R Module 4

Single Factor Analysis of Variance

Certificate in EnvIroStats (Non-Award)

This document is part of an online Certificate in EnviroStats (Non-Award) by the University of Canberra. Course enquiries can be directed to the address below. Expressions of interest in the course can be made online through:

http://aerg.canberra.edu.au/envirostats

Copies of this publication are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: **bernd.gruber@canberra.edu.au, georges@aerg.canberra.edu.au**

R is an open source statistical programme. It is developed by:
R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL **http://www.R-project.org**.

SPONSORED BY:

Australian Government
Australian Centre for
International Agricultural Research

WorldFish
CENTER

Australian Government
Fisheries Research and
Development Corporation

Materials development team:

Author:	Arthur Georges, 2002, 2006
Instructional designer:	Peter Donnan, 2002
Editor:	Loretta Barnard, 2002
Graphic Design:	Peter Delgado, 2002
Desktop Publishing:	Kristi McDonald, 2004 Sue Bebbington, 2004
FDDU Project Manager:	Deborah Veness, 2002
Dynamic Web Page Design:	TCNI Software Solutions PO Box 47 LATHAM ACT 2615 Australia

First prepared in January, 2002 for Semester 1, 2002.
Reprinted January 2003 for Semester 1, 2003.
Reprinted January 2004 for Semester 1, 2004.
Reprinted November 2004 for Semester 1, 2005.
Revised and reprinted, June 2006
Reprinted February 2007 for Semester 1, 2007
Revised combined with former SAS Module 4 and converted to R, February 2011

Published by Technology & Educational Design Services

(TEDS)
University of Canberra
ACT 2601, AUSTRALIA

Module 4: Single Factor ANOVA

Lesson 1: Key Concepts in ANOVA

Overview

Analysis of variance, or ANOVA for short, is fundamental for much of the application of statistics in biology. In its simplest form, it provides an extension of the t-test, enabling the simultaneous comparison of two or more means. In more complex designs it enables us to consider the effects of two or more factors simultaneously without disregarding the possible interaction between them.

Analysis of variance is more than just a technique. It provides insight into the nature of variation of natural events. It represents a major conceptual advance which, once understood, will guide you in the way you plan and execute much of your research. Analysis of variance is an indispensable conceptual and practical tool for the modern biologist.

In its simplest form, the analysis is used to understand the effects of a single-factor, acting alone. We might, for example, wish to discover if 'Method of Determination' has an effect on the concentration of copper detected in fish tissue. The methods available to us include flame atomic absorption spectroscopy, graphite furnace atomic absorption spectroscopy, anodic stripping voltametry and inductive coupled plasma mass spectroscopy (Table 4–1).

Table 4–1. Comparison of four methods for measuring levels of copper in fish tissue.

Flame AAS	Graphite furnace AAS	Anodic stripping voltametry	Inductive coupled plasma mass spec
25	23	25	18
24	18	25	26
25	22	20	22
26	28	18	28
	17	23	17
	25	19	16
	19	26	
	16		

Analysis of variance provides a basis for deciding between whether there is good evidence of a true difference between methods, or whether the differences we observe between methods arose by chance alone. This is a basic similarity shared with all hypothesis tests.

In the context of ANOVA, 'Method of Determination' is the Factor, with four discrete **factor classes**, and copper concentration is the **response variable**.

Single-factor analysis of variance, or one-way ANOVA as it is sometimes called, can be considered in a practical sense to be an extension of the t-test. Had there been only two methods involved in the above comparison, a t-test would have been appropriate. Indeed, such an analysis was undertaken in Module 3. The t-test is used to test hypotheses regarding the equality of two population means, whereas the single-factor ANOVA tests hypotheses regarding two or more population means.

Two models of single-factor ANOVA are recognised—the **fixed model** and the **random model**. In the fixed model, the objective of the analysis is not only to determine whether the factor under consideration has an overall effect, but which factor classes differ significantly from which others. The factor classes are chosen specifically and are therefore under the full control of the experimenter. In the case above, we might be interested both in whether the method of determination has an overall effect on the levels of copper detected and may wish to know which method gave the highest determination. The methods are fixed in the sense that if the study were to be repeated, exactly the same methods would be chosen again. Because we are interested in the performance of the four specifically chosen methods of analysis in the example above, the model is fixed.

A significant result in the fixed model single-factor ANOVA indicates significant variation among the means over and above that expected to occur by chance alone, but it does not provide information on which factor classes differ from which others. A significant result in the ANOVA must be followed by a set of comparisons to determine where the differences lie. The appropriate procedures to follow-up a significant result in a fixed model include testing the significance of differences between pairs of means using one of several **multiple comparison procedures**. R provides a wide range of options for undertaking multiple comparisons following a significant result in a fixed ANOVA.

In the random model, the factor classes are selected at random from a substantial population of possible choices. Variation among the observed means is due both to sampling error in the selection of factor classes and to variation among measurements within factor classes. The factor is not considered fixed because if the experiment or study were to be repeated, there would be no reason to expect that the same specific factor classes would be chosen again. In a random model, one is not interested in comparing specific means, only in whether the factor under consideration has had a differential effect. The appropriate follow-up procedure in the random model is to estimate the added variance component due to the effect of the factor. The example given above would be a random model if we had chosen the four methods listed in Table 4–1 at random from a large pool of possible choices. Had this been the case, we would not have been interested in specific comparisons among methods, only in whether choice of method added significantly to variation in the determinations of copper concentration.

Examples of fixed and random designs in single-factor ANOVA are given as worked examples later.

Rationale

Analysis of variance arose from the study of replicated samples, that is, from the study of samples drawn from the same or identical populations. Consider by way of example, a case where an experiment is designed to investigate the effect of cropping by macro-invertebrates on algal standing crop. Seven transects were established parallel to the banks of a small stream, in what appeared to be homogeneous riffle environment. Five paving bricks were placed along each transect and left for two months to accumulate a cover of algal growth. At the end of two months, they were removed from the stream and the accumulated algae was scraped from them, dried and weighed (Table 4–2).

This design has the factor TRANSECT with seven factor classes corresponding to the seven transects. Each transect comprises a sample of five replicated paving bricks and corresponding biomass measurements (the response variable). The transects themselves are laid in a homogeneous environment. The samples each of five pavers can therefore be considered to represent identical populations, and we should not expect significant differences among transects. Bear this in mind for the discussion that follows.

Table 4–2. Algal biomass (mg/m^2) from paving bricks laid along transects in shallow stream riffle.

Transect						
1	2	3	4	5	6	7
22.4	10.3	21.9	36.4	19.6	15.7	20.2
18.7	9.5	16.8	21.9	19.5	24.0	18.5
29.2	25.6	16.0	24.7	21.4	23.2	33.0
23.2	24.0	18.8	21.2	28.8	16.8	19.9
22.9	21.0	27.9	23.9	28.3	17.7	21.6

Statistics from these data reveal an ambiguous pattern (Table 4–3, Figure 4–1). Clearly the means vary from transect to transect, but then replicated means taken from the same population would be expected to vary by chance alone, through sampling error.

Table 4–3. Mean and variance in algal biomass (mg/m^2) collected from paving bricks.

	Transect						
	1	2	3	4	5	6	7
Mean	23.3	18.1	20.3	25.7	23.5	19.5	22.6
Variance	14.24	58.57	23.34	38.35	21.69	14.73	34.75

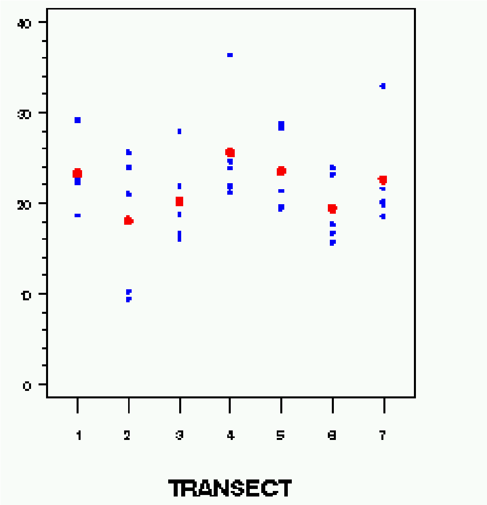


Figure 4–1. Variation in algal biomass (mg/m^2) collected from paving bricks laid along transects in shallow stream riffle. Means are shown as larger dots.

The central problem addressed by ANOVA is:

Do the observed differences among sample means provide evidence of true differences among the populations from which the samples were drawn, or did the observed differences arise by chance alone?

The basis for a decision on this comes from examination of variances:

- the variances among measurements within samples (ie the transects) and
- the variance of the sample means.

It is one of the ironies of the analysis that much time is spent considering variances in order to test a hypothesis on means.

If each sample of five pavers can be considered to have been drawn from the same or identical populations, they must have the same parametric mean and variance. Consider now how we might estimate the common population variance in biomass within transects, that is σ^2.

The obvious approach is to calculate the variance separately for each transect,

$$VAR[Y] = \frac{\sum^{n}\left(Y - \overline{Y}\right)^2}{n-1} \approx \sigma^2$$

for $n = 5$ pavers.

These are shown in Table 4–3. However, no single such estimate will give us the best estimate of the common population variance, σ^2, because each is based on only five values. Instead, a better estimate of the common population variance is obtained by averaging the a = 7 sample variances.

$$\overline{VAR[Y]} = \frac{1}{a}\sum^{a}\frac{\sum^{n}\left(Y - \overline{Y}\right)^2}{n-1}$$

$$= \frac{\sum^{a}\sum^{n}\left(Y - \overline{Y}\right)^2}{a(n-1)}$$

$$= 29.38 \approx \sigma^2$$

This estimate of the common population variance, calculated as the average within sample variance, is called **mean square within**. It is an estimate of the common population variance that is calculated solely from variation among measurements within samples (ie $Y - \overline{Y}$), in this case, within transects.

A second method for estimating the common population variance involves using information on variation among the seven replicated sample means. Statisticians have established a relationship between the variation exhibited by replicated sample means and the variation exhibited by measurements within samples. We have:

$$\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$$

This equation is most familiar when expressed in terms of standard deviations rather than variances. It is simply a restatement of the equation that equates the standard error as the standard deviation over the square root of the sample size *n*.

This equation makes intuitive sense. The variance of the means will depend inversely on the sample size:

$\sigma^2_{\overline{Y}}$ is inversely proportional to *n* because the larger the samples, the closer will each of the sample means be to the true parametric mean and the lower will be the variance among the sample means.

Also, the variance of the means will be directly proportional to the within sample variance:

$\sigma^2_{\bar{Y}}$ is directly proportional to σ^2

because the more variable the measurements themselves, the less representative each sample mean will be of the true parametric mean, for a given sample size.

So according to the equation

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$$

the greater the variance within samples, the greater will be the variance among the means and the greater the sample size, the smaller will be the variance among the means.

This equation can be rearranged to yield

$$n.\sigma^2_{\bar{Y}} = \sigma^2$$

which provides us with a second method of estimating the common population variance, this time based on observed variation among replicated population means.

$$n.VAR\left[\bar{Y}\right] \approx \sigma^2$$

Taking the variance of the means of Table 4–3 yields 6.99 and multiplying by the sample size (x 5) yields an estimate of the common population variance of 34.97. This estimate, because it is calculated from variation among the sample means, is called **mean square among**.

Substituting the formula for $VAR\left[\bar{Y}\right]$ to provide a formula in the same form as that for MS$_{\text{within}}$ yields:

$$MS_{among} = n.\frac{\sum^{a}\left(\bar{Y} - \bar{\bar{Y}}\right)^2}{a-1}$$

which can be rewritten

$$MS_{among} = \frac{\sum^{a}\sum^{n}\left(\bar{Y} - \bar{\bar{Y}}\right)^2}{a - 1} \approx \sigma^2$$

since the sum of the squared deviations of class means from the grand mean is constant for all *n*.

So now we have two independent estimates of the common population variance. One, called MS_{within}, is calculated as the average within sample variance and is independent of information on variation among sample means. The other, called MS_{among}, is calculated from the observed variation among sample means, and is independent of information on the variability of measurements within samples. Both estimate the parameter σ^2.

$$MS_{among} = n.VAR\left[\bar{Y}\right]$$

$$= \frac{\sum^{a}\sum^{n}\left(\bar{Y} - \bar{\bar{Y}}\right)^2}{a - 1} \approx \sigma^2$$

$$MS_{within} = \overline{VAR\left[\bar{Y}\right]}$$

$$= \frac{\sum^{a}\sum^{n}\left(Y - \bar{Y}\right)^2}{a(n - 1)} \approx \sigma^2$$

We could test to see if these two estimates of the common population variance differ significantly by performing an F-test:

$$F = \frac{MS_{among}}{MS_{within}} = \frac{34.97}{29.38} = 1.19$$

MS_{among} is a variance estimate based on *a*=7 values (the seven sample means) and so has *a*-1=6 degrees of freedom. MS_{within} is the average of *a*=7 variances, each with degrees of freedom, and so has *a*(*n*-1)=28 degrees of freedom. For reasons that will become apparent later, the F-test in ANOVA is a one-tailed test.

From tables:

$F_{0.05(1)[6,28]} = 2.45$

and the two estimates of the common population variance are not significantly different. This should come as no surprise because, for replicated transects, MS_{among} and MS_{within} independently estimate the same parameter, the common population variance σ^2. If you are unable to follow the argument at this point, refer back to hypothesis testing and the F-test in Module 3.

Consider now what happens when we apply a differential treatment to the transects. While MS_{among} and MS_{within} are independent estimates of the common population variance, they differ in one important property. If we introduce a factor that alters the mean algal biomass of some of the transects but not others, while maintaining the spread of biomass measures about the mean for each transect, then MS_{among} will be inflated by the differential treatment while MS_{within} will remain unbiased. This can be best visualised by considering Figure 4–2.

Table 4–4. Mean and variance in algal biomass (mg/m^2) collected from paving bricks laid along transects in shallow stream riffle. Means for transects 1, 2 & 3 are inflated by application of an insecticide (+ 15 mg/m^2).

	Transect						
	1	2	3	4	5	6	7
Mean	38.3	32.1	35.3	25.7	23.5	19.5	22.6
Variance	14.24	58.57	23.34	38.35	21.69	14.73	34.75

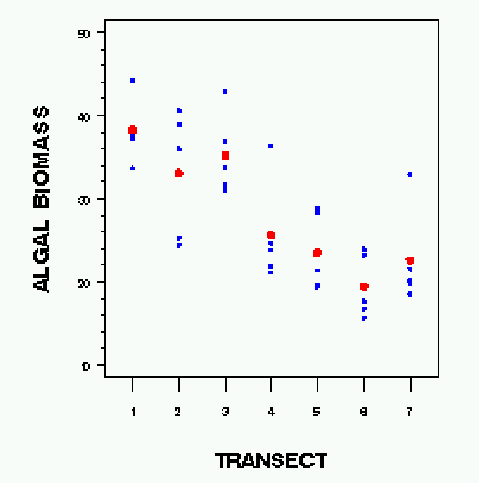


Figure 4–2. Variation in algal biomass (mg/m^2) collected from paving bricks laid along transects in shallow stream riffle. Means are shown as dots. Means for transects 1, 2 & 3 are inflated by application of an insecticide.

Table 4–4 and Figure 4–2 show the situation where a differential factor is applied to the transects. Algae reproduce and grow on each paver only to be consumed by a variety of invertebrate grazers. The biomass present at any one time reflects a balance between these two processes. Paving tiles in transects 1, 2 and 3 had the residual insecticide Pyrethrum applied to reduce grazing pressure upon them. The obvious effect of this was to increase the standing crop of algal biomass on the treated tiles compared with the untreated tiles, by 15 mg/m^2 in each case. In terms of the calculations presented above, the effect of the differential application of insecticide was to increase the variability among the sample means. Provided this effect was additive, the relative position of the transect means would be altered, but the spread of biomass measurements about those means would remain the same.

Hence MS$_{\text{among}}$ becomes an inflated, biased estimate of the common population variance, because of the differential effect of the insecticide, and MS$_{\text{within}}$ remains an unbiased estimate of the common population variance. This bias can be tested using the F-test.

Because the variance among the means is inflated by the differential effect of the factor applied, we have:

$$VAR\left[\overline{Y}\right] \approx \sigma_{\overline{Y}}^2 + \sigma_A^2$$

where $\sigma_{\overline{Y}}^2$ is the variance among means expected by chance alone and σ_A^2 is the added variance component due to the effect of the factor applied differentially to the various transects. For MS$_{\text{among}}$ we have:

$$MS_{among} = n.VAR\left[\overline{Y}\right] \approx n.\sigma_{\overline{Y}}^2 + n.\sigma_A^2$$

$$\approx \sigma^2 + n.\sigma_A^2$$

As MS$_{\text{within}}$ continues to estimate σ^2 without bias

$$F = \frac{MS_{among}}{MS_{within}} \approx \frac{\sigma^2 + n, \sigma_A^2}{\sigma^2}$$

The null hypothesis is

$$H_0 : \sigma^2 + n.\sigma_A^2 = \sigma^2$$

or in other words

$$H_0 : \sigma_A^2 = 0$$

The alternative hypothesis is

$$H_1 : \sigma^2 + n.\sigma_A^2 > \sigma^2$$

or in other words

$$H_1 : \sigma_A^2 > 0$$

since a variance component cannot be negative. Hence, the F-test of ANOVA is a one-tailed test.

Under the null hypothesis, $\sigma_A^2 = 0$ and so MS_{among} and MS_{within} both estimate the common population variance σ^2. The F ratio would be around 1, subject only to sampling error. If the added variance component σ_A^2 is non-zero, then the F ratio will, on average, be greater than 1 by more than would be expected by chance alone, and the result will, on average, be significant.

If the factor has no effect then, barring a Type I error, the results of the ANOVA will support the null hypothesis. If the factor has an effect then, barring a Type II error, the result of the ANOVA will be significant.

For the data at hand (Table 4–4), we have:

$$MS_{among} = n.VAR\left[\overline{Y}\right] = 250.19$$

$MS_{within} = \overline{VAR\left[Y\right]} = 29.38$ as before.

$$F = \frac{MS_{among}}{MS_{within}} = \frac{259.19}{29.38} = 8.82$$

When compared to the tabulated F of 2.45 (see page 4-11), the result is highly significant. In fact, the probability of getting an F ratio of 16.96, or one more extreme by chance alone, is less than 0.0001. We conclude that there is an effect of the differential application of pyrethrum on algal biomass.

Hence, by comparing two estimates of the common population variance, one biased by the effect of the factor under consideration and the other remaining unbiased whether or not the factor has an effect, we are able to decide whether or not the differential treatment had an effect of the means. We come to a decision on the equality of the sample means based on a comparison of variances.

An intuitive view

The above interpretation of F in ANOVA as the ratio of two independent estimates of the common population variance may be satisfying for those of us with a sound education in statistical theory, but for the rest of us, it is a difficult concept to retain in plain language thinking.

Recall that the observed variance of the sample means is given by

$$VAR\left[\overline{Y}\right] = \frac{\sum^{a}\left(\overline{Y} - \overline{\overline{Y}}\right)^2}{a - 1}$$

which can readily be calculated from the set of a sample means. This is how variable the means actually are.

But given the average variation within samples, MS_{within}, we would <u>expect</u> the means to vary by

$$\frac{MS_{within}}{n}$$

If we compare the observed with the expected in the form of an F ratio

$$F = \frac{VAR\left[\overline{Y}\right]}{MS_{within}/n} = \frac{n.VAR\left[\overline{Y}\right]}{MS_{within}} = \frac{MS_{among}}{MS_{within}}$$

which is the same F ratio that we use in the ANOVA.

Hence, the F in ANOVA can be viewed as follows.

F in single-factor ANOVA provides a comparison of how variable the sample means are with how variable they are expected to be, given observed variability within samples.

If the sample means are significantly more variable than expected by chance alone, we say that there is a significant difference among the means, or that the factor under consideration has a significant influence on the means.

Some Common Terms

At this point, it is worth reviewing the terms used so far and introducing some new terms.

In single-factor analysis of variance, two measurements are abstracted from each entity under consideration, whether the entities be points in a stream, rats in an experiment, paving bricks in a stream or water alloquots. One measurement is of a discrete variable or **factor** and the other is a continuous variate called the **response variable**. The factor takes on a set of discrete values called **factor classes**. The initial objective of ANOVA is to decide whether the factor has an effect on the response variable. Examples of factors, factor classes and response variables are given in Table 4–5.

