5.0 Introduction

In Chapter 2 we used two-sample t-tests based on sample data from two groups to decide whether two population means were the same or different. In this chapter, we extend the analysis to situations where there are more than two groups. First, we examine a statistical method which poses the question: are all the population means the same, or is at least one of them different from the others? If differences exist, then, in order to identify which groups are different, we carry out tests similar to two-sample t-tests which allow for the fact that many group comparisons, instead of a single comparison, are to be made. The method we use to make the overall comparison is called ‘Analysis of Variance’ or ANOVA for short.

Why not go straight to the two-sample comparisons? ANOVA is a general approach to data analysis and it is convenient to introduce it here in the simplest context of extending the analysis of two-sample studies to comparisons of more than two groups. It becomes the natural method of analysis for more complex structures. For example, suppose an occupational psychologist wanted to study training methods for overcoming resistance to the uptake of new technologies. Her study might involve employees of both sexes, four age/experience groups, and three training methods (these are referred to as ‘factors’ in ANOVA and her study would be described as a ‘three-factor study’). This means that 24 different subject-type/treatment-group combinations would be studied, which allows for \( \binom{24}{2} = 276 \) possible pairwise comparisons – a lot! The research questions will often not be posed in terms of comparisons between selected pairs of the set of study groups. Here, for example, the psychologist will very likely ask if there are differences between the sexes, and if these differences depend on age group; are the training methods equally effective for all age groups, and is any age dependency itself dependent on sex, and so on. Such questions are answered in a natural way using ANOVA. In this chapter, we will not arrive at the analysis of such multi-factor structures, but the introduction given here, to the simplest form of ANOVA, will prepare you to develop your knowledge of the methods, should you need them for your work.

5.1 Example 1: A laboratory comparison study

A multinational corporation makes adhesive products in which boron is an important trace element at the parts per million (ppm) level. Concerns had been expressed about the comparability of the analytical results for boron content produced by different laboratories in the corporation. The laboratories all use ICP-AES (ICP = inductively

\[ \text{Example 1 and 4 are based directly on Chapter 7 of Eamonn Mullins, Statistics for the Quality Control Chemistry Laboratory, Royal Society of Chemistry, Cambridge, 2003.} \]
coupled plasma spectrophotometer, AES = atomic emission spectrometer) systems for the analysis, but these systems are not all identical in their configurations.

<table>
<thead>
<tr>
<th></th>
<th>Lab 1</th>
<th>Lab 2</th>
<th>Lab 3</th>
<th>Lab 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st value</td>
<td>4.9</td>
<td>5.4</td>
<td>5.8</td>
<td>4.5</td>
</tr>
<tr>
<td>2nd value</td>
<td>5.7</td>
<td>5.5</td>
<td>6.0</td>
<td>4.9</td>
</tr>
<tr>
<td>3rd value</td>
<td>5.1</td>
<td>4.8</td>
<td>6.0</td>
<td>4.7</td>
</tr>
<tr>
<td>4th value</td>
<td>5.3</td>
<td>4.9</td>
<td>5.5</td>
<td>4.7</td>
</tr>
<tr>
<td>5th value</td>
<td>5.4</td>
<td>5.2</td>
<td>5.9</td>
<td>4.4</td>
</tr>
<tr>
<td>6th value</td>
<td>5.5</td>
<td>5.4</td>
<td>5.8</td>
<td>4.8</td>
</tr>
<tr>
<td>Mean</td>
<td>5.32</td>
<td>5.20</td>
<td>5.83</td>
<td>4.67</td>
</tr>
<tr>
<td>SD</td>
<td>0.29</td>
<td>0.29</td>
<td>0.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 5.1.1 Replicate measurements of the boron content of an adhesive (ppm).

An inter-laboratory study was conducted which involved a large number of laboratories each measuring several products over a period of weeks. Table 5.1.1 shows results for one product from four selected laboratories; the six replicates were measured on different days, so the variation between them reflects medium term chance measurement error within the laboratories. Figure 5.1.1 shows dotplots for the laboratory data: without any formal statistical analysis it is obvious that the average results in laboratories 3 and 4 are different; it is not clear, however, whether laboratories 1 and 2 differ by more than the chance analytical day-to-day variation which is clearly present in all laboratories.
Measuring variation

Variation between observations is usually measured by the standard deviation or, alternatively, by its square, the variance. Thus, suppose we obtain a random sample of \( n \) observations, \( y_1, y_2, y_3, \ldots, y_n \), from some population or process whose mean is \( \mu \) and whose variance is \( \sigma^2 \); the sample variance (the square of the standard deviation) is defined as:

\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.
\]  

(5.1)

This quantity, which estimates the unknown \( \sigma^2 \), is the average of the squared deviations of the data, \( y_i \), from their overall mean, \( \bar{y} \), where the divisor is \( n-1 \), the ‘degrees of freedom’. The term ‘degrees of freedom’ is used because only \( n-1 \) of the deviations \( (y_i - \bar{y}) \) are free to vary: they are constrained by the fact that they must sum to zero. Thus, if \( n-1 \) deviations sum to some arbitrary value, the last deviation is, necessarily, equal to minus that value.

When the data collection mechanism is more complicated than that assumed above, the method known as Analysis of Variance (ANOVA) may be used to break up both the total sum of squares (the numerator of 5.1) and the total degrees of freedom (the denominator of 5.1) into components associated with the structure of the data collection process.

The basic idea underlying the simplest form of ANOVA is that the variation between individual observations can be viewed as having two components: within-group and between-group. The within-group component is seen as pure chance variation – within-group conditions are held constant and there are no reasons why any two observations should be different, apart from the influence of the chance variation that affects the system being studied. In the case of the laboratories study, the same material was measured on each of six days using the same equipment, the same glassware, the same reagents etc., by the same team of analysts, within each laboratory. The key word here is ‘same’.

Between-group variation, on the other hand, is seen as (at least potentially) systematic: different groups are different, or are treated differently, in ways that may lead to higher or lower average responses. In the case of the laboratories, the equipment used was either different (made by different manufacturers) or was set up in different ways; there were different teams of analysts in the four laboratories, which were in different parts of the world with markedly different climates. There were, as a result, many possible factors which could lead to different average responses in the four laboratories.

---

\(^2\) If this is not obvious, check that it is true for a simple example, say three values 6, 7, 8.
ANOVA splits the variation, as measured by the corresponding sums of squares, into components associated with these two categories (between and within-group) and then averages the sums of squares, by dividing by the correspondingly partitioned degrees of freedom (see below). This gives two mean squares, which correspond directly to expression (5.1) – they are sample variances associated with within-group and between-group variation. If the between-group mean square is not substantially bigger than the within-group mean square, which measures purely chance variation, then there is no reason to believe that there are systematic differences between the responses of the different groups. The statistical test for between-group differences is based on the ratio of the two mean squares: MS(Between-group)/MS(Within-group) – if this is large, it suggests systematic between-group differences.

