UNIT 8 ANALYSIS OF VARIANCE: ONE-WAY CLASSIFICATION

8.1 INTRODUCTION

In Unit 6, you have learnt to compare two population means. Often, in practice, one is required to compare more than two population means. In this unit, we shall study a statistical procedure, called analysis of variance, which allows one to test a hypothesis comparing several normal population means.

The problem of comparing several population means arises quite naturally in practice. For instance, on the basis of sample data, one might wish to decide whether there is any real difference between three teaching methods of a foreign language. Quite often in agriculture, an experimenter is interested in comparing the yielding abilities of several varieties of a crop, say wheat. Similarly, there may be four different drugs for the control of blood pressure and it is of interest to know whether these four drugs are equally efficient in the control of blood pressure.

In each of the above examples, we have several populations (one for each method of teaching, each crop variety, each drug) and the hypothesis that is required to be tested is whether the means of these populations are equal. Analysis of variance, written in short as ANOVA is useful in such situations. We shall assume that each of the populations have a normal distribution with possibly different means but having the same variance.

We shall start this unit by acquainting you with real life problem in which you need to compare means of several populations simultaneously. We shall introduce the method of analysis of variance through this example. In this unit we shall confine our attention to study the effect of a single factor on a variable under study. Such study lead to one-way classification data. We shall then formulate the model for one-way classification for the problem considered and use this model to explain the procedure to

have better insight about the model parameters we have also discussed the estimation of model parameters and their pairwise comparison. We shall end our discussion in this unit by acquainting you with a concept of random effects.

Objectives

After reading this unit, you should be able to

- identify different sources of variation in a given problem;

- distinguish the one-way classification data from other types;

- write down the model for one-way classification problems;

- carry out the test of hypothesis on equality of all treatment means, equality of pairs of treatments means;

- estimate the model parameters and provide confidence intervals;

- draw inferences and conclusions from the analysis.

8.2 ANALYSIS OF VARIANCE : BASIC CONCEPTS

We shall start with a simple real life problem that many of us face. Now a days most of us use gas for cooking purposes. Most of the gas users are customers of big companies The customers get their refills (filled gas cylinders) through the agents of these companies. One of the customers, Mrs. Devi, who is attached to ABC agency, has faced a problem in the recent past. She observed that her cylinders were not lasting as long as they used to be in the past. So she suspected that the amount of gas in the refills was less compared to what she used to get in the past. She knew that she is supposed to get 14.2 kgs of gas in every refill. She explained her problem to the customers' redressal cell of the company.

Subsequently, the company made a surprise check on an ABC agent. They took 25 cylinders that were being supplied to customers from this agency and measured the amount of gas in each of these cylinders. The 25 observations were statistically analysed and through a simple test of hypothesis (you may recall how this is done) it was inferred that the mean amount of gas in the cylinders supplied by ABC agency was significantly lower than 14.2 kgs. On investigation, it was revealed that the agent was tapping gas from cylinders before they are being supplied to the customers.

There were five agents of the company in the town where Mrs.Devi was living. To protect customers' interests, the company decided to carry out surprise checks on all the agents from time to time. During each check, they picked up 7 cylinders at random from each of the five agents resulting in the data given in Table-1. Is it possible to test from this data whether the mean amount of gas per cylinder differs from agent to agent? Well, it is possible to carry out a simple test of hypothesis for each of the agents separately. But there is a better statistical procedure to do this simultaneously. We shall see how this can be done.

E1) What is the difference between the hypothesis you studied in Unit 6 and the hypothesis in the above problem?

Before we proceed further with the above problem, we introduce to you some concepts and terminology.

You know that variation is inevitable in almost all the variables (measurable characteristics) that we come across in practice. For example, the amount of gas in two refills is not the same irrespective of whether the gas is tapped or not. Consider the data in Table-1.

Table-1: Data on Gas Weights (kgs)

S. No.	Agent				
	1	2	3	4	5
1	14.36	13.51	13.74	14.09	13.51
2	14.15	13.60	13.94	14.52	14.03
3	14.20	13.71	14.10	14.58	13.84
4	14.12	13.94	13.93	14.08	13.29
5	14.05	13.95	13.61	13.66	13.97
6	14.15	13.67	13.97	13.90	13.44
7	14.17	13.79	13.88	14.13	13.94

We have the weights of gas in 35 cylinders taken at random, seven from each of the five agents. These 35 weights exhibit variation. You will agree that some of the possible reasons for this variation are one or more of the following:-

- The gas refilling machine at the company does not fill every cylinder with exactly same amount of gas.

- There may be some leakage problem in some of the cylinders.

- The agency/agents might have tapped gas from some of these cylinders.

- All the 35 cylinders are not filled by the same filling machine.

Thus, the variation in the 35 weights might have come from different sources. Though the variation is attributable to several sources, depending upon the situation, we will be interested in analysing whether most of this variation can be due to differences in one (or more) of the sources. For instance, in the above example, the company will be interested in identifying if there are any differences among the agents. So the **source of variation** of interest here is **agents**. In other words, we are interested in one factor or, one-way analysis of variance. Before continuing further with the discussion, try to identify the sources of variations in the following exercises.

E2) Five different fertilisers were tested on a particular crop to compare their performances. Each fertiliser was applied on 4 plots. All the twenty plots used here belong to the same location. Identify the sources of variation and mention which of these you suspect to have larger contribution to the variation.

E3) One of several components manufactured by an automobile company is the drive shaft. This component is manufactured on a special purpose machine which has 8 stations. Drive shafts are produced simultaneously on all the eight stations. Milling is one of the operations carried out at each station. Milling operation produces a slot on the drive shaft and the length of the slot, x, should be between 5.85 cms and 6.05 cms. Identify possible sources that contribute to variation in x.

