One-way Analysis of Variance (ANOVA)

Introduction

The hypothesis that the means of two groups are equal can be assessed by an appropriate *t*-test, or possibly by some distribution-free analogue such as a Mann-Whitney test. Analysis of variance, often abbreviated to ANOVA, is the technique that is employed when there are more than two groups to compare. Of course, just as with a *t*-test where there are paired and non-paired versions available for data with different structures, there are several versions of ANOVA. The analogue of the unpaired *t*-test is one-way ANOVA and this is the technique that will occupy most of the following: two-way ANOVA, which is the analogue of the paired *t*-test, will be introduced by way of example in the exercise. ANOVA is actually a very powerful technique and there are many versions which will not be mentioned here: the interested reader can pursue these in Armitage and Berry (chapters 7 and 8, 1994). As will be seen, most ANOVAs assume the data analysed have a Normal distribution: several distribution-free versions are available, the common ones being associated with the names Kruskal-Wallis and Friedman, neither of which will be considered here.

Dataset

The technique will be demonstrated by the analysis of data on liver weights of mice

MINITAB - Untitled Worksheet - [Data]

File Edit Manip Calc Stat Window Graph Editor Help

	C1	C2	C3	C4	C5	C6	C7
↓	low	control	high				
1	3.64	3.42	3.17	3.64	1		
2	3.77	3.96	3.63	3.77	1		
3	4.18	3.87	3.38	4.18	1		
4	4.21	4.19	3.47	4.21	1		
5	3.88	3.58	3.39	3.88	1		
6	3.93	3.76	3.41	3.93	1		
7	3.91	3.84	3.55	3.91	1		
8	3.96		3.44	3.96	1		
9				3.42	2		
10				3.96	2		
11				3.87	2		
12				4.19	2		
13				3.58	2		
14				3.76	2		
15				3.84	2		
16				3.17	3		
17				3.63	3		
18				3.38	3		
19				3.47	3		

Figure 1: example dataset in Minitab spreadsheet

given different diets. Three diets were given which differed in the amount of carbohydrate they contained: 8 mice received a low carbohydrate diet, 8 received a high carbohydrate diet and 7 a control diet with an intermediate amount of carbohydrate. The liver weights are expressed as percentage of body weight. The data, shown entered on a Minitab spreadsheet in figure 1, have been entered in two

ways. Perhaps the more natural way is as three separate columns, each for a different diet, whereas the second method is to put all the weights in a single column and enter a code to indicate the diet in a second column: the second method is shown because the Minitab command that uses this form of input has some extra facilities over the version that uses three columns.

Overview and Assumptions of one-way ANOVA

This method is used to compare three or more independent groups: the meaning of independent will become clearer in the example sheet: it has the same meaning as in the unpaired *t*-test where two *independent* groups are compared. In the example, the three groups are independent because different mice appear in each group. In the following the term ANOVA will be taken to mean one-way ANOVA, unless otherwise indicated.

There are two assumptions underlying the technique: first it is assumed that within each group to be compared the data follow a Normal distribution: second, it is assumed that these Normal distributions share a common SD: if it is necessary to make explicit reference to this quantity the symbol σ will be used. The position is exemplified for three groups in figure 2 below, where the unknown population means

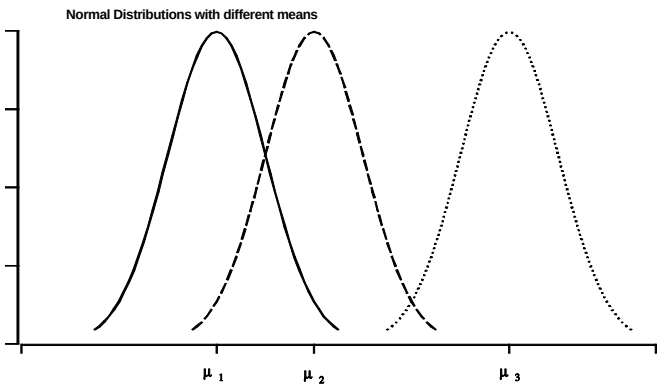

Figure 2: assumptions underlying one way ANOVA

in the groups are written as μ_1, μ_2, μ_3.

