X}_{.1}$	$\bar{X}_{.2}$	...	$\bar{X}_{.j}$	...	$\bar{X}_{.k}$	$\bar{X}_{..}$

Example 20.1 Given the data below, test the hypothesis that the means of the three populations are equal. Let $\alpha = 0.05$.

Sample 1	Sample 2	Sample 3
40	70	45
50	65	38
60	66	60
65	50	42

(i) We state our null and alternative hypotheses as

$H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.* all the three means are equal, and

H_1 : Not all three means are equal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_b^2}{s_w^2},$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$ and $v_2 = n-k$ degrees of freedom.

(iv) The compu...

	Sample 1 $X_{i1}\ (X_{i1}^2)$	Sample 2 $X_{i2}\ (X_{i2}^2)$	Sample 3 $X_{i3}\ (X_{i3}^2)$	Total
	40 (1600) 50 (2500) 60 (3600) 65 (4225)	70 (4900) 65 (4225) 66 (4356) 50 (2500)	45 (2025) 38 (1444) 60 (3600) 42 (1764)	--- --- --- ---
$T_{\cdot j}$	215	251	185	651
$T_{\cdot j}^2$	46225	63001	34225	143451
$\sum_i X_{ij}^2$	11925	15981	8833	36739

$$\text{Correction Factor } (C.F.) = \frac{T_{..}^2}{n} = \frac{(651)^2}{12} = 35316.75$$

$$\text{Total} \quad SS = \sum_i \sum_j X_{ij}^2 - C.F.$$

$$= 36739 - 35316.75 = 1422.25$$

$$\text{Between} \quad SS = \frac{\sum_j T_{\cdot j}^2}{r} - C.F.$$

$$= \frac{143451}{4} - 35316.75 = 546.00, \text{ and}$$

Within SS = Total SS – Between SS = 1422.25 – 546.00 = 876.25.

The Analysis of Variance table is:

Source of Variation	d.f.	Sum of Squares	Mean Square	Computed F
Between Samples	2	546.00	273.00	$\frac{273.00}{97.36} = 2.80$
Within Samples	9	876.25	97.36	----
Total Variation	11	1422.25	---	---