Table 4–5. Examples of the sort of data required for single-factor ANOVA. The factor is the variable whose effect is under scrutiny, the factor classes are the discrete values taken on by that factor, the entities are the objects measured, both with respect to the factor and the response variable, and the response variable is the variable that is affected by the level of the factor.

Factor	Factor classes	Entities	Response variable
Waterbody	Grayson Pond, Beaver Lake, Rock River	Water samples	Strontium $(\mu g/l)$
Sex	Male, Female	Feral Pigs	Grain intake (kg)
Site	A-E	Water Samples	Macro-invertebrate counts
Laboratory	LAB01-25	Biological samples	Chlorophyll a
Species	E. aurifrons, E. albifrons, E. tricolor	Birds	Duration of territorial chase
Method	Interpolated mapping, minimum convex polygon, Dirichlet tessellation	Badgers	Home range size
Topographic Position	Hilltop, Valley, Sth Slope, Nth Slope	Soil sample	Phosphorus

The ANOVA Table

It is customary to present the results of an ANOVA in the form of an ANOVA Table (Table 4–6). Only four items in this table should be familiar at this point — MS_{among}, MS_{within}, F and the probability under H_0.

Table 4–6. Summary of the analysis of variance used to compare algal biomass (mg/m^2) collected from paving bricks laid along transects in a shallow stream riffle. Means for three transects are inflated by application of an insecticide.

Source	Degrees of freedom	Sums of squares	Mean square	F value	Prob under H_0
Among transects	6	1555.11	259.19	8.82	P<0.0001
Within transects	28	822.58	29.38		
Total	34	2377.69			

The sums of squares and degrees of freedom are components of the mean squares. If we consider the general form of a variance:

$$VAR[Y] = S^2 = \frac{\sum^{n}\left(Y-\overline{Y}\right)^2}{n-1}$$

It can be seen to comprise two components. The numerator is the sum of the squared deviations of the Y values from their mean, that is the **sum of squares**. The denominator is called the number of **degrees of freedom**, representing the number of independent deviations upon which the sums of squares is based.

There are several ways of gaining an appreciation of why it is *n*-1 degrees of freedom and not *n* degrees of freedom, a difficult concept to grasp without a full appreciation of the mathematical basis of statistics. Essentially, ten independent values comprise ten independent pieces of information, because knowledge of one value provides no information per se on any other value. Knowledge of the sample mean uses up one piece of information, because if you know the mean and nine of the sample values, then the tenth sample value is uniquely determined (you have 10 equations in ten unknowns). Similarly, once you specify nine deviations from the sample mean, the tenth is uniquely determined, since the sum of deviations about the mean must be zero. Only nine of the ten deviations are free to vary — there are nine degrees of freedom.

Now consider the equations for MS_{among} and MS_{within}

$$MS_{among} = \frac{\sum^{a}\sum^{n}\left(\overline{Y} - \overline{\overline{Y}}\right)^2}{a-1}$$

$$MS_{within} = \frac{\sum^{a}\sum^{n}\left(Y - \overline{Y}\right)^2}{a(n-1)}$$

From these we can define:

$$SS_{among} = \sum^{a}\sum^{n}\left(\overline{Y} - \overline{\overline{Y}}\right)^2$$

with (*a* - 1) degrees of freedom, and

$$SS_{within} = \sum^{a}\sum^{n}\left(Y - \overline{Y}\right)^2$$

with *a*(*n* - 1) degrees of freedom.

The associated degrees of freedom make intuitive sense because MS_{among} is a variance based on a means, hence (*a*-1) degrees of freedom, whereas MS_{within} is a variance calculated as an average of *a* sample variances each with (*n* - 1) degrees of freedom.

Partitioning the sums of squares

The statistical model developed above considers the values taken by individual measurements of the response variable to be the sum of various effects:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

where Y_{ij} is the jth value in the ith sample, α_i is the grand mean for the combined data, μ_i is a deviation of class mean from the grand mean and e_{ij} represents the deviations of each measurement from its class mean. This equation is referred to as the ANOVA model. In plain English, it states that the value taken by each measurement of the response variable is composed of the grand mean of the population (centred), a deviation of its class mean from the grand mean, and its own deviation from its class mean.

We can rearrange this equation to yield:

$$Y_{ij} - \mu = \alpha_i + e_{ij} = (\mu_i - \mu) + (Y_{ij} - \mu_i)$$

which states that the deviations of the individual measurements from the overall mean comprise the sum of their deviations from their class mean and the deviations of the class mean from the overall mean. When we represent the parameters of this equation by their sample estimates, we have algebraic identity:

$$\left(Y_{ij} - \bar{\bar{Y}}\right) = \left(\bar{Y} - \bar{\bar{Y}}\right) + \left(Y_{ij} - \bar{Y}\right)$$

It is simple to prove, though it is not done here, that the same is true of the sums of squares.

$$\sum^{a}\sum^{n}\left(Y - \bar{\bar{Y}}\right)^2 = \sum^{a}\sum^{n}\left(\bar{Y} - \bar{\bar{Y}}\right)^2 + \sum^{a}\sum^{n}\left(Y - \bar{Y}\right)^2$$

$$SS_{total} = SS_{among} + SS_{within}$$

and for the corresponding degrees of freedom

$$an - 1 = a - 1 + a(n - 1)$$

Note that the SS_{total} is calculated by pooling the data from all samples and calculating the squared deviations of the individual sample values from the overall grand mean. There are a times *n* values in all, and therefore (*an* - 1) degrees of freedom.

The above equality is what is referred to as a **partition of the sums of squares**.

When dealing with the sums of squares, it is sensible to talk of partitioning the total sums of squares into two components, one representing variation of the individual *Y* values about their own mean, and one representing variation of the sample means about the grand mean. This is shown diagrammatically in Figure 4–3.

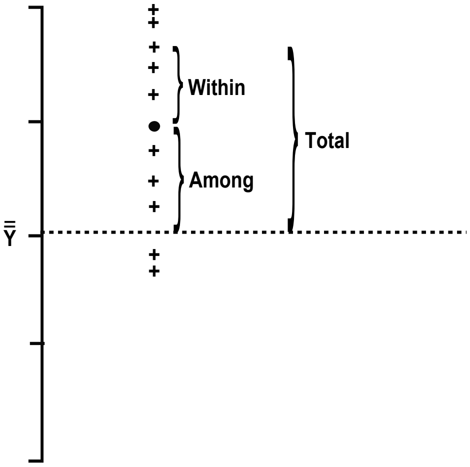


Figure 4–3. A diagrammatic representation of the partition of sums of squares in single-factor ANOVA.

Partitioning the mean squares is not possible, since

$$\frac{\sum^{a}\sum^{n}\left(Y-\bar{\bar{Y}}\right)^{2}}{an-1} \neq \frac{\sum^{a}\sum^{n}\left(\bar{Y}-\bar{\bar{Y}}\right)^{2}}{a-1} + \frac{\sum^{a}\sum^{n}\left(Y-\bar{Y}\right)^{2}}{a(n-1)}$$

$$MS_{total} \neq MS_{among} + MS_{within}$$

The variance obtained by pooling all the data cannot be decomposed neatly into MS_{among} and $\text{MS}_{\text{within}}$. It is not sensible to talk of partitioning the mean squares.

Because of their special additive properties, the sums of squares are included in the ANOVA table with the mean squares. It is also traditional to include the degrees of freedom for each (Table 4-6).

Where have we come?

In this lesson, we have covered the basics of single factor ANOVA. It is a very important lesson, because the understanding it conveys is essential for you to apply ANOVA with the flexibility you need to address the very great range of applications of the technique.

You should appreciate

- That the single factor ANOVA is an extension of the T-test, in that it enables the simultaneous comparison of two or more means.

That there is a simple relationship between the variation within samples and the expected variation among sample means, in the absence of any treatment effect, namely

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$$

- and that this relationship forms the basis of the analysis.
- That ANOVA can be understood from the point of view of two estimates of the population variance common to all samples (under the null hypothesis), one biased by any effect of the treatment and one unbiased.
- That alternatively, ANOVA can be understood from the point of view of comparing observed variation among the means with that expected, given the observed level of within sample variation.
- The terminology commonly used in ANOVA, and in particular, the **factor**, with several discrete **factor levels**, that influence the **response variable**.
- That the **Mean Square** is just another name for variance.
- That a Mean Square can be decomposed into two elements, the numerator **Sums of Squares** and the denominator **degrees of freedom**.
- That unlike the mean squares, the sums of square and the degrees of freedom are additive, so it makes sense to talk of carving up the cake (partitioning the sums of squares) into components attributable to within and among sample variation.
- That there is a formal basis for reporting the results of an ANOVA, called an **ANOVA Table**.

If little of this makes sense, you will need to review the above theory, or seek alternative explanations in a text of your choice. It is essential that you understand the above concepts before moving on.

Lesson 2: Multiple Comparison Tests

Where to from here?

In the study of algal growth on submerged pavers subjected to differential application of insecticide, the analysis of variance, summarised succinctly in the ANOVA table (Table 4–6), demonstrates a significant difference among the means (F = 16.96; df = 6,28; p < 0.0005).

What the ANOVA results do not tell us is which of the means differ from which others. This is critical information, because a range of possibilities exists. All means may differ significantly from all others, only one mean may stand out significantly from all the rest, or one of a range of possibilities in between may apply. Clearly, we must carry the analysis further.

A first guess at how to tackle this problem would be to apply a series of Student's t-tests to the set of means taken two at a time. This is inappropriate for several reasons. First, each t-test would use only a fraction of the available data, being based on only two variances at a time. The power of such an approach would be greatly reduced.

Second, applying a series of t-tests in a single connected analysis results in compounding of errors. Whenever a test is applied at the = 0.05 level of significance, there is a 5% probability that a significant result will emerge when in fact there is no real difference between means (ie a Type I error). If there are seven means to compare, then 21 comparisons will be required. The probability of at least one of these comparisons being significant when there is no real difference among the means is not 0.05 but closer to:

$$1 - (1 - 0.05)^{21} = 0.66$$

The majority of statistical practitioners find this unacceptable, and prefer procedures whereby the probability of obtaining any Type I error among the totality of pairwise comparisons is 0.05.

What we need is a test that:

- uses all the data as a basis for comparisons, and
- adjusts the critical value for significance such that there is a probability of 0.05 of obtaining any Type I error at all among the pool of related pairwise comparisons.

The probability of obtaining any Type I error at all in a related set of comparisons is called the **experimentwise error rate**, horrid jargon, but difficult to avoid. A wide range of possible tests that meet these requirements have been developed, and are classed under the broad heading of **multiple comparison tests**.

Multiple comparison tests

One of the simplest multiple comparison tests to understand is Tukey's Honestly Significant Difference for equal sample sizes, generalised to the **Tukey-Kramer Method** for unequal sample sizes. It begins by taking the formula for *t* of the student's t-test

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{1}{n}\left(S_1^2 + S_2^2\right)}}$$

and replacing S_1^2 and S_2^2 with MS$_{\text{within}}$. After all, MS$_{\text{within}}$ is the best available estimate of the population variance common to all samples.

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{2}{n} MS_{within}}}$$

Instead of having 2(*n* - 1) degrees of freedom, this t test based on MS_{within} has *a*(*n* - 1) degrees of freedom, and in that sense is more powerful (since the critical value of *t* declines toward 1.96 as the number of degrees of freedom increases).

The next step is to replace the standard critical values of *t* with more appropriate values, to compensate for the compounding of errors that occurs when multiple related comparisons are performed. The Tukey-Kramer procedure assumes that the investigator intends to perform exhaustive pairwise comparisons between all means used in the ANOVA. Tables with appropriate adjustment of the critical values are available in the form of the Studentised Range, written $Q_{\alpha[a,a(n-1)]}$ (see Table 18 of Sokal and Rohlf, 1994). These tables give the critical values for the difference between two of a means. The tables are based on a rearrangement of the above formula, noting that *n* and MS_{within} are constant for all comparisons. The test will be significant if:

$$t = \frac{\left|\overline{Y}_1 - \overline{Y}_2\right|}{\sqrt{\frac{2}{n} MS_{within}}} > \left(tabulated\_critical\_value\right)$$

that is, if

$$\left|\overline{Y}_1-\overline{Y}_2\right| > (critical\_value)\sqrt{\frac{2}{n}MS_{within}}$$

$$> Q_{\alpha[a,a(n-1)]}\sqrt{\frac{MS_{within}}{n}}$$

$$> MSR$$

MSR is the minimum significant range between two means. Thus, a table of differences between means is constructed and the differences are compared with the *MSR*. The *MSR* is calculated from MS_{within}, the sample size *n*, and the tabulated value of the studentised range $Q_{\alpha[a,a(n-1)]}$.

The adjustment of the critical value for significance in multiple comparison tests has important implications. In order to achieve an experimentwise error rate of 0.05, the level of significance of each individual comparison is necessarily much lower than 0.05. If there are many means in the ANOVA, the actual error for individual comparisons between means may be a tenth, a hundredth or even a thousandth of the experimentwise error. Multiple comparison procedures are typically conservative, compared with the conventional pairwise t-tests.

There is a plethora of multiple comparison procedures available, differing from the Tukey-Kramer Method, and from each other, in:

- the level of adjustment necessary to compensate for the number of comparisons possible under the experimental design, and
- the approach taken to make those adjustments.

Some options are described below. For a more detailed treatment of the subject of multiple comparison tests, refer to Keppel (1973) and Day and Quinn (1989).

Unplanned comparisons

Unplanned comparisons are a subset of possible comparisons chosen on the basis of information obtained from examination of the results of the ANOVA. In many cases, unplanned comparisons are exhaustive, comprising comparisons of all possible combinations of the available means. In this sense, unplanned comparisons are exploratory, looking for the most plausible explanation for the overall significant result obtained in the ANOVA. Of the techniques available for unplanned multiple comparisons, the Tukey-Kramer Method is recommended.

Tukey-Kramer Method

The Tukey-Kramer Method is suitable for exhaustive comparisons among means where it is assumed that the pool of possible comparisons includes only those between the means taken a pair at a time. This technique has fared extremely well in Monte Carlo simulations (Dunnett 1980a,b; Day and Quinn, 1989), and has been shown to control the experiment-wide error rate to the desired level (Hayter, 1984). The rationale for the test has been described above.

Bonferroni and Sidak Methods

The Bonferroni method depends upon the observation that if we set the significance level for each of *k* comparisons at

$$\frac{\alpha}{k} = \frac{0.05}{k}$$

then the probability of obtaining one or more Type I errors across all *k* comparisons (the experimentwise error rate) will be less than 0.05. Hence, by setting the level of significance of tests between a means taken a pair at a time at

$$\frac{\alpha}{k} = \frac{0.05}{a(a-1)/2}$$

we can be sure that the experimentwise error rate is less than 0.05.

The Sidak method depends upon the observation that if we set the significance level for each of k comparisons at

$$1-(1-\alpha)^{1/k} = 1-(1-0.05)^{1/k}$$

then the probability of obtaining one or more Type I errors across all *k* comparisons (the experimentwise error) will be less than 0.05. Hence, by setting the level of significance of tests between a means taken a pair at a time at

$$1-(1-\alpha)^{2/a(a-1)} = 1-(1-0.05)^{2/a(a-1)}$$

we can be sure that the experimentwise error rate is less than 0.05.

Of the two procedures, Sidak's is less conservative than Bonferroni's and is therefore preferred.

The advantage of these procedures is that they can provide simultaneous tests of more than one related hypothesis in a wide range of contexts. All that is required is knowledge of the number of comparisons in the total pool of comparisons possible (ie *k*) and the level of significance required for the experimentwise error rate (i.e. $\alpha = 0.05$). The pool of possible comparisons can vary; depending upon the experimental design, from exhaustive comparisons of all means against all other means taken singly or in combination, to exhaustive comparisons between pairs of means, to a restricted set of comparisons (see planned comparisons below). The disadvantage of the two techniques is that because they rely on inequalities, which merely set upper bounds to the experimentwise error rate, they are typically more conservative than other available procedures such as the Tukey-Kramer procedure.

Planned comparisons

Planned comparisons are a subset of possible comparisons, chosen before the experiment is done so that they are not suggested by the results. Planned comparisons are typically decided as part of the overall experimental design. The pool of possible comparisons is typically severely restricted by the experimental design.

Comparing all treatments to a control

A special case of multiple comparisons is where the only comparisons contemplated at the time of designing the experiment are between a single control and a set of treatments. In this case, improved power can be legitimately achieved by adjusting the level of significance on individual tests only for this restricted class of comparisons. Dunnett (1956) proposed a test to cater for this situation, and it is now widely used. **Dunnett's test** holds the experimentwise error to a level not exceeding the stated level of significance, usually 0.05.

The alternative of using Bonferroni's or Sidak's method is considered unnecessarily conservative in comparison with Dunnett's test, but may be the best option when faced with more complex designs involving several control and treatment samples.

A priori unrelated comparisons

There are rare instances where relatively few, completely independent comparisons are contemplated using the results of an ANOVA based on a more extensive data set than required to address the hypotheses of interest. Providing the investigator is willing to accept an error rate of 0.05 on each comparison, then it may be legitimate to apply t-tests independently to each hypothesis. It is still sensible to use the best available estimate of the common population variance, MS_{within}, in place of the individual sample variances, and a suitable test in this case is the Least Significant Difference Test. The occasions for its use are rare indeed, and it should definitely not be used for unplanned comparisons or for comparisons with a control.

Deciding a personal policy

Appropriate choice of multiple comparison test remains a controversial subject and you may wish to address the issues yourself. You need to:

- Consider whether you accept at all, the need to adjust the level of significance on each test to compensate for the number of comparisons in an experiment.

Not all researchers accept this need, though nowadays, they would have difficulty publishing.

- Decide the total pool of comparisons contemplated as part of the study design, in advance of collecting the data.

For five means, there is a pool of 5(5-1)/2 = 10 exhaustive pairwise comparisons possible, but only four comparisons are of interest if the object is to compare four treatment means against a control. This is the most important criterion in choosing a multiple comparison procedure.

- Choose a multiple comparison procedure that adjusts the per-comparison error rate appropriately for the pool of comparisons identified above. It should be the most powerful of appropriate alternatives available.
- Decide the level of significance required for the experimentwise error rate.

There is general agreement on the acceptable per-comparison error rate, usually 0.05, but certainly in the range 0.10 to 0.01—not so large as to allow a large number of false conclusions regarding the presence of differences between means, but not so small as to greatly reduce our chances of detecting differences when they are present. There is no general agreement on an acceptable experimentwise error rate for decision making—is it 0.05, 0.10 or 0.20 (Keppel 1973)? For the purposes of this Module, I have chosen 0.05, but this is subject to debate.

There are a few misconceptions associated with multiple comparison procedures. A common error is to argue for a reduced pool of comparisons after the data have been examined. Following an analysis of variance, you may choose only to compare the largest and smallest of the means. You have chosen to do a single comparison only, but this does not justify ignoring the need to adjust the per-comparison error rate for the overall pool of such comparisons from which you selected one.

A second error is commonly made when the response to the confusing array of options available is to choose the option that gives the most satisfactory results. The folly of this approach should be obvious. You may be confronted with a confusing and uninterpretable array of overlapping non-significant subsets of sample means, using, say, the Tukey-Kramer procedure. You need to appreciate that failure to reject the null hypothesis that two population means are equal, should not lead you to conclude that they are in fact equal. It implies only that the difference between population means, if any, is not large enough to be detected with the given sample sizes. Your approach should be to gather more data, or to redesign and repeat the experiment giving more thought to planned comparisons. It even may be appropriate to increase the experimentwise error rate to 0.10 or 0.20, provided this is openly stated. It is not appropriate to cast your eye around for a more powerful but inapplicable multiple comparison test such as the LSD test.

Confusion may arise from the non-transitive nature of non-significance. The mean of sample A may be significantly different from the mean of sample B, but neither may be significantly different from a mean lying in between. When the sample sizes are unequal, the results may be even more counter-intuitive. If we have four cells with means ranked A>B>C>D, the difference between B and C, each based on 1,000 values, may be significant while the difference between A and D, each based on 5 values, might not.

Where have we come?

The full analysis is now available to you. As a researcher, when you are faced with the problem of comparing the means of several samples, classified according to a single factor, the approach is as follows:

- perform a single-factor ANOVA to determine whether the means are more variable than can be considered by chance alone;
- if the ANOVA yields a significant result, perform an appropriate multiple comparison test to determine where the differences lie.

