Introduction to analysis of variance

What is analysis of variance?

Analysis of variance, often abbreviated to ANOVA, is a powerful statistic and a core technique for testing causality in biological data. Researchers use ANOVA to explain variation in the magnitude of a response variable of interest. For example, an investigator might be interested in the sources of variation in patients' blood cholesterol level, measured in mg/dL. Factors that are hypothesised to contribute to variation in the response may be categorical or continuous. A categorical factor has levels – the categories – that are each applied to a different group of sampling units. For example, sampling units of hospital patients may be classified as male or female, representing two levels of the factor 'Gender'. By contrast, a continuous factor has a continuous scale of values and is therefore a covariate of the response. For example, age of patients may be quantified by the covariate 'Age'. ANOVA determines the influence of these effects on the response by testing whether the response differs among levels of the factor, or displays a trend across values of the covariate. Thus, blood cholesterol level of patients may be deemed to differ among male and female patients, or to increase or decrease with age of the patient.

A factor of interest can be experimental, with sampling units that are manipulated to impose contrasting treatments. For example, patients may be given a cholesterol-lowering drug or a placebo, which represent two levels of the factor 'Drug'. Alternatively, the factor can be mensurative, with sampling units that are grouped according to some pre-existing difference. For example, patients may be classified as vegetarians or non-vegetarians, which represent two levels of the factor 'Diet'.

Biologists use ANOVA for two main purposes: prediction and explanation. In predictive studies, ANOVA functions as an exploratory tool to find

the best fitting set of response predictors. From a full model of all possible sources of variation in the response, procedures of model simplification allow the investigator to discard unimportant factors and so develop a model with maximum predictive power. This application of ANOVA is just one of many forms of exploratory analyses now available in standard statistics packages. ANOVA really comes into its own when it is used for hypothesis testing. In this case, the primary goal is to explain variation in a response by distinguishing a hypothesised effect, or combination of effects, from a null hypothesis of no effect. Any such test of hypothesised effects on a response has an analytical structure that is fixed by the design of data collection. Although this book provides some guidance on model simplification, its principal focus is on the hypothesis-testing applications of ANOVA in studies that have been designed to explain sources of variation in a response. More exploratory studies concerned with parameter estimation may be better suited to maximum likelihood techniques of generalised linear modelling (GLIM) and Bayesian inference, which lie beyond the scope of the book.

The great strength of ANOVA lies in its capacity to distinguish effects on a response from amongst many different sources of variation compared simultaneously, or in certain cases through time. It can identify interacting factors, and it can measure the scale of variation within a hierarchy of effects. This versatility makes it a potentially powerful tool for answering questions about causality. Of course tools can be dangerous if mishandled, and ANOVA is no exception. Researchers will not go astray provided they adhere to the principle of designing parsimonious models for hypothesis testing. A parsimonious design is one that samples the minimum number of factors necessary to answer the question of interest, and records sufficient observations to estimate all potential sources of variance amongst those chosen few factors. As you use this book, you will become aware that the most appropriate models for answering questions of interest often include nuisance variables. They need measuring too, even if only to factor them out from the effects of interest. One of the biggest challenges of experimental design, and best rewards when you get it right, is to identify and fairly represent all sources of variation in the data. True to the playful nature of scientific enquiry, this calls for building a model.

How to read and write statistical models

A statistical model describes the structure of an analysis of variance. ANOVA is a very versatile technique that can have many different

structures, and each is described by a different model. Here we introduce the concept of a statistical model, and some of the terminology used to describe model components. The meanings of terms will be further developed in later sections, and all of the most important terms are defined in a Glossary on page 271.

Analysis of variance estimates the effect of a categorical factor by testing for a difference between its category means in some continuous response variable of interest. For example, it might be used to test the response of crop yield to high and low sowing density. Data on yield will provide useful evidence of an effect of density if each level of density is sampled with a representative group of independent measures, and the variation in yield between samples can be attributed solely to sowing density. The test can then calibrate the between-sample variation against the residual and unmeasured within-sample variation. A relatively high between-sample variation provides evidence of the samples belonging to different populations, and therefore of the factor explaining or predicting variation in the response. The analysis has then tested a statistical model:

$$Y = A + \varepsilon$$

We read this *one-factor* model as: 'Variation in the *response* variable [Y] is explained by [=] variation between levels of a *factor* [A] in addition to [+] *residual* variation [ε]'. This is the *test hypothesis*, H_1, which is evaluated against a mutually exclusive *null hypothesis*, H_0: $Y = \varepsilon$.