8.2.2 One-Way Classification

Now that you know what is source of variation, you can think of different types of sources. In the gas company example, agents form one type of source. If the cylinders

under consideration were refilled by different filling machines, then filling machines is another type of source of variation. Similarly, in E2), supposing that 10 of the 20 plots belong to one location and the remaining 10 belong to another location, then there are two types of sources of variation, namely, 'fertilisers' and 'location'. **When the data are classified only with respect to one type of source of variation, we say that we have one-way classification data.** On the other hand, in the above fertiliser example, if we also have reasons to believe that there could be differences between locations, then we have a two-way classification data. In many situations, one conducts experiments to study the effect of a single factor on a variable under study. Such experiments, known as one-factor experiments, lead to one-way classification data. The following is an example of one such experiment.

Example 1: The tensile strength of synthetic fibre used to make cloth is of interest to the manufacturer. It is suspected that the strength is affected by the percentage of cotton in the fibre. The cloth was produced by varying the cotton percentage at five different levels, namely, 15%, 20%, 25%, 30% and 35%. Samples were drawn from the cloth produced at each of these levels and the tensile strengths were measured. In this way one-way classification data was obtained.

* * *

As we have already mentioned, in this unit, we will confine ourselves to one-way classification only. We shall now look at the theoretical model for the one-way classification. To make it easy for you to understand, we shall describe the model with the gas company example.

8.3 MODEL FOR ONE-WAY CLASSIFICATION

As you already know the first step in any modelling process is to identify the various parameters involved in the problem. We shall now specify these parameters for the problem under consideration. Consider the problem of the gas company example. It is suspected that the agents might tap the gas from each of the cylinders. Let us say that agent i, taps a_i kgs of gas from each of the cylinders supplied by him, $i = 1, 2, ..., k$. Here k is the number of agents. Recall that $k = 5$ for Mrs. Devi's town. We shall assume that the weight of each cylinder supplied by the company to the agents varies marginally around 14.2 kgs, so that the average weight of a cylinder is given by $\mu = 14.2$ kgs. Next, let us say that, according to the vigilance policy, the company picks up n_i cylinders at random from agent i and measures the weights of the gas. For the data in Table-1, n_is are all equal to 7. Let y_{ij} be the weight of the gas in j^{th} cylinder from agent i. Then, we would expect the value of y_{ij} to be 14.2 kgs provided there is no tapping of gas. However, since agent i taps a_i kgs of gas, we would expect y_{ij} to be 14.2-a_i kgs. and this should be true for $j = 1, 2, \ldots, n_i$. Yet, when we measure y_{ij}, it would not be exactly equal to 14.2-a_i kgs. There could be several reasons for this. When the cylinder was supplied to the agent, its weight might be slightly less than or more than 14.2 kgs. There could have been minor variation in reading the measurement. Due to these reasons, it is reasonable to assume that $y_{ij} = \mu - a_i + \epsilon_{ij}$, where ϵ_{ij}, called the error, is the difference between the observed weight and the true weight. Letting $\tau_i = -a_i$, for $1 \leq i \leq k$, we can write

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k. \tag{1}$$

In the above equation, μ is called the general (mean) effect and τ_i is called the effect of agent i. In general, μ and τ_is are unknown quantities but are assumed to be constants. **These are known as the model parameters.** On the other hand, the ϵ_{ij}s **are due to random fluctuations** and hence are assumed to be **random variables**. As a result, y_{ij}s are also random variables.

After specifying the model parameters we move on to the next step of the procedure.

Stating The Hypothesis

If all the agents are honest, then τ_is must all be equal to zero. Thus, the company would be interested in testing the hypothesis

$$H_0: \tau_1 = \tau_2 = \ldots = \tau_k = 0. \tag{2}$$

The alternative hypothesis here is that at least one of the τ_is is different from zero (or equivalently, $\tau_l \neq \tau_m$ for some l and m).

In Model (1) above, we make the following two assumptions:

i) **The ϵ_{ij} s are independent and identically distributed random variables with mean 0 and variance σ^2.**

ii) **ϵ_{ij}s follow normal distribution.**

The model is a typical example of a **model for one-way classification.**

In Eqn.(2), H_0 states that all τ_is are equal to zero. However, in general, we may not always be interested in testing the equality of means to zero. One may just want to test the equality of means only, that is, want to test the hypothesis

$$K_0 : \tau_1 = \tau_2 = \ldots = \tau_k$$

Fortunately, it can be shown that the test procedure is same for both H_0 and K_0.

On the same lines as above you may now try to construct the linear model for the fertiliser example.

E4) Consider E2). Assuming that all the twenty plots are homogeneous, construct a linear model for this problem and explain various quantities you define.

In the following section, we will learn the analysis of variance (ANOVA) technique to carry out the tests of hypothesis of the type mentioned in Eqn.(2) above.

8.4 TEST PROCEDURE

As the name suggests, in ANOVA, we analyse the variability in any given set of data and try to see if this variability is mainly due to certain specific reasons. We first try to understand the procedure through the gas company example.

8.4.1 Sums of Squares

Consider the gas company example and the model given in Eqn. (1). Look at seven weights under any agent in Table-1. You find there is variation. This variation arises from several sources such as measurement error etc. But these sources are common to all the agents. So it is reasonable to expect the variance of the observations under each agent is the same for all agents. If this common variance is represented by σ^2, then to test the null hypothesis that the k population means are all equal, we shall compare two estimates of σ^2— one based on the variation among the sample means, and one based on the variation within the samples.

A pooled estimate of σ^2 i.e., the mean of the sample variances s_i^2, is given by

$$\sigma^2 = \frac{\sum_{i=1}^5 s_i^2}{5} = \frac{\sum_{i=1}^5 \sum_{j=1}^7 (y_{ij} - \bar{y}_i)^2/6}{5} = \frac{1.5306}{30} = 0.05102$$

The quantity $\sum_{i=1}^{5} \sum_{j=1}^{7} (y_{ij} - \bar{y}_i)^2$ is called the **error sum of squares** (SS$_e$), the quantity that estimates random (or chance) error.

