

7 Differences between more than two samples

To investigate potential differences between more than two samples we need an extension of the tests described in Chapter 4. The tests are classified not only on their requirement for the data to be normally distributed, but also according to whether we are looking at more than one independent variable at a time. This chapter covers:

- One-way analysis of variance for normally distributed data
- Kruskal-Wallis one-way analysis of variance using ranks for data measured at least on an ordinal scale
- Two-way analysis of variance for normally distributed data
- Two-way analysis of variance using ranks for data measured at least on an ordinal scale
- Consideration of alternative models of analysis of variance for more complicated experimental and survey designs

In Chapter 4 we compared the means (or medians) of two samples using a t test (for normally distributed data) or a Mann-Whitney U test (for ordinal or non-normal data). However, many experiments or surveys involve the comparison of three or more samples. For example, in a survey of household energy bills, we might place houses in three categories depending on the type of insulation they have – loft insulation, double-glazing or minimal insulation. To investigate the effect of insulation type on household energy bill, it might seem logical to perform three t tests (comparing 'loft' with 'double-glazing', 'loft' with 'minimal' and 'double-glazing' with 'minimal'). This would certainly be possible. However, if we had more samples, the computation would become very long-winded and time-consuming (for seven samples, for example, we would need 21 separate tests). There is a more serious

objection to conducting multiple t tests: that is, when we compute several tests we increase our chances of obtaining a significant result by chance alone. Remember that the critical probability value is 0.05. This is the value at which the null hypothesis is rejected because there is only a 5% probability that the result occurred by chance. We can turn this around and say that on 5% of occasions the null hypothesis is rejected when in fact it is true (this is a type I error: see Chapter 3). It now becomes clear that if we do 20 tests, at



bird distribution, 36
 long-tailed tits and classified
 high forest (HF). The

the Presence or absence
 of long-tailed tits

A
 P
 P
 P
 A
 P
 A
 P
 P
 P
 A
 P
 P
 A
 P
 A
 P
 A

column and grand totals).
 check that the row, column
 sums.
 e.

tests in the different types

and was recorded as either
 36 lichens were found on
 north-facing quadrants and 45

of trees, use the external
 to the observed and
 goodness of fit test.
 of fit test.
 s?

least one may appear to be significant even if there were no real differences between the means.

To get around this problem, we need a test that will simultaneously compare three or more means (or medians) with an overall 5% probability of committing a type I error. For normally distributed data, we use a parametric **analysis of variance (ANOVA)**. There are several types of ANOVA depending on the experimental or survey design: the simplest type is used to compare two or more means and is called a **one-way analysis of variance**. In one-way ANOVA there is one independent variable (in our household energy example, insulation type) and one dependent variable (energy bill). An assumption of one-way ANOVA is that all the data are randomly collected and each data point is independent of all the others. This assumption would not be true in our energy bill example if half the houses had been sampled from London and half from Manchester, because there are likely to be differences between the two areas (e.g. in climate) that would affect energy bills. In this case, the survey would have two independent variables: insulation type and city. The appropriate way of analysing experiments or surveys with two independent variables is by **two-way analysis of variance**, which will be considered later in this chapter.

If the data cannot be used in a parametric analysis of variance, either because they are measured on an ordinal scale, or are not normally distributed and cannot be transformed, then a nonparametric version of one-way ANOVA, **Kruskal-Wallis one-way analysis of variance using ranks**, should be applied. When there is more than one independent variable, then an equivalent nonparametric two-way procedure is used.

This chapter first covers one-way analyses of variance (parametric and then nonparametric tests) and then two-way analyses (again both parametric and nonparametric tests). Extensions of the matched-pair tests of Chapter 4 follow. Finally, analyses that are appropriate for more complex experimental and survey designs are briefly considered.

Parametric one-way analysis of variance

As in many of the tests covered in this book, ANOVA compares the two major sources of variation in the data: variation that is explained by the design of the experiment or survey (explained variation) and that which is due to the inherent variation within the samples (unexplained variation). In the energy bill example, the aim is to investigate whether variation in energy bill can be explained by type of insulation. Of course, variation in energy bill could be caused by many factors, for example type of heating system and how many hours the house is heated. Although in the survey we would standardise as many factors as possible (for example by selecting houses of approximately similar size occupied by similar numbers of people), there would still be inherent variation in the energy bills. However, if

insulation type is an important factor in explaining variation in energy bills, we would expect the explained variation (due to insulation type) to be greater than the unexplained variation (inherent variation). One-way ANOVA effectively calculates the test statistic (F) as the ratio of explained to unexplained variation (the F statistic is sometimes called the variance ratio).

It is an assumption of ANOVA that the variances of the samples are not different, therefore the first step of the analysis is to check this. If the variances are significantly different, then transforming the data by one of the methods in Chapter 3 may equalise the variances (this is because the mean of some non-normal distributions is dependent on the variance, and transformation can simultaneously remove this dependence and normalise the data: see Chapter 3). If, following transformation, the variances still differ, then we would need a nonparametric test (the Kruskal–Wallis one-way analysis of variance using ranks, which is covered in the next section).

Testing for equality of variances

The test for equality of variances (also known as homoscedasticity) is simple. First, the variances are obtained for each sample. The largest variance is then compared with the smallest (because if the largest and smallest are not significantly different from each other then the others cannot be). The largest variance is divided by the smallest to obtain a value called $F_{\max}$, and this is compared to a table of critical values of $F_{\max}$ (an extract from which is given in Table 7.1) to find the probability of there being no significant difference between the variances. This procedure is demonstrated in Worked Example 7.1a on energy bill data from households with three different insulation types. If the calculated value equals or is lower than the table value at $P = 0.05$ (i.e. $P \geq 0.05$), we can accept the null hypothesis that there is no significant difference between the variances and proceed with the ANOVA. If our value is greater than the table value, then the variances are significantly different ($P < 0.05$) and we should either transform the data (if appropriate) and recheck, or use a nonparametric test. The table is consulted at two points: k , the number of samples we are comparing (in this case 3) and the degrees of freedom ($n - 1$) of the samples (in this case there are eight houses in each sample, so the degrees of freedom are 7). If we had an unequal number of data points in the two samples being compared by the $F_{\max}$ test then we would use the degrees of freedom of the sample with fewer data points. In our example, where $k = 3$ and $df = 7$, the critical value of $F_{\max}$ at $P = 0.05$ is 6.94 (see the shaded area in Table 7.1). From Worked Example 7.1a, the calculated $F_{\max}$ (1.109) is lower than this, so we can conclude that the variances are not significantly different ($P > 0.05$) and it is safe to proceed with the ANOVA.

Note that the data in Worked Example 7.1a are listed in three separate columns (one for each sample). If you use a computer program to calculate your ANOVA tests, this format of data will often not be appropriate. See Appendix B (Table B.6) for details of data entry for statistical analysis using computers.


