Chemometrics and Intelligent Laboratory Systems, 6 (1989) 259–272
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

Analysis of Variance (ANOVA)

LARS STÅHLE *

Department of Pharmacology, Karolinska Institutet, Box 60400, S-10401 Stockholm (Sweden)

SVANTE WOLD

Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, S-90187 Umeå (Sweden)

(Received 16 January 1989; accepted 27 July 1989)

CONTENTS

0169-7439/89/$03.50

ABSTRACT

Ståhle, L. and Wold, S., 1989. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, 6: 259–272.

Univariate ANOVA is reviewed from a user point-of-view with emphasis on understanding the model building and the assumptions underlying the method. Illustrative examples are taken from organic chemistry and analytical chemistry. The use of graphical techniques to visualize the ANOVA model as well as to analyse residuals is recommended. The main models of ANOVA are developed in some detail including one-factor ANOVA, crossed designs, nested designs, repeated measures ANOVA and variance components estimation. Hypothesis testing by *F*-tests and follow up by pairwise comparison methods is shown. The distinction between random effects and fixed effects is explained. Methods to handle non-linearities by transformations or by using response surface methodology are mentioned. Throughout the paper the importance of experimental design is emphasized. References are given to ANOVA methods for more complicated models.

1 INTRODUCTION

1.1 General

ANOVA (Analysis of Variance) is probably the most widely used statistical method for hypothesis testing currently in use. It encompasses a rich field and has the flexibility to cover a larger number of experimental designs [1–3]. Almost every introductory textbook in statistics contains one or more chapters on ANOVA (e.g. ref. 4). Nevertheless, many scientific papers employ ANOVA in a suboptimal or erroneous fashion. Our motive in writing this tutorial is to provide an intuitive understanding of some of the most essential features of ANOVA for practical purposes. We will focus upon properties of the method which we think should have a major influence upon the scientist's choice of statistical analysis while other aspects are left to be discovered in textbooks in mathematical statistics [5,6]. It is our belief that the reader most likely to benefit from this tutorial is the chemist who is willing to take the trouble to understand the principles behind ANOVA. Consequently, the formulae presented are given in a form aimed at facilitating an understanding of the subject rather than being given in an optimal form for computation, which is clearly unnecessary in this age of computers.

Four examples are given that represent typical applications of ANOVA. In the first example three catalysts are compared with respect to their effect on the purity of the product of a chemical reaction and the data are analysed by one-factor ANOVA. In the second example the influence of three catalysts on the reaction with four different reagents is analyzed by crossed two-factor ANOVA for replicated runs. In the third example three methods to measure prothrombin activity in plasma are compared by means of ANOVA of repeated measures. In the final example ANOVA is used to estimate the percentage contribution from two independent sources to the total variance in an assay for lead.

1.2 Terminology

It is necessary to define the meaning of some words used in the jargon of ANOVA. Investigations analyzed by ANOVA usually aim at an assessment of the effect of various factors (catalyst, solvent, procedures etc.) on some response (percentage purity, retention time in HPLC etc). There are two distinct types of factor effects: random effects and fixed effects. The random effect is the variability caused by, for instance, the batch of a chemical or by the individual column in HPLC. If it is found that HPLC columns have a considerable variation, much more for instance, than the loop injection device, than it seems reasonable to try to reduce the between-column variability. In such an investigation little attention is paid to each individual column. It is the effect of a randomly chosen column which is the concern of the chemist; hence it is a random effect. The

fourth example (Section 5, variance components) is a random effect experiment.

An example of the circumstances under which the fixed factor effect is studied is when comparing solvents for an organic synthesis. It is assumed that the experimenter has complete control over the properties of the factors in the sense that ethanol will always have the same effect on a given reaction and that this effect always will differ in the same way (under the same conditions) from the effect of another solvent. Other fixed effects may be pH, temperature, catalyst, etc. Examples 1 (Section 2.1) and 2 (Section 3.4) illustrate fixed effect experiments.