Performing an ANOVA has two stages: first an analysis to see if any differences at all exist. If it seems there may be differences, the second stage attempts to identify the nature of the differences. The reason for this is that the null hypothesis which ANOVA tests is simply that the group means are equal, that is:

Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$

Unlike the two group case, the null hypothesis can be false in several ways. All the means could differ, one could be different from the other two, they may follow a trend with another variable (e.g. in the example there may be a trend of liver weight with carbohydrate content). The second stage explores *how* the null hypothesis is contradicted. There are some circumstances when it is legitimate to enter the second stage even if the first stage provided no evidence against the null hypothesis.

However, such instances are very much the exception rather than the rule and it is best to regard the process as sequential, only entering stage two if stage one provides clear evidence that the means are not equal.

Stage I: the ANOVA table

The commands to perform one-way ANOVA in Minitab can be found under the **Stat** menu, as shown in figure 3. The command selected there is the one which assumes the data for the different groups are in different columns. The command **Oneway** is the version that assumes the responses are in a single column, with group membership indicated in a second column. This version is needed when certain quantities required for checking assumptions have to be computed; for the moment the

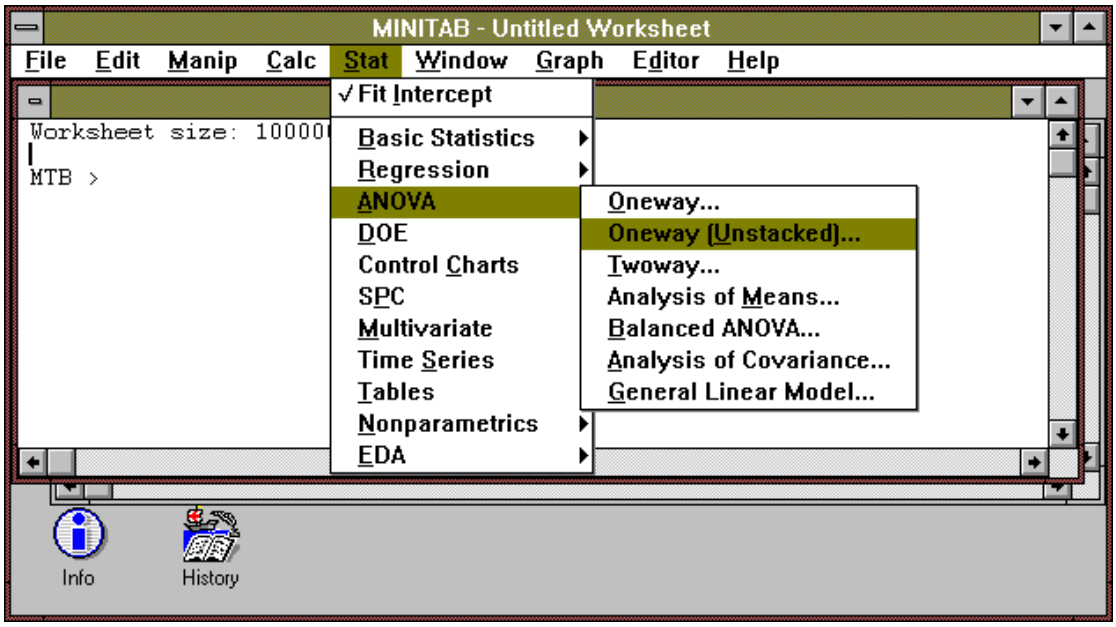

Figure 3: selecting one-way ANOVA in Minitab version 9

command in figure 3 will suffice and when applied to the data in figure 1 gives the output in figure 4.

The output in the session window in figure 4 has four components which need explanation, and these will be taken in turn.

1. Table of means

The table at the bottom left, just above POOLED STDEV, is simply the table of means and standard deviations for each of the groups. This has two uses: first, regardless of the sophistication of ANOVA, the sample means in the groups remain the most important summaries in the analysis; second, inspection of the standard deviations (SDs) in the groups allow the analyst to decide whether or not the assumption of a common population SD is tenable. No formal method of assessment is available and to an extent the decision whether or not the assumption is reasonable is a matter of judgment. SDs from small samples are very variable, so quite large apparent differences can be tolerated. Rather than looking at differences alone it can be informative to look for trends: a common way in which differing SDs can arise is when the SD of the response variable depends on the level of the response, e.g. there may be a tendency for larger SDs to be associated with larger means. In this case the

assumption of constant SD may be more reasonable when applied to the log of the response.