Views on the application of multiple comparison tests are varied, and the subject remains controversial. You need to come to a reasoned position yourself. The recommendations I make are :

- If unplanned comparisons are required as part of the experimental design, apply the Tukey-Kramer procedure.
- If planned comparisons involving a single control are required, apply Dunnett's test.
- If the design is more complex, but where the pool of comparisons possible under the design is restricted, then the Sidak procedure is recommended.

- In rare cases, where a few unrelated comparisons are decided independent of the data at hand, and independent of each other, the LSD procedure may be applied.

What you do not yet know is that there are different models in ANOVA, and that the choice of model affects the direction your analysis takes. That is the subject of the next module.

Lesson 3: Models in ANOVA

Fixed and Random Models

One more class of analysis is yet to be introduced. Two models are recognised in single-factor ANOVA—fixed models and random models. The two differ little in the computation leading to the final F statistic. Where they differ is in the direction taken during follow-up analysis.

Fixed model ANOVA

In the fixed model ANOVA, the criterion upon which the factor levels are chosen are fixed and repeatable. For example, each sample corresponding to a factor level may be drawn from individuals of specified ages, or subject to different specified treatments, or belong to specified genetic strains. A case might involve a pastoralist who wishes to compare the ability of breeds of cattle to produce butterfat in the milk. Here, the factor is breed and the response variable is butterfat concentration. If the objective of the study is to select the best breed for butter production, then the approach would be to take five specific breeds (the factor levels) known to be candidates for the new herd, probably including the breed that makes up the existing herd.

This is a fixed design because:

- the breeds chosen for investigation are fixed in the sense that if the experiment were to be repeated, the same breeds would be chosen again,
- the investigation is designed to detect significant differences among specific breeds, in this case, to find the breed with the highest butterfat content in the milk,
- it would be sensible to follow up the analysis with multiple comparison tests to determine where the differences among the means, if any, lie.

Random model ANOVA

In the random model ANOVA, the criteria upon which the factor levels are chosen are random—that is, the factor levels are chosen at random from a substantial pool of possible choices. NSW Agriculture may wish to address a similar problem to that of the pastoralist, but without a specific interest in any one breed. Instead, they may be interested in the more general problem of whether the factor breed has any effect on butterfat production. From their perspective, there may be 50 breeds available for experimentation, but financial and logistic constraints preclude using all of them. They choose five breeds at random.

The questions that would be asked of such a design are quite different from those asked of a fixed design. NSW Agriculture are not at all interested in whether there are significant differences between specific breeds, because they were selected at random. All they are interested in is whether or not there are differences among breeds overall, that contribute to variation in butterfat production. In statistical jargon, they are interested in whether there is an added variance component due to variations among breeds in butterfat production. If there is, they might proceed to measure the relative strength of the added variance component.

In the random design:

- the factor levels (breeds) are randomly selected from a large pool of possible choices. If the experiment were to be repeated, there is no reason to expect that the same factor levels (breeds) would be chosen second time around;
- the investigation is not designed to detect significant differences among specific breeds, but rather to detect an overall effect of the factor—breed—on the response variable, butterfat concentration;
- it is not sensible to follow up the analysis with multiple comparison tests to determine where the differences among the means, if any, lie. Rather, the added variance component due to the overall effect of the factor — breed — is estimated.

Estimating the added variance component

Recall from the discussion of rationale of ANOVA that the mean squares are estimates of the common population variance, but that one, MS_{among}, is biased by the effect of the factor.

$$MS_{among} \approx \sigma^2 + n.\sigma_A^2$$

$$MS_{within} \approx \sigma^2$$

To estimate the added variance component, we need an estimate of σ_A^2. Such an estimate is provided by:

$$S_A^2 = \frac{MS_{among} - MS_{within}}{n} \approx \sigma_A^2$$

There are two intuitive interpretations of S_A^2. First, it provides an estimate of how much of the variability among means cannot be explained by observed variability within samples. It is a measure of the **strength of the real effect** of the factor, which can be expressed as a percentage:

$$\frac{S_A^2}{\frac{MS_{within}}{n}+S_A^2}x100\%$$

Second, it provides a measure of how reproducible measurements are across samples relative to the repeatability of measurements within samples. Later in this Module, an example is presented where the National Association of Testing Agencies (NATA) was interested to know to what degree chlorophyl-*a* determinations were repeatable within labs and to what degree results were reproducible across labs. A random sample of 25 laboratories Australia-wide were asked to extract and determine the concentration of chlorophyl-*a* in three samples labelled A, B and C. The laboratories involved were not told that the three samples were simply replicates of the same batch.

We need to ask, how much more variable would be a set of 25 determinations, one from each laboratory, than a set of 25 determinations from a single laboratory?

Variation in single determinations across laboratories is given by:

$$VAR\left[\overline{Y}\right] \approx \sigma_{\overline{Y}}^2 + \sigma_A{}^2 = \frac{\sigma^2}{n} + \sigma_A^2 = \sigma^2 + \sigma_A^2$$

since, for single determinations, n = 1. Variation in single determinations within a laboratory is given by:

$$VAR\left[Y\right] \approx \sigma^2$$

Hence, of the total variation among single determinations taken one from each laboratory, an amount of:

$$\frac{\sigma_A^2}{\sigma^2+\sigma_A^2}.100\%$$

is *additional* to what would be expected of values taken from a single laboratory, namely:

$$\frac{\sigma^2}{\sigma^2+\sigma_A^2}.100\%$$

Thus

$$\left(\frac{S_A^2}{MS_{within} + S_A^2}\right).100\%$$

is an estimate of the variation in single determinations from laboratories *over and above* what would be expected if the laboratories had common methods and equipment and could perfectly reproduce each others' results.

It is the percentage contribution to variation in single determinations among laboratories that can be attributed to differences in their equipment and procedures.

On this basis, we can define an index of **reproducibility**, the ability to reproduce results across laboratories, as:

$$Reproducibility = \left(1 - \frac{S_A^2}{MS_{within} + S_A^2}\right).100\%$$

Reproducibility ranges from zero to 100%, the latter being desirable in interlaboratory comparisons.

An estimate of the expected variation in single determinations across laboratories derived from variation in determinations within laboratories is:

$$\left(\frac{MS_{within}}{MS_{within} + S_A^2}\right).100\%$$

It is the percentage contribution to variation in single determinations among laboratories that can be attributed to the inability of laboratories to repeat their own determinations. It is the variation among laboratories that would exist even if all laboratories had identical equipment and procedures.

Thus, **repeatability** can be defined as the ability to repeat results in a single laboratory, relative to the ability to reproduce results across laboratories.

$$Repeatability = \left(1 - \frac{MS_{within}}{MS_{within} + S_A^2}\right).100\%$$

If there is no added variance component due to differences among labs ($S_A^2 \approx \sigma_A^2 = 0$), then the results will be perfectly reproducible. If you had 10 replicate bottles of water to evaluate, it would not matter whether you sent them to 10 different laboratories or all 10 bottles to the one laboratory. This is an unachievable ideal for NATA, though progress toward the ideal can be achieved by setting standards for equipment and procedures to be adhered to by member laboratories.

If the results from laboratories were perfectly repeatable $\left(MS_{within} \approx \sigma^2 = 0\right)$, then it would not make sense to send more than one replicate bottle to each laboratory. Rather, you would spread the bottles across laboratories to obtain the best average determination.

Of course, there are many possibilities in between.

Where have we come?

In this lesson, the two different models of ANOVA were introduced. These are important in single factor ANOVA because they determine the direction the analysis takes following a significant result in the ANOVA. They are even more important in more complex ANOVA designs because they influence the calculations in the ANOVA itself.

You should now appreciate

- The distinction between fixed and random models in ANOVA.
- The procedure used to follow up a significant result in a random model ANOVA.
- The distinction between repeatability and reproducibility, and how to construct meaningful measures of them from the ANOVA table.

We now move on to an important topic, that of power and how to interpret a non-significant result in ANOVA.

Lesson 4: Power analysis

Planning of experiments

The effectiveness of ANOVA is greatly influenced by sample size. Sample sizes need to be adequate to be reasonably certain of detecting an important effect when it exists. At the same time, sample sizes should not be so large that the cost of the study becomes excessive, nor is there much value in having weak, unimportant effects becoming highly significant. Planning the intensity of sampling using a **prospective power analysis** can be important in the design of experimental and observational studies that use ANOVA as the means of analysis.

As with the t-test, optimal sample size for an ANOVA will be affected by:

- **The size of the smallest effect or difference that it is important to detect**. The smaller the effect, the larger will be the sample size required to detect it, all other things being constant.
- **The variability of the data.** The more variable the data within samples, the more difficult it will be to demonstrate a given effect or difference among sample means against the backdrop of that variability.
- **The acceptable probability of detection**. The more certain you want to be of detecting a difference of a given size, the larger will be the samples required to give you that greater certainty.
- **The level of significance of the test.** It will take larger samples to be reasonably sure of detecting a given difference at the 1% level of significance than at the 5% level of significance.

There are obvious parallels with the procedure for determining appropriate sample sizes for the t-test. ANOVA has the added complication of needing a decision on exactly what effect it is that is of interest to us. A number of possibilities exist. The smallest effect of interest might be defined in terms of:

- the **smallest average effect** across samples that is regarded to be of importance;
- the **smallest single difference** between one sample mean from the overall average that is regarded to be of importance;
- a **minimal scenario** or set of minimal scenarios involving several sample means that is regarded to be of importance.

Estimating sample size under all of these scenarios is complex and not well covered by statistical packages. Only one approach is covered in this Module, for the fixed-factor ANOVA. The approach recognises that on achieving significance in the ANOVA, you will follow with multiple comparisons among the sample means. The objective of the power analysis is to set a sample size to be 80% sure, say, of detecting a true difference of given minimal magnitude δ at the level of significance after Bonferroni correction α'.

To be 80% sure ($P = 1 - \beta = 0.8$) of detecting a given difference between any two of ten means (δ) at the 5% level of significance ($\alpha = 0.05$), you will require a sample size of:

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 \left(t_{\alpha'[\nu]} + t_{2(1-P)[\nu]}\right)^2$$

where n is the required size of each sample, σ is the true common parametric standard deviation, ν is the degrees of freedom for σ, $t_{\alpha'[\nu]}$ is the value from a two-tailed t-table with ν degrees of freedom and level of significance α' and $t_{2(1-\beta)[\nu]}$ is the value from a two-tailed t-table with ν degrees of freedom and level of significance $2(1-\beta)$. In this case, where we have 10 means, and therefore 45 potential pairwise comparisons,

$$\alpha' = \frac{\alpha}{45} = \frac{0.05}{45} = 0.001$$

The inequality above must be solved iteratively, as n is on both sides of the inequality, ν being a function of n.

To undertake such a prospective power analysis, you need also to make some hard decisions. First, you need to decide on what is the smallest difference between any two of your samples (δ) upon which you will place some importance. You must decide that differences smaller than that value are of little or no consequence.

Second, you need to estimate the common parametric standard deviation, σ, and this must often be estimated before you collect the data. You can use a ball-park figure based on experience, or you can undertake a pilot study to estimate σ using the root mean square error

$$\sqrt{MS_{within}} \approx \sigma$$

and then the desired sample size before expending resources on the major optimised study. You may find the ratio of δ to σ easier to estimate.

Third, you need to decide on the risk you are willing to take ($P = 1 - \beta$) in not finding an important difference when it actually exists. There is no general agreement on the value of P. The value of 0.80 seems to have currency in the same way as 0.05 has currency for α. Some would argue for higher values of 0.90 and 0.95, but ultimately it comes down to how important to you it is to detect a true difference of δ if it exists. What risk are you willing to take of missing it?

Finally, you need a clear idea of the experimental design and in particular, what comparisons will be made in following up a significant result. This is very important, as it will determine the level of Bonferonni correction that will be applied, and this in turn will affect the optimal sample sizes and expenditure on the project.

Regardless of the problems of its computation, the cost savings of this approach can be considerable, either through minimising the risk of undertaking the study only to find that no difference can be demonstrated (when it exists) or by avoiding the expense of collecting more data than is required for success.

Interpretation of non-significant results

Retrospective power analysis is used to decide if it is reasonable to accept the null hypothesis, that is, for dealing with the ambiguity of a non-significant result. There may well be no difference between samples, or the sample sizes may not be large enough to detect a difference that is there (you make a Type II error). The risk of such a false negative result cannot usually be quantified, unless the alternative hypothesis *H1* is known. A non-significant result, therefore, is difficult to interpret.

Strictly speaking, you would interpret a non-significant result equivocally, as having failed to demonstrate a difference. If however you do wish to draw a firm inference from a non-significant result, then a **retrospective power analysis** is mandatory.

In a retrospective power analysis, you ask, given your sample sizes, what might be the smallest difference (δ) you could be reasonably confident of detecting ($P = 1 - \beta = 0.80$). Using the MS_{within} as an estimate of σ^2, the formula for the smallest difference likely to be detected by an ANOVA, followed with Bonferoni comparisons, with sample sizes each of *n*, is:

$$\hat{\delta} \geq \left(t_{\alpha'[\nu]} + t_{2(1-P)[\nu]}\right)\sqrt{\frac{2MS_{within}}{n}}$$

where n is the size of each sample, MS_{within} is our best estimate of the true common parametric variance, ν is the degrees of freedom for MS_{within}, P is the intended power of the test, $t_{\alpha'[\nu]}$ is the value from a two-tailed t-table with ν degrees of freedom and level of significance α' and $t_{2(1-\beta)[\nu]}$ is the value from a two-tailed t-table with ν degrees of freedom and level of significance $2(1-\beta)$. In this case, where we have 10 means, and therefore 45 potential pairwise comparisons, again

$$\alpha' = \frac{\alpha}{45} = \frac{0.05}{45} = 0.001$$

If $\hat{\delta}$ is so small as to be of no consequence, then your interpretation of the negative result is acceptable. If, on the other hand, even a large difference would often go undetected with your sample sizes, you have nothing to report.

Retrospective power analysis is a controversial area, and the analyses for ANOVA have not adequately been incorporated into statistical packages. Many power analysis algorithms give you the optimal sample size to be reasonably sure of detecting the difference you observed in your analysis, which is not useful. Others differ on how to define the minimum effect size, with the decision having a strong effect on the outcome of the power analysis. The advice above is only one option available to you, and is over-engineered in the sense that the Bonferoni correction is one of the most conservative of multiple comparison procedures.

As with power analysis and the t-test, choice of the value for the probability of detecting the difference if it exists is controversial. It has been argued that just as a small α (Type I error) is required to declare a difference to be nonzero, so too a small β (Type II error) should be required to declare a difference to be zero. We have chosen $P = 1 - \beta = 0.80$ above, which is developing similar currency as $\alpha = 0.05$ as the defacto standard for the Type I error. Cogent arguments can be made for P = 0.84, 0.90 and 0.95.

Where have we come?

In this lesson, you learned about the concept of power, how to use it to plan experiments and how to use it to place interpretation on a non-significant result.

In particular, you should appreciate:

- The considerations necessary to determine the level of sampling intensity required to be reasonably sure of detecting a difference of a given magnitude when it exists.

- That non-significant results are ambiguous. It could be that there is no difference of any importance there to find, or it could be that there is an important difference to find, but that your sample sizes are too small to demonstrate it. Power analysis allows you to make a judgement and so report a non-significant result with some level of confidence.

We have now covered the theory of single factor ANOVA. There are a number of matters to consider in applying this theory to real world problems, not least of which are the assumptions of the technique. The next lessons deal with nuances in the application of ANOVA.

Lesson 5: Application

Assumptions of ANOVA

Up to this point, ANOVA has been presented without much attention paid to the assumptions of the technique. This is the approach adopted by Sokal and Rohlf (1994), in the belief that non-mathematical audiences learn better if they come to understand the structure and purpose of the analysis, without being distracted from the central theme by whether or not the data are strictly amenable to such an analysis. However, it is essential for the practitioner to verify that the assumptions are reasonable. If they are not, steps should be taken to ensure that the assumptions are met.

In this section, I describe the assumptions of ANOVA, how to check if they are reasonable, and how to proceed in the face of perceived violations. Alternatives to parametric ANOVA are briefly described, but their practical application is beyond the scope of this Module.

Single-factor ANOVA has four assumptions, namely, randomness and independence in sampling, equality of variances across samples, and normality. Each of these will be dealt with in turn.

Randomness in sampling

Randomness in the selection of entities for measurement

In inferential statistics, we study samples intensively in order to infer attributes of the populations from which those samples are drawn. It is critical to any experiment to ensure that the measurements taken are representative of the system under study, if our inferences are to have any validity. The entities selected for measurement must be representative (unbiased) of the population from which the entities are drawn.

Take for example a study of stream invertebrates at each of several sites. We may choose to take 10 replicate collections of invertebrates from each site, in order to characterise invertebrate abundance at the sites and as a basis for comparisons among sites. The mean invertebrate abundance must be representative of the true parametric mean for the each site, and their variances must be representative of the true parametric variances for each site. If they are not, we may as well close up shop. Randomness in the selection of the entities for measurement is by far the safest way of achieving this unbiased representation.

Non-randomness may manifest itself as lack of independence of the entities, or in unequal variances or in non-normality. Violation of the assumption of randomness in sampling cannot be overcome easily, and typically the data must be discarded, the sampling protocols redesigned and the data recollected. Adequate attention must be paid at the time of designing an experiment, or when sampling from natural populations, to ensure adequate randomness in the selection of entities.

Randomness in the allocation of entities across factor classes

Where the allocation of entities to factor classes is within the control of the investigator, ANOVA expects that the items, individuals or entities are allocated to each of the factor classes at random. In a field trial, plots should be allocated at random to the treatments they are to receive. If we fail to do so, we run the risk of introducing a systematic bias that will confound our interpretation of a significant result. For example, if we consciously or unconsciously allocate the better looking plots first and so preferentially to a control factor class, and the poorer looking plots last and so preferentially to our manipulated factor classes, what are we to make of a significant result? We will not know if the significance is a result of our manipulations in comparison to the control, or a result of differences between better-looking plots and poorer-looking plots. Our experiment will have been **confounded**.

Often it is not possible to randomly allocate entities across factor classes. Having selected sites in a river for investigation, it is not possible to then randomly allocate invertebrate collections to them – the invertebrate collections are constrained to be those that are collected from the individual sites. But here the entities (invertebrate collections) are integrally connected with the characteristics of the site — confounding of the sort outlined above is not an issue. Provided we ensure that the entities within sites are selected at random, the ANOVA can proceed.

Independence

Independence requires that knowledge of the value of one measurement provides no information on the value of one of the other measurements, relative to its expected value. **Note that we demand that the residual values under the working model are independent, not the measurements themselves.** Human heights are not absolutely independent, for none of us are the size of gnats. However, randomly selected humans may have heights that are independent, in the sense that knowledge of the above average height of one individual provides no information on the height of the next, relative to the average height. Heights of twins are not independent, because if one twin is known to be of above average height, then the other is odds on to be of above average height.

Because independence is an attribute of the residuals (or if our focus is on the parent populations, the errors ε_i) independence is a relative concept. Dependence considered in one context (say ANOVA) may become independence in another (say regression).

Independence within factor classes

Measurements within factor classes must be independent. In other words, if you were to arrange the measurements in one factor class in some logical order independent of their magnitude (say in the order of collection), then their deviations from the average value, the residuals, should follow a random sequence. A run of large values followed by a run of small values would be cause for suspicion (see the runs test, Sokal and Rohlf, 1994). Adjacent plots on the ground in an experiment using plots spread out across a field, adjacent trees in a forest, identical twins among unrelated individuals, eggs from a single clutch in an experiment where eggs are taken from many clutches, successive hourly measurements of algal abundance over a four day period, are unlikely to be independent in their response to any experimental treatment applied to them.

The consequences of dependence within factor classes can be devastating. If replicates are more alike than randomly selected entities in the population (that is, pseudoreplicated), as a consequence of their interdependence, then MS_{within} will be deflated. Significant results will emerge from the analysis without basis. If replicates are less alike than randomly selected entities in the parent population, as a consequence of their interdependence, then MS_{within} will be inflated. Power of the test will be compromised.