We now turn our attention to the technical details.

**Decomposing Sums of Squares**

A slightly more elaborate notation, which reflects the within-group/between-group data structure, is helpful when dealing with data from several groups. Thus, for the data of Table 5.1.1, each observation $y_{ij}$ is uniquely labelled by two subscripts: i refers to the group (laboratory) in which the observation was generated ($i=1,2,3,4$) while j labels the replicates within laboratory ($j=1,2,\ldots,6$). For example, $y_{13}=5.1$ and $y_{31}=5.8$ in Table 5.1.1. Figure 5.1.2 shows schematically how the deviation of any observation $y_{ij}$ from the overall mean $\bar{y}$ can be regarded as the sum of two components – its deviation from its own group mean $^3(y_{ij} - \bar{y}_i)$ and the deviation of its group mean from the overall mean $(\bar{y}_i - \bar{y})$. Thus we can write:

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}). \quad (5.2)$$

---

$^3$ The mean of the six results in laboratory i is $\bar{y}_i$; the dot indicates that we have summed over the j subscript; the bar that we have averaged, i.e., $\bar{y}_i = \frac{1}{6} \sum_{j=1}^{6} y_{ij}$. In general there are I groups and J replicates within each group. It is not necessary that the number of replicates should be the same for each group, but adjustments to the calculations are required if this is not the case; the statistical software will take care of these.
When both sides of equation (5.2) are squared and summed over all the data points, the cross-product of the two terms on the right-hand side sums to zero and we get:

\[
\sum_{\text{all data}} (y_{ij} - \bar{y})^2 = \sum_{\text{all data}} (y_{ij} - \bar{y}_{i.})^2 + \sum_{\text{all data}} (\bar{y}_{i.} - \bar{y})^2
\]  

(5.3)

\[
\text{SS(total)} = \text{SS(within groups)}^4 + \text{SS(between groups)}  
\]

(5.4)

where SS stands for ‘sum of squares’.

For the laboratory study the decomposition of sums of squares is:

\[
5.2996 = 1.1750 + 4.1246
\]

---

^4 Note that statistics packages almost invariably refer to the component associated with purely random variation as the ‘error’ component. In the current context this is the within-group component, so ‘within-group’ and ‘error’ sums of squares and degrees of freedom will be used interchangeably.
Decomposing Degrees of Freedom

The degrees of freedom can be decomposed in a similar manner. Degrees of freedom (df) in ANOVA are typically calculated as the number of objects being considered minus one. Thus, the df(total) is \( IJ - 1 \), where there are \( J \) replicate results in each of \( I \) groups. For the laboratory study this is \( 24 - 1 = 23 \). There are \( I = 4 \) groups, so there are \( I - 1 = 4 - 1 = 3 \) degrees of freedom for the SS(between groups). Within each group there are \( J = 6 \) observations, so there are \( J - 1 = 6 - 1 = 5 \) degrees of freedom. When the within-group degrees of freedom are combined for the \( I = 4 \) groups, we get \( df(\text{within groups}) = I(J - 1) = 4(6 - 1) = 20 \).

In summary, for the degrees of freedom decomposition we have:

\[
df(\text{total}) = df(\text{within groups}) + df(\text{between groups})
\]

(5.5)

\[
IJ - 1 = I(J - 1) + (I - 1).
\]

\[
24 - 1 = 4(6 - 1) + (4 - 1).
\]

The sums of squares and degrees of freedom are usually presented in an Analysis of Variance table – this is shown in Table 5.1.2, page 8. When the sums of squares are divided by the degrees of freedom, we get ‘Mean Squares’, as shown in column 4 of the table. Mean squares are sample variances of exactly the same form as a single sample variance (see page 3, expression 5.1). Columns 2-4 of the table simply summarise the calculations required to split up the data into elements associated with within-group and between-group variation. To understand the implications of the last three columns, we need a statistical model for our data.

The Statistical Model

The decomposition of sums of squares discussed above reflects an underlying statistical model for the data. This model assumes that:

- the data are generated independently both within and between the various groups;
- the data come from distributions with the same standard deviation, say \( \sigma \);
- data variation within-group is Normal in all cases;
- the long-run group means means, \( \mu_i \), may vary from group to group.
The model is illustrated schematically in Figure 5.1.3.

![Figure 5.1.3: The ANOVA for four (arbitrary) groups](image)

Note that this model is identical to that which underlies the two-sample t-test analysis of Chapter 2: data independence, within-group Normality, common standard deviation, and possibly different means. The only new aspect here is that the number of groups will typically be greater than two.

Clearly, it could be that the (long-run) chance variation could be different in different groups. Such a situation is more likely to arise in a purely observational study (such as the laboratory study, or for the many types of observational data that are collected in the social sciences), than in experimental studies. Where subjects or experimental units are randomly assigned to the experimental treatments the within-group variability will be determined largely by the many non-controlled factors affecting the responses, and random allocation will tend to balance these. There is, of course, the possibility that one or more of the treatments could change the response variability as well as changing the mean. If this is the case, then our simple model will not be correct. Our reaction to this might be to transform the data (e.g., by taking logarithms) to create constant standard deviation, and then carry out the analysis in the transformed scale. Alternatively, we might have to turn to a more sophisticated analysis which allows for different variability in different groups. Here we focus only on the the simplest case, i.e., constant standard deviation.

Later, we will consider ways of checking that our model assumptions are correct. For the laboratory data, the assumptions appear reasonable, as we will see, so we may proceed with our analysis.
The ANOVA Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Squares</th>
<th>Expected Mean Squares</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>4.1246</td>
<td>3</td>
<td>1.3749</td>
<td>$\frac{\sum (\mu_i - \mu)^2}{I - 1}$</td>
<td>23.40</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Within-groups or Error</td>
<td>1.1750</td>
<td>20</td>
<td>0.0588</td>
<td>$\sigma^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.2996</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1.2 The ANOVA table for the laboratory data

Columns 2 and 3 of Table 5.1.2 give the sums of squares and degrees of freedom, respectively, associated with between-group and within-group variation. Column 4 gives the 'mean squares', i.e., column 2 divided by column 3. The mean squares are sample variances and the fifth column of the table shows what their 'expectations' are. Expectation is the statistical terminology for long-run average values. Thus, the expression $E(y_{ij}) = \mu_i$ indicates that the observations in group $i$ vary around a long-run mean $\mu_i$.