It is not widely appreciated that violating the assumption of equal SDs is more serious than violating that of Normality, so it is important that some attention is paid to

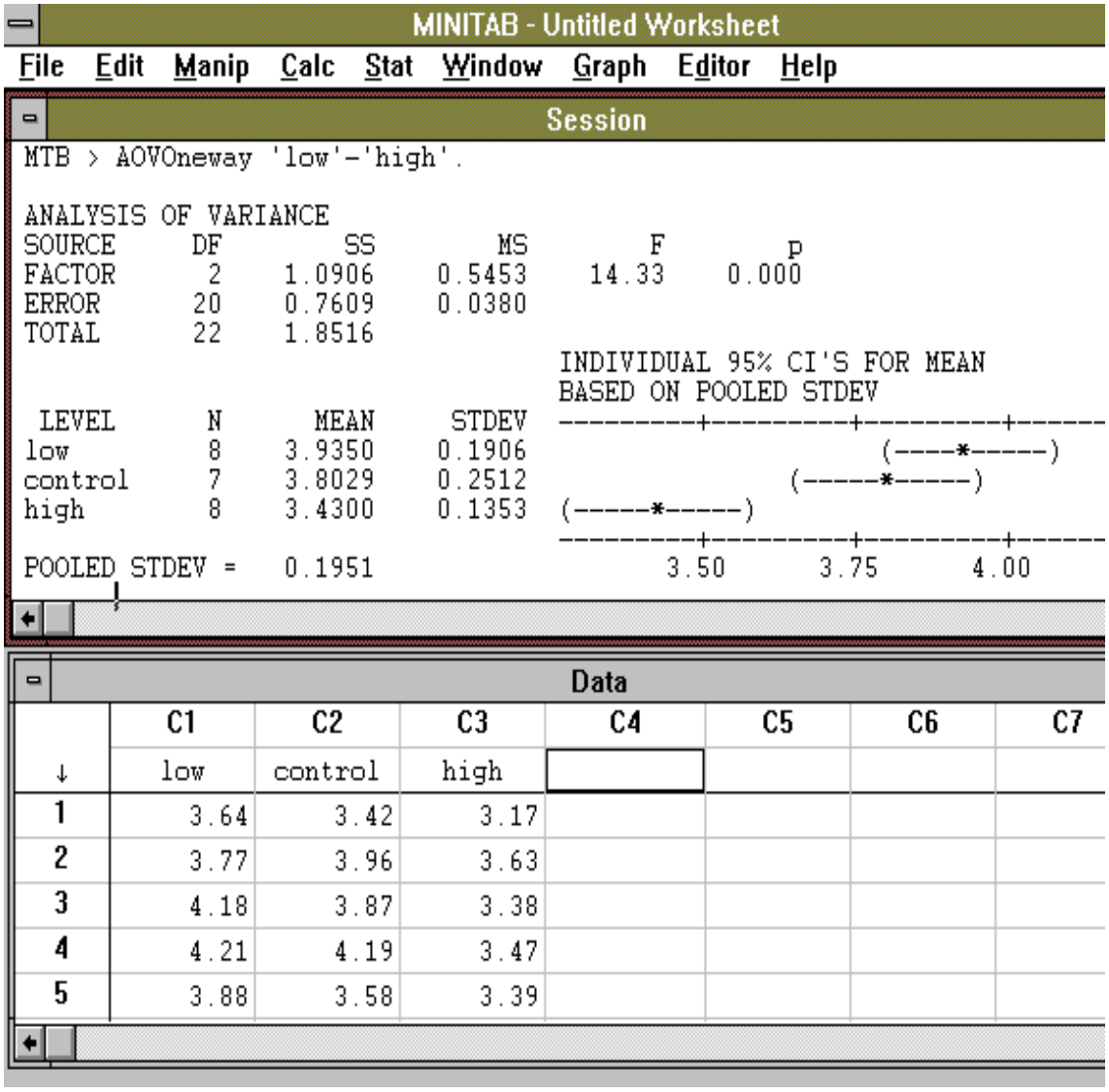

Figure 4: results of output of AOVO command

this point at the start of an analysis.

2. Pooled Standard Deviation

If the assumption that the SDs in the groups are the same appears reasonable, then a more precise estimate of σ is obtained by pooling information on variability across the groups. Under these circumstance the POOLED STDEV is the best estimate of σ, because it is based on information from all the groups. It is a form of average of the SDs from each group. There are two ways in which the average differs from a simple mean. First, it is the square of the POOLED STDEV that is the average of the squares of the group SDs. Second, the average is a weighted mean, so that the SDs from larger groups have more influence on the POOLED STDEV: for technical reasons the weights are one less than the sample size of the group.