Independence across factor classes

Measurements within factor classes must also be independent. If the value of a measurement in one factor class is more like the value in another factor class, by virtue of interdependence between the two measurements, then the effect will be to draw their respective means together, with consequential loss in power to detect a true difference between the factor classes. If the reverse is true, and the measurements have a negative dependence, then the effect will be to draw their respective means apart artificially, and a significant result may emerge without basis. Either way, the validity of the ANOVA is compromised.

A special case of repeated measures

A common scenario where measurements are dependent across factor classes is in repeated measures designs. The following data comprise the number of eggs in a clutch (clutch size) for successive clutches for a species of marine turtle. A numbered titanium tag attached to the front flipper established the female's identity. The measurements (clutch size) are repeated for each female as she returns to the nesting beach repeatedly in a single season.

While an inexperienced analyst might be tempted to undertake a single-factor ANOVA on these data, treating the females as replicates, this is not valid. Knowledge that the size of the first clutch of female X16017 is above average for first clutches provides information on the likelihood that her second, third and fourth clutches are likely to be above average. This invalidates the analysis, its power potentially dramatically reduced. The correct approach is to undertake a two-factor ANOVA (possible even in the absence of replication) with Clutch Number as one factor and Female as the second factor (see Module 5).

FEMALE	**CLUTCH NUMBER**			
Tag	**1st**	**2nd**	**3rd**	**4th**
X16005	136	.	133	129
X16013	111	105	107	113
X16017	192	191	164	188
X16024	111	121	111	122
X16029	138	.	123	135

Independence in random models

In random model ANOVA, the factor levels are selected at random from a large or infinite pool of possibilities. The object of such an analysis is to estimate the added variance component resulting from the influence of the factor. If there is dependence among the factor levels because of deficiencies in our experimental design (say pseudoreplication of factor classes), our estimate of the added variance component will be compromised.

Clearly the impact of dependence among factor classes in the random model, or among entities across or within factor classes in the fixed model, is profound.

Essentially, as with randomness in sampling, violations of the assumption of independence of errors cannot be overcome easily, and typically the data must be discarded, the sampling protocols redesigned and the data recollected. Adequate attention must be paid at the time of designing an experiment, or when sampling from natural populations, to ensure independence. Independence is achieved through appropriate experimental design.

Equality of variances

In developing the rationale of ANOVA, it was argued that the effect of the factor across samples should act differentially to increase or decrease the sample means, but not to differentially alter the sample variances. An assumption of the analysis is that the individual sample variances estimate a common population variance, that is, that the population variances are equal.

There are two approaches to deciding whether the assumption of equality of variances is tenable and taking remedial action if necessary. The first approach is quantitative, relying upon formal tests of equality of variances followed by transformation of the data, if necessary, and retesting. A test for the equality of two variances was introduced in Module 3, based on the two-tailed F-test. A variety of tests for equality of variances also are available where the number of means exceeds two, including Bartlett's, the F_{max} and the Scheffe-Box tests discussed by Sokal and Rohlf (1994). The difficulty with these tests, to varying degrees, is that they have assumptions of their own, and may well be more sensitive to violations of those assumptions than is the ANOVA itself. For example, Bartlett's test is particularly sensitive to departures from the assumption of normality, and a significant result may indicate non-normality rather than unequal variances.

The second approach, and the one adopted in this series, is more qualitative. It is based on visually examining the scatter of sample values about their predicted values, the factor class means. Ideally, this scatter should be random across the classes. There should be no systematic trend or difference in the scatter of values about their respective means.

The visual examination is achieved by constructing a plot of residuals, that is, by plotting the observed deviation of each sample value from its factor class mean against the predicted value for that class, namely the class mean. Often, the residuals are scaled by dividing by their standard errors, and referred to as studentised residuals. The method of plotting residuals is demonstrated in the worked examples that follow later in this Module.

If substantial heterogeneity of variances is revealed in the residual plot, then a transformation may be applied to bring the variance of the residuals closer to equality, or the original data may be scrutinised for a single suspect outlier in case an error has been made. Common transformations are described in a section below.

Normality

A single-factor ANOVA assumes that the individual measurements in each sample are drawn from a normally distributed population.

This assumption cannot be verified simply by pooling the data and applying a visual or statistical test for departures from normality, because the data for each sample may well be centered on quite different means. For example, two samples drawn from populations with perfectly normal distributions but with quite different means will, when the data are pooled, yield a bi-modal distribution with serious departures from normality. What we must do first is centre the data on the factor class means (that is, consider the residuals), examine the distribution of the residuals, and apply tests of normality.

This approach is conditional on the assumption of equality of variances. By a similar argument to the above, two samples drawn from populations with perfectly normal distributions but with quite different variances will, when the data are pooled, yield a distribution with serious departures from normality. It will be leptokurtic.

Pooling the residuals for a test of normality is only valid provided there is no evidence of a departure from the assumption of equality of variances.

Analysis of residuals

As outlined above, residual analysis is the recommended approach for assessing the validity of the assumptions of homogeneity of variances and normality. It is a graphical approach, and less sensitive than the range of statistical tests for homogeneity of variances and normality, but this is its strength. The ANOVA is a robust procedure, especially if the design is balanced, and violations of its assumptions of a magnitude likely to affect the outcome of the ANOVA will be evident in the graphical residual analysis. Less serious violations, likely to be detected by hypothesis testing (Bartlett's Test, Shapiro-Wilks Test), are unlikely to impact on the outcome of the ANOVA.

Analysis of residuals involves the following steps (Figure 4-4):

- Apply the working model (in our case, a single-factor ANOVA), and save the raw or studentised residuals;
- Plot the residuals against the predicted values of the working model (in our case, against the means for each sample);
- Transform if necessary to address any heterogeneity in the sample variances;

- Re-run the analysis, re-examine the residuals and apply another transformation if necessary;
- Once homogeneity of variances is achieved, construct a histogram of the residuals for an assessment of the assumption of normality.

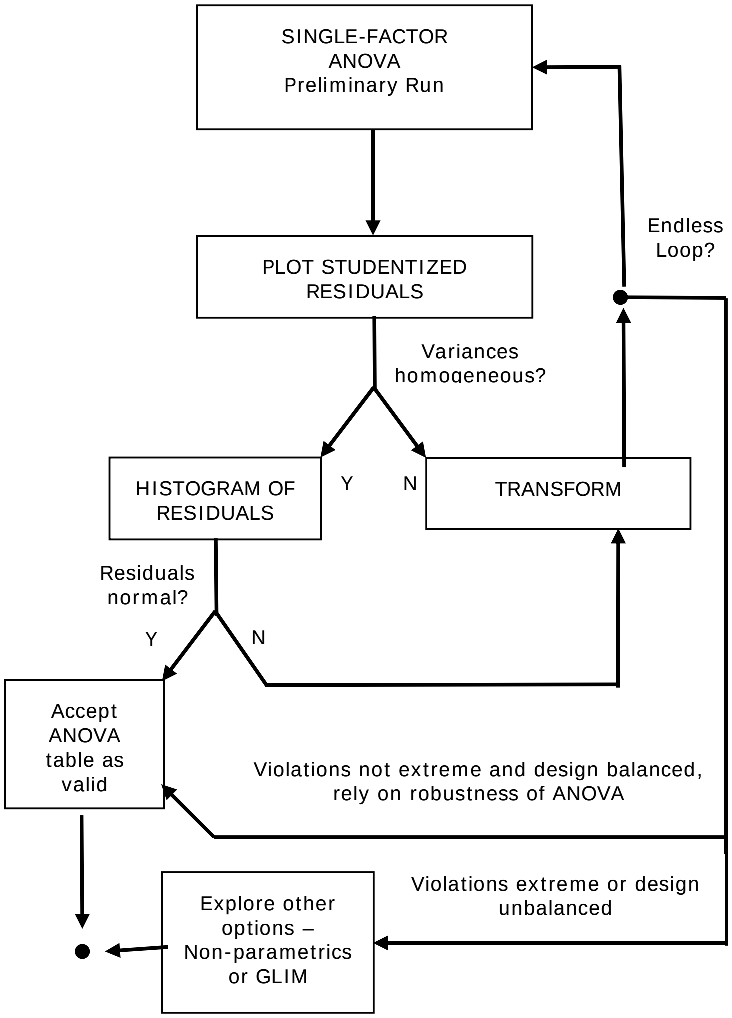
When limited data are at hand, such as is available for two-sample comparisons (Module 3), it is seldom possible to undertake a satisfactory examination of the validity of the assumptions. There are simply too few data to support a residual analysis. For this reason, you are advised to draw upon experience with the type of data at hand or on more extensive studies reported in the literature to decide an appropriate transformation. It is traditional to transform counts, for example, with either a square root (counts of independent entities) or log transformation (counts of aggregated entities).

With ANOVA, the sum total of data available is typically greater than for two-sample comparisons. It is reasonable therefore to conduct an ANOVA, examine the residuals, apply a transformation if deemed necessary, repeat the ANOVA and re-examine the residuals, and apply alternative transformations if the results of the first transformation are unsatisfactory.

Remember, though, the objective of transformation is to render normal, or of equal variance, the populations from which the samples were drawn, not the samples themselves. When it comes to the samples themselves, some deviation from normality and inequality of variances is acceptable. When small samples are involved, quite large apparent deviations from normality or homogeneity of variances can occur even when these assumptions are met for the parent populations.

Finally, it is important to appreciate that residual analysis allows an assessment of the validity of homogeneity of variances and normality. It does not necessarily address the assumptions of randomness or independence, or the adequacy of the ANOVA model to the data at hand. For example, pseudoreplication may result in deflation of the within-sample variance across the experiment. The variances may well remain homogeneous, and so the violation would not be detected. A residual analysis is not an unqualified ticket to proceed with the ANOVA.

Figure 4–4. A Decision Tree for Residual Analysis.


SINGLE-FACTOR ANOVA Preliminary Run
PLOT STUDENTIZED RESIDUALS
Variances homogeneous?
Y
N
HISTOGRAM OF RESIDUALS
TRANSFORM
Endless Loop?
Residuals normal?
Y
N
Accept ANOVA table as valid
Violations not extreme and design balanced, rely on robustness of ANOVA
Violations extreme or design unbalanced
Explore other options – Non-parametrics or GLIM

Summary of the assumptions of ANOVA

The assumptions of randomness and independence in sampling must be ensured by paying adequate attention to the random selection and allocation of items to the experimental classes or, if the design is constrained by the logistics of working with natural populations, by paying adequate attention to the random selection of items from within the experimental classes. If the assumptions of randomness or independence are violated, the results of the analysis can be profoundly affected, and the only recourse is to discard the data, redesign and repeat the experiment.

Departures from the assumption of equality of variances can be detected in a qualitative way by examining a plot of residuals, and a suitable transformation might be suggested by the pattern of scatter of those residuals. The effectiveness of the transformation may be evaluated by examination of the residuals following the transformation.

Having convinced yourself that the assumption of equality of variances is tenable, the assumption of normality may be tested by examining a histogram of the pooled residuals or a probability plot, depending upon your preference. Applying one or more of the tests introduced in Module 2 (Shapiro-Wilks test, probability plots etc) to the pooled residuals is likely to detect departures from normality that are of no practical consequence to ANOVA, especially where the design is balanced.

The residuals cannot be pooled for an assessment of the assumption of normality until homogeneity of variances across the factor classes has been achieved. A residual analysis will not necessarily detect violations of the assumptions of randomness and independence.

Transformations

Transformations are applied to either render the populations from which the samples are drawn normal, or to break a relationship between the mean and variance in order to comply with the assumption of homogeneity of variances. Very often both of these departures from the assumptions are simultaneously cured by the same transformation. The most commonly used transformations were first introduced in Module 3.

The following transformations are routinely used in the biological sciences.

The log transformation

$$Y' = \log_{10}(Y+1)$$

is applied in cases where the standard deviation is proportional to the mean, or when the distribution of the parent population is skewed to the right.

The square root transformation

$$Y = \sqrt{Y+0.5}$$

is applied when the data are counts for which a Poisson distribution is a satisfactory model, say for counts of organisms that are randomly distributed in the environment. It is usually unnecessary for such counts unless the mean count is less than 10.

The arcsine transformation

$$Y = SIN^{-1}\sqrt{p}$$

where *p* is a proportion. This transformation is appropriate for percentages and proportions. Data of this form are seldom satisfactorily modeled by a normal distribution if values occur outside the range 0.3 to 0.7 (30% to 70%). There is no need to apply the transformation if all values fall within this range.

There are other transformations, but they are beyond the scope of this Module. Refer to Sokal and Rohlf (1994, Chapter 13, section 6) or Zar (1984, Chapter 14) for further information. You can of course invent your own.

Robustness of ANOVA

The approach to checking assumptions recommended in this Module, that is, through qualitative examination of residuals, is not particularly rigorous. It relies in part on a general belief that analysis of variance is robust to moderate violations of the assumptions of normality and equality of variances. All but moderate violations would be evident on examination of the residuals in the manner described.

The foundation for this belief lies in Monte Carlo simulations undertaken in the middle of the twentieth century and reported by Lindquist (1953:78) and Keppel (1973). These studies show that moderate violations of normality do not constitute a serious problem and that, provided the samples sizes are equal or nearly so, nor do moderate departures from the assumption of equality of variances. If you are to rely heavily upon the robustness of ANOVA to violations of the assumption of equality of variances, in designing experiments, it is important to balance the design, that is, to ensure that the size of samples in each factor class are the same.

Non-parametric alternatives to ANOVA

If the assumptions of ANOVA are not met by the data, and no suitable transformation can be found to rectify the problem, or if the data are not measured at the interval or ratio level of measurement but rather the ordinal level, then we may decide to resort to a non-parametric alternative to ANOVA. One widely used non-parametric test is the Kruskal-Wallis one-way ANOVA by Ranks followed by multiple comparisons using a modified version of the Wilcoxon sum-rank test (Siegel and Castellan 1988; Sokal and Rohlf, 1994: Chapter 13). The functions `kruskal.test()` and `wilcox.test()` are available for non-parametric ANOVA, but the subject will not be covered further in this Module. One word of caution though. The

Kruskal-Wallis and Wilcoxon rank-sum tests may be non-parametric, but when used to address hypotheses specifically directed at means or medians, they are not assumption free. In addition to the universal assumptions of randomness in sampling and independence, these tests assume that the populations under consideration differ only in location, that is, that they have equal variances and that the shape of the distributions is the same.

ANOVA procedure in a nutshell

You should now have a grasp of one of the most important statistical concepts of use to biologists—the analysis of variance. Its fundamental objective is to determine whether the observed variation among a set of means is greater than would be expected by chance alone. It does this by comparing the observed variation among means with that expected on the basis of observed variation within samples.

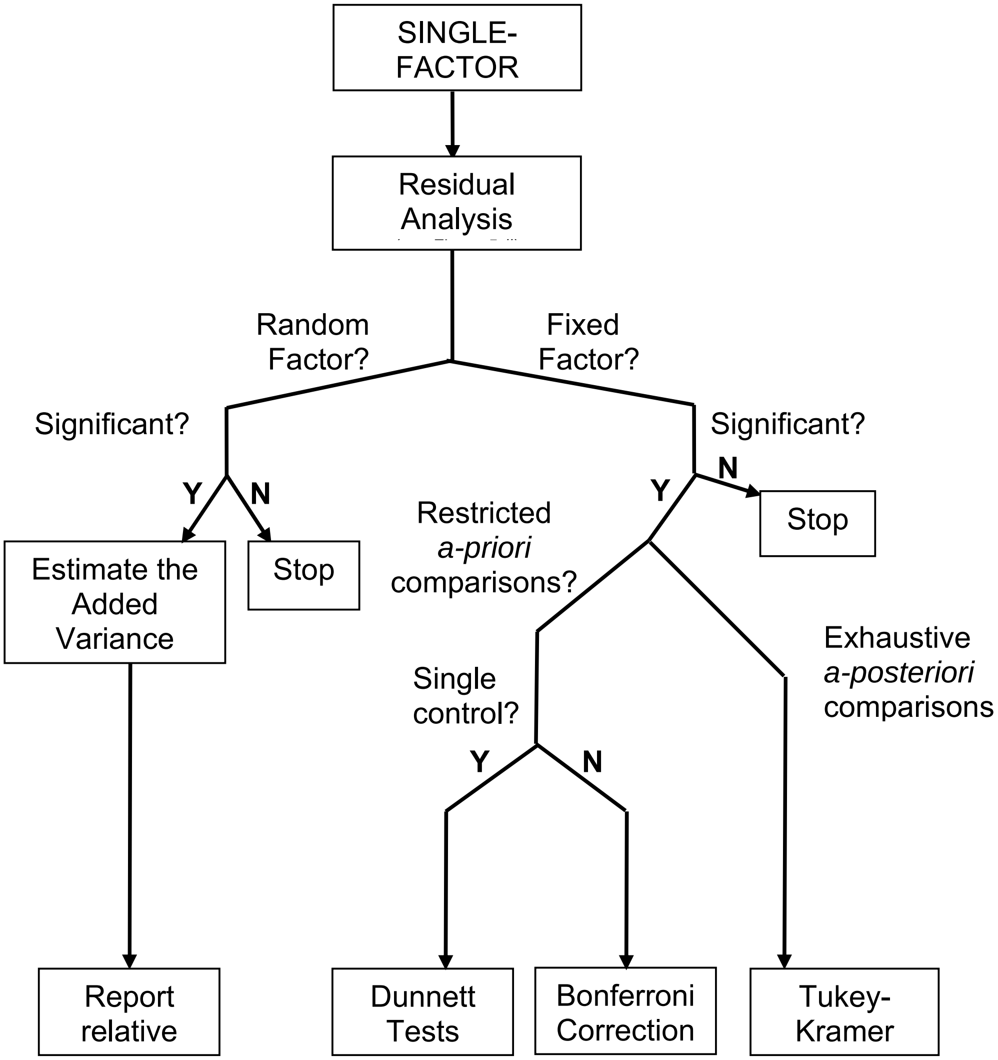
The general procedure for undertaking a study involving single-factor ANOVA is summarised as follows:

- Decide the research question you wish to address.
- Carefully select your factor and factor classes so that differences among the factor classes will unambiguously address the research question. This would normally mean holding all other potentially influential factors constant.
- Design your experiment and sampling protocols to ensure that the entities to be measured are either randomly allocated to factor classes. If class membership is beyond your control, ensure that the entities are selected at random from the populations represented by each class. Ensure independence of the entities selected within and across factor classes.
- If the model is fixed, plan your comparisons in advance of beginning the experiment if at all possible. This will greatly increase power.
- Collect the data.
- Undertake an exploratory analysis, based on graphical techniques, preliminary runs of the ANOVA and examination of residuals to verify that the assumptions of ANOVA are tenable. Transform the data where necessary.
- Perform the final ANOVA, and follow by multiple comparison tests if the model is fixed or by estimating the added variance components if the model is random — see the Decision Tree of Figure 4-4.
- Interpret results statistically. If the ANOVA is not significant, and you wish to conclude that there is no effect, undertake a

retrospective power analysis to support this conclusion. Otherwise conclude that you were unable to demonstrate a difference.

- Interpret results in the context of the initial question that you wished to address.

Figure 4–5. A Decision Tree for single-factor ANOVA.



Where have we come?

We have now covered all of the basics of analysis of variance where there is only a single driving factor. You should appreciate:

- an intuitive meaning for MS_{within}, MS_{among}, their sums of squares and degrees of freedom, and the F ratio itself;

- the difference between fixed and random models in ANOVA, and the practical consequences of these differences;
- the issues central to choosing an appropriate multiple comparison procedure, and a sensible position on a workable set of procedures to cover the common circumstances;
- the meaning and practical value of estimating the added variance component in the random model ANOVA;
- the assumptions of ANOVA, how to detect violations and how to overcome them, with emphasis on displaying and interpreting residuals.
- power analysis, and how to use it to interpret a non-significant result.

It is now appropriate to put this knowledge to use in worked examples and exercises. The practical application of the technique is much more straight-forward than you might think.

Lesson 6: Step-through Examples

Example 4-1: Macro-invertebrates of Crackenback River

This is an example of a one-way ANOVA where the factor is fixed, followed by unplanned (*a posteriori*) comparisons using the Tukey-Kramer Procedure.