Now look at the agent wise averages $\bar{y}_i$s in Table-2. If the agents are uniform, we expect the variance of $\bar{y}_i$s to be small. The sample variance of $\bar{y}_i$s is given by

$$\frac{\sum_{i=1}^{5}(\bar{y}_i - \bar{y})^2}{4} = \frac{1.2976}{4} = 0.3244$$

where $\bar{y}$ is the average of the $\bar{y}_i$s. The quantity $\sum_{i=1}^{5}(\bar{y}_i - \bar{y})^2$ is called the **sum of squares due to agents** (SS$_a$).

The sample variance of all the 35 weights is given by

$$\frac{\sum_{i=1}^{5} \sum_{j=1}^{7}(y_{ij} - \bar{y})^2}{34} = \frac{2.8282}{34}.$$

The quantity $\sum_{i=1}^{5} \sum_{j=1}^{7}(y_{ij} - \bar{y})^2$ is called the **total sum of squares** (TSS).

Observe that 2.8282 = 1.2976 + 1.5306. That is, TSS = SS$_a$ + SS$_e$. Therefore, if SS$_a$ is large compared to SS$_e$, we can infer that τ_is are indeed different. However, the term 'large' or 'small' are subjective. In order to quantify the largeness of SSa related to SSe, we compute the ratio

$$F = \frac{SS_a/4}{SS_e/30} = 6.35.$$

This ratio, under the hypothesis, is known to have an F-distribution on 4 and 30 degrees of freedom. Comparing this ratio with the table value of F-distribution with 4 and 30 degrees of freedom at say, 5 % level of significance allows one to decide whether to reject the hypothesis of equality of means. We shall talk more about it later.

Let us now consider a general case. Throughout this unit we shall denote the total number of observations by N, i.e., $N = \sum_{i=1}^{k} n_i$. Let these N observations be

$$y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}, \ldots y_{k1}, y_{k2}, \ldots, y_{kn_k}.$$

The **total sum of squares** (TSS), of all these observations is given by

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2,$$

where $\bar{y}$ is the average of all the y_{ij}s. Let $\bar{y}_i$ be the average weight of the gas in cylinders from agent i, i.e., $\bar{y}_i = (\sum_{j=1}^{n_i} y_{ij})/n_i$. Using simple algebra it can be shown that

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 \tag{3}$$

We are leaving this for you to verify yourself.

E5) Show that $\sum_{i=1}^{k} \sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2.$

Using $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, we see that

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(\epsilon_{ij} - \bar{\epsilon}_i)^2,$$

where $\bar{\epsilon}_i = (\sum_{j=1}^{n_i} \epsilon_{ij})/n_i$.

You may note here that $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ involves only the errors ϵ_{ij}s. We call this sum of **squares due to error** and denote it by SS_e. Similarly, $\sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$ indicates the dispersion among the agents' averages. This is called the **sum of squares due to agents** (or **sum of squares due to treatments** in general) and we denote this by SS_a (SS_{tr} in general).

When all the n_is are equal, say equal to n, the data are said to be **balanced**. In our further discussion we shall assume that the data are balanced, unless mentioned otherwise. Thus we have from Eqn.(3) above

$$TSS = SS_{tr} + SS_e.$$

i.e., **the total sum of squares is decomposed into sum of squares due to error and sums of squares due to treatments.** When all the agents are honest, we expect all the $\bar{y}_i$s to be close to $\bar{y}$ and hence SS_a to be small. Under Assumption (i) ϵ_{ij}s are independent and identically distributed (iid) random variables with mean 0 and variance σ^2, it can be shown that the statistical expectation of $SS_{tr}/(k-1)$, called the **mean sum of squares due to treatments**, and denoted by MS_{tr}, is given by

$$E(MS_{tr}) = \sigma^2 + \frac{\sum_{i=1}^{k} n_i \tau_i^2}{(k-1)}. \tag{4}$$

Also, as per Assumption (ii) above ϵ_{ij}s are normally distributed, then under H_0, SS_{tr}/σ^2 follows χ^2-distribution with $(k-1)$ degrees of freedom (df).

On the other hand, the mean squares for error, $SS_e/(nk-k)$ denoted by MS_e, is a measure of natural variability present in the data due to non-assignable causes and is an unbiased estimator of σ^2.

Recall the sampling distributions you have studied in Unit 4.

Thus, under Assumptions (i) and (ii), it follows that SS_e/σ^2 has χ^2-distribution with $\nu = (nk-k)N-k$ df. In fact, MS_e acts as yardstick to decide whether SS_{tr} is large or small. To decide whether to reject H_0 or not, we need to compute the ratio

$$F_0 = \frac{MS_{tr}}{MS_e} = \frac{SS_{tr}/(k-1)}{SS_e/\nu}.$$

This ratio follows **F-distribution** with **degrees of freedom** $k-1$ **and** ν. Recall that type I error, α, is the chance of rejecting H_0 when it is true.

Suppose we decide to carry out the test at (5% level of significance), all that we have to do is to get the tabulated F-value with respective df from the F-table(ref. Table-1 in the Appendix given at the end of the block.) and reject H_0 if F_0 is greater than this tabulated F-value.

Let us now apply this test to our gas company example. We first prepare the ANOVA table for gas company example.

8.4.2 Preparation of ANOVA Table

Consider the data in Table-1. Let us first summarise these data by computing the agent-wise totals and averages. Let y_i denote the total of y_{ij}s under agent i. Then $\bar{y}_i = (\sum_{i=1}^{7} y_{ij})/7, i = 1, 2, \ldots, 5$. These values are summarised in Table-2.

Table-2: Summary of Data on Gas Weights

Statistic	Agent				
	1	2	3	4	5
Total (y_i)	99.20	96.17	97.17	98.96	96.02
Average $(\bar{y}_i)$	14.1714	13.7386	13.8814	14.1371	13.7171
Sample variance, $s_i^2 = \sum_{j=1}^{7} \frac{(y_{ij}-\bar{y}_i)^2}{6}$	0.00911	0.02748	0.02585	0.10536	0.08812
$6 \times s_i^2$	0.05469	0.16489	0.15509	0.63214	0.52874

To carry out the test of hypothesis H_0 given in Eqn. (1), we prepare a table called the ANOVA table. For the one-way classification, the ANOVA Table typically looks like Table-3.