For example, in the example the group SDs are 0.1906, 0.2512, 0.1353 from groups of size 8,7 and 8 respectively. The POOLED STDEV is found as the square root of:

$$\frac{(8-1)\times 0.1906^2+(7-1)\times 0.2512^2+(8-1)\times 0.1353^2}{(8-1)+(7-1)+(8-1)},$$

which equals 0.03805 and the POOLED STDEV is the square root of this, which is 0.1951, as given in figure 4.

3. Sketch Graph of Means

To the right of the table of means in figure 4 there is a sketch graph of the means, with confidence interval indicated. The latter are calculated using the pooled SD and assuming Normality, so they are only valid if the assumptions of the method are tenable.

4. Analysis of Variance Table

The first structure printed in the output is the ANOVA table and this is where the test of the null hypothesis is performed. A full explanation of the table would be very extensive, so only the essential aspects are considered here: a fuller explanation can be found in Armitage and Berry (1994), chapter 7 or Altman (1991) p.205-.

Consideration will be restricted to the columns headed MS, F and P: for details about the other columns, see the above references.

Even if the null hypothesis is true and the population means are the same for each group, the *sample* means will differ because of sampling variation. The purpose of the ANOVA table is to assess whether the discrepancies between the sample means are commensurate with differences that could arise because of the underlying variation (with SD σ) that is a feature of the responses. In a *t*-test the difference is simply the difference between the two means, but with more than two groups a more sophisticated approach is needed. The figure in the column headed MS (for Mean Square) and the row marked FACTOR (0.5453) measures the difference between the three group means. It is a form of squared SD based on the three means and suitably weighted to allow for differing precisions in the means due to different sample sizes.

If the null hypothesis is true then the only reason why the group means differ is the underlying variability which has SD σ. The MS for groups is defined so that if the null hypothesis is true then its value will be approximately σ^2. However, if the null hypothesis is false then the MS for groups will be inflated because it will then also reflect the difference between the group population means. ANOVA works by comparing this estimate of σ^2 with the estimate found from the variability *within* each sample, that is from the square of the POOLED STDEV. The latter is an estimate of σ^2 regardless of the truth of the null hypothesis, so if that based on the group MS greatly exceeds that based on the POOLED STDEV then there is evidence that the null hypothesis is false.

The square of the POOLED STDEV appears in the ERROR row of the MS column of the ANOVA table. The ratio of the group MS and the ERROR MS is given in the column headed F (note 0.5453/0.0380=14.33 to within rounding error): if this quantity is large then there is evidence that the null hypothesis is false. This ratio, known as the variance ratio or the F-statistic, is the analogue in ANOVA of the *t* value in the *t*-test.

This leaves open the question of how large is large? This is answered using the same approach as in the *t*-test, that is by working out how large the ratio can be when the null hypothesis is true. The final column of the table, headed p, shows how rare it is for an F value as large as that observed to occur if the null hypothesis is true. The interpretation is that *either* the null hypothesis is false *or* an event of probability p has occurred, so if p is small this provides evidence the null hypothesis is false. In the example an F value of 14.33 is so large that the p-value is printed as zero to 3 decimal places (more accurately it is 0.00014). So if the null hypothesis is true, the example is an event that would occur only about 3 times in every 20000 analyses, so it is sensible to interpret the result as providing strong evidence against the null hypothesis.

Assessing the Normality Assumption.

In point 1. above the importance of checking the assumption of equal SDs in the groups was stressed. The other assumption, namely that the data follow a Normal distribution, should also be scrutinised. Perhaps the most sensitive way to assess Normality is the Normal Plot, a plot which gives an approximate straight line if the data plotted have a Normal distribution: this can be produced using the specific Normal plot command under the **Graph** menu. Clearly the data cannot simply be plotted in this way because responses from different groups will have different means. To get round this the *residuals* are plotted, the residual being simply the observed value minus its group mean. The residuals all have zero mean and if the assumption of equal group SDs holds they share the same Normal distribution. Residuals can be found automatically if the stacked version of the one way ANOVA command is used. The result of a Normal plot of the example is shown in figure 5.

Stage II: Exploration of Differences.