WORKED EXAMPLE 7.1A ANOVA for energy bills (£ per annum) in houses with different types of insulation: calculating the $F_{\max}$ test

Loft insulation, x_L ($n = 8$)	Double glazing, x_D ($n = 8$)	Minimal, x_M ($n = 8$)
149	134	171
198	181	184
179	198	191
158	187	213
134	151	198
155	174	186
181	178	234
161	169	199
$\Sigma x_L = 1\,315$	$\Sigma x_D = 1\,372$	$\Sigma x_M = 1\,576$
$\Sigma x_L^2 = 219\,073$	$\Sigma x_D^2 = 238\,212$	$\Sigma x_M^2 = 313\,104$
$\bar{x}_L = 164.4$	$\bar{x}_D = 171.5$	$\bar{x}_M = 197$
$s_L^2 = 417.1$	$s_D^2 = 416.3$	$s_M^2 = 376$

$$F_{\max} = \frac{\text{Largest variance}}{\text{Smallest variance}}$$

$$F_{\max} = \frac{s_L^2}{s_M^2} = \frac{417.13}{376.00} = 1.109$$

At $df = 7$ and $k = 3$, from the shaded area of Table 7.1, we can see that $P > 0.05$

Table 7.1 Selected values of $F_{\max}$ for testing equality of variances.
 Reject the null hypothesis if the calculated value of $F_{\max}$ is greater than the table value. Values are given for $P = 0.05$. Shaded lines indicate the critical values for the example referred to in the text. A more comprehensive table of $F_{\max}$ values is given in Table D.9 (Appendix D)

		Number of samples being compared (k)			
		2	3	4	5
Degrees of freedom	6	5.82	8.38	10.4	12.1
of the smallest	7	4.99	6.94	8.44	9.70
sample ($n - 1$)	8	4.43	6.00	7.18	8.12

Calculating the F statistic for ANOVA

Analysis of variance involves first finding the total variation in the data (measured as the **sums of squares**: first encountered in step 3 in Box 2.3) and dividing this variation into explained variation (i.e. that which is accounted for by the category or treatment – in this case insulation type – which is also called between variation, or SS_{between}) and unexplained variation (the inherent variation within the samples, also called within variation, or SS_{within}). Thus:

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$$

The formulae for calculating the sums of squares are shown in Box 7.1, and the calculations for the energy example are shown in Worked Example 7.1b.

Box 7.1 Formulae for the sums of squares for ANOVA

The sum of squares of all the data points is:

$$SS_{\text{total}} = \sum x_T^2 - \frac{(\sum x_T)^2}{n_T}$$

where:

$\sum x_T$ is the sum of all values in all samples;

n_T is the total number of values in all samples.

The within variation is the sum of the variation in the individual samples. When comparing the means of k samples, the within sum of squares is:

$$SS_{\text{within}} = SS_1 + SS_2 + SS_3 + \dots + SS_k$$

where:

$$SS_1 = \sum x_1^2 - \frac{(\sum x_1)^2}{n_1}$$

The between sample sum of squares is calculated by:

$$SS_{\text{between}} = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} + \dots + \frac{(\sum x_k)^2}{n_k} - \frac{(\sum x_T)^2}{n_T}$$

Here $\sum x_1$ is the sum of all values in sample 1, $\sum x_2$ the sum of all values in sample 2, and so on up to sample k ; n_1 is the number of values in sample 1, n_2 the number of values in sample 2, and so on up to sample k .

As a check, the total sum of squares should be equal to the between plus the within sums of squares:

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$$

Note that these formulae are the same irrespective of whether sample sizes are equal or unequal.

er annum) in houses

Minimal, x_M
($n = 8$)

171
184
191
213
198
186
234
199

$$\begin{aligned}\sum x_M &= 1\,576 \\ \sum x_M^2 &= 313\,104 \\ \bar{x}_M &= 197 \\ s_M^2 &= 376\end{aligned}$$

can see that $P > 0.05$

SS.
er than
cate
e
dix D)
ared (k)

5
12.1
9.70
8.12



WORKED EXAMPLE 7.1B ANOVA for energy bills (£ per annum) in houses with different types of insulation: calculating the sums of squares using the data from Worked Example 7.1a

Loft ($n = 8$)	Double-glazing ($n = 8$)	Minimal ($n = 8$)	Total ($n = 24$)
$\sum x_L = 1315$	$\sum x_D = 1372$	$\sum x_M = 1576$	$\sum x_T = 4263$
$\sum x_L^2 = 219\,073$	$\sum x_D^2 = 238\,212$	$\sum x_M^2 = 313\,104$	$\sum x_T^2 = 770\,389$
$\bar{x}_L = 164.4$	$\bar{x}_D = 171.5$	$\bar{x}_M = 197$	$\bar{x}_T = 177.625$
$s_L^2 = 417.1$	$s_D^2 = 416.3$	$s_M^2 = 376$	$s_T^2 = 572.766$

First we compute the total sum of squares:

$$SS_{\text{total}} = \sum x_T^2 - \frac{(\sum x_T)^2}{n_T} = 770\,389 - \frac{(4263)^2}{24} = 131\,73.625$$

and then the sum of squares for each individual sample:

$$SS_{\text{loft}} = \sum x_L^2 - \frac{(\sum x_L)^2}{n_L} = 219\,073 - \frac{1315^2}{8} = 2919.875$$

$$SS_{\text{double-glazing}} = \sum x_D^2 - \frac{(\sum x_D)^2}{n_D} = 238\,212 - \frac{1372^2}{8} = 2914$$

$$SS_{\text{minimal}} = \sum x_M^2 - \frac{(\sum x_M)^2}{n_M} = 313\,104 - \frac{1576^2}{8} = 2632$$

Now we calculate SS_{within} and SS_{between} :

$$SS_{\text{within}} = SS_{\text{loft}} + SS_{\text{double-glazing}} + SS_{\text{minimal}} = 2919.875 + 2914 + 2632 = 8465.875$$

$$\begin{aligned}
 SS_{\text{between}} &= \frac{(\sum x_L)^2}{n_L} + \frac{(\sum x_D)^2}{n_D} + \frac{(\sum x_M)^2}{n_M} - \frac{(\sum x_T)^2}{n_T} \\
 &= \frac{1315^2}{8} + \frac{1372^2}{8} + \frac{1576^2}{8} - \frac{4263^2}{24} = 761\,923.125 - 757\,215.375 = 4707.75
 \end{aligned}$$

As a check on your calculations, you can confirm that:

$$SS_{\text{total}} = SS_{\text{within}} + SS_{\text{between}}$$

$$13173.625 = 8465.875 + 4707.75$$

per annum) in houses
squares using the data

Total ($n = 24$)		
1576	$\sum x_T =$	4263
104	$\sum x_T^2 =$	770 389
197	$\bar{x}_T =$	177.625
376	$s_T^2 =$	572.766

Table 7.2 Producing an ANOVA table

Source of variation	1	2	Column 3	4	5
	df ^a	Sums of squares (SS)	Mean squares (MS)	F statistic	P
Between	$k - 1$	SS_{between}	$\frac{SS_{\text{between}}}{df_{\text{between}}}$	$\frac{MS_{\text{between}}}{MS_{\text{within}}}$	Look up in tables using the between df and the within df
Within	$(n_T - 1) - (k - 1)$	SS_{within}	$\frac{SS_{\text{within}}}{df_{\text{within}}}$		
Total	$n_T - 1$	SS_{total}			

^a k is the number of samples being compared; n_T is the number of data points in the whole analysis.