The evidence for an *effect* of factor A on variation in Y is determined by testing H_0 with a *statistic*, which is a random variable described by a probability distribution. Analysis of variance uses the F statistic to compute the probability P of an effect at least as big as that observed arising by chance from a true null hypothesis. The null hypothesis is rejected and the factor deemed to have a significant effect if P is less than some predetermined threshold α, often set at 0.05. This is known as the *Type I error rate* for the test, and $\alpha = 0.05$ means that we sanction 5% of such tests yielding false positive reports as a result of rejecting a true null hypothesis. The analysis has a complementary probability β of accepting a false null hypothesis, known as the *Type II error rate*. The value of β gives the rate of false negative reports, and a lower rate signifies a test with more power to distinguish true effects. We will expand on these important issues in later sections (e.g., pages 13 and 248); for the purposes of model building, it suffices to think of the factor A as having a significant effect if $P < 0.05$.

Analysis of variance can also estimate the effect of a continuous factor. This is done by testing for a trend in the response across values of the

covariate factor. The analysis is now referred to as *regression*. For example, one might wish to test the response of crop yield to sowing density measured on a continuous scale of seeds/m^2. A single sample of independent measurements of yield over a range of sowing densities allows the effect of sowing density to be tested with a statistical model having the same structure as the one for the categorical factor:

$$Y = A + \varepsilon$$

We read this *simple linear regression* model as: 'Variation in the response variable [Y] is explained by [=] variation in a *covariate* [A] in addition to [+] residual variation [ε]'. The process of distinguishing between the test hypothesis and a null hypothesis of no effect is exactly the same for the covariate as for the categorical factor. The null hypothesis is rejected and the covariate deemed to cause a significant linear trend if $P < \alpha$.

Users of statistics employ a variety of terminologies to describe the same thing. One-factor designs may be referred to as *one-way* designs. The response may be referred to as the *data* or *dependent* variable; each hypothesised effect may be referred to as a *factor*, *predictor* or *treatment*, or *independent* or *explanatory* variable; categories of a factor may be referred to as *levels*, *samples* or *treatments*; and the observations or measures within a sample as *data points*, *variates* or *scores*. Each observation is made on a different *sampling unit* which may take the form of an individual *subject* or *plot* of land, or be one of several repeated measures on the same subject or block of land. The residual variation may be referred to as the *unexplained* or *error* variation. The precise meanings of these terms will become apparent with use of different models, for some of which residual and error variation are the same thing and others not, and so on. A summary of the standard notation for this book can be found on page 44, and further clarification of important distinctions is provided by the Glossary on page 271.

The full versatility of ANOVA becomes apparent when we wish to expand the model to accommodate two or more factors, either categorical or continuous or both. For example, an irrigation treatment may be applied to a sample of five maize fields and compared to a control sample of five non-irrigated fields. Yield is measured from a sample of three randomly distributed plots within each field. Thus, in addition to differences between plots that are the result of the irrigation treatments, plots may differ between fields within the same treatment (due to uncontrolled variables). This design has an Irrigation factor A with two levels: treatment and control, and a Field factor B with five levels per

level of A. Factor B is *nested* in A, because each field belongs to only one level of A. This *two-factor nested* model is written as:

$$Y = A + B'(A) + \varepsilon$$

The model equation is read as: 'Variation in growth rate [Y] is explained by [=] variation between treatment and control fields [A], and [+] variation between fields nested within each treatment level [B′(A)], in addition to [+] residual variation between plots within each field [ε]'. This model has two test hypotheses: one for each factor. At the cost of greater design complexity, we are now able to test the region-wide applicability of irrigation, given by the A effect, even in the presence of natural variation between fields, given by the B′(A) effect.