Within Groups

In column 5, we see that the expected value for MSE is $\sigma^2$, $E(MSE)=\sigma^2$, which indicates that the sample mean square error (i.e., within-group mean square) has as its long-run average the within-group variance $\sigma^2$; this measures the chance day-to-day measurement variation in each laboratory, and is assumed to be the same for all four laboratories. This suggests that we should use MSE when we want to estimate $\sigma^2$. 
Between Groups

Column 5 of Table 5.1.2 shows that the expectation of the mean square between-groups is $\sigma^2 + \frac{6}{4-1} \sum_{i=1}^{4} (\mu_i - \mu)^2$. The quantity $\mu$ that appears in the expectation is simply the average of the four long-run group means, $\mu_i$:

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4}. \quad (5.6)$$

Note that, although in calculating the between-groups mean square we are explicitly trying to quantify the variation between the different group means, in fact, the quantity we calculate includes in its expectation a measure ($\sigma^2$) of within-group chance variation, also. This results from the fact that the sample means are the averages of the observed values, each of which is subject to within-group chance variation, as well as possible between-group variation.

The expected value of the mean square between groups depends on the sum of the squared deviations of the long-run group means around their average, $\frac{1}{2} \sum_{i=1}^{l} (\mu_i - \mu)^2$. If all the group means are the same ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$) then this sum will be zero. If they are not all the same, this quantity will be positive (due to the squaring); the greater the differences between the means ($\mu_i$) the larger the term, $\frac{1}{2} \sum_{i=1}^{l} (\mu_i - \mu)^2$ will become.

The outcome of the above discussion is that column 5 provides the basis on which a statistical test for the equality of the long-run means ($\mu_i$) is constructed.

The F-test

The last two columns of Table 5.1.2 describe a test of the null hypothesis of no differences between the long-run group means, i.e., that there are no relative biases between the laboratories. The null and alternative hypotheses for this test are specified as:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

$$H_1: \text{not all laboratory means } (\mu_i) \text{ are the same.}$$

The appropriate test statistic is the F-ratio:

$$F = \frac{MS(\text{between-groups})}{MS(\text{within-groups})} = \frac{MS(\text{between-groups})}{MS(\text{error})}.$$
If the null hypothesis is true, then both the numerator and denominator of the F-ratio have the same expectation, \( \sigma^2 \), (since all the \((\mu_i - \mu)\) terms are zero when the null hypothesis is true) and the ratio should be about one.

In practice, the ratio is subject to chance variation. When the null hypothesis is true, the sample F ratio has as its sampling distribution an F-distribution with degrees of freedom corresponding to those of the numerator and the denominator, i.e., 3 and 20, respectively, for the current example.

![Figure 5.1.4: The F-distribution with 3 and 20 degrees of freedom](image)

A large F-ratio suggests that the term \( \sum_{i=1}^{I} (\mu_i - \mu)^2 \) is non-zero, i.e., that not all the long-run group mean (laboratory) deviations are zero and, hence, in this case, that not all the laboratory measurement systems vary around the same mean \((\mu)\). Accordingly, the null hypothesis should be rejected. If a significance level of \( \alpha=0.05 \) is selected, then the critical value will be \( F_c=3.1 \), as shown in Figure 5.1.4 (Table ST-3 is a table of critical values for the F-distribution). Since the F-value in the ANOVA table (F=23.4) is larger than this, we reject the hypothesis of no relative biases and conclude that not all the laboratories produce comparable results. Statistical packages usually produce a p-value associated with the F-value - the p-value corresponding to our F-ratio of 23.4 is less than 0.0005. The p-value has its usual interpretation: the probability of obtaining a sample F value of 23.4 or larger would be less than 0.0005, if the null hypothesis were true. Our study has produced a highly statistically significant result.

Note that for the F-test in this case (as opposed to that of Chapter 2, where two sample standard deviations were compared) the rejection region is in the right-hand tail only. This is so because the \( \sum_{i=1}^{I} (\mu_i - \mu)^2 \) terms are squared in the expression for the between-
groups expected mean square; hence, the numerator of the F statistic is expected to be bigger than the denominator and the F-ratio is expected to be bigger than 1, when there are differences between the groups. A small value of F would be regarded a purely chance event.

Note that the sample size is the same for each group in this and all the other examples of the chapter. This is purely for simplicity of presentation – the calculations are a bit messier if the number of observations changes from group to group, but conceptually nothing changes. Suitable software (such as Minitab) allows for this. In multi-factor designs, where different sized groups of subjects or experimental units are studied, the non-constant sample size can cause problems. Such issues are, however, outside the scope of the current discussion.

**Model validation**

The F-test has established that not all the laboratory means are the same, but we still need to establish which ones are different and by how much they differ from each other. Before doing this, we will turn our attention to checking that the assumptions underlying the analysis are valid. If they are valid, then it makes sense to carry on to a more detailed examination of the data. If they are not valid, then the F-test may not be valid and the conclusions drawn from it will be suspect. The validation is based on the analysis of residuals, as was the case for two-sample t-tests.

When the laboratory means (the fitted values) are subtracted from the corresponding data points within each laboratory, we are left with the residuals – by construction these vary about a mean of zero. If our model assumptions of Normality and constant within-laboratory standard deviation are correct, these residuals should have the properties that would be expected from four random samples from a single Normal distribution, with zero mean.

Figure 5.1.5 shows the residuals plotted against the fitted values. The spread of values is not identical for all laboratories, but this could be the result of the purely chance variation which is inevitably present due to measurement error; here, there is no reason to believe that the variation is other than random. Note, in particular, that there is no tendency for the spread of values to increase with the magnitude of the results. The assumption of constant standard deviation appears reasonable.\(^5\)

---

\(^5\) Formal tests for the equality of the (long-run) standard deviations are available in Minitab and other statistical packages.
To assess the Normality assumption, we combine all four sets of residuals into a single set – this is valid as the four sets of residuals have the same mean (zero) and a common standard deviation, if our reading of Figure 5.1.5 is correct. This will give us a single plot and Anderson-Darling test, based on 24 values, which will be much more powerful than four separate analyses, each based on only six values. Figure 5.1.6 shows a Normal plot of the residuals; the plotted points are close to a straight line, as would be expected from Normal data. The Anderson-Darling test gives a p-value of 0.192 which supports the assumption that the data come from a Normal distribution.
Figures 5.1.5 and 5.1.6, although based on only relatively small numbers of data points, give us some assurance that the statistical model on which our ANOVA F-test was based is acceptable. The same assumptions are required for the tests and confidence intervals discussed next.

Comparing Means - Least Significant Difference

The F-test is a global test that asks if all the long-run group means are the same or not. Once, as here, the F-test indicates that differences exist, we need to carry out some further analysis to investigate the pattern of differences. A natural approach to comparing any pair of sample means is to carry out a t-test of the hypothesis that their long-run values do not differ. We saw in Chapter 2 that the difference between any pair of means \( \bar{y}_i \) and \( \bar{y}_j \) is statistically significant if:

\[
t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{\frac{2s^2}{n}}} > t_c
\]

that is if

\[
\bar{y}_i - \bar{y}_j > t_c \sqrt{\frac{2s^2}{n}} \tag{5.7}
\]

where the sample size is \( n \) in each case, \( s^2 \) is the combined within-group variance, \( t_c \) is the critical value for a two-sided test with a significance level of \( \alpha=0.05 \), and \( \bar{y}_i > \bar{y}_j \).