The sums of squares values are now used to calculate the test statistic (F). These calculations can be arranged in an ANOVA table as in Table 7.2 (this is also the usual output from statistics programs). The degrees of freedom are entered in column 1: the total degrees of freedom are the total number of data points in the whole analysis minus 1; the between degrees of freedom are the number of samples minus 1; and the within degrees of freedom are the total df minus the between df. The sums of squares that we calculated in Worked Example 7.1b are entered in column 2. Note that the between and within degrees of freedom add up to the total degrees of freedom, and the between and within sums of squares add up to the total sum of squares. In column 3, the **mean squares** (abbreviated to MS) are simply the sums of squares divided by their respective degrees of freedom to standardise the variation. In column 4, the test statistic, F , is calculated as MS_{between} divided by MS_{within} . This is the standardised between variation (explained) divided by the standardised within variation (unexplained). The test statistic is then compared to table values of F (an extract of which is shown in Table 7.3) using two values of degrees of freedom (the between and the within) to give the probability of there being no significant difference between the means (column 5). Several alternative terms for the components of the ANOVA table are used in statistics textbooks and statistical software. The between variation is sometimes called treatment, factor or group variation, the within variation is sometimes called error (from the Latin *errare* – to wander) or residual (remaining variation). The ANOVA table for our energy example, including the calculations, is laid out in Worked Example 7.1c.

Once we have obtained the F value, we need to decide whether it is significant. Remember that the F value is the ratio of the explained to the unexplained variation. The null hypothesis is that the explained variation (between variation) is not significantly greater than the unexplained variation (within variation). The alternative hypothesis is that the explained variation is greater than the unexplained variation (i.e.

5

4

$$4 + 2632 = 8465.875$$

$$-757\,215.375 = 4707.75$$

Table 7.3 Selected values of the *F* distribution for ANOVA. Reject the null hypothesis if the calculated values of *F* are greater than the table values. The upper (bold) figures are for $P = 0.05$, while the lower figures are for $P = 0.01$. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of *F* values is given in Table D.7 (Appendix D)

		df for the between MS			
		1	2	3	4
df for the within MS	19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50
	20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43
	21	4.32 8.02	3.47 5.78	3.07 4.87	2.84 4.37
	22	4.30 7.95	3.44 5.72	3.05 4.82	2.82 4.31
	23	4.28 7.88	3.42 5.66	3.03 4.76	2.80 4.26



WORKED EXAMPLE 7.1C ANOVA for energy bills (£ per annum) in houses with different types of insulation: calculating the *F* statistic using the sums of squares from Worked Example 7.1b

	df	Sums of squares (SS)	Mean squares (MS)	<i>F</i> statistic	<i>P</i>
Between	$3 - 1 = 2$	4 707.75	$\frac{4707.75}{2} = 2353.875$	$\frac{2353.875}{403.1369} = 5.839$	Look up in tables at df of 2 and 21
Within	$(24 - 1) - (3 - 1) = 21$	8 465.875	$\frac{8465.875}{21} = 403.1369$		
Total	$(24 - 1) = 23$	13 173.625			

the *F* value is greater than 1, as is the case for our energy bill data – see Worked Example 7.1c, where the *F* value is 5.839). To see whether the explained variation is significantly larger than the unexplained variation, we compare our calculated *F* value to a table of *F* values, using the degrees of freedom for between ($df = 2$) and within ($df = 21$), as shown in the shaded area of Table 7.3. In cases where the explained variation is lower than the unexplained variation (i.e. the *F* value is less than 1), we can automatically accept that the null hypothesis is correct (i.e. there is no significant difference between the means) and $P > 0.05$. Looking at Table 7.3, for 2 and 21

hypothesis if the calculated F value is greater than the table value for both $P = 0.05$, while the lower value is referred to in the text. A

Mean MS

3	4
3.13	2.90
5.01	4.50
3.10	2.87
4.94	4.43
3.07	2.84
4.87	4.37
3.05	2.82
4.82	4.31
3.03	2.80
4.76	4.26

in houses with different insulation types from Worked Example 7.1

F statistic	P
$\frac{2353.875}{403.1369} = 5.839$	Look up in tables at df of 2 and 21

1 data – see Worked Example 7.1. The explained variation is our calculated F value (between df = 2) and within (df = 21) where the explained variation is less than 1), we conclude that there is no significant difference (Table 7.3, for 2 and 21).

degrees of freedom, our calculated F value (5.839) exceeds the table value for both $P = 0.05$ ($F = 3.47$) and $P = 0.01$ ($F = 5.78$). We therefore reject the null hypothesis and accept the alternative hypothesis that the variation between insulation categories is greater than the variation within insulation categories. The significant difference between the means could be expressed in the statement:

Insulation category (whether loft insulation, double-glazing or minimal insulation) has a highly significant effect on household energy bills ($F_{2,21} = 5.839$, $P < 0.01$).

Do not forget to quote both values of the degrees of freedom: notice the abbreviated way of expressing the degrees of freedom as a subscript to the F symbol with the between degrees of freedom first, followed by the within degrees of freedom. If you prefer, an abbreviated version of the ANOVA table (such as that shown in Table 7.4) can be placed in the results section of a report. Notice that we do not need to put in the sums of square values or the total degrees of freedom because the reader has enough information to calculate them.



Table 7.4 ANOVA for the effect of insulation type on energy bills

Source	df	Mean squares	F	P
Between	2	2353.875	5.839	< 0.01
Within	21	403.137		

We have demonstrated ANOVA on the comparison of three means. The test can be used to compare any number of means, including just two means: i.e. the situation where we used a t test in Chapter 4. If F is

computed instead of t we obtain exactly the same probability, because the t and F tests are mathematically related (in fact, for comparisons of two samples $F = t^2$).

After obtaining a significant F value in an ANOVA on more than two means, we can conclude that there is a significant difference between the means, but we cannot say where these differences lie. With two samples we could simply examine the means (if there is a significant difference, the larger mean must be significantly larger than the smaller one). With three or more samples, we can view the means and get an impression: in our example, houses with minimal insulation have the most expensive bills (£197.00 – see Worked Example 7.1a) and those with loft insulation the least expensive (£164.40) with those with double-glazing being somewhere in the middle (£171.50), but we cannot say that all of the means differ significantly from each other. In fact, there are several possibilities:

Minimal > Double-glazing > Loft
 Minimal > (Double-glazing, Loft)
 (Minimal, Double-glazing) > Loft

where the samples in brackets are not significantly different from each other.

To tell where the differences lie we need a further test, known as a multiple comparison test. Where there are equal sample sizes (as in this example), the Tukey

test is appropriate. If sample sizes differ then we need to use a Tukey–Kramer test. Note that some computer programs do not calculate Tukey or Tukey–Kramer tests.

Multiple comparison tests with equal sample sizes: the Tukey test

Multiple comparison tests are used only after an overall significant difference has been demonstrated. The test is done on all possible pairs of samples. To calculate how many tests need to be computed for k samples, the formula is:

$$\frac{k(k-1)}{2}$$

There are many methods for calculating multiple comparisons which all attempt to get around the problem of conducting several non-independent tests (for three samples there are $(3 \times 2) / 2 = 3$ tests, for four samples there are $(4 \times 3) / 2 = 6$ tests, etc.). This is a problem because the more tests we do, the more chance we have of falsely rejecting a null hypothesis and accepting a difference where one does not exist (type I error: see Chapter 3). Ideally we should compare the samples at an overall significance level of 5%. However by doing so, we increase the chances of falsely rejecting significant differences (type II error). The test we will use now is called the Tukey test (or HSD – honestly significant difference test), and appears to be an acceptable compromise between the two types of error. Box 7.2 shows the formula. The first step is to calculate the minimum significant difference (MSD) between samples. The MSD is then compared to the absolute differences between the pairs of means and those which are equal to or greater than the MSD are significantly different.