Table-3: Model ANOVA Table For One-Way Classification

SV	DF	SS	MS	F-Ratio
(source of variation)	(degrees of freedom)	(sum of squares)	(mean squares)	
Treatments	$k-1$	SS_{tr}	$MS_{tr} = \frac{SS_{tr}}{k-1}$	$\frac{MS_{tr}}{MS_e}$
Error	$N-k$	SS_e	$MS_e = \frac{SS_e}{k(n-1)}$	
Total	$N-1$	TSS		

To prepare the ANOVA table for the gas company example, we need to compute TSS, SS_a and SS_e. We shall now give the formulas for computation of these quantities for the general (unbalanced) case.

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - CF, \tag{5}$$

$$SS_{tr} = \sum_{i=1}^{k} \frac{y_i^2}{n_i} - CF, \tag{6}$$

$$\text{and } SS_e = TSS - SS_{tr}, \tag{7}$$

where CF is the correction factor and is given by $CF = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2}{N} = N\bar{y}^2$ and n_is are as defined in model (1). For the problem in question which is a balanced problem with $n = n_i$ for $i = 1, 2, \ldots, 7$, we have

$$\bar{y} = 13.9291, \ CF = 6790.7357,$$

$$\sum\sum y_{ij}^2 = 6793.5698, \ TSS = 2.8341,$$

$$SS_{tr} = \frac{(99.20)^2 + (96.17)^2 + (97.17)^2 + (98.96)^2 + (96.02)^2}{7} - CF$$
$$= 6792.0343 - 6790.7357 = 1.2986,$$

and
$$SS_e = TSS - SS_{tr} = 2.8341 - 1.2986 = 1.5355$$

The ANOVA table for the problem is given in Table-4 below.

Table-4: ANOVA For GC Example Data

SV	DF	SS	MS	F-Ratio
Agents	4	1.2986	0.3247	6.35
Error	30	1.5355	0.0511	
Total	34	2.8341		

The tabulated value of F with 4 and 30 degrees of freedom at 5% level of significance is equal to 2.69. Since the calculated F-value from the ANOVA table is larger than the

tabulated value, we reject the null hypothesis H_0, inferring thereby that the agents indeed do differ in respect of average weight of gas cylinder. In the case of balanced data above calculation can be simplified by using the special tips given in the following remark.

Remark: In case of balanced data if you compute the standard deviation, then you can compute the sum of squares TSS and SS_{tr} directly without computing the correction factor, CF. To get TSS, input all the y_{ij}s and get their standard deviation (SD). Then TSS is equal to square of this SD multiplied by the number of observations. For the above problem, this standard deviation is equal to 0.28426 and
TSS $= 35 \times (0.28426)^2 = 2.8281$. To get SS_{tr}, compute the SD of the totals (y_is), square it and then multiply it by $\dfrac{k}{n}$. For the above problem, SD of the totals is equal to 1.3483, so that
$SS_{tr} = (1.3483)^2 \times \dfrac{5}{7} = 1.2985$ (the discrepancy is due to rounding of errors). Next, SS_e is obtained by subtraction. Thus, $SS_e = 2.8281 - 1.2985 \doteq 1.5296$.

Here SD is the population SD, not the sample SD.

Remember these special tips apply only if the data are balanced, i.e., all the n_is are equal. Do not apply this to unbalanced data.

Before we continue with gas company example, you must get some practice on preparing the ANOVA table and carrying out the test of hypothesis for equality of means. You can do that while trying the following exercises.

E6) Continuing with E(3), it was reported that 5% of the drive shafts were being rejected due to non conformance of slot length, x, to its specifications. In order to diagnose the problem, ten drive shafts were collected from each of the 8 stations and their lengths were measured. These data are presented in Table 5. To reduce your efforts on computation some summary statistics are also incorporated in the table. If the settings of the special purpose machine were proper, the mean shaft lengths should be same for all the 8 stations. Now do the following.

This study was carried out in a leading automobile industry in Bangalore.

a) Frame the null hypothesis for equality of all means by introducing the necessary notation.

b) State the alternative hypothesis .

c) Prepare the ANOVA table and carry out the test of hypothesis.

Table 5: Slot Lengths (At the Investigation Stage)

S. No.	Station							
	1	2	3	4	5	6	7	8
1	5.85	5.92	5.87	6.01	6.02	5.87	5.94	6.02
2	5.89	5.95	5.91	5.95	5.92	5.90	5.91	6.07
3	5.90	5.91	5.94	5.90	6.00	5.92	5.95	6.00
4	5.92	5.88	5.92	5.93	5.94	5.95	5.96	6.03
5	5.95	5.92	5.98	5.97	5.97	5.96	5.98	6.11
6	5.91	5.92	5.92	5.94	5.98	5.94	5.95	6.00
7	5.88	5.95	5.90	5.96	5.96	5.92	5.88	5.99
8	5.95	5.90	5.89	5.98	6.00	5.91	5.93	6.01
9	5.92	5.93	5.95	5.89	5.94	5.95	5.97	5.98
10	5.84	5.94	5.95	5.97	5.97	5.98	5.98	6.02
Total	59.01	59.22	59.23	59.50	59.70	59.30	59.45	60.23
Average	5.901	5.922	5.923	5.950	5.970	5.930	5.945	6.023
$\sum_{j=1}^{10} y_{ij}^2$	348.2305	350.7052	350.8289	354.037	356.4178	351.6584	353.4393	362.7793

E7) After analysing the data in E6), it was found that the 8 stations were not uniform. Investigation by the concerned personnel revealed that there were discrepancies in the settings of some fixtures used in the stations. So the settings were adjusted and data on the slot lengths were collected again. These data are presented in Table-6.