Thus, the quantity on the R.H.S. of (5.7) is the ‘least significant difference’ (LSD) – the smallest difference between two sample means that will be declared statistically significant.

For our laboratory comparison study we have four means, each based on \( J=6 \) observations. We have a combined within-laboratory variance of \( s^2=MSE=0.0588 \), which is based on all \( I=4 \) laboratories. We use this as our estimate of \( \sigma^2 \); thus, when we compare two means, say for labs 1 and 2, we ‘borrow’ degrees of freedom from the other laboratories. The degrees of freedom for the t-distribution used to determine the critical values for the t-test are the same as those for \( s^2=MSE \), and so they are 20 in this case. Thus, by using MSE, rather than \( s^2 = (s_1^2 + s_2^2) / 2 \) (based on the two laboratories being compared), we gain an extra 10 degrees of freedom for the test and it becomes, accordingly, more powerful.

As stated above, the quantity on the right-hand side of expression (5.7) is the smallest difference between two sample means which will lead us to conclude that the corresponding population means should be considered different. Note that since \( \sigma^2 \) is assumed the same for all groups and each mean is based on \( J=6 \) observations, the LSD applies to all possible comparisons between pairs of means, not just to laboratories 1 and 2. Here we have:
Least Significant Difference (LSD) = $t_c \sqrt{\frac{2MSE}{J}} = 2.09 \sqrt{\frac{2(0.0588)}{6}} = 0.29$.  \hfill (5.8)

The six possible comparisons between the four laboratory means can be carried out very quickly using the LSD.  In Figure 5.1.7, below, the results are shown in ascending order of magnitude and a line is drawn under pairs of means that do not differ by at least 0.29, i.e., that are not statistically significantly different.

<table>
<thead>
<tr>
<th>Lab 4</th>
<th>Lab 2</th>
<th>Lab 1</th>
<th>Lab 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.67</td>
<td>5.20</td>
<td>5.32</td>
<td>5.83</td>
</tr>
</tbody>
</table>

Figure 5.1.7: Comparisons of laboratory means

Figure 5.1.7 indicates that the results from laboratory 4 are statistically significantly lower and those from Laboratory 3 are higher than those from the other laboratories.  The difference between laboratories 1 and 2 is not statistically significant.  This picture summarises concisely the results of the inter-laboratory study.

A 95% confidence interval for the difference between any pair of laboratory measurement process means can be obtained in the usual way using:

$$\left(\bar{y}_i - \bar{y}_j\right) \pm t_c \sqrt{\frac{2MSE}{J}}.$$  \hfill (5.9)

Filling in the various quantities gives:

$$\left(\bar{y}_i - \bar{y}_j\right) \pm 2.09 \sqrt{\frac{2(0.0588)}{6}}.$$  

$$\left(\bar{y}_i - \bar{y}_j\right) \pm 0.29$$

Thus, the error bounds on any difference $\bar{y}_i - \bar{y}_j$ are obtained by adding and subtracting 0.29 from the calculated difference.  The confidence interval measures the relative bias between the two laboratories.

**Multiple Comparisons**

Our investigation of the pattern of differences between the laboratory means essentially involves carrying out six t-tests: we can select six pairs from four groups.  The number of possible comparisons grows rapidly with the number of groups involved.  Thus, comparing the means for six groups involves $\frac{6.5}{1.2} = 15$ pairs of means, while ten groups
would allow \( \frac{10.9}{1.2} = 45 \) comparisons to be made. Multiple comparisons present us with a difficulty: the statistical properties of t-tests, as discussed in Chapters 2 and 4, hold for a single test, but will change radically if many tests are carried out simultaneously. This is very often ignored when only a small number of tests are carried out, but it becomes increasingly important when many comparisons are made. To see where the problem lies, we note that for any one test the significance level is the probability of rejecting the hypothesis of no difference between the long-run means, when, in fact, there is no difference, i.e., when the null hypothesis is true. However, the probability is much higher than, say, \( \alpha = 0.05 \) that one or more of the six comparisons for the laboratory study will produce a statistically significant result, when in fact there are no systematic differences.

To understand why this is so, consider a simple coin tossing experiment. If a fair coin is tossed, the probability of the result being a head is 1/2. However, the probability of one or more heads in six tosses is considerably higher. Thus, the probability that all six results are tails is \((1/2)^6 = 1/64\), so that the probability of at least one head is \(1 - (1/2)^6 = 63/64\). The corresponding calculation for the comparisons between the laboratories cannot be carried out in this simple way (since the comparisons involving pairs of means are not all independent of each other, as was assumed for the coin tossing experiment) but the underlying ideas are the same. However, the analogy suggests that use of multiple statistical tests simultaneously can lead to much higher type 1 error rates (falsely rejecting the null hypothesis of equal long-run group means) than the individual significance levels would suggest.

Note that the LSD method discussed above makes no allowance for the fact that multiple tests are being carried out (or, equivalently, multiple simultaneous confidence intervals are being calculated). For this reason, I do not recommend the use of LSD for the analysis of real data. It was introduced here simply to provide a basis on which to develop the idea of Tukey’s HSD method, which is discussed in the next paragraph and illustrated thereafter.

Various strategies are adopted to deal with the multiple comparison problem, but they all involve making the individual tests less powerful, i.e., less likely to detect a small but real difference between the means. One simple approach is to reduce the significance level of the individual tests (say to \( \alpha = 0.01 \)), so that the error rate for the family of comparisons is reduced to a more acceptable level. Minitab offers other multiple comparison alternatives. One of these is Tukey’s ‘Honestly Significant Difference’ (HSD) method. The derivation of the HSD interval is based on the range of a set of means, i.e., the difference that we might expect to see between the largest and smallest values in the set, if the null hypothesis of equal long-run means for all groups were true. If the largest and smallest means in a set differ by at least this amount they will be statistically significantly different. Since all other differences between pairs of means are, by definition, smaller than that between the largest and smallest means in the set, any pair of means that differ by at least HSD are considered statistically significantly different. Consequently, the significance level chosen applies to the whole family of
possible comparisons. The application of the HSD method is very similar to that of the LSD method – it simply requires the replacement of the critical t-value ($t_c$) by another multiplier ($T_c$), which is given by:

$$T_c = \frac{1}{\sqrt{2}} q(1-\alpha, I(J -1))$$  \hspace{1cm} (5.10)

where $q$ is the Studentised range distribution, $\alpha$ is the significance level for the family of comparisons, $I$ means are being compared and MSE has $I(J-1)$ degrees of freedom.