The problem

David Tiller of the University of Canberra undertook a study of the effects of human disturbance on the benthic macro-invertebrate fauna of the Crackenback River in Kosciusko National Park. The river passes by the Thredbo Village, which discharges its sewage effluent, after treatment, 1.5 kilometres downstream. To assess the effects of this potential source of pollution on the fauna, Tiller chose the following sampling stations or sites:

- **Site 1:** 1 kilometre upstream of the village;
- **Site 2:** 1 kilometre below the village;
- **Site 3:** 1.5 kilometres below the village, ie immediately above the sewage effluent outflow;
- **Site 4:** 0.2 kilometres below the sewage effluent outflow;
- **Site 5:** 1 kilometre below the sewage effluent outflow;
- **Site 6:** 3 kilometres below the sewage effluent outflow;
- **Site 7:** 4.5 kilometres below the sewage effluent outflow;
- **Site 8:** 8 kilometres below the sewage effluent outflow.

Ten replicate collections of benthic invertebrates were made at each site using a Surber Sampler. A Surber Sampler is a square frame, 30 cm on a side, attached to a net. The sampler was placed on the bottom and all rocks to a depth of 10 cm lying within the square frame were washed and removed. Invertebrates dislodged by this process are carried by the flowing water into the collecting net. The collections were fixed in 4% formalin and returned to the laboratory for sorting, identification and counting. The data are raw counts pooled across species (Table 4–7).

David was interested to know if there were significant differences among the sites with respect to invertebrate abundance, and if so, where those differences lay.

Table 4–7. Raw counts of benthic invertebrates from the Crackenback River, upstream and downstream of Thredbo Village, Kosciusko National Park. The counts are pooled over species.

SITE							
1	2	3	4	5	6	7	8
286	325	496	1065	411	207	121	207
669	703	798	1539	1116	104	197	246
142	332	989	1174	681	153	292	468
65	265	640	880	1281	283	208	435
304	351	931	2113	1102	156	243	291
185	516	495	1172	578	386	260	246
210	350	469	1291	361	120	408	225
119	496	1160	1054	309	262	168	200
254	600	1139	1423	701	141	190	291
255	850	1072	1030	1242	294	110	174

The analysis

Data entry and exploratory examination

To analyse these data using R, they need to be set up in the form of two columns, one containing the measurements (in this case counts of benthic invertebrates) and the other containing a character string or number indicating from which site the measurement was taken. The data file THREDBO.DAT has been provided in your data file folder. Make sure you are using the latest version of the data files and check the moodle site for the latest version.

Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui). Check if Tinn-R is installed properly.

We must first read the data and assign them to an object `threadbo`.

```
> setwd("d:\\bernd\\biometryworkbook\\data")
> threadbo <- read.table("thredbo.dat", header=TRUE)
```

We check the data using some of the following functions: `names()`, `str()`, `dim()`, `head()`, `tail()`, `summary()`. How are they organized, how many cases are there (80). Was that to be expected? E.g.

```
> dim(threadbo)

[1] 80  2
```

The `head()`function shows by default the first six lines of a data.frame.

```
> head(threadbo)

  site count
1    1   286
2    1   669
3    1   142
4    1    65
5    1   304
6    1   185
```

Extra task

Try to find out, how you can display the first 10 lines of the data.frame by using the `head()` function. Hint: You may want to look at the help page `?head`

```
> names(threadbo)

[1] "site"  "count"

> str(threadbo)

'data.frame':   80 obs. of  2 variables:
 $ site : int  1 1 1 1 1 1 1 1 1 1 ...
 $ count: int  286 669 142 65 304 185 210 119 254 255 ...
```

The summary function gives some summary statistics for each variable in the `data.frame.`

```
> summary(threadbo)

      site          count
 Min.   :1.00   Min.   :  65.0
 1st Qu.:2.75   1st Qu.: 221.2
 Median :4.50   Median : 350.5
 Mean   :4.50   Mean   : 538.4
 3rd Qu.:6.25   3rd Qu.: 811.0
 Max.   :8.00   Max.   :2113.0
```

After you checked the data you should know that there are two variables “`site`” and “`counts`” within the data.frame `threadbo`. The summary command shows that `site` is not recognized as a factor. This is because site is coded by numerical values from 1-8 and R had no idea when you used the `read.table()` function that these are not numbers, but rather a coded factor. Therefore we need to convert `site` into a factor using the `as.factor` function.

```
> threadbo$site <- as.factor(threadbo$site)
```

We now should check if this has changed `site` into a factor. There are various ways to do this. One way is to type `summary()` again. If site is a factor then the summay statistics such as mean, medina er not calculated, but a frequency table for the first six levels of teh factor is shown.

```
> summary(threadbo)

      site         count
 1      :10   Min.   :  65.0
 2      :10   1st Qu.: 221.2
 3      :10   Median : 350.5
 4      :10   Mean   : 538.4
 5      :10   3rd Qu.: 811.0
 6      :10   Max.   :2113.0
 (Other):20
```

Another way is to use the `is.factor()` function.

```
> is.factor(threadbo$site)

[1] TRUE
```

To get a feel for the data, and the possible effects of the various potential impacts on the stream, a good idea is to plot the counts against sites in the stream.

```
> boxplot(count~site, data=threadbo, col="honeydew")
```

Here we used the formula version of the `boxplot()` function. That is we asked to plot `count` against `site` using the "~". In addition we specified which data the function should use by specifying `data=threadbo`. The formula interface comes in handy if the data are organised in the form as in the example above, i.e. one column codes for the factor (site)and the other the response variable (count).

To make the plot a bit more informative we want to label the axes. This is done by using the `xlab` and `ylab` arguments of the function.

```
> boxplot(count~site, data=threadbo, col="honeydew",
  xlab="sites", ylab="counts of invertebrates")
```

The resulting graph is shown in Figure 4–6. The text and arrows will be added later on (see next page)

Extra task

To check more options how to change the plot, e.g. change the width of the boxes, see ?boxplot.

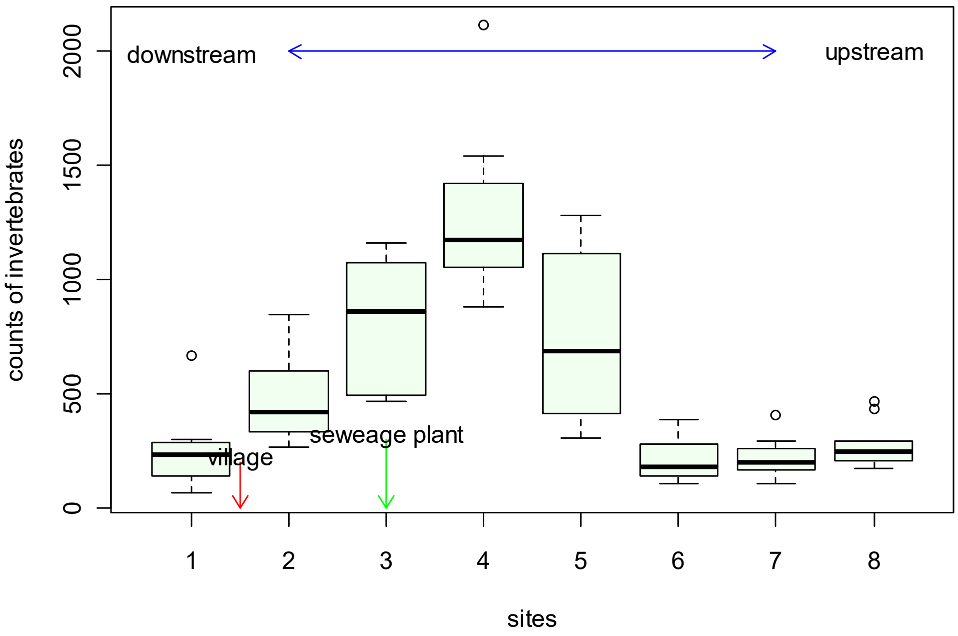
A much more tedious way would have been to specify each box seperatelty. If you are eager you can type in:

```
> boxplot(threadbo$count[1:10],threadbo$count[11:20] ,
threadbo$count[21:30], threadbo$count[31:40],
threadbo$count[41:50], threadbo$count[51:60],
threadbo$count[61:70], threadbo$count[71:80],
col="honeydew", xlab="sites", ylab="counts of
invertebrates")
```

This results in the same graph, but there are no number below the x-axis, so there is really no point doing it the hard way.

Submit the above commands for execution.

Figure 4–6. A graph showing counts of benthic invertebrates, for ten samples from each of eight sites in the Crackenback River. Vertical bars are ranges, boxes are interquartile distances on either side of the median.



At this point, you should mark on the plot, the locations of the various potential impacts. Remember the village is between site 1 and 2 and the seweage plant is directly at site 3.

R has great graphic capabilities and we want to use some of them to make the plot a bit more informative. First we want to indicate where the village is situated. We know it is between site 2 and 3, therefore we want to plot an arrow and a label between the boxes of these sites. This is done by the arrows() and text() functions. Both of them require coordinates to draw the items. The arrows() function requires to x-y coordinates, the beginning and the end of the arrow and the text function requires one x-y coordinate for the center of the text box.

The coordinates need to be guessed a bit. We may start with the arrow at x0=1.5 (between 1 and 2) and y0=200, a bit below the box of site two. The

```
> arrows(x0=1.5, y0=200, x1=1.5, y1=0, length=0.1,
col=”red”)
> text(x=1.5,y=230,labels="village")
```

And now we add the arrow and text for the seweage plant.

```
> arrows(x0=3, y0=300, x1=3, y1=0, length=0.1, col="green")
> text(x=3,y=330,labels="seweage plant")
```

Extra task

To see how the `arrows()` function works check ?arrows and then try to add the two labels “downstream” and “upstream” to the plot (see figure 4.6). How to you get a two-sided arrow? The label for the village is not really nicely placed (it touches the handles of the boxes, can you more it to somewhere else? What is the length argument for?

It would appear that the village itself has had an impact on the stream, perhaps aggravated by the sewage treatment plant, but at this point in the analysis, any interpretation would be pure supposition. The variation observed among the sites could well have happened by chance alone, and may not reflect a true difference in invertebrate abundance at the different sites. Tiller chose to perform an analysis of variance to determine if the observed differences among sites were significant, that is, if they reflected true differences at the sites from which they were drawn.

The within-sample variances differ considerably from site to site, and in fact appear to be correlated with the mean (higher abundances associated with higher variances). This is of some concern, as the analysis of variance assumes homogeneity of variances, but we will deal with this later.

Analysis of variance

To perform the ANOVA, we use the `aov()` function in R. There is another function called `lm()`, which does the same and is more general, hence can be applied to a wider array of data (e.g. unbalanced designs with unequal sample sizes). First we store our analysis in a new object called `anova.threadbo`. We do this because we can then reuse it for different version of outputs.

```
> anova.threadbo <- aov(count~site, data=threadbo)
```

After you have type in this command, you may expect an output such as an anova table, but disappointingly there is nothing. So let us try by typing the name of the anova.threadbo into the R-console.

```
> anova.threadbo

Call:
   aov(formula = count ~ site, data = threadbo)

Terms:
                    site Residuals
Sum of Squares  10414757   3893704
Deg. of Freedom        7        72

Residual standard error: 232.5494
Estimated effects may be unbalanced
```

This is a good start, but actually we would like to have a complete anova table including an F and p value. To achieve this type:

```
> summary(anova.threadbo)
```

Submit the above commands for execution.

The outpu should be as in Box 4–1.

Box 4-1. output of `summary(anova.threadbo)` used to analyse macro-invertebrate abundances in Crackenback River.

```
Df   Sum Sq Mean Sq F value    Pr(>F)
site         7 10414757 1487822  27.512 < 2.2e-16 ***
Residuals   72  3893704   54079
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Before launching into an interpretation of the ANOVA table, it would be wise to apply some diagnostic tests to determine if the ANOVA model is appropriate to the data at hand. We do this through an examination of residuals, that is, residual variation in the data after we have set the means for each sample to zero. A good way to look at them is to plot the distribution of residuals using the `hist()` function.

```
> hist(resid(anova.threadbo))
```

Another way to test the normality of the residuals and hence the appropriateness of the ANOVA approach is to use the inbuilt plots in R. If you type plot(anova.threadbo) you get four different plots, whereas the first one is the one we are looking here at this stage. You need to click on the plot area to see the plots actually, as the R console is waiting until you click with the mouse on the graphics window.

```
> plot(anova.threadbo)
```

Submit the above commands for execution.

The resulting plots are shown in Figure 4–7.

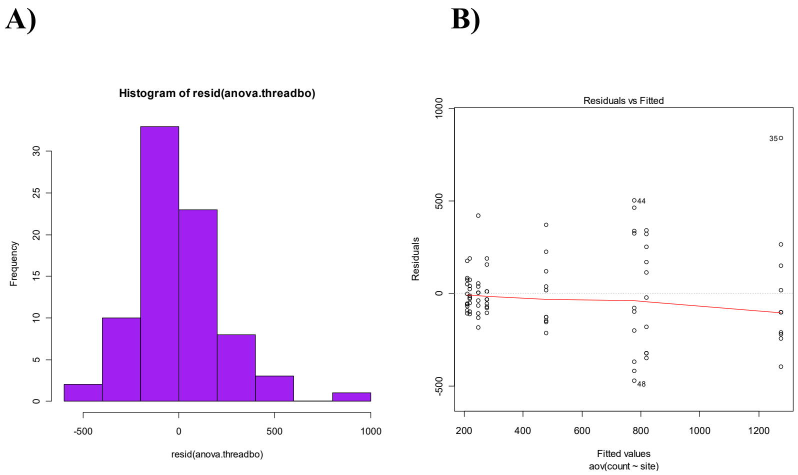


***Figure 4–7**. A) Histrogramm plot of residuals for counts of benthic invertebrates from eight sites in the Crackenback River.*

B) Plot of residuals versus predicted values. A correlation between the variance and mean is clearly indicated.

If the data conformed to the assumptions of homogeneity of variances and normality, we would expect the scatter of points to vary at random about the reference line (B). Instead, the scatter of points increases with increasing magnitude of the predicted value (mean). This correlation between the variance and mean is quite usual for counts of benthic invertebrates, and was not totally unexpected by Tiller. Tiller thought it prudent, on examination of the residual plot, to transform the counts before analysis. Two transformations would be appropriate to rectify the tendency for the variance to increase with the mean:

$$Y' = \sqrt{Y + 0.5}$$

or the stronger

$$Y' = \log_{10}(Y + 1)$$

Let us try the square root transformation, and re-examine if the residuals are normally distributed and show now increase in variance with increasing mean.

```
> threadbo$sqrt.count <- sqrt(threadbo$count)
> head(threadbo)

  site count sqrt.count
1    1   286  16.911535
2    1   669  25.865034
3    1   142  11.916375
4    1    65   8.062258
5    1   304  17.435596
6    1   185  13.601471
```

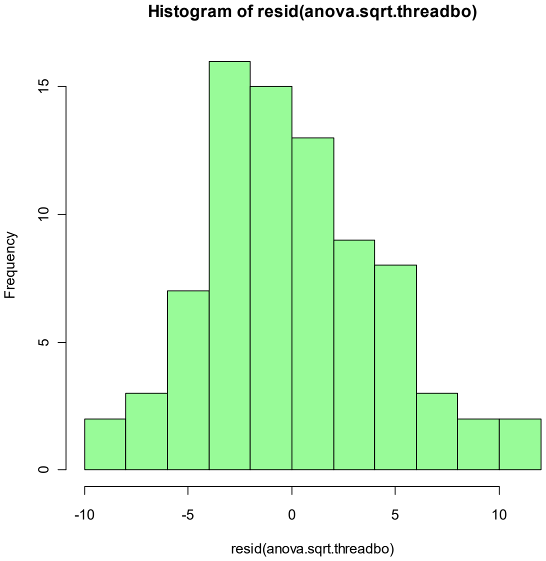
We added `sqrt.count` to the data.frame and redo now the whole excerise.

```
> anova.sqrt.threadbo <- aov(sqrt.count~site,
data=threadbo)
> summary(anova.sqrt.threadbo)

            Df Sum Sq Mean Sq F value    Pr(>F)
site         7 4377.7  625.38  31.072 < 2.2e-16 ***
Residuals   72 1449.1   20.13

> hist(resid(anova.sqrt.threadbo), col="palegreen")
> plot(anova.sqrt.threadbo)
```

Submit the above commands for execution.

The resulting residual plots are shown in Figure 4–8 and 4–9 . By inspection, it can be concluded that the residuals are not particularly skewed (Figure 4–8). The distribution of the residuals about the reference line is vastly improved. There is no longer an obvious trend in the residual variances (Figure 4–9).

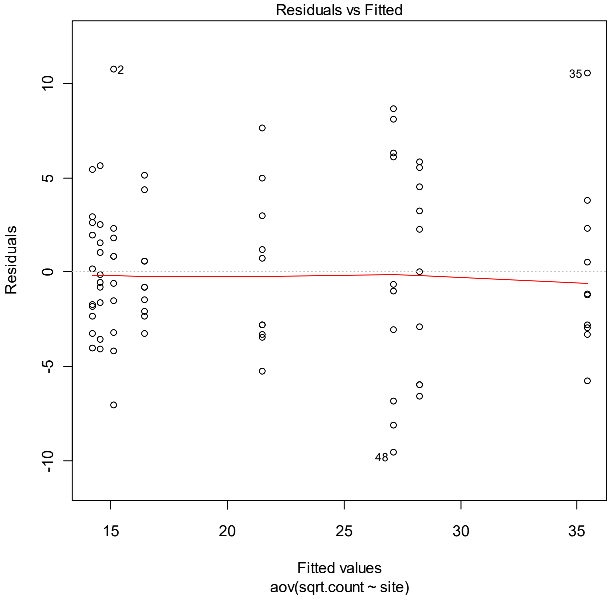
Figure 4–8. The distribution of residuals for counts of benthic invertebrates following transformation by square root.



Box 4–2. Output of summary(anova.sqrt. threadbo) used to analyse macro-invertebrate abundances in Crackenback River, following a square root transformation.

```
            Df Sum Sq Mean Sq F value    Pr(>F)    
site         7 4377.7  625.38  31.072 < 2.2e-16 ***
Residuals   72 1449.1   20.13                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4–9. *Residuals for counts of benthic invertebrates from eight sites in the Crackenback River. The counts were transformed by square root before analysis.*



We can now interpret the analysis of variance table based on the square rooted counts with confidence (Box 4–2).

The implications of the ANOVA of Box 4–2 are clear. The F value of 31.07 had a probability of occurring through chance alone of only <0.0001. As this is much less than the conventional cut-off value of 0.05, we conclude that the observed variation among sites is significant.

But where do the differences lie? To answer this question, Tiller chose to perform Tukey-Kramer multiple comparisons. To do so type simply:

```
> TukeyHSD(anova.sqrt.threadbo)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sqrt.count ~ site, data = threadbo)

$site
           diff          lwr         upr     p adj
2-1   6.4022766   0.1388780  12.6656752 0.0415675
3-1  13.1134643   6.8500657  19.3768629 0.0000002
4-1  20.3108457  14.0474471  26.5742443 0.0000000
5-1  12.0003141   5.7369155  18.2637127 0.0000021
6-1  -0.9022485  -7.1656471   5.3611500 0.9998165
7-1  -0.5406971  -6.8040957   5.7227015 0.9999944
8-1   1.3598262  -4.9035724   7.6232248 0.9973440
3-2   6.7111877   0.4477891  12.9745862 0.0271508
4-2  13.9085691   7.6451705  20.1719677 0.0000000
5-2   5.5980375  -0.6653611  11.8614361 0.1137063
6-2  -7.3045251 -13.5679237  -1.0411266 0.0113516
7-2  -6.9429737 -13.2063723  -0.6795752 0.0194695
8-2  -5.0424504 -11.3058490   1.2209482 0.2066807
4-3   7.1973814   0.9339828  13.4607800 0.0133535
5-3  -1.1131502  -7.3765488   5.1502484 0.9992646
6-3 -14.0157128 -20.2791114  -7.7523142 0.0000000
7-3 -13.6541614 -19.9175600  -7.3907628 0.0000001
8-3 -11.7536381 -18.0170366  -5.4902395 0.0000035
5-4  -8.3105316 -14.5739302  -2.0471330 0.0022483
6-4 -21.2130942 -27.4764928 -14.9496957 0.0000000
7-4 -20.8515428 -27.1149414 -14.5881442 0.0000000
8-4 -18.9510195 -25.2144181 -12.6876209 0.0000000
6-5 -12.9025626 -19.1659612  -6.6391641 0.0000003
7-5 -12.5410112 -18.8044098  -6.2776126 0.0000007
8-5 -10.6404879 -16.9038865  -4.3770893 0.0000319
7-6   0.3615514  -5.9018472   6.6249500 0.9999997
8-6   2.2620747  -4.0013238   8.5254733 0.9485753
8-7   1.9005233  -4.3628753   8.1639219 0.9800864
```

This looks quite confusing, but it is a table of all combinations of levels. Significant differences are these where the p value is below 0.05. A bit better is to have a plot on this.