Box 7.2 Formula for the Tukey multiple comparison test for use after parametric ANOVA (where sample sizes are equal)

Here, the minimum significant difference is calculated as:

$$\text{MSD} = q \times SE$$

where the standard error is:

$$SE = \sqrt{\frac{MS_{\text{within}}}{n}}$$

and:

MS_{within} is the mean square for the within variation taken from the ANOVA table;

n is the number of values in each sample;

q is the value obtained from q distribution tables (see Table D.10, Appendix D) using the number of samples involved (k) and the degrees of freedom for the within sample category.

e a Tukey–Kramer test.
or Tukey–Kramer tests.

Tukey test

nificant difference has
samples. To calculate how
is:

ons which all attempt to get
t tests (for three samples
 $\times 3) / 2 = 6$ tests, etc.). This
e we have of falsely
e one does not exist (type I
les at an overall
e the chances of falsely
will use now is called the
and appears to be an
x 7.2 shows the formula.
rence (MSD) between
ences between the pairs of
SD are significantly

Comparison test for use s are equal)

he ANOVA table;

0, Appendix D) using the
for the within sample

The calculation of the MSD involves a new value (q) which is multiplied by a standard error term. This standard error term is the square root of the within mean square from the ANOVA table (in our example, the value of 403.137 from Table 7.4) divided by the number of data points in each sample being compared (in our example there are eight houses in each sample). The appropriate value of q is taken from a table (an extract of which is shown in Table 7.5) against the number of samples being compared (in our example, 3) and the degrees of freedom for the within sample category in the ANOVA table (in this case 21 – see Table 7.4). Since the exact number of degrees of freedom for our example (21) are not covered by the q table (see the shaded area of Table 7.5), Worked Example 7.1d shows a method of estimating it, called **interpolation**. An alternative, conservative method (i.e. one that increases the chances of a type II error) is simply to take the q value for the next lowest degrees of freedom (i.e. 20).



WORKED EXAMPLE 7.1d ANOVA for energy bills (£ per annum) in houses with different types of insulation: calculating the minimum significant difference between means

To calculate the minimum significant difference (MSD) between our samples, we first calculate the SE :

$$SE = \sqrt{\frac{MS_{\text{within}}}{n}} = \sqrt{\frac{403.137}{8}} = 7.0987$$

The MSD is simply this value multiplied by q for degrees of freedom (= 21) and number of means (= 3) from Table 7.5. Since we cannot find the exact q value for degrees of freedom of 21, we interpolate between the values for 20 and 24 degrees of freedom. From Table 7.5 the value for $df = 21$ must lie between 3.58 (the value for $df = 20$) and 3.53 (the value for $df = 24$). For the 4 degrees of freedom between the table values, the value of q decreases by $3.58 - 3.53 = 0.05$ units. If we assume that there are equally sized increments between the values on the table, then there is a drop of $0.05 / 4 = 0.0125$ units for each degree of freedom between 20 and 24. To find the value at 21 degrees of freedom, one increment is removed from the value at $df = 20$ (i.e. $3.58 - 0.0125 = 3.5675$). We can now calculate the MSD:

$$MSD = q \times SE = 3.5675 \times 7.0987 = 25.32$$

For the means to differ significantly from one another, they must differ by at least 25.32. We now compile a table giving the differences between the mean energy bill per insulation type (mean energy bills for each sample – in £ per annum – are given in brackets; differences between means that are greater than the MSD are highlighted with an asterisk):

	Loft (164.375)	Double-glazing (171.500)
Double-glazing (171.500)	7.125	–
Minimal (197.000)	32.625 *	25.500 *

Table 7.5 Selected values of q for the Tukey test. Table values are for $P = 0.05$. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of q values is given in Table D.10 (Appendix D)

		Number of samples being compared		
		3	4	5
Degrees of freedom for within MS	19	3.59	3.98	4.25
	20	3.58	3.96	4.23
	24	3.53	3.90	4.17
	30	3.49	3.85	4.10

The goal of the test is to determine whether any of the absolute differences between the means are greater than the MSD. These comparisons for our energy and insulation data are shown in Worked Example 7.1d. Two of the comparisons exceed the MSD and are therefore significantly different from each other. In a results section of a report, this can be expressed as:

Insulation category (whether loft insulation, double glazing or minimal insulation) has a highly significant effect on household energy bills ($F_{2,21} = 5.839$, $P < 0.01$). Multiple comparisons using the Tukey test revealed that the minimally insulated households have significantly higher bills than those with either loft insulation or double glazing ($P < 0.05$). However, there is no significant difference between the energy bills of houses with loft insulation and those with double glazing ($P > 0.05$).

Alternatively the following code could be used (this is a useful shorthand if summarising many results in one table):

Minimal > (Loft insulation, Double-glazing)

where items in brackets indicate no significant difference between them. Another method of displaying the Tukey test results is to add them to a graphical illustration of the means using symbols, as shown in Figure 7.1.

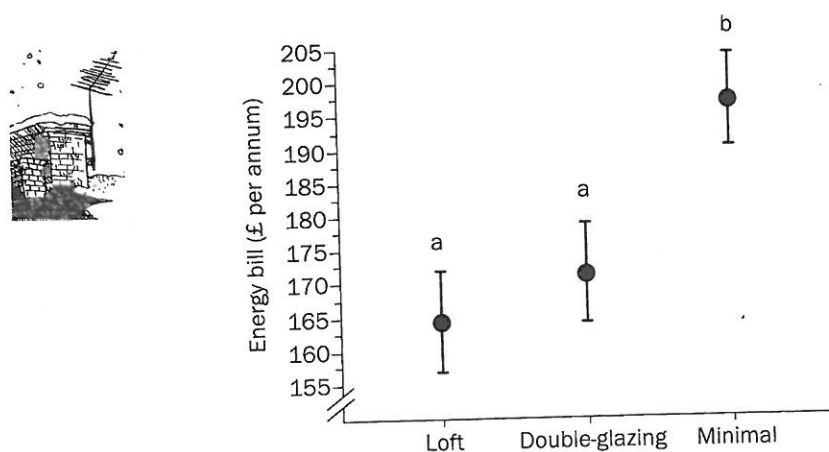


Figure 7.1 Annual energy bills (mean $\pm$ SE) in households with different types of insulation ($n = 8$ in each sample). Means labelled with the same letter are not significantly different from each other following the Tukey test for multiple comparisons

goal of the test is to
mine whether any of the
lute differences between
means are greater than
MSD. These comparisons
our energy and insulation
are shown in Worked
nple 7.1d. Two of the
parisons exceed the
D and are therefore
ificantly different from
other. In a results
ion of a report, this can
xpressed as:

nimal insulation) has a
 $P < 0.01$). Multiple
ulated households have
ouble glazing ($P < 0.05$).
lls of houses with loft

ul shorthand if

ween them. Another
a graphical illustration of

ifferent types of insulation
ot significantly different from

Note that it is possible to find a significant F value (meaning that at least one mean differs from the others) and yet find no significant multiple comparisons. This may seem odd, but it arises because multiple comparisons are not very sensitive tests.

Multiple comparisons with unequal sample sizes: the Tukey–Kramer test

For multiple comparisons of samples of unequal sizes, the Tukey–Kramer method should be used. The formula is very similar to the Tukey test (in fact if sample sizes are equal, the formula simplifies to the Tukey formula). It is more laborious to calculate by hand, because the minimum significant difference between each pair of means is calculated for each pair in turn using the formula in Box 7.3.