For the laboratory study, Table ST-6 shows that requiring a family significance level of $\alpha=0.05$ will lead to a critical $T_c$ value of:

$$T_c = \frac{1}{\sqrt{2}} q(1-\alpha, I(J -1)) = \frac{1}{\sqrt{2}} q(0.95,4,20) = \frac{1}{\sqrt{2}} (3.96) = 2.80$$  \hspace{1cm} (5.11)

The honestly significant difference is then given by:

$$\text{Honestly Significant Difference (HSD)} = T_c \sqrt{\frac{2\text{MSE}}{J}} = 2.80 \sqrt{\frac{2(0.0588)}{6}} = 0.39.$$

Thus, HSD is 0.39 whereas LSD was 0.29. Accordingly, the HSD method is more conservative in that results need to be further apart before they are declared statistically significantly different. Often, of course, widening the interval means that small, but real, long-run differences, which would have been declared statistically significant using the narrower LSD interval, will now not be declared different. In other words, using the wider HSD interval results in more Type II errors (failing to detect real differences). Protecting ourselves from Type I errors increases the risk of Type II errors.

For the laboratories’ data the conclusions drawn, based on the HSD method, are exactly the same as those shown in Figure 5.1.7, which was based on the LSD method: laboratories 1 and 2 are not significantly different from each other, but taking these as a pair they are significantly different from both 3 and 4, which are also significantly different from each other. Note that the two methods do not always lead to the same conclusions (see Example 4 where this occurs).

**Exercise 5.1.1**

An individual’s critical flicker frequency (cff) is the highest frequency (cycles per second - cps) at which the flicker in a flickering light source can still be detected. At frequencies above the cff the light source appears to be continuous even though it is actually flickering. A preliminary investigation, carried out to see if average cff depends on iris colour, yielded the following cff values for eighteen subjects.

The data are shown in Table 5.1.3. Note that Total Sum of Squares = 52.08 and the MSE=1.96. You may assume that the model assumptions hold.
• Create the full ANOVA table, carry out the F-test and interpret the result.
• Carry out a least significant difference analysis on the sample means and decide how best to report your results.
• Carry out an honestly significant difference analysis on the sample means and decide how best to report your results. Compare your results to those obtained from the least significant difference analysis.

<table>
<thead>
<tr>
<th>Iris Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
</tr>
<tr>
<td>26.8</td>
</tr>
<tr>
<td>26.9</td>
</tr>
<tr>
<td>23.7</td>
</tr>
<tr>
<td>25.0</td>
</tr>
<tr>
<td>26.3</td>
</tr>
<tr>
<td>24.8</td>
</tr>
</tbody>
</table>

| Mean       | 25.6       | 26.9       | 28.3     |
| SD         | 1.283      | 1.649      | 1.231    |

Table 5.1.3: Critical flicker frequencies for the three groups of subjects

5.2 Example 2: A dietary study in rats

The data in Table 5.2.1 were presented by Snedecor and Cochran [2] as coming from a study of the weight gains (grams) in male rats fed on six diets, in a fully randomised experiment. The diets were formulated by the level of protein and also the source of the protein; but we will consider the results as coming from six unstructured diets (as shown below).

<table>
<thead>
<tr>
<th>Diet</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>73</td>
<td>98</td>
<td>94</td>
<td>90</td>
<td>107</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>74</td>
<td>79</td>
<td>76</td>
<td>95</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>118</td>
<td>56</td>
<td>96</td>
<td>90</td>
<td>97</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>111</td>
<td>98</td>
<td>64</td>
<td>80</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>95</td>
<td>102</td>
<td>86</td>
<td>98</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>107</td>
<td>88</td>
<td>102</td>
<td>51</td>
<td>74</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>82</td>
<td>108</td>
<td>72</td>
<td>74</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>77</td>
<td>91</td>
<td>90</td>
<td>67</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>86</td>
<td>120</td>
<td>95</td>
<td>89</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>92</td>
<td>105</td>
<td>78</td>
<td>58</td>
<td>82</td>
</tr>
</tbody>
</table>

| Mean | 100  | 85.9 | 99.5 | 79.2 | 83.9 | 78.7 |
| SD   | 15.1 | 15.0 | 10.9 | 13.9 | 15.7 | 16.5 |

Table 5.2.1: Weight gains (grams) of male rats under six diets
Figure 5.2.1 shows side-by-side dotplots of the data:

The picture suggests that diets 1 and 3 may result in higher weight gains than the other diets, but given the chance variation within the six groups, it is as well to check this out with formal statistical tests. Also, it is not clear whether or not there are differences between the other diets.

Table 5.2.2 shows a one-way\(^6\) ANOVA analysis for the data. The model assumptions of constant within group variability (as measured by the standard deviation) and of data Normality will be investigated later.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet</td>
<td>5</td>
<td>4613</td>
<td>923</td>
<td>4.30</td>
<td>0.002</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>11586</td>
<td>215</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>16199</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(S = 14.65\)

Table 5.2.2: ANOVA analysis of weight gains

\(^6\) The analysis is often described as ‘one-way’ as there is only one classifying factor - diet. If the data were reported in a two-way table, say rows labelled ‘source of protein’ and columns ‘levels of protein’, then the more elaborate data structure would require a ‘two-way ANOVA’. Three-way (the study described in Section 5.0 is a three-way structure – sex, age and training method) and higher order structures may also be analysed by ANOVA.
Note that there are $6-1=5$ degrees of freedom for diets and $6(10-1)=54$ degrees of freedom for error: always check the degrees of freedom, especially in more complicated cases - it is easy to mis-specify the model when using statistics package (especially when fitting more complicated ANOVA models). However, if the degrees of freedom are correct, it generally means your model specification is correct.

The ANOVA F-test addresses the question "are the data consistent with the six population means being the same?" (i.e., if six equivalent populations of rats were fed on the six diets, would the long-run means be the same?). We write this formally as:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu$$

$$H_1: \text{not all diet means (} \mu_i \text{) are the same.}$$

The F-statistic shown in Table 5.2.2 has an $F$-distribution with degrees of freedom 5, 54, when the null hypothesis is true. The critical value for a significance level of $\alpha=0.05$ is $F_c= 2.39$. The observed F-value of 4.30 greatly exceeds this, so we reject the null hypothesis and conclude that, in the long-run, not all six diets result in the same average weight gain. Instead of comparing our test statistic to the critical value, we could use the p-value produced by the package as a measure of departure from the null hypothesis: the small p-value (0.002) associated with the F-test makes such an hypothesis implausible – at least one of the diets is likely to produce weight gains that show a different mean to the others.

**Model Validation**

The scatterplot of residuals versus fitted values for the rat dietary study, shown in Figure 5.2.2, does not suggest that the variability of the residuals changes in any systematic way with the mean yield level. The assumption of constant standard deviation appears valid.