Problem was rectified and the rejections due to slot length non conformance were eliminated.

Table-6: Slot Lengths (After Corrective Action)

S. No.	Station							
	1	2	3	4	5	6	7	8
1	5.87	5.86	5.97	5.94	5.88	5.93	5.88	5.87
2	5.94	5.93	5.88	5.92	5.93	5.90	5.90	5.92
3	5.92	5.96	5.92	5.90	5.92	5.96	5.92	5.92
4	5.91	5.86	5.89	5.95	5.95	5.95	5.98	5.89
5	5.89	5.92	5.95	5.91	5.91	5.92	5.92	5.95
6	5.95	5.94	5.91	5.91	5.96	5.94	5.93	5.94
7	5.88	5.86	5.94	5.98	5.89	5.92	5.85	5.90
8	5.91	5.88	5.84	5.88	5.92	5.97	5.91	5.87
9	5.97	5.90	5.94	5.93	5.95	5.98	5.94	5.92
10	5.86	5.92	5.92	5.97	5.95	5.95	5.93	5.93
Total	59.10	59.03	59.16	59.29	59.26	59.42	59.16	59.11
Average	5.910	5.903	5.916	5.929	5.926	5.942	5.916	5.911
$\sum_{j=1}^{10} y_{ij}^2$	349.29	348.47	350.01	351.53	351.18	353.08	350.01	349.41

Do (a), (b) and (c) as given in E6).

In the next section we shall consider comparison of pairs of treatment means.

8.5 PAIRWISE COMPARISONS

Having rejected the hypothesis of equality of means, one might wish to compare pairs of treatment means. Though there are several methods to do this, we shall confine our attention to the **least significant difference (LSD)** method.

Supposing we want to compare the treatment means τ_l and τ_m. An unbiased estimator of $\tau_l - \tau_m$ is $\bar{y}_l - \bar{y}_m$ which has variance $(\frac{1}{n_l} + \frac{1}{n_m})\sigma^2$. Furthermore, under the normality assumptions, $\bar{y}_l - \bar{y}_m$ is normally distributed and is independent of MS_e. Consequently, under $H_0 : \tau_l - \tau_m = 0$,

$$t_0 = (\bar{y}_l - \bar{y}_m)/\sqrt{MS_e/(\frac{1}{n_l} + \frac{1}{n_m})} \tag{8}$$

follows t-distribution with ν df where ν is the error df as before. The hypothesis H_0 is rejected if the absolute value of $(\bar{y}_l - \bar{y}_m)$ is larger than the

$$LSD = t_{\alpha/2,\nu} \times \sqrt{MS_e/(\frac{1}{n_l} + \frac{1}{n_m})}.$$

You may recall that in the gas company example, we have rejected the hypothesis that the treatment means are equal. Looking at the averages in Table-2, we have reasons to suspect that agents 2,3 and 5 have possibly different means. Before investigating these agents, let us see if the other two agents 1 and 4, have the same mean. First let us examine whether $\tau_1 = \tau_4$. Since all n_is are equal to 7 in this example, we have LSD = $1.96 \times \sqrt{2 \times 0.051/7} = 0.2366$. Since $\bar{y}_1 - \bar{y}_4 = 14.1714 - 14.1371 = 0.0343$ is less than LSD, we cannot reject the hypothesis that $H_0 : \tau_1 = \tau_4$. We therefore conclude that there is no substantial evidence to suspect that agents 1 and 4 do differ in respect of their means.

It is now time for you to try the following exercises to make sure that you understand what is going on.

E8) Consider the data in Table-5. Frame the hypothesis (ie., write down H_0 and H_1) for each of the following cases and carry out the tests at 5% level of significance.

a) Test whether there is any difference between stations 1 and 6 with respect to mean slot length

b) Do the similar exercise as in (a) for stations 4 and 6.

c) Assuming that there is no difference between stations 4 and 7, test whether the common mean slot length for these two stations is equal to the target (since the specifications are 5.85 to 6.05, the target is the middle point which is 5.95).

d) Test whether the mean slot length for station 3 is significantly lower than the target.

E9) Consider the data in Table-6. Assuming that mean slot length is same for all the 8 stations, test whether slot length is set at the target.

Let us resume our analysis of gas company example and examine the agents 2, 3 and 5. Arranging the 5 averages in the descending order and computing the differences, we get

$$\bar{y}_1 = 14.1714$$
$$\bar{y}_4 = 14.1371 \quad \bar{y}_1 - \bar{y}_4 = 0.0343$$
$$\bar{y}_3 = 13.8814 \quad \bar{y}_4 - \bar{y}_3 = 0.2557$$
$$\bar{y}_2 = 13.7386 \quad \bar{y}_3 - \bar{y}_2 = 0.1428$$
$$\bar{y}_5 = 13.7171 \quad \bar{y}_2 - \bar{y}_5 = 0.0215.$$

Since LSD = 0.2366, we find that there is significant difference between τ_4 and τ_3.

However, pairwise comparisons among τ_2, τ_3 and τ_5 do not show up significant differences. From these analyses we can infer that customers get a bad deal from agents 2,3,and 5.

In many situations, the data are such that the number of observations under a treatment is not the same for all treatments. This might happen due to constraints or due to accidents. For instance, in the fertiliser example (E2), if we only had 18 plots available for the study, then it would not be possible to try the five fertilisers on the same number of plots. Two of the fertilisers may have to be applied on 3 plots each, while the other 3 on four plots each. Or it might be that 20 plots were available for the study and fertilisers were applied on the same number of plots, but due to unforeseen conditions, crops in two of the plots got damaged. In the next section, we give details of the analysis for unbalanced data.

8.6 UNBALANCED DATA

In the unbalanced case, the sum of squares for the ANOVA table can be computed using the formula given by (5), (6), and (7). We shall now work out an example to point out the minor differences in the analysis.

Example 2: An experiment was conducted to determine whether four specific firing temperatures affect the density of a certain type of brick. The data are presented in Table-7 along with some summary statistics.