The site factor B′ is conventionally written as B-prime in order to identify it as a *random* factor, meaning that each treatment level is assigned to a random sample of fields. Factor A is without prime, thereby identifying it as a *fixed* factor, with levels that are fixed by the investigator – in this example, as the two levels of treatment and control. We will return again to fixed and random factors in a later section (page 16), because the distinction between them underpins the logic of ANOVA. A nested model such as the one above may be presented in the abbreviated form: 'Y = B′(A) + ε', which implies testing for the main effect A as well as B′(A). Likewise, the abbreviated description: Y = C′(B′(A)) + ε implies testing for A and B′(A) as well as C′(B′(A)).

As an alternative or a supplement to nesting, we use designs with *crossed* factors when we wish to test independent but simultaneous sources of variation that may have additive or multiplicative effects. For example, seedlings may be treated simultaneously with different levels of both a watering regime (A) and a sowing density (B). This is a *factorial* model if each level of each factor is tested in combination with each level of the other. It is written as:

$$Y = A + B + B*A + \varepsilon$$

The model equation is read as: 'Variation in growth rate [Y] is explained by [=] variation in watering [A], and independently [+] by variation in sowing density [B], and also [+] by an inter-dependent effect [B*A], in addition to [+] residual within-sample variation [ε]'. This model has three test hypotheses: one for each factor and one for the interaction between them. We are now able to test whether A and B act on the response as independent *main effects* A and B additively, or whether the effect of each factor on Y depends on the other factor in an *interaction* B*A. An

interaction means that the effects of A are not the same at all levels of B, and conversely the effects of B differ according to the level of A. This factorial model can be written in abbreviated form: 'Y = B|A + ε', where the vertical separator abbreviates for 'all main effects and interactions of the factors'. Likewise, the description of a three-factor model as: Y = C|B|A + ε abbreviates for all three main effects and all three two-way interactions and the three-way interaction:

$$Y = A + B + B*A + C + C*A + C*B + C*B*A + \varepsilon$$

For any ANOVA with more than one factor, the terms in the model must be entered in a logical order of main effects preceding their nested effects and interactions, and lower-order interactions preceding higher-order interactions. This logical ordering permits the analysis to account for independent components in hierarchical sequence.

This book will describe all the combinations of one, two and three factors, whether nested in each other or crossed with each other. For example, the above cross-factored and nested models may be combined to give either model 3.3 on page 98: Y = C|B(A) + ε, which is also described with an example on page 51, or model 3.4 on page 109: Y = C(B|A) + ε. Throughout, we emphasise the need to identify the correct statistical model at the stage of designing data collection. It is possible, and indeed all too easy, to collect whatever data you can wherever you can get it, and then to let a statistical package find the model for you at the analysis stage. If you operate in this way, then you will have no need for this book, but the analyses will certainly lead you to draw unconvincing or wrong inferences. Effective science, whether experimental or mensurative, depends on you thinking about the statistical model when designing your study.

What is an ANOVA model?

Any statistical test of pattern requires a model against which to test the null hypothesis of no pattern. Models for ANOVA take the form: Response = Factor(s) + ε, where the response refers to the data that require explaining, the factor or factors are the putative explanatory variables contributing to the observed pattern of variation in the response, and ε is the residual variation in the response left unexplained by the factor(s). For each of the ANOVA designs that we describe in Chapters 1 to 7, we express its underlying model in three ways to highlight different features of its structure. For

example, the two-factor nested model introduced above is described by its:

- Full model, packed up into a single expression: $Y = B'(A) + \varepsilon$;
- Hierarchical nesting of sampling units in factors: $S'(B'(A))$;
- Testable terms for analysis, unpacked from the full model: $A + B(A)$.

A statistics package will require you to specify the ANOVA model desired for a given dataset. You will need to declare which column contains the response variable Y, which column(s) contain the explanatory variable(s) to be tested, any nesting or cross factoring of multiple factors (these are the 'testable terms' above), whether any of the factors are random (further detailed on page 16) and whether any are covariates of the response (page 29). On page 259, we describe a typical dataset structure and associate it with various models.