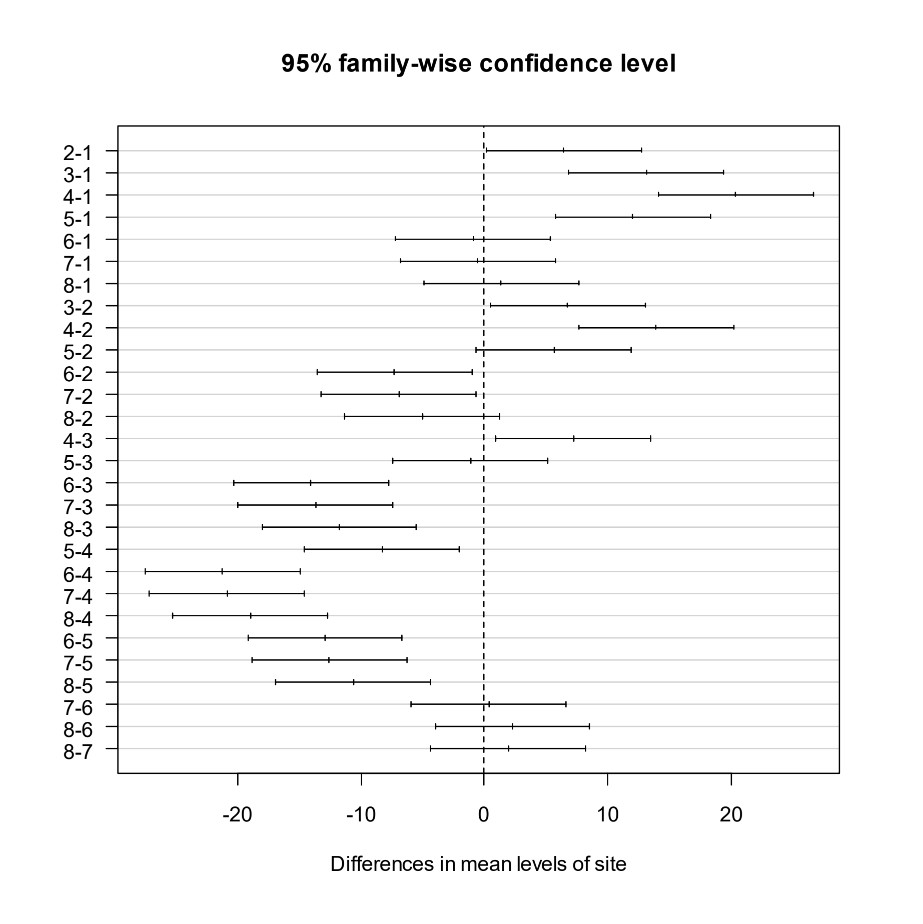
```
> plot(TukeyHSD(anova.sqrt.threadbo), las=1)
```

Type in and submit the above statements.

There are two points to note at this stage of the analysis. First, Tukey-Kramer multiple comparisons are only appropriate following a significant result in the ANOVA (it is a so-called protected test). Had the result not been significant, then the analysis would have been complete with the production of the analysis of variance table. Hence, it is important not to routinely run the Tukey comparisons with the ANOVA—the Tukey analysis is conducted only after a significant result is demonstrated by the ANOVA.

The output is shown in Box 4-3.

Box 4–3. Output from a Tukey Multiple Comparisons Test to determine which sites differed significantly from which others, following a significant result in the analysis of variance. The data are square rooted abundances of macro-invertebrates from Crackenback River.



The Tukey tests lists all differences in the means between all combination of levels. We have 8 levels in the factor site, hence there are 8 * (8-1)/2=28 comparisons.If the confidence interval in a comparison of two levels includes zero, then the difference between two levels is not significant. For example, Sites 1, 6, 7 and 8 are not significantly different from each other but are significantly different from Sites 3, 4 and 5. Site 4 stands out alone as significantly different from all other sites. It is below suspected sources of pollution, and shows a significant increase in invertebrate abundance compared with that of neighbouring sites upstream of the sources. This makes sense if you compare these results with the boxplot of figure 4–6. It appears that both the village and the sewage outlet have an impact on the stream, but the good news is that the effect has dissipated 3 kilometres downstream. Overlapping non-significant subsets are common in analyses of this sort, and are often difficult to interpret. They can only be resolved by increasing the sample sizes so that true differences, if any, become more apparent.

Results summary

The results of the analysis could be reported in the Results section of a report or paper along the following lines.

The differences among the eight sites in the abundance of benthic invertebrates (Figure 4–10) was significant, as demonstrated by a single-factor ANOVA applied to the square-root transformed counts (F = 31.07; df = 7,72; p < 0.0005) (Table 4–8).

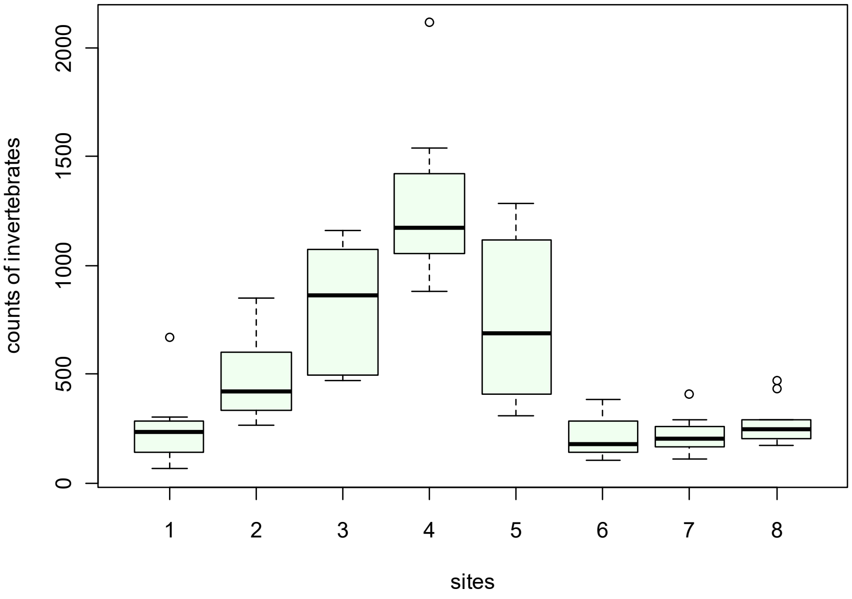


***Figure 4–10**. A graph showing counts of benthic invertebrates, for ten samples from each of eight sites in the Crackenback River. Vertical bars are ranges, boxes are interquartile distances on either side of the median..*

Multiple comparisons based on the Tukey-Kramer procedure ($\alpha = 0.05$) revealed that macro-invertebrate abundance was greatest at site 4, immediately downstream of the sewage outflow. Site 3, immediately above the outflow, and Site 5, 1 kilometre below the outflow were similar in terms of macro-invertebrate abundance, second only to Site 4. The upstream Site 1 and the downstream sites

6, 7 and 8 were not significantly different, suggesting that any impact the village and its sewage discharges may have had on the stream had dissipated by the time the stream water had flowed 3 kilometres beyond the impacts.

Table 4–8. *Summary of the analysis of variance used to compare macro-invertebrate abundance at eight sites on the Crackenback River near Thredbo Village.*

Source	Degrees of freedom	Sums of squares	Mean square	F value	Prob under H_0
Among sites	7	4377.7	625.38	31.07	P<0.0001
Within sites	72	1449.1	20.13		
Total	79	5826.8			

Discussion

The analysis can now be discussed in the context of the reasons for conducting the study. What advice can you give to the managers of the village or to the authorities responsible for the health of the Crackenback River and Lake Jindabyne downstream? How do the results of this study compare with those elsewhere, and what can be concluded about the effectiveness of the treatment system at Thredbo? These are the sorts of points that would be covered in a discussion of the results.

Adequacy of the design

At this point, it is constructive to consider the adequacy of the experimental design. First, the conclusions drawn from this experiment strictly relate only to the time of year that the experiment was undertaken, for there is no evidence that the pattern of differences observed in this study is repeated in all months of the year. Nor with a single experiment such as this are we able to gain insight into variation among years, and the pattern of differences among sites may vary from year to year. With an experiment such as the one described here, we are very limited in the degree to which we can draw inferences on the impact of human disturbance in the river beyond the period in which we conducted the study. David Tiller addressed these concerns in his broader study, with data collected every month for each of three years. Having replicated his treatments across years, Tiller was able to assess the level of variability between years and the interaction between the effect of location in the stream and the year or month chosen for study. However his broader analysis is beyond the scope of single-factor ANOVA.

The second point relates to the adequacy of the upstream site as a control. It may well have been that Site 1 was different from Sites 2, 3, 4 and 5 quite irrespective of the human impact, and this may have been so even before Thredbo Village had been built. Our interpretation that the development has had an impact on the stream, is predicated on the assumption that Site 1 resembles what Sites 2, 3, 4 and 5 would have been like had the development of the village not gone ahead. Quite an assumption, and the only support for it in this case is indirect, provided by the observation that Site 1 is not significantly different from sites below Site 5.

A third point on design is that the upstream site was not replicated. Tiller had only one site upstream of the suspected impacts. Without some measure of the variation among replicated sites upstream, we cannot tell if Site 1 was typical of sites upstream of the village. The whole analysis is vulnerable to the chance influences that might make Site 1 untypical of upstream sites.

A final point relates to Tiller's choice of follow-up analysis. The power of the analysis could have been substantially increased if Tiller had planned follow-up comparisons as part of the overall design before undertaking the study. By restricting the pool of possible comparisons under the study design, and using Sidak's tests in place of the exhaustive comparisons of Tukey-Kramer procedure, less adjustment of the experimentwise error would have been necessary. The resulting comparisons would have been less conservative, that is, more sensitive, with a corresponding increase in resolution of differences among sites. Planned comparisons are preferable to ad hoc unplanned comparisons in this regard.

All of these considerations must come into play, preferably when designing the study, and certainly when interpreting the results in preparation for publication.

Example 4-2: Rock lobster in aquaculture

This is a one-way ANOVA with a fixed factor and a single control treatment.

The Problem

The western rock lobster is recognized as a valued fishery estimated at 300 million dollars a year. To protect this valuable commodity and increase the production, the idea of culturing the lobster has become increasingly attractive. The aquaculture of these lobsters is still in its infancy. To achieve progress in this area, the development of an artificial diet and culturing environment is required as well as methods to allow the assessment of condition of cultured animals.

Tsvetnenko et al. (1999) developed a set of body condition factors to evaluate condition of cultured rock lobsters in conjunction with a nutritional study of the western rock lobster juvenile phase (post-pueruli or after settlement). A large sample of juvenile lobsters were collected at various locations and held in a tank at the Fisheries Marine Research Lab for two months as an acclimation period. During the acclimatisation period the lobsters were fed a diet of mussels and prawn pellets.

Table 4-1. Proximate composition of the experimental diets fed to rock lobster.

Parameter	D2*	D3*	D4**	D5**
Dry matter (%)	60.0	60.0	90.0	90.0
Dig. energy (MJ/kg)	9.0	10.0	13.6	15.1
Crude protein (%)	30.1	33.8	45.3	50.6
Crude fat (%)	4.8	7.4	7.2	11.1

* Semi moist; ** Dry

Subsequently, juvenile lobsters were fed for nine weeks on either fresh mussel diet (D1) or one of four artificial diets, two in semi-moist form (D2 and D3) and two in dry pelleted form (D4 and D5) (Table 4-1). The central hypothesis to be addressed was whether the lobsters performed as well on the artificial diets (D2-D5) as they did on the natural diet (D1). In this context, the natural diet of mussel was regarded as the control treatment, and the artificial diets were regarded as the experimental treatments.

A number of indices of condition were measured for comparison across treatments. One was the weight of wet tail muscle expressed as a percentage of the weight of the whole animal, wet. At the conclusion of the experiment, 10 animals were sampled at random from each diet treatment and their MSI condition measured.

Table 4-2. *Muscle somatic index (MSI) of lobster maintained on five dietary treatments.*

D1 (Control)	D2	D3	D4	D5
18.825	15.289	15.949	16.079	18.787
17.086	15.348	14.569	16.411	14.583
18.321	15.709	19.855	14.873	15.319
13.783	14.568	17.304	14.090	18.821
22.698	15.067	15.179	17.377	21.979
19.974	13.719	14.936	12.592	20.188
23.706	14.706	16.749	14.133	17.080
15.839	14.692	13.772	15.958	15.539
18.584	13.233	16.021	14.479	16.002
15.850	18.983	14.920	15.932	12.210

Data entry and exploratory examination

As always we have to read in the data. The name of the data file is lobdiet.dat and you can find it in your data folder.

```
> setwd("to\\your\\data\\folder")
> lobster <- read.table("lobdiet.dat", header=TRUE)
> lobster
```

```
       D1     D2     D3     D4     D5
1  18.825 15.289 15.949 16.079 18.787
2  17.086 15.348 14.569 16.411 14.583
3  18.321 15.709 19.855 14.873 15.319
4  13.783 14.568 17.304 14.090 18.821
5  22.698 15.067 15.179 17.377 21.979
6  19.974 13.719 14.936 12.592 20.188
7  23.706 14.706 16.749 14.133 17.080
8  15.839 14.692 13.772 15.958 15.539
9  18.584 13.233 16.021 14.479 16.002
10 15.850 18.983 14.920 15.932 12.210
```

So let us check whether the data were read in correctly and are in the right format

```
> dim(lobster)

[1] 10  5

> str(lobster)

'data.frame':   10 obs. of  5 variables:
 $ D1: num  18.8 17.1 18.3 13.8 22.7 ...
 $ D2: num  15.3 15.3 15.7 14.6 15.1 ...
 $ D3: num  15.9 14.6 19.9 17.3 15.2 ...
 $ D4: num  16.1 16.4 14.9 14.1 17.4 ...
 $ D5: num  18.8 14.6 15.3 18.8 22 ...
```

So we have five columns (i.e. the five treatments) and for each treatment 10 meassurements (i.e. MSI which is the ratio of the weight of the tail mussle for each lobster to the whole lobster).

Submit the above commands for execution.

At this point, we might like to define some value labels to the diets, as D1-D5 is not very informative and may lead to confusion later. We do this by assigning a new vector of names to the names of the data.frame `lobster`.

```
> names(lobster) <- c("control","wet1","wet2","dry1",
  "dry2")
>lobster

   control   wet1   wet2   dry1   dry2
1   18.825 15.289 15.949 16.079 18.787
2   17.086 15.348 14.569 16.411 14.583
3   18.321 15.709 19.855 14.873 15.319
4   13.783 14.568 17.304 14.090 18.821
5   22.698 15.067 15.179 17.377 21.979
6   19.974 13.719 14.936 12.592 20.188
7   23.706 14.706 16.749 14.133 17.080
8   15.839 14.692 13.772 15.958 15.539
9   18.584 13.233 16.021 14.479 16.002
10  15.850 18.983 14.920 15.932 12.210
```

That looks good, but before we can do the ANOVA we have to rearrange our data, as R is expecting the data in the format of only two columns. The first is the treatment, hence the various levels of the factor diet (e.g. control, wet1, wet2, etc...), and the second column should be the MSI. The easier task is to convert the five columns into a single by simply putting each column at the end of the previous one.

So we need to somehow access the column of the lobster data.frame separately. This is done by indexing, e.g.:

```
> lobster[,1]

 [1] 18.825 17.086 18.321 13.783 22.698 19.974 23.706
15.839 18.584 15.850

> lobster[,2]

 [1] 15.289 15.348 15.709 14.568 15.067 13.719 14.706
14.692 13.233 18.983
```

To combine each column into a single one we use the `c()` function and assign the new column to a new object called MSI.

```
> MSI <- c(lobster[,1], lobster[,2], lobster[,3],
lobster[,4], lobster[,5])

> MSI

  [1] 18.825 17.086 18.321 13.783 22.698 19.974 23.706 15.839 18.584
 [10] 15.850 15.289 15.348 15.709 14.568 15.067 13.719 14.706 14.692
```

```
[19] 13.233 18.983 15.949 14.569 19.855 17.304 15.179 14.936 16.749
[28] 13.772 16.021 14.920 16.079 16.411 14.873 14.090 17.377 12.592
[37] 14.133 15.958 14.479 15.932 18.787 14.583 15.319 18.821 21.979
[46] 20.188 17.080 15.539 16.002 12.210
```

```
> length(MSI)
```

```
[1] 50
```

So MSI is a new vector consisting of 50 values. If you compare the values to the `lobster` data.frame you will see that the first ten values are the ones from the control column, then another ten values for wet1, then 10 for wet2, 10 for dry1 and finally 10 for dry2.

Now we need to create a factor with the following content: 10 times "control", 10 times "wet1" etc until all 50 values have a treatment next to them. This factor then should be combined with the MSI vector into a new data.frame.

As always there are several ways to achieve this. We will use the `rep()` function, which is a helper function that creates based on repeated values (you should type `?rep` to see how this function works. We start with the following idea. We have the names of the diet treatment already stored somewhere, namely in the heading of lobster.

```
> names(lobster)
```

```
[1] "control" "wet1"    "wet2"    "dry1"    "dry2"
```

The general idea is to tell R: "please repeat these names 10 times". So let us try:

```
> rep(names(lobster), times=10)
```

```
 [1] "control" "wet1"    "wet2"    "dry1"    "dry2"    "control"
 [7] "wet1"    "wet2"    "dry1"    "dry2"    "control" "wet1"
[13] "wet2"    "dry1"    "dry2"    "control" "wet1"    "wet2"
[19] "dry1"    "dry2"    "control" "wet1"    "wet2"    "dry1"
[25] "dry2"    "control" "wet1"    "wet2"    "dry1"    "dry2"
[31] "control" "wet1"    "wet2"    "dry1"    "dry2"    "control"
[37] "wet1"    "wet2"    "dry1"    "dry2"    "control" "wet1"
[43] "wet2"    "dry1"    "dry2"    "control" "wet1"    "wet2"
[49] "dry1"    "dry2"
```

This is kind of good but not exactly, what we wanted. If you have read the help file of `?rep()` then you would have come across the argument `each`, so let us try this.

```
> rep(names(lobster), each=10)
```

```
 [1] "control" "control" "control" "control" "control" "control"
 [7] "control" "control" "control" "control" "wet1"    "wet1"
[13] "wet1"    "wet1"    "wet1"    "wet1"    "wet1"    "wet1"
[19] "wet1"    "wet1"    "wet2"    "wet2"    "wet2"    "wet2"
[25] "wet2"    "wet2"    "wet2"    "wet2"    "wet2"    "wet2"
```

```
[31] "dry1"    "dry1"    "dry1"    "dry1"    "dry1"    "dry1"
[37] "dry1"    "dry1"    "dry1"    "dry1"    "dry2"    "dry2"
[43] "dry2"    "dry2"    "dry2"    "dry2"    "dry2"    "dry2"
[49] "dry2"    "dry2"
```

Much better, `each` causes R to repeat each element in the vector separately in turn. The final step is now to convert this into a factor and then put both new columns together.

```
> diet <- as.factor(rep(names(lobster), each=10))
> diet
```

```
 [1] control control control control control control control
 [8] control control control wet1    wet1    wet1    wet1
[15] wet1    wet1    wet1    wet1    wet1    wet1    wet2
[22] wet2    wet2    wet2    wet2    wet2    wet2    wet2
[29] wet2    wet2    dry1    dry1    dry1    dry1    dry1
[36] dry1    dry1    dry1    dry1    dry1    dry2    dry2
[43] dry2    dry2    dry2    dry2    dry2    dry2    dry2
[50] dry2
Levels: control dry1 dry2 wet1 wet2
```

Finally we create a new `data.frame`, using the `data.frame()` function, and call it `lob2`, with `diet` and `MSI` as the only columns.

```
> lob2 <- data.frame(diet, MSI)
> head(lob2)
```

```
     diet    MSI
1 control 18.825
2 control 17.086
3 control 18.321
4 control 13.783
5 control 22.698
6 control 19.974
```

```
> str(lob2)
```

```
'data.frame':   50 obs. of  2 variables:
 $ diet: Factor w/ 5 levels "control","dry1",..: 1 1 1  ...
 $ MSI : num  18.8 17.1 18.3 13.8 22.7 ...
```

Submit the above commands for execution.