Box 7.3 Formula for the Tukey–Kramer multiple comparison test for use after parametric ANOVA (where sample sizes are not equal)

Here the minimum significant difference is calculated for each pair of means in turn:

$$MSD_{A,B} = q \times SE_{A,B}$$

where the standard error is:

$$SE_{A,B} = \sqrt{\frac{MS_{\text{within}}}{2} \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}$$

and:

$MSD_{A,B}$ is the minimum significant difference between the means of samples A and B;

n_A and n_B are the number of items in samples A and B;

q is the value obtained from q distribution tables (see Table D.10, Appendix D) using the total number of samples involved (k) and the degrees of freedom for the within sample category.



To illustrate the Tukey–Kramer test, Worked Example 7.2 shows measurements of organochlorine residues in gull eggs from colonies situated in four lakes (A, B, C and D). Lakes A, B and D have eight replicate measurements, but lake C has only six. The data are analysed using ANOVA and, since the probability is less than 0.05, the analysis is completed by performing multiple comparisons using the Tukey–Kramer method for unequal sample sizes. From these analyses we find that $C > (D, A) > B$. We can report this as:



WORKED EXAMPLE 7.2 ANOVA followed by the Tukey–Kramer test for minimum significant differences on organochlorine residues (mg kg⁻¹) in gull eggs

Lake A (n = 8)	Lake B (n = 8)	Lake C (n = 6)	Lake D (n = 8)	Total (n = 30)
3.6	2.5	5.8	3.0	
3.0	2.4	6.2	3.3	
3.4	2.6	6.5	3.7	
2.9	2.6	6.1	3.6	
2.9	2.8	5.6	3.5	
3.1	2.5	6.3	3.6	
3.5	2.7		3.2	
3.3	2.6		3.4	
$\Sigma x_A = 25.7$	$\Sigma x_B = 20.7$	$\Sigma x_C = 36.5$	$\Sigma x_D = 27.3$	$\Sigma x_T = 110.2$
$\Sigma x_A^2 = 83.09$	$\Sigma x_B^2 = 53.67$	$\Sigma x_C^2 = 222.59$	$\Sigma x_D^2 = 93.55$	$\Sigma x_T^2 = 452.9$
$\bar{x}_A = 3.2125$	$\bar{x}_B = 2.5875$	$\bar{x}_C = 6.083\ 333$	$\bar{x}_D = 3.4125$	$\bar{x}_T = 3.673\ 333$
$SS_A = 0.528\ 75$	$SS_B = 0.108\ 75$	$SS_C = 0.548\ 333$	$SS_D = 0.388\ 75$	$SS_T = 48.098\ 67$
$s_A^2 = 0.075\ 54$	$s_B^2 = 0.015\ 536$	$s_C^2 = 0.109\ 667$	$s_D^2 = 0.055\ 536$	$s_T^2 = 1.658\ 575$

The variances were not significantly different (checked using an $F_{\max}$ test), therefore we can proceed with the ANOVA (using the formulae in Box 7.1 for sums of squares and Table 7.2 for the other components):

	df	SS	MS	F	P
Between	3	46.524 083 333	15.508 027 778	256.073	< 0.01
Within	26	1.574 583 333	0.060 560 897		
Total	29	48.098 666 667			

There is an overall significant difference between the lakes. To find where the differences lie, multiple comparisons are needed. There are $k(k-1)/2 = (4 \times 3)/2 = 6$ possible comparisons. First we calculate the standard error (SE) for each pair:

$$SE_{A,B} = \sqrt{\frac{MS_{\text{within}}}{2} \left[\frac{1}{n_A} + \frac{1}{n_B} \right]} = \sqrt{\frac{0.060\ 560\ 897}{2} \left[\frac{1}{n_A} + \frac{1}{n_B} \right]} = \sqrt{0.030\ 280\ 448 \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}$$

For each comparison the SE is as follows:

A (n = 8) vs B (n = 8)

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{8} + \frac{1}{8} \right]} = 0.087\ 006\ 391$$

A (n = 8) vs C (n = 6)

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{8} + \frac{1}{6} \right]} = 0.093\ 977\ 643$$

continued . .

Worked Example 7.2 continued . . .

Cramer test for
(mg kg⁻¹) in gull eggs

= 8) Total (n = 30)

3.0
3.3
3.7
3.6
3.5
3.6
3.2
3.4

27.3 $\Sigma x_T = 110.2$
 3.55 $\Sigma x_T^2 = 452.9$
 4125 $\bar{x}_T = 3.673\ 333$
 8.75 $SS_T = 48.098\ 67$
 536 $s_T^2 = 1.658\ 575$

in $F_{\max}$ test), therefore
 1 for sums of squares

P
 56.073 < 0.01

find where the
 $t - 1/2 = (4 \times 3) / 2 = 6$
 2) for each pair:

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}$$

$$= 6) \quad \sqrt{\frac{1}{8} + \frac{1}{6}} = 0.093\ 977\ 643$$

continued . . .

A (n = 8) vs D (n = 8)

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{8} + \frac{1}{8} \right]} = 0.087\ 006\ 391$$

B (n = 8) vs C (n = 6)

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{8} + \frac{1}{6} \right]} = 0.093\ 977\ 643$$

B (n = 8) vs D (n = 8)

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{8} + \frac{1}{8} \right]} = 0.087\ 006\ 391$$

C (n = 6) vs D (n = 8)

$$\sqrt{0.030\ 280\ 448 \left[\frac{1}{8} + \frac{1}{6} \right]} = 0.093\ 977\ 643$$

Next we calculate the minimum significant difference (MSD) for each pair:

$$MSD = q \times SE$$

Looking up q for number of samples (k) = 4 and df for the within-sample category = 26 (Table D.10) (interpolating between the values for $df = 24$ (3.90) and $df = 30$ (3.85)) we find that q is 3.88. We multiply the standard errors throughout by 3.88, to obtain the MSD:

	A	B	C
B	0.3376	—	—
C	0.3646	0.3646	—
D	0.3376	0.3376	0.3646

Finally, we calculate the differences between the means in the table below and compare them to the equivalent MSD in the table above (means are given in brackets; differences between means that are greater than the MSD are highlighted with an asterisk):

	A (3.2125)	B (2.5875)	C (6.0833)
B (2.5875)	0.6250*	—	—
C (6.0833)	2.8708*	3.4958*	—
D (3.4125)	0.2000	0.8250*	2.6708*

The only comparison not significant is that of lake D compared with lake A.

There are significant differences in the levels of organochlorine residues in gull eggs between the four lakes ($F_{3,26} = 256.07$, $P < 0.01$). Tukey–Kramer multiple comparisons indicate that eggs from lake C have the highest and lake B the lowest residues ($P < 0.05$), but lakes A and D do not differ significantly from each other.

The results could also be presented using a letter code above each bar of a histogram as in Figure 7.2.