Factor effects may be studied in various combinations. Depending upon the arrangement of an investigation there is a distinction to be made between nested and crossed designs. When one level of the experimental design is subordinate to the next level we have a nested design. The most common example is measurements on samples that are taken in triplicate. Here the three measurements are subordinate to the sample. Furthermore, the first measurement has no more in common with the first measurement from another sample than with the second and third measurements from the other sample. The fourth example (Section 5) is an example of a nested arrangement.

Specific treatments are combined in the crossed design. For example solvents A, B and C can be combined with catalysts 1 and 2 giving the combinations A1, A2, B1, B2, C1, C2. Hence, all combinations are tested and each has a specific meaning and the catalysts are treated on the same level as the solvents.

Crossed and nested designs can be combined. The same is true for fixed and random effect ANOVA models, which are referred to as mixed models. There is also an alternative terminology for the classification of effects, models of type I are fixed models and type II models are random effect models.

2 ONE-FACTOR ANOVA

2.1 A common-sense development

Consider the simulated data in Table 1 in which 3 catalysts (A, B and C) are compared with respect to their ability to increase the purity of the yield from a chemical reaction. In the ANOVA jargon this is a fixed effect experiment (the distinction between nested and crossed models is unnessecary for one-factor ANOVA). Catalyst A was used in four experiments, B was employed in five experiments while C was used in four experiments. The question that the chemist obviously has in mind is "are there any differences between A, B and C, and, if there are any differences, which of the catalysts is best?". From intuition we may reason that if the difference between the mean purity from the experiments using the respective catalysts is large compared to the differences within the experiments using the same catalyst, then there is probably a true difference between the effects of the catalysts on the purity (see also Fig. 1).

TABLE 1

Simulation data

Percentage purity of yield from a chemical reaction using three different catalysts. The lower table is the so-called ANOVA table containing the sum of squares (*SSQ*), the degrees of freedom (*df*) and appropriate *F*-tests. In our table we have omitted total *SSQ* and mean squares which are occasionally included by others.

	A	B	C
	91	85	96
	95	88	98
	92	87	97
	90	86	97
		89	

ANOVA table

	SSQ	*df*	*F*
Catalyst	150	2	7.6
Residual	99	10	

Thus, what we need are measures of the between and within group dispersions. An obvious starting point is to take differences between the individual value and the mean in question. Using the notation in Table 2, one choice is:

$$x_{ij} - x_{.j} \tag{1}$$

However, it turns out that the sum of all differences from the mean in one group becomes 0

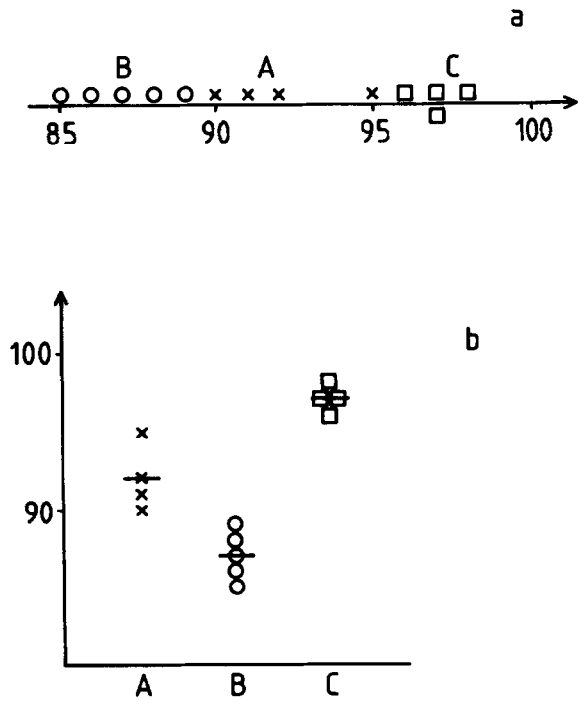

Fig. 1. Plot of the data in Table 1 in two ways; (a) along an x-axis and (b) in a 'bi'-variate plot with groups as discrete points on the horizontal axis.