The Normal plot, shown in Figure 5.2.3, and the Anderson-Darling test (p-value=0.43) are consistent with an underlying Normal distribution for the data.

There is no reason, therefore, to doubt the validity of the F-test or of the further tests that will be carried out next.
Figure 5.2.2: Scatterplot of residuals versus fitted values (group means) for the rat dietary study

Figure 5.2.3: Normal plot of residuals for the rat dietary study
Comparison of Means

Table 5.2.3 (page 23) shows Minitab output giving a Tukey Honestly Significant Difference comparison of the diet means, in the form of simultaneous confidence intervals for mean differences between diets. The family confidence coefficient for all comparisons is 0.95. These simultaneous intervals are set up as confidence intervals for long-run differences between group means, based on the formula for the standard two-sample confidence interval, but using the $T_c$ multiplier, instead of the corresponding Student's $t_c$ critical value (which would give Fisher’s Least Significant Difference intervals).

A single confidence interval between two means (labelled I and j) is given by:

$$(\bar{y}_i - \bar{y}_j) \pm t_c \sqrt{\frac{2\text{MSE}}{J}}.$$ 

where each mean is based on J replicates; here $J=10$.

To get the simultaneous Tukey intervals, we replace the t-multiplier by:

$$T_c = \frac{1}{\sqrt{2}} q(1-\alpha, I, I(J-1))$$

where $I=6$ is the number of means to be compared and $J=10$ is the number of replicates in each group. Our Table ST-6 does not give Studentised Ranges (q) with degrees of freedom $\nu=54$; we will use instead the nearest value $q(6,60) = 4.16$ [the value for $\nu=54$ is, in fact, 4.17]. The corresponding $T_c$ value is found by multiplying by $1/\sqrt{2}$: for degrees of freedom $\nu=60$ we get:

$$T_c = \frac{1}{\sqrt{2}} q(1-\alpha, I, I(J-1)) = \frac{1}{\sqrt{2}} q(0.95, 6, 60) = \frac{1}{\sqrt{2}} (4.16) = 2.94$$

The estimated standard error of the difference between two means is:

$$\sqrt{\frac{2\text{MSE}}{J}} = \sqrt{\frac{2(215)}{10}} = 6.56$$

which, when multiplied by the $T_c$ value, gives us the corresponding Tukey interval.

$$(\bar{y}_i - \bar{y}_j) \pm T_c \sqrt{\frac{2\text{MSE}}{J}}.$$ 

$$(\bar{y}_i - \bar{y}_j) \pm 2.94(6.56).$$
Using the value of $q=4.17$ (for $n=54$) would have given $\pm 19.36$, which is used in the Minitab output of Table 5.2.3.

When inspecting Table 5.2.3, we note that any confidence interval that covers zero corresponds to a significance test that fails to reject the null hypothesis of no difference between the corresponding long-run means. Thus, in the first sub-table that compares Diet 1 to all the others, the tests/confidence intervals cannot separate Diet 1 from Diets 2, 3 or 5, but it differs significantly (statistically) from Diets 4 and 6.

Figure 5.2.4 summarises the series of comparisons carried out in Table 5.2.3: any two means that are connected by a line cannot be separated statistically, while those not so connected are considered to be statistically significantly different. The size of the long-run mean difference between any two diets is estimated by the confidence interval in Table 5.2.3.

<table>
<thead>
<tr>
<th>Diets</th>
<th>6</th>
<th>4</th>
<th>5</th>
<th>2</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>78.7</td>
<td>79.2</td>
<td>83.9</td>
<td>85.9</td>
<td>99.5</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Figure 5.2.4: Multiple comparisons results

The multiple comparisons analysis suggests that Diets 1 and 3 give statistically significantly higher weight gains than Diets 6 and 4, but each of these pairs cannot be separated from Diets 5 and 2. Unlike the laboratory comparison study, the comparison of the means does not give a 'clean' result. Unfortunately, this often occurs: the chance variation in the study produces results which are ambiguous in this way.
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Diet

Individual confidence level\(^7\) = 99.54%

### Diet = 1 subtracted from:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Lower</th>
<th>Center</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-33.46</td>
<td>-14.10</td>
<td>5.26</td>
</tr>
<tr>
<td>3</td>
<td>-19.86</td>
<td>-0.50</td>
<td>18.86</td>
</tr>
<tr>
<td>4</td>
<td>-40.16</td>
<td>-20.80</td>
<td>-1.44</td>
</tr>
<tr>
<td>5</td>
<td>-35.46</td>
<td>-16.10</td>
<td>3.26</td>
</tr>
<tr>
<td>6</td>
<td>-40.66</td>
<td>-21.30</td>
<td>-1.94</td>
</tr>
</tbody>
</table>

### Diet = 2 subtracted from:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Lower</th>
<th>Center</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-5.76</td>
<td>13.60</td>
<td>32.96</td>
</tr>
<tr>
<td>4</td>
<td>-26.06</td>
<td>-6.70</td>
<td>12.66</td>
</tr>
<tr>
<td>5</td>
<td>-21.36</td>
<td>-2.00</td>
<td>17.36</td>
</tr>
<tr>
<td>6</td>
<td>-26.56</td>
<td>-7.20</td>
<td>12.16</td>
</tr>
</tbody>
</table>

### Diet = 3 subtracted from:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Lower</th>
<th>Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-39.66</td>
<td>-20.30</td>
</tr>
<tr>
<td>5</td>
<td>-34.96</td>
<td>-15.60</td>
</tr>
<tr>
<td>6</td>
<td>-40.16</td>
<td>-20.80</td>
</tr>
</tbody>
</table>

### Diet = 4 subtracted from:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Lower</th>
<th>Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-14.66</td>
<td>4.70</td>
</tr>
<tr>
<td>6</td>
<td>-19.86</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

### Diet = 5 subtracted from:

<table>
<thead>
<tr>
<th>Diet</th>
<th>Lower</th>
<th>Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>-24.56</td>
<td>-5.20</td>
</tr>
</tbody>
</table>

Table 5.2.3: Honestly significant differences between diets

\(^7\) This is the confidence level for **individual** comparisons which give confidence intervals of the same width as the Tukey intervals which have a **family confidence coefficient** of 0.95 or 95%.
5.3 Example 3: A study of melon yields

Mead and Curnow (3) report an experiment to compare the yields of four varieties of melon. Six plots of each variety were grown; the varieties were allocated to plots at random. The yields (in kg) and some summary statistics are shown in Table 5.3.1. Note that the SD(total)=8.4926.