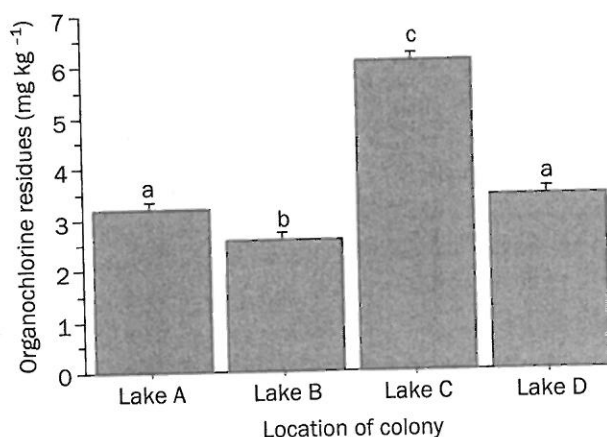


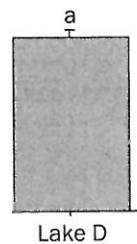
Figure 7.2 Mean organochlorine residues (mg kg⁻¹) in gull eggs from four colonies. Bars indicate standard errors. Samples labelled with the same letter are not significantly different following Tukey–Kramer's multiple comparisons

Kruskal–Wallis one-way analysis of variance using ranks

Since parametric ANOVA assumes that the data are normally distributed and the variances are equal, for data that either do not fit these requirements (and cannot be transformed to fit them), or are measured on an ordinal scale, a nonparametric alternative may be required. In the same way that a Mann–Whitney U test provides a nonparametric alternative to the t test for two samples, the Kruskal–Wallis one-way analysis of variance using ranks is a nonparametric alternative used where more than two samples are compared. As in many such nonparametric tests, the calculations are made on the ranks of the data, rather than on the actual data values themselves. Exactly as for the Mann–Whitney U test in Chapter 4, the data are ranked across the whole table (i.e. by placing all the data in ascending order, ignoring the separate samples for the present and giving mean values to tied data – see Box 4.4). The sums of the ranks are then calculated for each sample separately (as for the Mann–Whitney U test). The statistic calculated by the Kruskal–Wallis test is called H and the formula for calculating it is given in Box 7.4. Like the test statistic X^2 (Chapter 6), H approximates to the chi-square distribution. We compare our calculated H value to a chi-square table (an extract of which is given in Table 7.6) at $k - 1$ degrees of

ine residues in gull eggs between
multiple comparisons indicate that
differences ($P < 0.05$), but lakes A and

above each bar of a histogram



eggs from four colonies. Bars
are not significantly different

Rank

nally distributed and the
requirements (and cannot be
scale, a nonparametric
Mann-Whitney U test provides a
the Kruskal-Wallis one-way
native used where more than
parametric tests, the calculations are
data values themselves.
The data are ranked across the
er, ignoring the separate
data – see Box 4.4). The sums
ily (as for the Mann-Whitney
test is called H and the formula
tic X^2 (Chapter 6), H
our calculated H value to a
(.6) at $k - 1$ degrees of

Box 7.4 Formulae for Kruskal-Wallis one-way analysis of variance using ranks

To compare k samples, calculate the test statistic:

$$H = \left[\frac{12}{n_T(n_T + 1)} \times \left(\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \frac{(\sum R_3)^2}{n_3} + \dots + \frac{(\sum R_k)^2}{n_k} \right) \right] - 3(n_T + 1)$$

Here k is the number of samples being compared; n_T is the total number of observations across all of the samples; n_1 the number of observations in sample 1, n_2 the number of observations in sample 2, and so on up to sample k ; $\sum R_1$ the sum of ranks of observations in sample 1, $\sum R_2$ the sum of ranks of observations in sample 2, and so on up to sample k .

Then compare this H value to chi-square with $k-1$ degrees of freedom (see Table D.8, Appendix D), and reject the null hypothesis if the calculated value is greater than the table value.

Correction for ties

Carry out a parametric ANOVA (see Box 7.1 and Table 7.2) on the ranks of the data, but instead of finding an F value, take SS_{between} and MS_{total} to find the H value:

$$H = \frac{SS_{\text{between}}}{MS_{\text{total}}}$$

where SS_{between} is the sum of squares for the between variation, and MS_{total} is the SS_{total} divided by the total degrees of freedom.

freedom, where k is the number of samples being compared. If the calculated value of H is greater than the table chi-square value, then there are significant differences between the means. The procedure differs when there are three samples all with fewer than five observations (see Siegel and Castellan, 1988). In this case, the statistic does not approximate to the chi-square distribution and alternative tables of H should be consulted (see texts such as Siegel and Castellan 1988).

Although it may not seem so from the formula in Box 7.4, the Kruskal-Wallis test is the same as a one-way parametric ANOVA performed on the ranks of the data (although a different test statistic, an H value rather than an F value, is calculated at the end). The formula in Box 7.4 is much easier to use than a parametric ANOVA of the ranks because it is a simplified version that assumes an absence of ties. However, if there are tied ranks then the test is slightly too conservative (i.e. a type II error is more likely). If your computer program displays H corrected for ties and its associated P value, then it is these you should record. The presence of ties is usually only important if you obtain a marginally non-significant result (i.e. one where the calculated value of H is just less than the tabled value of chi-square with $k - 1$ degrees of freedom). In this instance, it may be worth recalculating it by parametric ANOVA

of the ranks and finding H by dividing the SS_{between} by MS_{total} (see the formula at the bottom of Box 7.4). Under these circumstances, you may find that your result is significant after all.

To demonstrate the Kruskal–Wallis test, we will analyse the responses of three samples of people who were asked about their feelings regarding the dangers to the environment of a large oil refinery. One set of people work in the refinery and live nearby, the second set also live nearby but work elsewhere, and the final set neither work at the plant nor live nearby.



WORKED EXAMPLE 7.3A *Kruskal–Wallis analysis of perceptions of the environmental dangers of an oil refinery by people with differing living and working connections to the plant: calculating the H statistic*

Living nearby and working in the oil refinery (A)		Living nearby and working elsewhere (B)		Living and working elsewhere (C)	
Scores	Rank	Scores	Rank	Scores	Rank
2	7	3	15	3	15
3	15	4	24	4	24
1	2	4	24	2	7
2	7	5	29.5	1	2
2	7	3	15	3	15
1	2	4	24	4	24
3	15	2	7	2	7
4	24	4	24	3	15
3	15	5	29.5	3	15
2	7	4	24	4	24
$\Sigma R_A = 101$		$\Sigma R_B = 216$		$\Sigma R_C = 148$	

The ranked values are placed next to the original scores. Note that the accuracy of the ranking process can be checked by adding up the sums of the ranks ($\Sigma R_A + \Sigma R_B + \Sigma R_C$) which should equal:

$$\frac{n_T(n_T + 1)}{2}$$

where n_T is the total number of observations across all samples. So:

$$\Sigma R_T = \Sigma R_A + \Sigma R_B + \Sigma R_C = 101 + 216 + 148 = 465$$

continued . . .

(see the formula at the
id that your result is



perceptions of the
ffering living and
ic

iving and working
lsewhere (C)

cores Rank

15

24

7

2

15

24

7

15

15

24

$\sum R_C = 148$

ote that the accuracy
ns of the ranks

ples. So:

continued . . .

Worked Example 7.3a continued . . .

and:

$$\sum R_T = \frac{n_T(n_T + 1)}{2} = \frac{30 \times 31}{2} = 465$$

Therefore the ranking is correct. Next we calculate H :

$$H = \left[\frac{12}{n_T(n_T + 1)} \times \left(\frac{(\sum R_A)^2}{n_A} + \frac{(\sum R_B)^2}{n_B} + \frac{(\sum R_C)^2}{n_C} \right) \right] - 3(n_T - 1)$$

$$= \left[\frac{12}{30 \times 31} \times \left(\frac{101^2}{10} + \frac{216^2}{10} + \frac{148^2}{10} \right) \right] - 3 \times 31$$

$$= (0.012\,903\,225 \times 7876.1) - 93 = 8.627$$

This value is then compared to chi-square with 2 degrees of freedom ($k - 1$). The calculated value (8.627) is larger than the table value at $P = 0.05$ ($\chi^2 = 5.991$) and smaller than the table value at $P = 0.01$ ($\chi^2 = 9.210$), and is therefore significant at $P < 0.05$.