In the event that the analysis indicates a real effect, this outcome can be described succinctly (detailed on page 260) and illustrated with a graph. Figure 1(a) shows a typical illustration of differences between group means for a model $Y = A + \varepsilon$, with three levels of A. The significance of the pattern is evident in the large differences between the three means relative to the residual variation around the means. A non-significant effect of factor A would result from larger sample variances, or sample means all taking similar values.

General principles of ANOVA

Analysis of variance tests an effect of interest on a response variable of interest by analysing how much of the total variation in the response can be explained by the effect. Differences among sampling units may arise from one or more measured factors making up the effect(s) of interest, but it will certainly also arise from other sources of unmeasured variation. Estimating the significance of a hypothesised effect on the response requires taking measurements from more than one sampling unit in each level of a categorical factor, or across several values of a covariate. The sampling units must each provide independent information from a random sample of the factor level or covariate value, in order to quantify the underlying unmeasured variation. This random variation can then be used to calibrate the variation explained by the factor of interest.

For example, we can use ANOVA to test whether gender contributes significantly to explaining variation in birth weights of babies. To assess the effect of gender as a factor in the birth-weight response, it makes sense to weigh one sample of male babies and another of female babies, with each baby picked at random from within the population of interest (perhaps a geographical region or an ethnic group). These babies serve as the replicate sampling units in each of the two levels (male and female) of the factor gender. The babies must be chosen at random from the defined population to avoid introducing any bias that might reinforce a preconceived notion, for example by selecting heavier males and lighter females. They should also contribute independent information to the analysis, so twins should be avoided where the weight of one provides information about the weight of the other. The ANOVA on these samples of independent and random replicates will indicate a significant effect of gender if the average difference in weight between the male and female samples is large compared to the variation in weight within each sample.

ANOVA works on the simple and logical principle of partitioning variation in a continuous response Y into explained and unexplained components, and evaluating the effect of a particular factor as the ratio between the two components. The method of partitioning explained from unexplained variation differs slightly depending on whether the ANOVA is used to compare the response among levels of a categorical factor or to analyse a relationship between the response and a covariate. We will treat these two methods in turn.

Analysis of variance on a categorical factor tests for a difference in average response among factor levels. The total variation in the response is given by the sum of all observations, measured as their squared deviations from the response grand mean $\bar{\bar{y}}$. This quantity is called the total sum of squares, SS_{total} (Figure 1). The use of squared deviations then allows this total variation to be partitioned into two sources. The variation explained by the factor is given by the sum of squared deviations of each group mean $\bar{y}_i$ from the grand mean $\bar{\bar{y}}$, weighted by the n values per group (where subscript i refers to the i-th level of the factor). This quantity is called the explained sum of squares, $SS_{explained}$. The residual variation left unexplained by the model is given by the sum of squared deviations of each data point y_{ij} from its own group mean $\bar{y}_i$ (where subscript ij refers to the jth observation in the i-th factor level). This quantity $SS_{residual}$ is variously referred to as the residual, error, or unexplained sum of squares.

Each sum of squares (SS) has a certain number of degrees of freedom (d.f.) associated with it. These are the number of independent pieces of