<table>
<thead>
<tr>
<th>Varieties</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.1</td>
<td>17.3</td>
<td>26.4</td>
<td>16.1</td>
</tr>
<tr>
<td>2</td>
<td>40.3</td>
<td>35.3</td>
<td>32.0</td>
<td>36.5</td>
</tr>
<tr>
<td>3</td>
<td>18.3</td>
<td>22.6</td>
<td>25.9</td>
<td>15.1</td>
</tr>
<tr>
<td>4</td>
<td>28.1</td>
<td>28.6</td>
<td>33.2</td>
<td>31.7</td>
</tr>
</tbody>
</table>

Mean | 20.5 | 37.4 | 19.5 | 29.9 |
SD   | 4.682| 3.942| 5.561| 2.214|

Table 5.3.1: Yields (kg) for four melon varieties

Side by side dotplots of the yields from the four varieties are shown in Figure 5.3.1. It is clear from this picture that the four melon varieties do not produce equal long-run yields. We will, however, obtain the ANOVA table, nevertheless, as this provides us with the MSE which is required for the comparisons among the sample means.
Table 5.3.2 shows the completed ANOVA table. In analysing this dataset we focus first on some of the calculations (not that I would recommend carrying out ANOVA calculations 'by hand' – always use a statistics package in practice) in order to make connections back to two-sample t-tests.

### One-way ANOVA: Yield versus Variety

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variety</td>
<td>3</td>
<td>1292.4</td>
<td>430.8</td>
<td>23.51</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>366.4</td>
<td>18.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>1658.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3.2: The ANOVA table for the melon study

The degrees of freedom for varieties are 4–1=3, those for error are 4(6–1)=20.

The standard deviation of all 24 data values is SD(total)=8.4926. If we square this and multiply by the degrees of freedom we get the Total Sum of Squares:

\[ \text{SSTO} = 23 \times (8.4926)^2 = 1658.9. \]

For each variety the sample SD is an estimate of the common population standard deviation \( \sigma \). In Chapter 2, where we had two groups, we averaged the two SD\(^2\) values to get a combined estimate. This was then used to estimate the standard error of the difference between the two means, which was required for carrying out two-sample t-
tests and calculating confidence intervals. Here we do the same thing, the only difference being that we have four groups.

\[ S^2 = \frac{s_1^2 + s_2^2 + s_3^2 + s_4^2}{4} = MSE \]

The average of the within-group variances gives us the Mean Square(within-group) which is MSE=18.32. This has \(4(6-1)=20\) degrees of freedom, so the sum of squares within-groups, or error sum of squares, SSE, is:

\[ SSE = 20MSE = 366.4. \]

Since the within-group and between-group sums of squares add to the total we can get the between-group sum of squares by subtraction.

\[ SS(\text{between-group}) = 1658.9 - 366.4 = 1292.5 \]

We can now complete the ANOVA table, which is given in Table 5.3.2.

The F-test addresses the null hypothesis that the long-run means associated with the four varieties are all the same (i.e., if the four varieties were planted on many thousands of similar plots, the average yields would be the same). The alternative hypothesis is that they are not all the same, i.e., at least one is different.

The critical value for a significance level of \(\alpha=0.05\) for an F-distribution with 3 and 20 degrees of freedom is 3.1. Since the observed F-value greatly exceeds this (and has a correspondingly small p-value \(p<0.0005\)) we reject the null hypothesis and conclude that not all the four varieties result in the same long-run yield.

**Model Validation**

Figures 5.3.2 and 5.3.3 suggest that the standard assumptions underlying the ANOVA analysis hold in the case of the melon yield study. Obviously, the scatter corresponding to the fitted mean of approximately 30 (variety 4) is a bit tighter than the others (but one of the set has to be the least variable!), but no systematic relationship between residual variation and average yield is suggested by Figure 5.3.2.

Since the model assumptions appear valid, we can have confidence in the result of our F-test and can proceed to making comparisons between the sample means: the exercise immediately following asks you to do this.
Exercise 5.3.1

For the melons study:

- Carry out a least significant difference analysis on the sample means and decide how best to report your results.
- Carry out an honestly significant difference analysis on the sample means and decide how best to report your results. Compare your results to those obtained from the least significant difference analysis.
5.4 Example 4: An analytical method development study

Four different configurations of a Gas Chromatograph (GC) were investigated when developing a method for the analysis of trace compounds in distilled spirits (1). The objective was to maximise the instrumental response, peak area (arbitrary units). Each configuration (A, B, C, D) was run three times; the twelve runs were carried out in a randomised order. The results for the analysis of the propanol content of test portions from a single bottle of whiskey are shown in Table 5.4.1 below.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>39.87</td>
<td>49.10</td>
<td>54.17</td>
<td>55.57</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>2.41</td>
<td>2.02</td>
<td>2.78</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Table 5.4.1: Peak areas (arbitrary units) for the GC study

The dotplots of the data, Figure 5.4.1, suggest that not all the system means are the same. Configuration A clearly gives smaller peak areas than the others; the differences between configurations B, C and D are less marked and will require further analysis, though B appears to give smaller responses than C and D. The ANOVA table, Table
5.4.2, confirms that the long-run configuration means are unlikely to be the same; the F-value of 31.66 is highly statistically significant (p<0.0005).

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC-config</td>
<td>3</td>
<td>454.26</td>
<td>151.42</td>
<td>31.66</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>38.26</td>
<td>4.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>492.52</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4.2: The ANOVA table for the GC study

Figure 5.4.2 shows the results of LSD and HSD analyses. The individual significance levels were 0.05 for the LSD analysis, while the family significance level was set as 0.05 for the HSD analysis. The values used in the analysis were LSD = 2.31 \sqrt{\frac{2(4.78)}{3}} = 4.12 and HSD = \frac{1}{\sqrt{2}} \cdot 4.53 \sqrt{\frac{2(4.78)}{3}} = 5.71.

<table>
<thead>
<tr>
<th>LSD analysis:</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>39.87</td>
<td>49.10</td>
<td>54.17</td>
<td>55.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HSD analysis:</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>39.87</td>
<td>49.10</td>
<td>54.17</td>
<td>55.57</td>
</tr>
</tbody>
</table>

Figure 5.4.2: Comparisons of configuration means

The LSD analysis suggests that that configurations C and D give about the same peak areas, on average, and this average is greater than that of configuration B, which in turn is greater than that for A. The HSD comparisons indicate that A gives smaller peak areas than the other configurations. It also shows D to be statistically significantly bigger than B. However, D is not statistically significantly bigger than C, and similarly, C is not statistically significantly bigger than B. On purely logical grounds, this set of conclusions might appear odd. However, what we have here is a summary of the statistical evidence rather than a set of logical propositions: although B and D can be
separated statistically, the chance variation in the data is such that B and C cannot be separated, and neither can C and D. Such ambiguities are commonplace when many pairwise comparisons are carried out. It is also not unusual that the more conservative HSD method should fail to distinguish between pairs of means that are reported as statistically significantly different when the LSD method is used.

**Model Validation**

The number of data points available for each of the four configurations of the GC system is small. Consequently, the standard residual analyses are of limited value – only gross departures from the model assumptions could be expected to be detected. For what they are worth, these are shown in Figures 5.4.3 and 5.4.4. Both graphs appear well behaved.