Table 7.6 Selected values of chi-square (χ^2). Compare H at $df = \text{number of samples minus one}$. Reject the null hypothesis if the calculated value of H^2 is greater than the table chi-square value. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of chi-square values is given in Table D.8 (Appendix D)

df	χ^2	
($n - 1$)	$P = 0.05$	$P = 0.01$
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086

Table B.6 (Appendix B) for details of data entry for statistical analysis using computers.

From the data in Worked Example 7.3a, the H statistic is calculated as 8.63. When compared with chi-square tables at $k - 1$ (here $3 - 1 = 2$) degrees of freedom (see the shaded area of Table 7.6), we find that the probability that the null hypothesis is true

The respondents are asked to score their attitudes on a five-point scale (where 1 indicates no dangers perceived, and 5 many dangers perceived). The data, ranks and calculation of H are given in Worked Example 7.3a. Ranks are calculated as shown in Box 4.4. Although our example has equal sample sizes, it is not necessary for this test.

Note that the data in Worked Example 7.3a are listed in three separate columns (one for each sample). If you use a computer program to calculate your Kruskal-Wallis tests, this data format will not usually be appropriate. See

is between $P = 0.05$ ($\chi^2 = 5.991$) and $P = 0.01$ ($\chi^2 = 9.210$). Therefore we can reject the null hypothesis and say:

There is a significant difference in the perceptions of the environmental dangers of a large oil refinery between people with differing connections (living and working) to the plant ($H = 8.627$, $df = 2$, $P < 0.05$).

There were a large number of ties in our data, therefore our H value is conservative. Since we have rejected the null hypothesis anyway, it does not alter the major outcome. However, if we had taken the tied values into account, H would have been 9.25 (calculated using a parametric ANOVA of the ranks – see the formula at the bottom of Box 7.4), and we would have rejected the null hypothesis with greater confidence ($P < 0.01$).

We can now say there is a difference in the perceptions of the three groups of people towards the dangers to the environment of the oil refinery, but we cannot say definitively where the differences lie. For this we need suitable multiple comparison tests. As with the parametric equivalents, there is a different test for equal (Nemenyi's test) and unequal (Dunn's test) sample sizes. Note that most computer programs do not calculate either Nemenyi or Dunn's tests.

Multiple comparison tests with equal sample sizes: the Nemenyi test

Where sample sizes are equal, the Nemenyi test is used to compare each possible pair of rank means. A rank mean is the sum of the ranks divided by the sample size (i.e. for sample A, the rank mean of A is $\sum R_A / n_A$). First, the MSD between rank means is calculated (see Box 7.5). This MSD is then compared to the absolute differences

Box 7.5 Formula for the Nemenyi multiple comparison test for use after Kruskal–Wallis analysis of variance using ranks (where sample sizes are equal)

The minimum significant difference is:

$$MSD = q \times SE$$

where the standard error is:

$$SE = \sqrt{\frac{k(n_T + 1)}{12}}$$

and:

q is the value obtained from the q distribution table using the number of samples involved (k) and degrees of freedom of infinity;

k is the number of samples being compared;

n_T is the total number of observations across the analysis.

)). Therefore we can reject

onmental dangers of a large oil
d working) to the plant

ir H value is conservative.
s not alter the major
count, H would have been
- see the formula at the
hypothesis with greater

the three groups of people
, but we cannot say
table multiple comparison
ent test for equal (Nemenyi's
ost computer programs do

ie Nemenyi test

o compare each possible pair
ed by the sample size (i.e.
MSD between rank means is
he absolute differences

comparison test for e using ranks (where

umber of samples involved (k)

(i.e. the differences ignoring any negative signs) between pairs of rank means. If an absolute difference is greater than the MSD, then the difference is significant. As in the Tukey test, this test also uses q and a standard error term, but this time q is consulted at k (the number of samples) and infinite degrees of freedom (given the symbol ∞). Using the rank means of the perceptions of the environmental dangers of a large oil refinery among people with different living and working connections to the plant, the Nemenyi test is calculated in Worked Example 7.3b.



WORKED EXAMPLE 7.3B *Kruskal-Wallis analysis of perceptions of the environmental dangers of an oil refinery by people with differing living and working connections to the plant: calculating the Nemenyi test*

For multiple comparisons on our example data from Worked Example 7.3a:

$$SE = \sqrt{\frac{k(n_T + 1)}{12}} = \sqrt{\frac{3 \times 31}{12}} = 2.783\,882\,181$$

From Table 7.7, q is 3.31, and:

$$MSD = q \times SE = 3.31 \times 2.783\,882\,181 = 9.215$$

This value is compared to the absolute difference between the rank means, given in the table below (rank means are given in brackets; differences between rank means that are greater than the MSD are highlighted with an asterisk):

	<i>Living close and working in refinery, A (10.1)</i>	<i>Living close and working elsewhere, B (21.6)</i>
Living close and working elsewhere, B (21.6)	11.5 *	—
Living and working elsewhere, C (14.8)	4.7	6.8

Table 7.7 Selected values of q for the Nemenyi test. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of q values is given in Table D.10 (Appendix D)

	Number of samples being compared (k)		
	3	4	5
Degrees of freedom	∞	3.31	3.63
			3.86

From Worked Example 7.3b, the minimum significant difference is 9.215. The only case where the difference between the rank means is greater than the MSD is the comparison of people living close to and working at the plant, with those living close to but working elsewhere:

Living close and working elsewhere > Living close and working in refinery.

However, 'Living and working elsewhere' lies somewhere between the above two; the multiple comparison test is not able to place it. This ambiguity occurs because multiple comparisons are not powerful tests, tending to increase the chances of type II errors (i.e. accepting the null hypothesis when in fact it is false): a further survey (perhaps incorporating larger samples) would be necessary to gain a better picture of where the scores from that group of people lie. We could describe this result in the following way:

There was an overall significant difference in perception of the environmental dangers of the oil refinery for people with different living and working connections to the refinery ($H = 8.63$, $df = 2$, $P < 0.05$). Using Nemenyi multiple comparison tests, those living near to and working elsewhere had significantly higher perception scores than those living near and working at the refinery ($P < 0.05$).

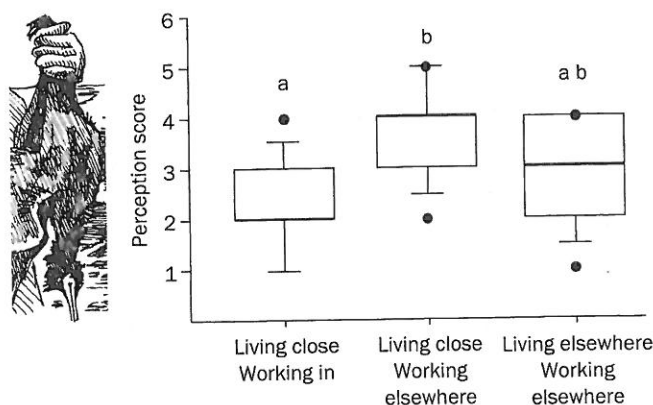


Figure 7.3 Median perception scores of the environmental dangers of a large oil refinery by people with differing work and living connections to the plant ($n = 10$ for each group). Scores range from 1 (no perceived dangers) to 5 (many perceived dangers). Boxes labelled with the same letter are not significantly different following the Nemenyi test for multiple comparisons

The medians and quartiles (see Boxes 2.2 and 2.7), along with the multiple comparison results, could be displayed in an annotated box and whisker plot as in Figure 7.3.