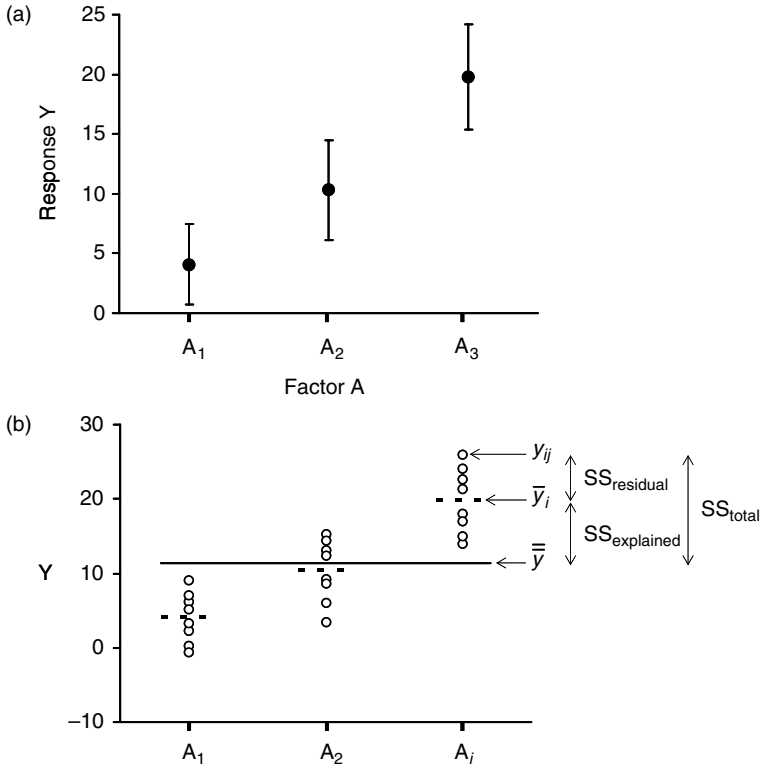

Figure 1 Dataset of three samples (a) summarised as group means and standard deviations, and (b) showing the $j = 8$ observations in each of the $i = 3$ groups. Total variation in the dataset, measured by the sum of squared deviations of each observation (y_{ij}) from the grand mean ($\bar{\bar{y}}$), is partitioned into an explained component that measures variation among the group means ($\bar{y}_i$), and an unexplained or residual component that measures variation among the data points within each group. The deviations indicated for the mean of group i and its j-th data point are summed across all data to obtain the model sums of squares.

information required to measure the component of variation, subtracted from the total number of pieces contributing to that variation. The total variation always has degrees of freedom that equal one less than the total number of data points, because it uses just the grand mean to calculate variation among all the data points. A one-factor model with n observations in each of a groups has $a - 1$ d.f. for the explained component of variation, because we require one grand mean to measure between-group variation among the a means; it has $na - a = (n - 1)a$ d.f. for the residual component, because we require a group means to measure

Table 1 *Generalised ANOVA table for testing a categorical factor, showing explained and residual (unexplained) sums of squares (SS), degrees of freedom (d.f.) and mean squares (MS),* F*-ratio and associated* P*-value. Subscript* i *refers to the* i*th group, and* j *to the* j*th observation in that group.*

Component of variation	SS	d.f.	MS	*F*-ratio	*P*
Explained	$\sum_{i=1}^{a} n \cdot (\bar{y}_i - \bar{\bar{y}})^2$	$a-1$	$SS_{expl}/d.f._{expl}$	MS_{expl}/MS_{res}	<0.05?
Residual	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$	$(n-1)a$	$SS_{res}/d.f._{res}$		
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2$	$na-1$			

within-group variation among all na data points. These explained and residual degrees of freedom sum to the $na-1$ total d.f.

Dividing each SS by its d.f. gives each component a mean square (MS) which is a measure of the variation per degree of freedom explained by that source. The explained component of variation is judged to contribute significantly to total variation in the response if it has a high ratio of its MS to the MS for the unexplained residual variance. This ratio is the estimated F-value from the continuous probability distribution of the random variable F. The F distribution for the given explained and residual d.f. is used to determine the probability P of obtaining at least as large a value of the observed ratio of sample variances, given a true ratio between variances equal to unity. Researchers in the life sciences often consider a probability of $\alpha = 0.05$ to be an acceptably safe threshold for rejecting the null hypothesis of insignificant explained variation. An effect is then considered significant if its F-value has an associated $P < 0.05$ (Table 1), indicating a less than 5% probability of making a mistake by rejecting a true null hypothesis of no effect (the Type I error rate). This is reported by writing $F_{a-1,(n-1)a} = \#.\#\#$, $P < 0.05$, where the subscript '$a-1$, $(n-1)a$' are the numbers of test and error d.f. respectively. Every F-value must always be reported with these two sets of d.f. (further detailed on page 260) because they provide information about the amount of replication, and therefore the power of the test to detect patterns.

The validity of the ANOVA test depends on three assumptions about the residual variance: that the random variation around sample means has the same magnitude at all levels of the factor, that the residuals contributing to this variation are free to vary independently of each