If an ANOVA provides evidence against the null hypothesis, the process of

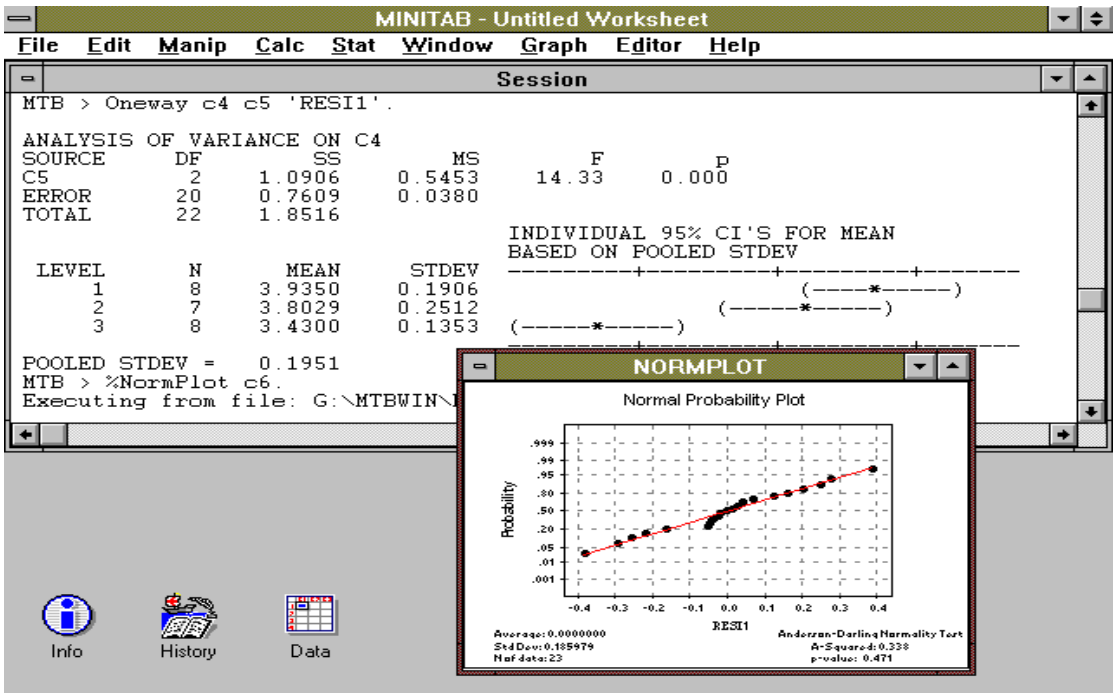

Figure 5: Normal plot of residuals from example

elucidating exactly *how* the hypothesis was contradicted has two distinct aspects. On

the one hand there are the arithmetical procedures which need to be performed to compare means or combinations of means and on the other there is the business of which comparisons should be made and why. The former is much more straightforward and can be explained briefly by way of example.

Post hoc Comparison of Means

Suppose there is interest in whether the means of the control and high carbohydrate diets differ. This is resolved by testing the sub-hypothesis $\mu_{high} = \mu_{control}$ using a form of *t*-test. The usual *t*-test involves the calculation of the quotient:

$$\frac{m_{control} - m_{high}}{se(m_{control} - m_{high})}$$

and it is the same here, the difference is in the way the denominator is computed. The numerator is 3.8029-3.4300=0.3729. The denominator is, following the usual formula for the standard error of a difference (Altman, 1991, p. 160-1),

$$\sqrt{se(m_{control})^2 + se(m_{high})^2}$$

and $se(m_{either}) = \sigma / \sqrt{n_{either}}$. It is at this point that the calculation departs from that for a standard *t*-test because the ANOVA was based on the assumption that the SD is the same in each group so σ is estimated from the SD pooled across all the groups (not just the high carbohydrate and control groups), that is POOLED STDEV. So in this case the standard errors are 0.1951/√7 and 0.1951/√8 and the denominator to the *t* statistics is:

$$0.1951 \times \sqrt{\tfrac{1}{7} + \tfrac{1}{8}} = 0.1010$$

and the *t* statistic is 0.3729/0.1010=3.69. The p-value corresponding to this *t* value is found in the same way as in the *t*-test, although because the denominator was based on the pooled SD, the degrees of freedom for the *t* distribution is that given in the ERROR row of the DF column in the ANOVA table, that is 20. This gives p=0.0014, and as the 95% point of the *t*-distribution on 20 degrees of freedom is 2.086, a 95% confidence interval for the difference is 0.3729±2.086×0.101=0.1619, 0.5839. Thus there appears to be good evidence that there is a difference between the percentage liver weights of those on high carbohydrate and control diets.