Table-7: Density of Bricks and Summary Statistics

	Temperature			
	100°C	125°C	150°C	175°C
	21.8	21.7	21.9	21.9
Density	21.9	21.4	21.8	21.7
	21.7	21.5	21.8	21.8
	21.6	21.4	21.6	21.4
	21.7	-	21.5	-
n_i	5	4	5	4
Total, y_i	108.7	86.0	108.6	86.8
$\bar{y}_i$	21.74	21.50	21.72	21.7
$\sum_{j=1}^{n} y_{ij}^2$	2363.19	1849.06	2358.90	1883.70

Let τ_1, τ_2, τ_3 and τ_4 be the effects of temperature on density at $100°C$, $125°C$, $150°C$ and $175°C$ respectively. Let μ be the general effect. To test the hypothesis, $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4$, we compute the sum of squares and form the ANOVA Table. We have

$$\sum_{i=1}^{4}\sum_{j=1}^{n_i} y_{ij} = 108.7+86.0+108.6+86.8 = 390.1, \quad \sum n_i = 18, \quad \bar{y} = \frac{390.1}{18} = 21.6722,$$

$$\sum_{i=1}^{4}\sum_{j=1}^{n_i} y_{ij}^2 = 2363.19 + 1849.06 + 2358.90 + 1883.7 = 8454.85,$$

$$CF = \frac{(390.1)^2}{18} = 8454.3338, \quad TSS = 8454.85 - 8454.3338 = 0.5161,$$

$$SS_a = \sum_{i=1}^{4} \frac{(y_i)^2}{n_i} - CF = \frac{(108.7)^2}{5}+\frac{(86.0)^2}{4}+\frac{(108.6)^2}{5}+\frac{(86.8)^2}{4}-8454.3338 = 0.1562,$$

and $SS_e = TSS - SS_a = 0.5161 - 0.1562 = 0.3599$. We can now write down the ANOVA Table.

Table-8: ANOVA for the Density Data

SV	DF	SS	MS	F
Temperature	3	0.1562	0.0520	2.02
Error	14	0.3599	0.0257	
Total	17	0.5161		

Remember that the df for error is always equal to total number of observations minus the number of treatments.

Since the observed F-ratio, 2.02, is less than the tabulated F-value at 5 % confidence level with 3 and 14 df which is 3.34, we cannot reject the null hypothesis.

Let us remember that not being able to reject H_0 does not mean that the treatment means (τ_is) are equal. Let us test if τ_1 and τ_2 are significantly different. In other words, we wish to test

$$H_0 : \tau_1 = \tau_2 \qquad Vs \qquad H_1 : \tau_1 \neq \tau_2.$$

We shall now compute t_0 given by Eqn.(8) which has t-distribution with 14 df. Substituting the values, we get $t_0 = 2.231$, whereas the tabulated $t_{0.025, 14}$ is equal to 2.145 (See Table-2 in the Appendix given at the end of the block.). Hence, we conclude that there is significant difference (at 5% level of significance) between τ_1 and τ_2.

* * *

You may now try the following exercises to reinforce your learning for the unbalanced case.

E10) In the above example, test if τ_2 and τ_3 are significantly different (at 5% level).

E11) A tailoring shop owner has 3 tailors, A, B and C, working under him who stitch only men's shirts. During a particular week, the owner tried to study their efficiencies with regard to productivity and obtained the following data.

Table-9 : Tailors' Productivity Data

Day	Number of shirts stitched per shift		
	A	B	C
1	6	7	9
2	8	8	8
3	6	7	*
4	5	5	6
5	*	8	*

* Tailor was on leave.

Prepare a summary table, present the ANOVA table and test (at 5% level) whether all the 3 tailors are equally productive.

In the following section, we will acquaint you with a concept known as random effects. Our purpose here is not give you the details but to make you understand broadly the differences between fixed effects and random effects. Again here, we shall concentrate on the balanced data.

8.7 RANDOM EFFECTS MODEL

In all the examples that we have seen so far in this unit, the treatment levels were all fixed. In the gas company example, the agents were five specific agents. They are not picked up at random from a group of agents. Similarly, in the fertiliser example, we were interested in the effects of five specific fertilisers. Since the treatment levels are fixed, we assumed that their effects, τ_is, are also fixed constants. Therefore, these effects are called **fixed effects** and in this case the model (1) is called a **fixed effects model.**

In some situations, the levels of treatment are picked up at random and then the data are collected at these random levels. As an example, let us once again look at the gas company example, but this time we will confine ourselves to the factory where the cylinders are refilled. One such factory has 500 filling stations. Cylinders are refilled simultaneously at each of these stations.

One of the things that the quality control engineer has to ensure is that the stations are homogeneous. That is, he must ensure that the mean amount of gas filled by a station is the same for all the stations. To do this, the engineer selects, from time to time, 5 stations at random, and from each station he picks up five cylinders, again at random, and measures the amount of gas in each of these cylinders. One such set of data is presented in Table-10.

Table-10 : Gas Weights Data

S.No.	Station Number				
	5	47	193	301	398
1	14.24	13.92	14.16	14.19	14.06
2	14.24	13.95	14.18	14.20	14.05
3	14.26	13.96	14.17	14.19	14.06
4	14.25	13.97	14.14	14.19	14.03
5	14.26	13.95	14.15	14.21	14.06
$\sum y_i$	71.25	69.75	70.8	70.98	70.26
$\sum \bar{y}_i$	14.25	13.95	14.16	14.2	14.05
$\sum y_{ij}^2$	1015.3129	973.0139	1002.529	1007.6324	987.2942

Let τ_i be the effect of station i on the gas weight. Since stations are selected at random, we may assume that τ_is are random variables. **Note that** the data are one-way classification data. We use the model given in Eqn.(1) in this case also but interpret the τ_is as random variables. In fact, we assume that the τ_is are iid with mean 0 and variance σ_τ^2. The ϵ_{ij}s play the same role as in the fixed effects model. The new model with τ_is as random variables is called the **random effects model.**

The engineer tries to test the homogeneity of all the 500 stations. This is done by testing

$$H_0 : \sigma_\tau^2 = 0 \qquad vs \qquad H_1 : \sigma_\tau^2 > 0.$$

The test procedure is exactly the same as the one for the fixed effects model. That is, prepare the ANOVA table and check if the observed F-value is larger than the tabulated F-value. If this is the case, then reject H_0 and infer that the stations are not homogeneous.