Extra task (a hard one)

There is another way to create the factor diet. Look at the function `?expand.grid()` and try to figure out how this can be achieved. A good start is to try:

```
> expand.grid(names(lobster))
```

A plot of the data reveals a clear trend (Figure 4-19), at least in the sample data. The boxes are only one standard error on either side of the mean, so the likely significance of the trend is in doubt. We will have to wait and see what the ANOVA tells us.

```
> boxplot(MSI ~ diet, data=lob2, col="paleturquoise2")
```

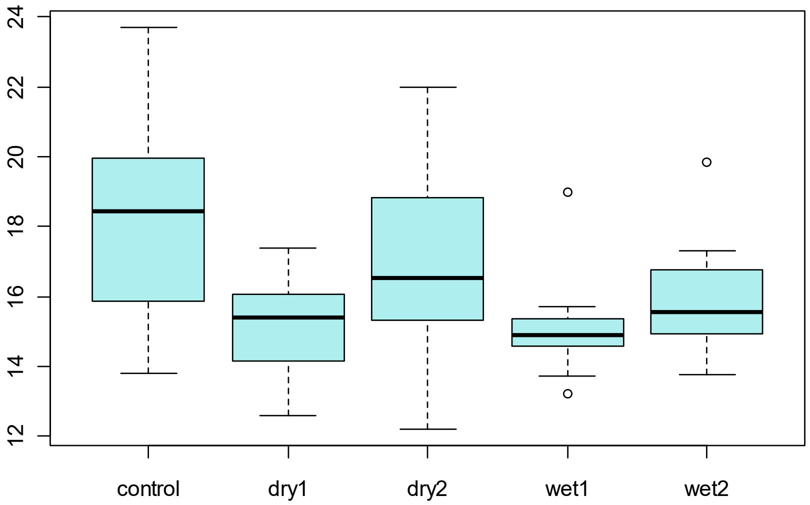


***Figure 4–19**. A box plot showing variation in lobster MSI for different diets. Note, the boxes show two standard errors on either side of the mean. The vertical bars are ranges.*

🏃 Generate the plot by entering the above command.

Preliminary examination of the plot indicates some promising variation among the treatments. The ANOVA will determine which differences we can regard as significant. To perform an ANOVA is quite simple once the data are in the right format. We use the formula notation with the "~" again. MSI is our response variable and diet our predictor (in this case it is a factor with five levels). The ANOVA will tell us if the factor diet can explain the significantly the variation in the data.

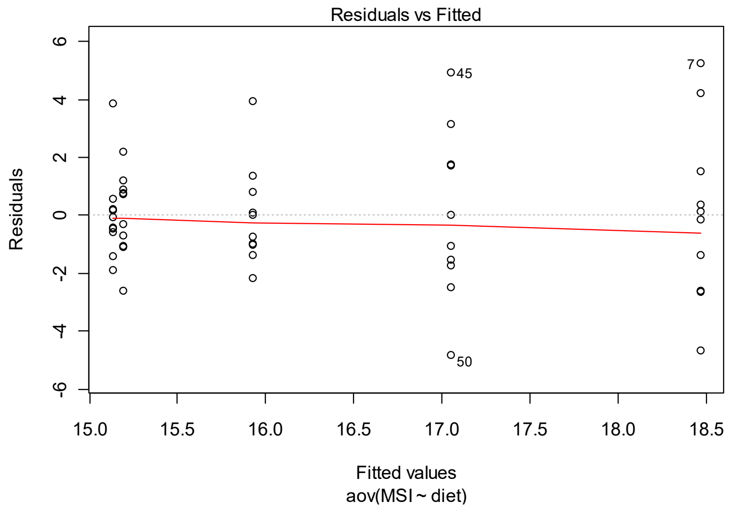
```
> anova.lobster <- aov(MSI ~ diet, data=lob2)
> summary(anova.lobster)

            Df  Sum Sq Mean Sq F value  Pr(>F)
diet         4  79.764 19.9409  3.9349 0.00801 **
Residuals   45 228.047  5.0677
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So in overall we see that the factor diet seems to be significantly explaining the variation in the data (p<0.01). Before we get too excited we need to test the assumption of the ANOVA by looking at the residuals. As you will recall an ANOVA is only valid if the resulting residuals are normally distributed.

```
> plot(anova.lobster)
```

Figure 4–2. *A residual plot for lobster MSI across different diets.*



The residuals look quite respectable, with only a slight trend in variance with increasing magnitude of the response variable. We could try to rectify this with a transformation, but in this case we will follow the authors lead and run with the raw data.

```
> hist(resid(anova.lobster), col="palegreen")
```

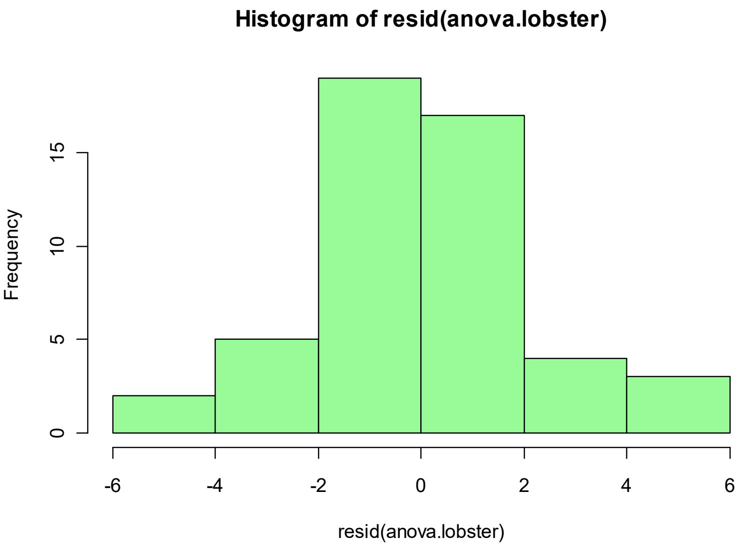
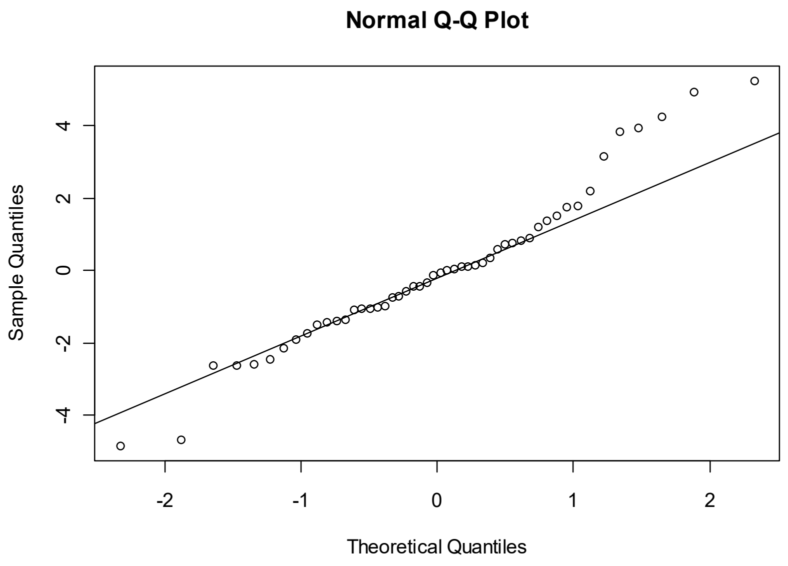


Figure 4–3. A histogram of the residuals from and ANOVA on lobster MSI across different diets.

There is no indication of substantial deviation from normality by the residuals in the histogram (Figure 4-3), the probability plot or in terms of significance of the Shapiro-Wilkes test (Box 4-1). We can now proceed to interpret the ANOVA printout.

```
> qqnorm(resid(anova.lobster))
> qqline(resid(anova.lobster))
> shapiro.test(resid(anova.lobster))
```

Box 4–1. Tests of Normality of the residuals from and ANOVA on lobster MSI across different diets.



```
        Shapiro-Wilk normality test

data:  resid(anova.lobster)
W = 0.9638, p-value = 0.1286
```

Box 4–2. *Results of an ANOVA comparing lobster MSI values for different diets.*

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
diet         4  79.764 19.9409  3.9349 0.00801 **
Residuals   45 228.047  5.0677
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lobsters raised on different diets differed significantly in condition as indicated by the muscle-somatic index (MSI). The question that now must be addressed is whether the artificial diets promote or retard body condition compared with the more natural diet (control).

This question can be addressed using Dunnett's Test, which is designed to adjust the level of significance to compensate for multiple comparisons where experimental treatments are each compared against a single control. Tukey's multiple comparison procedure would be far too conservative, as it tests for differences in all combinations of levels, provided we are not interested in comparing among the artificial diets.

Admittedly doing the Dunetts Test is a bit of a pain in R. We need to load a new package called multcomp (for multiple comparisons) and then use the function glht in this package to specify the comparisons we want.

```
> install.packages("multcomp",  lib="H:\\biometry")
> libary(multcomp)
> dunnett.comp <- glht(anova.lobster, linfct = mcp(diet =
"Dunnett"))
> summary(dunnett.comp)
```

The output is shown in Box 4-3.

All artificial diets yielded lower wet muscle-somatic index than lobsters fed control diet (control), however this result was significant only for lobsters fed diets that included formula dry1 and wet1. (Box 4-3).

Box 4–3. *Results of a Dunnett's multiple comparison procedure to*

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts
```

follow a significant result in an ANOVA comparing lobster MSI values for different diets.

```
Fit: aov(formula = MSI ~ diet, data = lob2)

Linear Hypotheses:
                        Estimate Std. Error t value Pr(>|t|)
dry1 - control == 0   -3.274      1.007   -3.252   0.0078 **
dry2 - control == 0   -1.416      1.007   -1.406   0.4360
wet1 - control == 0   -3.335      1.007   -3.313   0.0067 **
wet2 - control == 0   -2.541      1.007   -2.524   0.0507 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
(Adjusted p values reported -- single-step method)
```

Extra task

The Dunnett test does a comparison between control and each level only, hence is doing only four tests in this example. A Tukey test compares the combination of all levels (5*(5-1)/2 = 10), hence a Tukey test does 10 tests and therefore we the significance levels are adjusted differently. Do a Tukey test and compare the significane levels with the Dunnett test above. Which p values are lower (more significant)?

Results Summary

A fixed-effects, single factor ANOVA found a significant difference in wet muscle somatic index among diets fed to Rock Lobsters (F=3.93; df=4,45; p <0.01). Lobsters fed diets based on formula 1, whether semi-moist or in dry pellet form had significantly lower wet muscle somatic index, 3.3 g lower in each case, than lobsters fed a more natural Mussel control diet (Dunnett's multiple comparison test, α = 0.05). No difference could be demonstrated between the muscle somatic index of lobsters fed on formula 2 and those fed on the more natural mussel diet.

Discussion

Lobsters fed the natural mussel diet grew significantly faster than those fed the artificial diets, but this difference was statistically significant only in the case of formula 1 diets, whether wet or dry. If muscle somatic index were the only consideration, then the artificial diet based on Formula 2 would be the preferred choice among the two artificial diets.

Western rock lobsters naturally eat molluscs in the wild and supplement this feed with coralline algae. It should be noted that when all analyses were considered, the results agreed with previous research in that natural diets produce significantly better growth rates in lobsters than artificial feeds. Tsvetneko et al. (1999) later determined that the artificial feeds may have had too much lipid for good dietary performance, which seems to be born out in the partial results given here for MSI.

Source

The data that form the basis of this example were kindly provided by Dr. Elena Tsvetnenko, Muresk Institute of Agriculture, Curtin Univeristy of Technology, Suite 3, Enterprise Unit 1, 11 Brodie Hall Drive, Technology Park, Bentley WA 6102.

Tsvetneko, E., J. Brown, B. D. Glencross and L.H. Evans. 1999. Measures of condition in dietary studies on western rock lobster post-pueruli. Pp. 100-109 in Proceedings, International Symposium on Lobster Health Management, Adelaide, September 1999.

Example 4-3: Atmospheric SO_2 and Soybean Growth

This is a one-way ANOVA with a fixed factor followed by a retrospective power analysis.

The problem

The economy of China has been developing rapidly, supported largely by coal-fired power stations. Many regions regularly experience high atmospheric concentrations of sulphur dioxide emitted by the power stations. In the highly industrial area of Shenyang in the Lioaning Province, the average atmospheric SO_2 concentration during the summer is 38 ppb.

China is the world's third largest producer of soybeans. Soybean is an important cash crop, and vegetable oil produced from soybean is of high quality. Soybean (*Glycine max*) is sensitive to atmospheric SO_2, decreasing in yield by up to 4% at SO_2 concentrations as low as 0.05 ppm.

Open-top chambers (cylinders of 2.4 m high and 3 m in diameter) were used to fumigate soybean plants with high (488.6 ppb) and low (97.3 ppb) concentrations of SO_2. A control chamber was also established with background levels of SO_2 (1.2 ppb). Seeds of soybean were planted in 2 L plastic pots filled with a 2L an artificial potting material and randomly allocated to fumigation chambers.

The response variable was growth of soybean measured as total plant mass at the end of the experiment (g/pot).

Table 4-9. *Total mass (g) of soybean following treatment with atmospheric SO2 at background levels (control), low concentrations (97.3 ppb) and high concentrations (488.6 ppb).*

Replicate	Control	Low SO_2	High SO_2
1	79.1	61.8	61.2
2	64.7	58.9	65.2
3	67.0	76.3	51.0

The Analysis

Data entry and exploratory examination

We need to load our data. The name of the data file is "soybean.dat". Luckily this time the data are arranged as two columns, one containing the treatment and the other containing the measurement of soybean biomass.

```
> setwd("c:\\bernd\\data\\")
> soy <- read.table("soybean.dat", header=TRUE)
> soy
```

```
    treatment mass
  1   control 79.1
  2   control 64.7
  3   control 67.0
  4      so2+ 61.8
  5      so2+ 58.9
  6      so2+ 76.3
  7     so2++ 61.2
  8     so2++ 65.2
  9     so2++ 51.0
```

You should check the data using the usual functions such as (dim(), str() etc.)

A plot of the data reveals a clear trend (Figure 4-11), at least in the sample data. The boxes are only one standard error on either side of the mean, so the likely significance of the trend is in doubt. We will have to wait and see what the ANOVA tells us.

```
  > boxplot(mass ~ treatment, data=soy, col=" deepskyblue2")
```

Enter and submit the above commands.

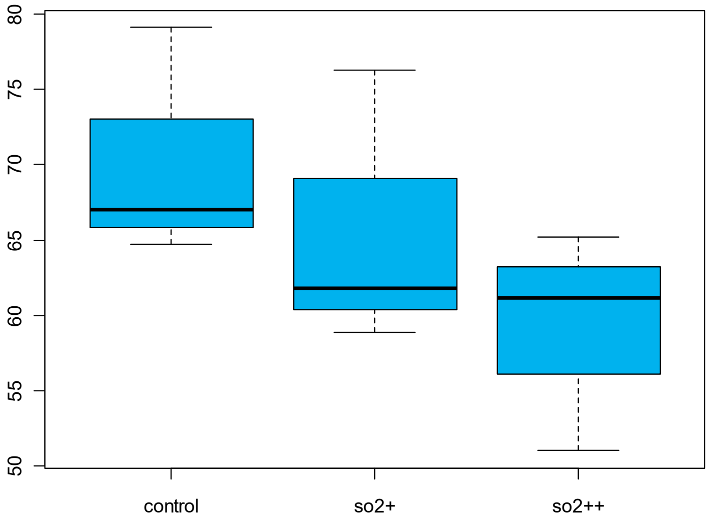


Figure 4–11. A box plot showing the results of treatment of soybean with three concentrations of atmospheric SO₂. Note, the boxes show only one standard error on either side of the mean, and so are not 95% confidence limits.

Although the sample sizes are exceptionally small (n=3), we should undertake a residual analysis before proceeding with the full ANOVA.

```
  > anova.soy <- aov(mass ~ treatment, data=soy)
```

```
> plot(anova.soy)
> hist(resid(anova.soy), col="lightskyblue2")
```

Generate the residual plots by entering and submitting the above commands.

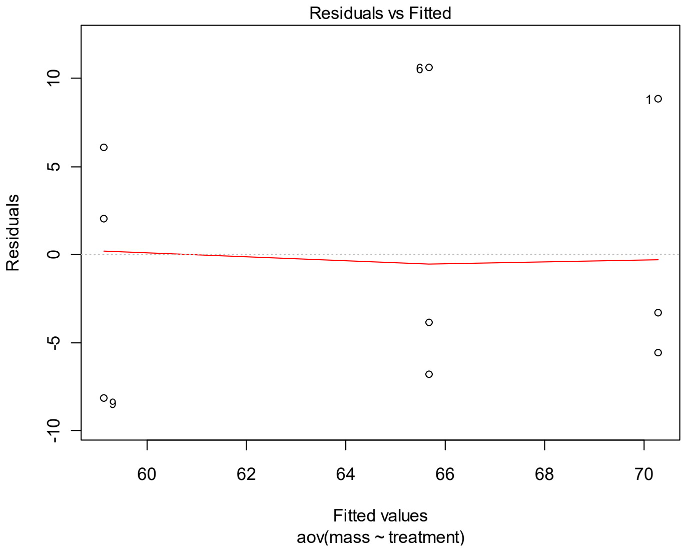


Figure 4–12. *A residual plot for an ANOVA of soybean growth against a treatment of three concentrations of atmospheric SO$_2$.*

The spread of the residuals looks good, so we should now consider normality.

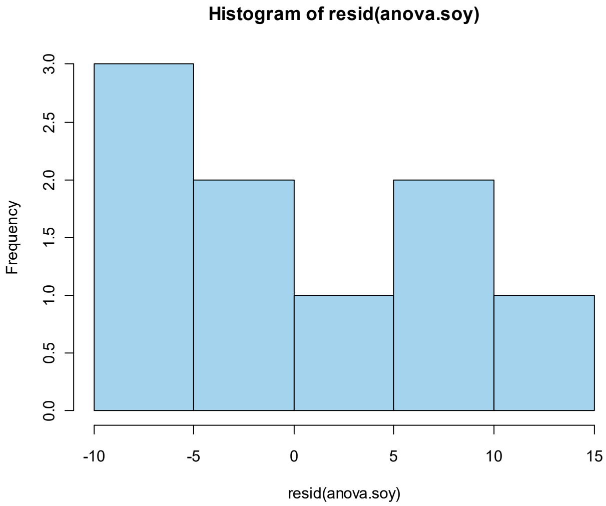


Figure 4–13. A *histogram* of residuals *from* an ANOVA of soybean growth against a treatment of three concentrations of atmospheric *SO$_2$.*

There is a indication of a problem with the distribution of the residuals, though with only nine values, there is not much to be gained by transformation.

We can now proceed to interpret the ANOVA printout.

```
> summary(anova.soy)
```

Box 4–4. *Output from summary(anova(()) used to undertake an ANOVA of soybean growth against a treatment of three concentrations of atmospheric SO_2.*

```
               Df Sum Sq Mean Sq F value Pr(>F)
treatment       2 187.80  93.898  1.4059 0.3157
Residuals       6 400.72  66.787
```

The analysis of variance is not significant (F=1.41; df=2,6; p =0.3157) so the trend, which appeared so clear in the original box plot diagram, is not supported. The differences we observed in the sample data may well have occurred by chance.

Had we got a significant difference, we would have proceeded to perform a Dunnett Test to compare each manipulated SO_2 treatment against the control.

We can report that we were unable to demonstrate an effect of SO_2 on the growth of soybean, which may have resulted either because SO_2 does not influence soybean growth, or because our sample sizes were so small that any effect could not be detected by the ANOVA. Because of our small samples, we cannot say that SO_2 does not affect soybean growth.

To proceed further, we need to do a **retrospective power analysis**. Such an analysis will tell us one of two things. First, it might tell us that we have sufficient power to detect all but insubstantial differences between treatment and control. If so, then our non-significant result can be regarded as evidence that SO_2 has no substantial effect on soybean growth. Alternatively it might tell us that our samples are far too small to be reasonably sure of detecting even a substantial effect of SO_2 on soybean growth, in which case we will have to carry our studies further with larger sample sizes.