In cases like this, where only small numbers of observations are available within the study itself, we must rely on our prior knowledge of the properties of the measurements under study. The GC being investigated here, for example, was in daily use in the laboratory, and the data available from routine quality control measurements would allow the assumptions to be validated. Most experimental studies are relatively small. Consequently, it makes sense to accumulate data about the systems being studied, both for the purpose of validating model assumptions and to provide the information required to plan future studies, as discussed in Chapter 4.

![Figure 5.4.3: Scatterplot of residuals versus fitted values for GC study](image-url)
References


Text © Eamonn Mullins, 2014; data, see references
Outline Solutions

Exercise 5.1.1 (Iris Colour – critical flicker frequency study)

The ANOVA table for the critical flicker frequency study is shown below.

**One-way ANOVA: CFF versus Iris Colour**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris Colour</td>
<td>2</td>
<td>22.69</td>
<td>11.35</td>
<td>5.79</td>
<td>0.014</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>29.39</td>
<td>1.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>52.08</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1.1.1: ANOVA table for the critical flicker frequency data

The critical value for an F test with a significance level of $\alpha=0.05$ is 3.7 (the F distribution has 2, 15 degrees of freedom). The value given in the table (5.79) exceeds this, so we reject the null hypothesis that all the corresponding population means are the same, and conclude that at least one of them differs from the other two. The p-value of 0.014 is an alternative measure of statistical significance – it tells us that our F value of 5.79 has an area of only 0.014 to its right, which indicates that it would be an unusually large value if the null hypothesis were true.

**LSD Analysis**

The Least Significant Difference (LSD) is given by:

$$\text{Least Significant Difference (LSD)} = t_c \sqrt{\frac{2\text{MSE}}{J}} = 2.13 \sqrt{\frac{2(1.96)}{6}} = 1.72.$$ 

The t-value used in the calculation has 15 degrees of freedom – the same as those of the MSE. The results of the LSD comparisons are shown in Figure 5.1.1.1.

<table>
<thead>
<tr>
<th>Brown</th>
<th>Green</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.58</td>
<td>26.32</td>
<td>28.33</td>
</tr>
</tbody>
</table>

Figure 5.1.1.1: LSD comparison of cff means

The LSD is 1.72. Brown and Blue differ by more than this amount and so they are declared statistically significantly different. Neither Brown and Green nor Green and Blue are different by 1.72 so we cannot separate them statistically.
This was a small study – only six observations in each group – so we might not be surprised that the outcome is not clear-cut. However, there is evidence that differences may exist and we now have a basis for designing a larger study. The square root of the MSE is an estimate of the standard deviation of person-to-person variation and it could be used to decide on an appropriate sample size which would allow us to distinguish between Brown and Green or Green and Blue (if differences truly exist), if the outcome of this study was considered sufficiently interesting.

**HSD Analysis**

The Honestly Significant Difference (HSD) is given by:

\[
\text{Honestly Significant Difference (HSD)} = T_c \sqrt{\frac{2\text{MSE}}{J}}
\]

Table ST-6 shows that requiring a family significance level of \( \alpha=0.05 \) will lead to a critical \( T_c \) value of:

\[
T_c = \frac{1}{\sqrt{2}} q(1 - \alpha, I, I(J - 1)) = \frac{1}{\sqrt{2}} q(0.95, 3, 15) = \frac{1}{\sqrt{2}} (3.67) = 2.6
\]

We then get:

\[
\text{HSD} = T_c \sqrt{\frac{2\text{MSE}}{J}} = 2.6 \sqrt{\frac{2(1.96)}{6}} = 2.1
\]

Since Brown and Blue differ by more than this amount, the HSD analysis leads us to the same conclusions as the LSD analysis and these are summarised in Figure 5.1.1.1

**Exercise 5.3.1 (Melons study)**

**LSD Analysis**

The ANOVA F-test in Example 3 established that differences exist between the mean yields of the four melon varieties involved in the study. We now need to establish which means reflect higher long-run yields. We will first use Fisher’s Least Significant Difference (LSD) and then Tukey’s HSD method to make the required comparisons between means. A confidence interval for the difference between any pair of means is given by:

\[
(\bar{y}_i - \bar{y}_j) \pm t_c \sqrt{\frac{2\text{MSE}}{J}}
\]
where $J = 6$ is the number of replicates in each group, MSE is the mean square error (the estimate of the square of the common long-run standard deviation) and the critical t-value is $t_c = 2.09$ since there are 20 degrees of freedom for error; the quantity after the ± sign is the $\text{LSD} = 2.09 \sqrt{\frac{2(18.3)}{6}} = 2.09(2.47) = 5.16$, the amount by which any pair of means must differ in order to be statistically significantly different. Figure 5.3.1.1 shows the results of the multiple comparisons.

$$
\begin{array}{cccc}
V3 & V1 & V4 & V2 \\
19.5 & 20.5 & 29.9 & 37.4 \\
\end{array}
$$

Figure 5.3.1.1: Summary of LSD multiple comparisons ($V =$ variety)

The two means joined by the underscoring cannot be separated statistically, whereas V4 is statistically significantly bigger than V3 or V1, and V2 is bigger than V4. For highest yields we would recommend Variety 2.

**HSD Analysis**

The LSD method involves calculating individual confidence intervals, each with a confidence coefficient of 0.95, for all possible comparisons, here 6. It makes no allowance for the inflation in the probability of type I error (falsely concluding that a difference exists) when many comparisons are carried out simultaneously. The Tukey method does this by calculating confidence intervals which are wider, but which have a family coefficient of 0.95, i.e., all the intervals are expected to cover the corresponding true mean differences, simultaneously, with this level of confidence. The method simply involves replacing the critical t-value by:

$$
T_c = \frac{1}{\sqrt{2}} q(1-\alpha, I, I(J-1))
$$

where $q$ is the Studentised Range critical value where I means are to be compared and the MSE has $I(J-1)$ degrees of freedom. Here $I = 4$ and $I(J-1) = 4(6-1) = 20$.

$$
T_c = \frac{1}{\sqrt{2}} q(1-\alpha, I, I(J-1)) = \frac{1}{\sqrt{2}} q(0.95,4,20) = \frac{1}{\sqrt{2}} (3.96) = 2.80
$$

This gives $\text{HSD} = 2.80(2.47) = 6.92$. Since V4 is more than this distance above V1, and V2 is more than this distance above V4, our conclusions from the LSD analysis still stand.

A 95% confidence interval for the long-run yield from the best variety (Variety 2) is:
\[ \bar{y} \pm t_c \sqrt{\frac{MSE}{J}} \]

\[ 37.4 \pm 2.09 \sqrt{\frac{18.3}{6}} \]

37.4 ± 3.7 kg.