In random effects model, σ^2 and σ_τ^2 are called the **variance components**. In case of the **balanced data**, they can be estimated unbiasedly by

$$\hat{\sigma}^2 = MS_e \qquad and \qquad \hat{\sigma}_\tau^2 = \frac{MS_{tr} - MS_e}{n}.$$

Now let us analyse the data of Table-10.

We have

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_i = 353.04, \quad \sum n_i = 25, \quad \bar{y} = 14.1216$$

$$\sum_{i=1}^{5} \sum_{j=1}^{5} y_{ij}^2 = 4985.7824$$

$$C.F. = \frac{(353.04)^2}{25} = 4985.4897$$

$$TSS = 0.2927$$

$$SSa = \sum \frac{y_i^2}{n_i} - CF$$

$$= 4985.7786 - 4985.4897$$

$$= 0.2889$$

$$SSe = TSS - SSa$$

$$= 0.0038$$

The ANOVA table is given below.

Table-11: ANOVA For Gas Weights Data

SV	DF	SS	MS	F
Stations	4	0.2889	0.07223	380.16
Error	20	0.00380	0.00019	
Total	24	0.29273		

We see that the observed F-value is very large while the tabulated F-value at 5 % level of significance is equal to 2.87. If you observe the data in Table-10, you find there is some problem with station 47. Probably there was some setting problem or something else. Since the 5 stations are selected at random from the 500 stations, we conclude that there may be problems with other stations as well. So this calls for further investigation.

The estimates of σ^2 and σ_τ^2 are

$$\hat{\sigma}^2 = 0.00019 \qquad and \qquad \hat{\sigma}_\tau^2 = \frac{0.07223 - 0.00019}{5} = 0.0144.$$

The engineer's aim must be to reduce σ_τ^2 as much as possible by improving the uniformity among the stations. This might perhaps be achieved by setting the faulty stations right.

And now an exercise for you.

E12) In textile mill there are 300 looms. It is known that the performance of the looms affect the strength of the fabric. Four looms were selected at random, and three samples were tested from the fabric produced from each of these looms. Test if the looms in the company are homogeneous. Estimate the variance components.

Table-12 : Fabric Strength Data

Loom	Observations				Average
1	98	98	99	96	97.75
2	91	90	93	92	91.50
3	96	95	97	95	95.75

We now end this unit by giving a summary of what we have covered in it.

8.8 SUMMARY

In this unit we have learnt the following important points.
1) In real life situations, we often encounter problems in which we need to study several populations.

2) Analysis of variance is a useful technique to test the equality of means of several populations and carry out subsequent analysis.

3) One-way classification data arise when we are interested in studying the effect of a single source of variation on a variable.

4) Building a linear model for the one-way classification data.

5) Preparation of the ANOVA table and testing the equality of all treatment means.

6) Estimation of model parameters.

7) The least significant difference method for comparison of all pairs of treatment means.

8) Analysis of the unbalanced one-way classification data.

9) The concept of random effects model.

8.9 SOLUTIONS/ANSWERS

E1) In Unit 6, the tests of hypotheses were concerning either one population mean being equal to a constant or equality of two population means. That is, they were of the type $H_0 : \mu = 14.2$ or $\tau_1 = \tau_2$. In gas company example, the hypothesis is concerning five populations, that is, $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5$.

E2) Here fertilisers will be the main source of variation of interest. If the fertility of the soil of the 20 plots varies widely, then the plots should also be considered as a major source of variation. However, since it is given that all the twenty plots belong to the same location, it is expected that their effect on the yield is same. In this case, the plots may not be treated as a major source of variation.

E3) The main source of variation of interest here is 'stations'. We should be, in the first place, interested in examining if the output from all the eight stations is uniform. Besides stations, if the machine is operated by different operators at different times, then the operators could be another source of variation.

E4) Let y_{ij} be the crop yield of jth plot on which fertiliser i is applied. The general mean μ may be interpreted as the average yield per plot if no fertiliser is applied. Let τ_i be average additional yield per plot if fertiliser i is applied. Then the

expected value of y_{ij} is $\mu + \tau_i$. Since y_{ij}s are subjected to random fluctuations, we can write $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $j = 1, 2, .., 4$, $i = 1, 2, .., 5$, where ϵ_{ij}s are the deviations from the expected means. The variation in ϵ_{ij}s is due to several causes that are common to all yields. For these reasons, ϵ_{ij}s are assumed to have mean zero and variance σ^2. Therefore, the linear model is given by

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad j = 1, 2, .., 4, \ i = 1, 2, .., 5,$$

where ϵ_{ij}s are iid random variables with mean 0 and variance σ^2.

E5) We can write

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

squaring both sides and summing on i and j, we have

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2$$

$$+ 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \ldots \qquad (9)$$

Now

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = \sum_{i=1}^{k} (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$$

since $\bar{y}_i$ is the mean of the ith sample, and hence $\sum_{j=1}^{n} (y_{ij} - \bar{y}_i) = 0 \ \forall \ i$

Also, in (Eqn.9) above

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

E6) (a) Let μ be the general mean and let τ_i be the effect of station i on x. Let x_{ij} be the slot length of jth drive shaft from i^{th} station, $j = 1, 2, .., 10$, $i = 1, 2, ..8$. The linear model is

$$x_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

The null hypothesis is $H_0 : \tau_1 = \tau_2 = \ldots = \tau_8$.