Multiple comparison tests with unequal sample sizes: Dunn's test

Where multiple comparisons are required, but the sample sizes are not equal, Dunn's method is appropriate. Here (like the Tukey–Kramer method for parametric ANOVA) a separate MSD must be calculated for each pair of samples. The formula, given in Box 7.6, is similar to that of the Nemenyi test in Box 7.5, but a

different critical value is used (Q , an extract of a table of Q values is shown in Table 7.8). As an example, data on stock damage in commercial forests from three regions of Britain after a series of winter storms are given in Worked Example 7.4. The data are the percentage of the stock from each of 20 forests (8 in region A, 5 in region B and 7 in region C) which show substantial foliar damage (where over 10% of their foliage, relative to a perfect tree, is lost). The overall H value is significant, and

ence is 9.215. The only
than the MSD is the
it, with those living close

king in refinery.

etween the above two; the
ty occurs because multiple
chances of type II errors
further survey (perhaps
better picture of where the
result in the following way:

environmental dangers of the
tions to the refinery ($H = 8.63$,
ose living near to and working
living near and working at the

ians and quartiles (see
2 and 2.7), along with the
comparison results, could
yed in an annotated box
ker plot as in Figure 7.3.

le comparison tests nequal sample sizes: test

multiple comparisons are
, but the sample sizes are
1, Dunn's method is
ate. Here (like the
Kramer method for
ric ANOVA) a separate
ust be calculated for each
amples. The formula, given
7.6, is similar to that of the
i test in Box 7.5, but a
values is shown in Table
forests from three regions
ed Example 7.4. The data
n region A, 5 in region B
where over 10% of their
lue is significant, and

Dunn's multiple comparisons demonstrate where the differences lie, thus we can say that:

There is a significant difference between the percentage of stock showing foliar damage in three regions ($H = 13.913$, $df = 2$, $P < 0.01$, $n_A = 8$, $n_B = 5$, $n_C = 7$). Dunn's multiple comparison tests show that the mean rank of region A is lower than the mean ranks of either region B or C, and that regions B and C did not differ significantly from each other ($P < 0.05$).

Box 7.6 Formula for Dunn's multiple comparison test for use after Kruskal–Wallis analysis of variance using ranks (where sample sizes are not equal)

The minimum significant difference is:

$$MSD_{A,B} = Q \times SE_{A,B}$$

where the standard error (SE) is:

$$SE_{A,B} = \sqrt{\frac{n_T(n_T+1)}{12} \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}$$

and:

$MSD_{A,B}$ is the minimum significant difference between samples A and B;

Q is the value obtained from the Q distribution table (an extract of which is shown in Table 7.8) using the number of samples (k) – note that this is different to the q table used previously;

n_T is the total number of observations across all the samples;

n_A and n_B are the number of observations in samples A and B.



WORKED EXAMPLE 7.4 Kruskal–Wallis analysis followed by Dunn's test for stock damage to commercial forests in three regions of Britain

Region A ($n = 8$)		Region B ($n = 5$)		Region C ($n = 7$)	
% stock	Rank	% stock	Rank	% stock	Rank
72	1	89	12.5	86	10
76	2	85	9	90	14
80	5	92	15	88	11
82	7	95	19	95	19
84	8	89	12.5	95	19
81	6			93	16
78	3			94	17
79	4				
$\sum R_A = 36$		$\sum R_B = 68$		$\sum R_C = 106$	
Mean rank of A = 4.5		Mean rank of B = 13.6		Mean rank of C = 15.143	

continued . . .

Worked Example 7.4 continued . . .

First check the ranking procedure:

$$\sum R_T = \sum R_A + \sum R_B + \sum R_C = 36 + 68 + 106 = 210$$

and

$$\sum R_T = \frac{n_T(n_T + 1)}{2} = \frac{20 \times 21}{2} = 210$$

The ranking is correct. Next, calculate H :

$$H = \left[\frac{12}{n_T(n_T + 1)} \times \left(\frac{(\sum R_A)^2}{n_A} + \frac{(\sum R_B)^2}{n_B} + \frac{(\sum R_C)^2}{n_C} \right) \right] - 3(n_T + 1)$$

$$= \left[\frac{12}{20 \times 21} \times \left(\frac{36}{8} + \frac{68}{5} + \frac{106}{7} \right) \right] - 3 \times 21$$

$$= (0.028\,571\,428 \times 2\,691.942\,857) - 63 = 13.913$$

From Table D.8, Appendix D, H is significant ($P < 0.01$) at $df = k - 1 = 3 - 1 = 2$. We carry out Dunn's multiple comparisons, first finding the standard error (SE) for each comparison using:

$$SE_{A,B} = \sqrt{\frac{n_T(n_T + 1)}{12} \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}$$

A ($n = 8$)B ($n = 5$)

$$B\ (n = 5)\ SE_{A,B} = \sqrt{\frac{20 \times 21}{12} \left[\frac{1}{8} + \frac{1}{5} \right]} = 3.372\,684$$

$$C\ (n = 7)\ SE_{A,C} = \sqrt{\frac{20 \times 21}{12} \left[\frac{1}{8} + \frac{1}{7} \right]} = 3.061\,862 \quad SE_{B,C} = \sqrt{\frac{20 \times 21}{12} \left[\frac{1}{5} + \frac{1}{7} \right]} = 3.464\,10$$

Multiplying throughout by $Q = 2.394$ (see the shaded area of Table 7.8), at $P = 0.05$ (where k is 3) to find MSD values:

	A	B
B	8.07	—
C	7.33	8.29

Worked Example 7.4 continued . . .

Differences in rank means between regions (rank means are given in brackets; differences between rank means that are greater than the MSD are highlighted with an asterisk):

	A (4.5)	B (13.6)
B (13.6)	9.1*	—
C (15.1)	10.6*	1.5

Table 7.8 Selected values of Q for Dunn's test. Values are for $P = 0.05$. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of Q values is given in Table D.11 (Appendix D)

	Number of samples (k)			
	3	4	5	6
	2.394	2.639	2.807	2.936



Parametric two-way analysis of variance

So far we have investigated differences between categories of one independent variable. However, with analysis of variance it is possible to simultaneously examine the differences due to more than one independent variable. For two independent variables we use a two-way analysis of variance. This also allows us to test for the presence of interactions between the variables: an interaction occurs when a combination of variables produces an effect greater (or less) than would have been expected by examining each variable alone. The design of a two-way experiment or survey can be visualised as in Figure 7.4, where the two independent variables form the r rows and c columns of a two-way design, and each cell is one of $r \times c$ samples, each one representing a particular combination of categories of the independent variable. Here, which of the two independent variables is allocated as the row variable and which is the column variable is arbitrary. If there is only one data point in each cell, then we need a different type of analysis (discussed at the end of this section). In this book, we only consider cases for two-way analysis of variance where the sample sizes are equal. If sample sizes are unequal, the computation and significance testing of the ANOVA are more complex – see texts such as Zar (1999) and Sokal and Rohlf (1995).

Note that Figure 7.4 is a convenient way to visualise two-way designs. However, if you use a computer program to calculate your ANOVA tests, this format will not be

 $t_T + 1)$

t $df = k - 1 = 3 - 1 = 2$. We
standard error (SE) for each

B ($n = 5$)

$$\sqrt{\frac{20 \times 21}{12} \left[\frac{1}{5} + \frac{1}{7} \right]} = 3.464\ 102$$

a of Table 7.8), at $P = 0.05$

3

3.29