Given our sample sizes (n=3), what is the smallest difference that we would be 80% sure of detecting at the 0.05 level of significance? We have:

$$P = 1 - \beta = 0.80\,;\ \alpha = 0.05\,;\ MS_{within} = 66.79\,;\ n = 3\,;\ \nu = 6$$

We have decided a priori that we are only interested in comparing the two manipulated SO2 treatments with the control so applying the Bonferroni correction, we have:

$$\alpha' = \frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

From tables, we have $t_{0.025[6]}$ =2.9687 and $t_{2(1-P)[\nu]}$ =0.9057. If you do not have tables, you can obtain these figures from R using the qt() function (equivalent to one-tailed tables):

```
> qt(0.025/2,6, lower.tail=F)

[1] 2.968687

> qt(0.4/2,6, lower.tail=F)

[1] 0.9057033
```

The figure we require is given by

$$\hat{\delta} \geq \left(t_{\alpha'[\nu]} + t_{2(1-P)[\nu]}\right)\sqrt{\frac{2MS_{within}}{n}} = (2.9687 + 0.9057)\sqrt{\frac{2\,x\,66.79}{3}} = 25.8532$$

So the smallest difference we could be reasonably sure of detecting with samples of size 3 is 26 g, or about a 37% reduction from the average for the growth of soybean in the experimental control (70 g). Clearly, our experiment is not nearly sensitive enough.

So if we are to run the experiment again, what sample sizes should we use to be reasonably sure of detecting an effect? To answer this question, we need to do a **prospective power analysis**.

Taking a 10% drop in yield as our smallest important difference, and striving for 80% chance of detecting such a difference, we have for our first iteration

$$\delta = 0.10 * 70 = 7.0\ g\,;\ P = 1 - \beta = 0.80\,;\ \alpha = 0.05\,;\ \alpha' = 0.025\,;$$
$$MS_{within} = 66.79\,;\ n = 3\,;\ \nu = 6$$

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 \left(t_{\alpha'[\nu]} + t_{2(1-P)[\nu]}\right)^2 = 2\frac{66.79}{7.0^2}(2.9687 + 0.9057)^2 = 40.92$$

and for our second iteration

$$n = 40\,;\nu = 117$$

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 \left(t_{\alpha'[\nu]} + t_{2(1-P)[\nu]}\right)^2 = 2\frac{66.79}{7.0^2}(2.2706 + 0.8447)^2 = 26.45$$

and for our third iteration

$n = 26$; $\nu = 75$

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 \left(t_{\alpha'[\nu]} + t_{2(1-P)[\nu]}\right)^2 = 2\frac{66.79}{7.0^2}(2.2873 + 0.8464)^2 = 26.77$$

Hence, the minimum sample size for us to be 80% sure of detecting a difference in yield as small as 10% at the 0.05 level of significance in an ANOVA with one control and two treatments is n=26. Our current sample size of n=3 is definitely inadequate, but served the purpose of allowing us to estimate how many measurements per factor level we need to take.

If a sample size of 26 is impracticable, we have the choice of revising our decision on the minimum effect that is important to us, or to in some way reduce the variability in growth between plants within treatments. We might do this by taking germination as our starting point in the experiment rather than time of planting of the seed, or by reducing the genetic variability of the seeds we use, but with loss of generality.

Results summary

A fixed-effects, single-factor ANOVA failed to detect any significant differences in plant growth among fumigation treatments with SO_2 either at high (488.6 ppb) or low (97.3 ppb) concentrations (F=1.41; df=2,6; p>0.05) compared to the experimental controls (1.2 ppb).

A retrospective power analysis indicated that a difference in soybean growth between the experimental control and one of the SO_2 treatments would have to be 15 g (a 20% reduction) or more to have a reasonable chance (80%) of being detected at the 5% level of significance. Clearly sample sizes of n=3 are not sufficient to detect all differences of importance, so we cannot conclude that SO_2 has no impact on soybean growth.

Sample sizes of 62 or more would be needed to have an 80% probability of detecting a difference between control and treatment of 4%.

Discussion

The exposure of soybean to high and low concentrations of SO_2 (488.6 and 97.3 ppb respectively) was expected to reduce the production of soybean as measured by the total biomass at the end of the experiment (g/pot). That we failed to demonstrate such an effect leaves us with an inconclusive result. Either SO_2 has no effect on soybean growth under the conditions we applied, or the effect is less than likely to be detected with three replicates per treatment.

The recommendation is that the work be expanded, and that the researchers consider alternative designs for increasing the power of the analysis. Such options include:

- re-evaluating the size of the smallest difference to be considered important, perhaps defining the smallest important difference as the difference that will trigger management intervention;
- considering whether it can be argued that SO_2 will either have no effect or will impede growth, allowing for a series of one-tailed tests following a significant result in the ANOVA;
- increasing the number of SO_2 treatments to enable analysis by regression rather than ANOVA;
- increasing the sample size substantially.

Source

The data that form the basis of this example were kindly provided by Frank Stagnitti and Xianzhe Xiong of Deakin University. They are preliminary data from a pilot study of the effects of metals and sulphur dioxide on the production of primary crops.

Lesson 7: Some Challenging Exercises

Assignment 1: Turbidity in Lake Burley Griffin

Turbidity in lakes and storages is of interest to water scientists because it has a profound affect on aquatic biota, and it is especially implicated with the switch from systems where rooted aquatic plants dominate *(e.g. Vallisneria)* to systems where planktonic algae dominate (e.g. *Microcystis).*

Turbidity is relatively easy to measure, and may be used as a surrogate for phosphorus in programmes monitoring water quality.

In a pilot study, Kurt Hammerschmidt collected ten replicate samples of water from each of ten sites in Lake Burley Griffin. The sites were specifically chosen at set intervals along the main channel leading from the inflow to the Scrivener Dam wall so that they could be revisited if necessary. Turbidity (in ntu) was measured for each replicate sample taken at each site, and the data are shown in the table below. The data file "turbid.dat" can be found in your data file folder.

Figure 4-15. *A map of Lake Burley Griffin showing the location of sampling sites.*

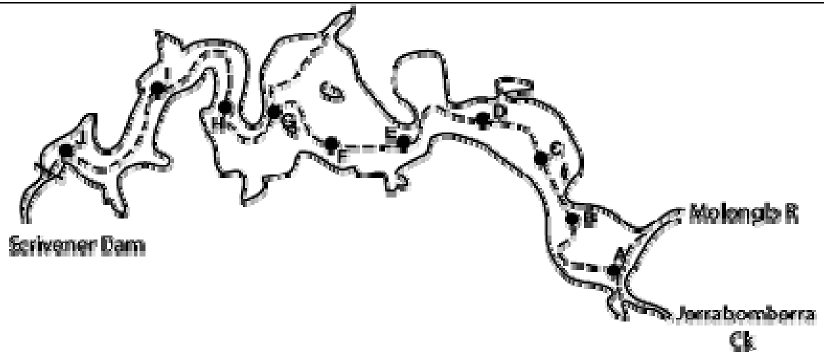


Table 4-11. *Turbidity measurements taken from each of ten sites along the channel in Lake Burley Griffin.*

SITE									
A	B	C	D	E	F	G	H	1	J
43	25	23	32	17	23	14	13	15	13
28	28	24	32	21	21	18	26	15	15
43	28	30	32	18	17	14	18	14	14
28	25	32	33	17	18	16	15	12	13
42	25	25	32	25	19	14	15	17	16
43	25	28	29	17	24	9	14	19	19
40	26	23	26	18	14	14	17	15	16
35	25	25	38	14	17	26	15	14	15
42	23	26	27	15	17	10	11	16	11
43	25	27	29	15	18	15	14	14	15

Kurt was interested to know whether there were significant differences among the sites in mean turbidity, and if so, where those differences lay. The study was conducted to determine whether it is sensible to take measurements of turbidity in urban lakes and storages at only a few sites, as is current practice.

Describe as completely as possible an appropriate analysis and give reasons for your choice. Be sure to specify the nature of the Factor (s) involved, and to state clearly the null hypotheses to be addressed.

Enter the data in a form suitable for the nominated analysis, and conduct an exploratory analysis based on graphical presentations with box diagrams. Include the box diagrams below. What would you anticipate the results of an appropriate ANOVA to be?

Perform the Analysis of Variance and summarise the results in the form of an ANOVA table.

Before interpreting the ANOVA table, examine a plot of the residuals to determine whether the assumptions of the analysis are tenable. If not, try some potential remedies, and repeat the ANOVA. Please include any graphs from your residual analysis below.

If the above analyses demonstrate a significant difference among the mean turbidity values, perform an appropriate follow-up analysis to determine where the differences lie. Present your results below.

Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to an ANOVA table and a figure showing the variation among sites (box diagrams). Include in your results, a statement of any clear and statistically significant trends in turbidity, but do not at this stage attempt to explain them.

Discuss the analysis in the context of the reasons for conducting the study. What might be the causes of the observed variation in mean turbidity among sites? What advice can you give to the water scientists charged with the responsibility of monitoring turbidity in Lake Burley Griffin?

Exercise 4-2: Point Impact Assessment

Phosphorus is an important nutrient in aquatic ecosystems, which can be changed dramatically through artificial discharges into streams and lakes. Phosphorous loads can be increased by agricultural activities of adjacent catchment areas, from discharge of household effluents, especially detergents, and from discharge of industrial wastes.

The following data are for concentrations of phosphorus (μg/l) in samples of water taken at various distances up and downstream of an industrial wastewater outlet. The figures in the body of the table are replicates taken from their respective locations at the one time.

Table 4-12. *Phosphorus concentration with distance downstream.*

		Distance Downstream (km)		
- 0.5	0.0	1.0	2.0	3.0
4.86	6.16	6.82	5.86	5.31
4.86	5.83	6.67	5.73	4.98
5.19	6.93	6.34	5.62	4.98
4.31	6.16	6.08	4.83	5.46
4.99	6.93	5.73	5.49	4.66

The water scientist wishes to know whether the mean phosphorous levels differed among the sites and if so, were they significantly lower or higher downstream from the effluent discharge compared to the upstream "control" site. He or she also needed to assess if the impact, if any, could be considered local or if it persisted well downstream.

Describe as completely as possible an appropriate analysis and give reasons for your choice. Be sure to specify the nature of the Factor (s) involved, and to state clearly the null hypotheses to be addressed.

Enter the data in a form suitable for the nominated analysis, and conduct an exploratory analysis based on graphical presentations with box diagrams. Include the box diagrams below. What would you anticipate the results of an appropriate ANOVA to be?

Perform the Analysis of Variance. Before interpreting the ANOVA table, examine a plot of the residuals to determine whether the assumptions of the analysis are tenable. If not, try some potential remedies, and repeat the ANOVA. Please include any graphs from your residual analysis below.

Once you have the residuals in an acceptable form, repeat if necessary the Analysis of Variance and summarise the results in the form of an ANOVA table.

If the above analyses demonstrate a significant difference among the mean phosphorus values, perform an appropriate follow-up analysis to determine where the differences lie. Present your results below.

Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to an ANOVA table and a figure showing the variation among sites (box diagrams). Include in your results, a statement of any clear and statistically significant trends in turbidity, but do not at this stage attempt to explain them.

Discuss the analysis in the context of the reasons for conducting the study. What impact did the industrial effluent have on stream phosphorus levels, and how persistent was that impact. If the impact is moderated by distance downstream, what biological processes might you suggest to explain this? What advice would you give to the Environmental Protection Authorities charged with responsibility for monitoring and maintaining water quality in our streams?

Discuss the adequacy of the experimental design, especially with regard to the use of an upstream site as a control.

Exercise 4-3: Duration of chase in Australian Chats

Three species of Australian Chat (*Epthianura*) can be found in micro-sympatry in mesic coastal, semi-arid and xeric arid regions of Western Australia. *Epthianura aurifrons* is the most physiologically competent to handle aridity, *Epthianura albifrons* is the least physiologically competent and *Epthianura tricolor* is intermediate in competence.

Territorial behaviour is expensive in terms of maintaining water balance, so data was collected for each species in the arid zone to see if physiological competence has a bearing on the duration of the territorial chase. A bird is said to engage in a territorial chase when it sees another bird off its territory. The data comprise a variable giving the species and a variable giving the duration of chase (in seconds).

Table 4-13. *Duration of chase in three species of Australian chat.*

SPECIES	DURATION OF CHASE			
Albifrons	48	24	32	39
Tricolor	66	66	54	51
Aurifrons	72	74	76	70

Analyse the data using an appropriate ANOVA model to address hypotheses on differences in duration of chase.

Describe as completely as possible an appropriate analysis and give reasons for your choice. Be sure to specify the nature of the Factor (s) involved, and to state clearly the null hypotheses to be addressed.

Enter the data in a form suitable for the nominated analysis, and conduct an exploratory analysis based on graphical presentations with box diagrams. Include the box diagrams below. What would you anticipate the results of an appropriate ANOVA to be?

Before preparing an ANOVA table, examine a plot of the residuals to determine whether the assumptions of the analysis are tenable. If not, try some potential remedies, and repeat the ANOVA. Please include any graphs from your residual analysis below.

Perform the Analysis of Variance and summarise the results in the form of an ANOVA table.

If the above analyses demonstrate a significant difference among the mean chase durations, perform an appropriate follow-up analysis to determine where the differences lie. Present your results below.

Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to an ANOVA table and a figure showing the variation among species (box diagrams). Include in your results, a statement of any clear and statistically significant trends in chase duration, but do not at this stage attempt to explain them.

Discuss the analysis in the context of the reasons for conducting the study. What might be the causes of the observed variation in chase duration among species taking into account differences in their physiological tolerance to aridity?

Exercise 4-4: Home range Estimation in Badgers

Badgers are widespread in Britain. In 1988, there were estimated to be around 42,000 social groups of badgers, and just under 200,000 adult badgers. By 1997 this had risen to just over 50,000 social groups and 310,000 adult badgers. The population is now probably stable. However, badgers are a high profile species, and their biology and conservation is of great interest.

Badgers live in social groups of four to 12 adults. They are a focus system for development of behavioral models in social grouping behavior that are uniquely relevant to carnivores. One of the prevailing models is known as the Resource Dispersion Hypothesis (RDH). The RDH hypothesis suggests that the dispersion and richness of resources in the environment provide a passive mechanism for the formation of groups, even without the direct benefits of group living. However, few studies have empirically tested the RDH in the field.

Dominic Johnson and his colleagues tested several hypotheses about RDH with their data on badgers in Wytham Woods, near Oxford, UK. In order for them to make this a meaningful test though, they had to standardise on an accepted method of describing home range. A preferred method of calculating home range is a contentious issue ebated among wildlife biologists, so Dominic wished to compare the estimates of home range area by three calculation methods: the interpolated mapping method (INT), the minimum convex polygon method (MCP), and the Dirichlet tessellation method (TES).

The data reside in the file badger.dat and comprise a class variable containing the method of home range area estimation and the home range areas (km^2) as data pairs.

Describe as completely as possible an appropriate analysis and give reasons for your choice. Be sure to specify the nature of the Factor (s) involved, and to state clearly the null hypotheses to be addressed.

Conduct an exploratory analysis based on graphical presentations with box diagrams. Include the box diagrams below. What would you anticipate the results of an appropriate ANOVA to be? Do you think it likely that the assumptions of the analysis will be upheld. If not, why not?

Perform the Analysis of Variance. Before interpreting the ANOVA, examine a plot of the residuals to determine whether the assumptions of the analysis are tenable. If not, try some potential remedies, and repeat the ANOVA. You may need to hit the data with a sledge hammer to meet the assumption of normality – Y' = LOG10(LOG10(Y+1). If there is a conflict in meeting the two assumptions of normality and homogeneity of variances, which takes precedence in this case? Why? Please include any graphs from your residual analysis below.

After selecting your final transformation, repeat the ANOVA. If the above analyses demonstrate a significant difference among the home range areas, perform an appropriate follow-up analysis to determine where the differences lie. Present your results below.

Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to an ANOVA table and a figure showing the variation among sites (box diagrams). Include in your results, a statement of any clear and statistically significant differences or trends in home range area, but do not at this stage attempt to explain them.

Discuss the analysis in the context of the reasons for conducting the study. What are the management implications of these results?

Exercise 4-5: T-tests versus ANOVA

You would appreciate by now that the single factor fixed design analysis of variance can be applied to compare several means, but you may not have realised that it can be used for the two sample case in place of the Student's t-test. Under such circumstances, when two statistical procedures are equally appropriate, it is satisfying to learn that they are mathematically equivalent and so always yield the same result. In fact, there is a simple relationship between F and t.

$$F_{0.05(1)[1,\nu]} = t^2_{0.05(2)[\nu]}$$

so the *F* value for the one-tailed test of analysis of variance, with 1 and ν degrees of freedom, is equal to the square of the *t* value for the equivalent t-test, with ν degrees of freedom.

This exercise is designed to demonstrate the point empirically. Consider again the study of *Antechinus* conducted by Geoff Smith at Cooloola in 1977 and 1978. Recall that he was interested to see if the weights of male marsupial mice *Antechinus flavipes* differed for the two years.

Table 4-14. *Weights of Antichinus flavipes in two consecutive years.*

YEAR			
1977		1978	
66	66	52	52
72	57	52	49
53	53	54	54
62	61	54	50
59	59	45	61
		58	

Analyse the data using a student's t-test to determine if the weights of *Antechinus flavipes* differ significantly between years. Record the sample value of t, the degrees of freedom, and the level of significance (*p* value). Refer to statistical tables to obtain the tabulated value of t at the 95% level of significance.

Now analyse these data as a single factor fixed model analysis of variance to address the same hypothesis. Complete the following table of results.

Does the above relationship hold for the sample values of t and F?

What do you conclude about the likely outcomes of the two approaches to the same problem?

Where have we come?

Lesson 7 is where the real learning occurs. In earlier lessons, you have read and understood written material and been led through worked examples. In Lesson 7 you were required to recall and integrate the information to complete some challenging real-world exercises. Recall in the context of problem solving is one of the best ways of achieving lasting learning.

In completing this module successfully, you will have achieved a number of core competencies, namely,

- Knowledge of the options available to you for analysing a single sample in terms of reporting its mean and the level of precision associated with it – the confidence limits.
- Knowledge of the options available to you for comparing two samples to see if the differences you observe are sufficient to conclude that the difference is real.
- Practical skills in the operation of R to undertake the necessary analyses.
- The ability and confidence to to interpret the results of the analyses in a biological context based on demonstrated understanding of the analyses.
- The ability to present findings in a style appropriate to the scientific literature.
- Appropriate attitudes and efficient strategies for extending your abilities to conduct analyses and solve problems beyond the scope of this module, by using resource materials such as statistical texts, software manuals, and your colleagues.

References

Day, RW & Quinn, GP (1989). Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs* 59:433-463.

Dunnett, CW (1980a). Pairwise multiple comparisons in the homogeneous variance, unequal sample-size case. *Journal of the American Statistical Association* 75:789-795.

Dunnett, CW (1980a). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association* 75:789-795.

Hayter, AJ (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics* 12:61-75.

Johnson, D. D. P., D. W. McDonald, C. Newman, and M. D. Morecroft. 2001. Group size versus territory size in group-living badgers: a large-sample field test of the resource dispersion hypothesis. *Oikos* 95: 265-274.

Keppel, G (1973). *Design and Analysis. A Researcher's Handbook*. Prentice Hall, New Jersey.

Lindquist, EF (1953). Design and Analysis of Experiments in Psychology and Education. Houghton Mifflin, Boston, p393.

Siegal S & Castellan Jr. NJ (1988). *Nonparametric statistics for the Behavioral Sciences.* 2nd ed. McGraw Hill Book Company New York.

Sokal & Rohlf (1994). *Statistical Tables*. 2nd ed, W.H. Freeman and Company, San Francisco, USA.

Sokal & Rohlf (1994). *Biometry*. *The Principles and Practice of Statistics in Biological Research*. 3rd ed, W.H. Freeman and Company, San Francisco, USA.

Tsvetneko, E., J. Brown, B. D. Glencross and L.H. Evans. 1999. Measures of condition in dietary studies on western rock lobster post-pueruli. Pp. 100-109 in Proceedings, International Symposium on Lobster Health Management, Adelaide, September 1999.

Zar, JH (1999). Biostatistical Analysis. 4th Ed, Prentice Hall, New Jersey.