(b) The alternative hypothesis is

$$H_1 : \tau_l \neq \tau_m \text{ for some l and m.}$$

(c) $N = 80$, $\bar{x} = \dfrac{59.01 + 59.22 + .. + 60.23}{8} = 5.9455,$
$CF = 80 \times (5.9455)^2 = 2827.918$, $\sum \sum x_{ij}^2 = 2828.096$,
$TSS = 2828.096 - 2827.918 = 0.178$,
$SS_a = \dfrac{(59.01)^2 + .. + (60.23)^2}{8} - 2827.918 = 0.099.$

Table-13: ANOVA Slot Length Data

SV	DF	SS	MS	F
Stations	7	0.099	0.014	14
Error	72	0.079	0.001	
Total	79	0.178		

The tabulated F-value with 7 and 72 df is 2.13. Therefore, H_0 is rejected and we conclude that station means are not equal.

E7) Answers to (a) and (b) are same as in E6)

(c) $N = 80$, $\bar{x} = \dfrac{59.1 + 59.03 + .. + 59.11}{8} = 5.9191$,

$CF = 80 \times (5.9191)^2 = 2802.883$, $\sum\sum x_{ij}^2 = 2802.97$,

$TSS = 2802.97 - 2802.883 = 0.087$, $SS_a =$

$\dfrac{(59.1)^2 + .. + (59.11)^2}{8} - 2802.883 = 0.011$.

Table-14: ANOVA Slot Length Data

SV	DF	SS	MS	F
Stations	7	0.011	0.0015	1.5
Error	72	0.076	0.0010	
Total	79	0.087		

The tabulated F-value with 7 and 72 df is 2.13. Therefore, we cannot reject H_0. There is no substantial evidence to conclude that station means are different.

E8) (a) $H_0 : \tau_1 = \tau_6$ Vs $H_1 : \tau_1 \neq \tau_6$.

$LSD = t_{0.025, 72} \times \sqrt{2 \times MS_e / 10} = 1.99 \times \sqrt{2 \times 0.001 / 10} = 0.028$.

Since the difference between $\bar{x}_1$ and $\bar{x}_6$ ($= 0.290$) is more than LSD, there is significant difference between the effects of station 1 and station 6.

(b) $H_0 : \tau_4 = \tau_6$ Vs $H_1 : \tau_4 \neq \tau_6$.

Since the difference between $\bar{x}_4$ and $\bar{x}_6$ ($= 0.200$) is more than LSD, we reject H_0 and conclude that the effects of station 4 and station 6 are significantly different.

(c) Assuming $\tau_4 = \tau_7 = \tau$, the estimate of $\mu + \tau$ is given by

$$u = \frac{\bar{x}_4 + \bar{x}_7}{2} = 5.9475.$$

The hypothesis to be tested is

     $H_0 : \mu + \tau = 5.95$ Vs $H_1 : \mu + \tau \neq 5.95$.

Under normality assumptions, $\mu + \tau$ follows normal distribution with mean $\mu + \tau$ and variance $\sigma^2 / 20$. Under H_0,

$$t_0 = \frac{u - 5.95}{\sqrt{MS_e / 20}}$$

follows t-distribution with 72 df. Substituting the values, we get $t_0 = 0.353$. Since this is less than the tabulated t-value, we conclude that there is no substantial evidence to say that $\mu + \tau$ is away from the target.

(d) Let $\mu_3 = \mu + \tau_3$. We wish to test

     $H_0 : \mu_3 = 5.95$ Vs $H_1 : \mu_3 < 5.95$.

Here we use the one-sided t-test and reject H_0 if

$$t_0 = \frac{\bar{x}_3 - 5.95}{\sqrt{MS_e / 10}}$$

is less than the lower 5% point of t-distribution with 72 df. Substituting the values,

$$t_0 = \frac{5.923 - 5.95}{\sqrt{0.001 / 10}} = -2.7.$$

Since the tabulated t-value is -1.66, we reject H_0 and conclude that μ_3 is significantly lower than the target.

E9) (a) When all τ_is are equal, say equal to τ, $\bar{x}$ is an estimate of $\mu + \tau$ and has normal distribution with mean $\mu + \tau$ and variance $\frac{\sigma^2}{80}$. To test

$$H_0 : \mu + \tau = 5.95 \qquad Vs \qquad H_1 : \mu + \tau \neq 5.95$$

we compute the t-statistic

$$t_0 = \frac{\bar{x} - 5.95}{\sqrt{MS_e/80}} = \frac{5.919 - 5.95}{0.0035} = -8.76.$$

Conclude that the special purpose machine setting is significantly lower than the target value.

E10) To test

$$H_0 : \tau_2 = \tau_3 \qquad Vs \qquad H_1 : \tau_2 \neq \tau_3$$

compute

$$t_0 = \frac{21.5 - 21.72}{\sqrt{0.3599/(\frac{1}{4} + \frac{1}{5})}} = -0.275.$$

Since the tabulated t-value is 2.145, there is no substantial evidence to reject H_0.

E11)

Table-15 : Summary of Tailors' Productivity Data

Summary	Tailor			Total
	A	B	C	
n_i	4	5	3	12
Total	25	35	23	83
Average	6.25	7	7.66	21.41
$\sum y_{ij}^2$	151	251	161	593

$n = 12$, $CF = 574.0833$.

Table-16: ANOVA For Tailors' Data

SV	DF	SS	MS	F
Tailors	2	3.50	1.75	1.5
Error	9	15.4166	1.7129	
Total	11	18.9166		

Tabulated F-value with 2 and 9 df at 5% level is equal to 4.256. There is no substantial evidence to say that the the tailors are not equally productive.

Table-17: ANOVA For Fabric Strength Data

SV	DF	SS	MS	F
Tailors	2	81.5	40.75	29.34
Error	9	12.5	1.3888	
Total	11	94.0		

The estimates of variance components are given by

$$\hat{\sigma}^2 = 1.3888 \text{ and } \sigma_\tau^2 = \frac{40.75 - 1.3888}{4} = 9.8403.$$