Do We Need a New Technique?

Given that two of the three groups have just been compared using a variant of the *t*-test, it may be thought that there is no need for ANOVA at all, surely the analysis could be accomplished by repeated applications of ordinary *t*-tests between pairs of groups? There are at least four relevant comments that indicate that ANOVA is needed:

a) it is usually only sensible to start pairwise comparisons when the overall ANOVA test is significant, and even then not all pairwise tests should be done;

b) if many hypothesis tests are performed it becomes difficult to judge which effects have arisen because of genuine differences and which simply because so many tests have been performed;

c) if, as is often the case, the assumption of equal SDs in the different groups holds, then basing the comparisons on a pooled estimate of SD makes better use of the available data;

d) use of ANOVA makes for more coherent inference. For example, if groups A, B and C are compared using pairwise comparisons, it is possible to find that A and B do not differ significantly and neither do B and C but there is a significant difference between A and C. The nature of 'a significant difference' means this is not actually contradictory but it can be puzzling and may be avoided if an ANOVA is used instead of pairwise comparison.

Which Comparisons to Make?

If an ANOVA yields a significant F-test, then it will almost certainly be of interest to discover how the difference has arisen. Proceeding to use the variant of the *t*-test described above to compare all possible pairs of groups will cause problems of the type just described in b) and d). There are various ways round this problem.

i) Various procedures, such as the Newman-Keuls test, Duncan's multiple range test, the least significant difference test, Scheffé's test and Bonferroni correction have been proposed to allow all comparisons to be made without running into some of problems of uncontrolled multiple comparisons. If the analyst has to embark on making all possible comparisons to find differences that may explain why the null hypothesis is not true then one of these procedures must be used; some are described in Armitage and Berry (3rd edn. 1994, p226-227). However, the results from these procedures can be difficult to interpret and are seldom satisfactory.

ii) The procedures described in i) are seldom needed, and indeed may be over-used in practice. The reason for this is that certain comparisons of groups arise naturally out of the design of the experiment and these can be tested using the variant of the *t*-test without recourse to any multiple comparison procedure. For example, in a cancer trial three treatment groups A, B and C are, respectively, palliative care, chemotherapy and radiotherapy+chemotherapy; then A vs B measures the effect of chemotherapy whereas B vs C measures the effect of adjuvant radiotherapy.
It is not necessary to compare just pairs of groups, more involved comparisons (known as contrasts) can be considered. For example, if groups A and B are two different dialyser membranes made by different firms but using the same technology and C is a membrane made using a new process, then it may be sensible to compare the old and new technologies by comparing the mean of groups A and B with group C. Another possibility that is useful when groups are defined by some factor implies an ordering, is to see if there is a trend across the groups. For example, if the carbohydrate contents of the diets in the example were known then a contrast representing the trend with carbohydrate content could be assessed. Details of the implementation of these techniques can be found in Armitage and Berry (3rd edn. 1994, p. 224-226). In all cases, it is helpful if comparisons to be made in an analysis are specified in the protocol of the experiment.

iii) A special way in which certain comparisons can arise is when the groups have a *factorial structure*. A trial of fluid and electrolyte regimens for neonates has four groups i) high fluid, high Na ii) high fluid, low Na iii) low fluid, high Na iv) low fluid, low Na. The groups are made up of the four combinations of the two *factors*, i.e. sodium and fluid, each at two *levels*, i.e. low and high. When treatment groups have this form certain comparison arise naturally: (i + ii) vs (iii + iv) measures the difference between the fluid levels whereas the sodium effect comes from the comparison (i + iii) vs (ii + iv). It is also possible to ascertain whether the effect of sodium depends on the fluid regimen in use from the

comparison (i + iv) vs (ii + iii), know as the *interaction* between sodium and fluids. If a factorial experiment has been used then the ANOVA should be extended to permit these comparisons to be made readily: see Armitage and Berry (1994, p. 249-259).

The matter of which comparisons should be made is a subtle matter and further discussion can be found in Armitage and Berry (1994, section 7.4) or Altman (1991, section 9.8.4).

References

Altman, DG (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London.

Armitage, P and Berry, G (1994) *Statistical Method in Medical Research* (3rd edn.). Blackwell, Oxford.