(which follows from the definition). Hence a better choice would be:

$$|x_{ij} - x_{.j}| \tag{2}$$

Absolute values are unfortunately not easy to treat mathematically. Instead their squares turn out to be more natural. Because variances of independent experiments are additive, the squared difference gives a practical measure of dispersion in ANOVA. Thus what we compare are the squared differences between groups (in our example the mean purity using the jth catalyst minus the mean total purity)

$$(x_{.j} - x_{..})^2 \tag{3}$$

and within groups (the ith experiment using the jth catalyst minus the mean purity using the jth catalyst)

$$(x_{ij} - x_{.j})^2 \tag{4}$$

If we take the sum of all squared differences we get

$$SSQ_w = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2 \tag{5}$$

and

$$SSQ_b = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{.j} - x_{..})^2 = \sum_{j=1}^{p} n_j (x_{.j} - x_{..})^2 \tag{6}$$

The sum of squares SSQ_b and SSQ_w are not directly comparable but they should be of the same size if we take their mean values i.e. take SSQ_b/p (where p is the number of groups) and $SSQ_w/\Sigma n_j$. However, statisticians have found that these quantities are biased (the concept of bias is non-trivial, but it will not be dealt with here see e.g. ref. 6). Instead the divisors should be

$$p - 1 \tag{7}$$

which is the number of degrees of freedom between groups and

$$\sum_{j=1}^{p} (n_j - 1) \tag{8}$$

which is the number of degrees of freedom within the groups. The estimate of the variance of a population from a sample with mean $x_{.}$ is Because the variance of the sum of independent

TABLE 2

Notation used in the equations in the text

x_{ij}	ith observation in the jth group
x_{ijk}	ith observation in the jth group in the kth block
n, n_j	number of objects in the jth group
n_{jk}	number of objects in the jth group in the kth block
p, p_a	number of groups
p_b	number of blocks
$x_{.j}$	mean value in the jth group ($\Sigma_{i=1}^{n} x_{ij}/n_j$)
$x_{..}$, $x_{...}$	total mean value
$x_{.jk}$	mean value in the jth group in the kth block
$x_{.j.}$	mean value of the jth group over all blocks
$x_{..k}$	mean value of the kth block over all groups
μ	population mean
α, β	population factor effects
δ	population interaction effect

samples is the sum of their variances, the mean has the variance

$$\sum_{i=1}^{n} (x_i - x_.)^2/(n-1) \tag{10}$$

So the relation between the variance of the individual observations and the variance of the mean is

$$\sigma^2 = n\sigma_\mu^2 \tag{11}$$

therefore (5) divided by (8) is an estimate of the residual variance and (6) divided by (7) is another, independent, estimate of the same variance under the null hypothesis. The ratio (6) to (7) is not an estimate of the residual variance when the null hypothesis is false. This is the reason why these ratios are often referred to as 'mean squares'. We are now in a position to devise the ANOVA F-test which is constructed so that a value of the F-statistic that is sufficiently larger than 1.0 indicates a difference between the groups. This is accomplished by the following formula for F

$$F = \frac{\sum_{j=1}^{p} n_j(x_{.j} - x_{..})^2/(p-1)}{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - x_{.j})^2 \Big/ \sum_{j=1}^{p}(n_j - 1)} \tag{12}$$

which will be treated below in Section 2.5. We will now give a slightly different development of one-factor ANOVA based on a geometrical viewpoint.

2.2 A geometrical view of ANOVA

In Fig. 1 the data of Table 1 are represented as points along a real line (a) and as point scatters in a X–Y plot with a discrete X-axis (b). Returning to the common-sense view we may reformulate the statements of Section 2.1 in terms of distances along the real line. Thus, if the distance between the points in Fig. 1a within each group is small compared to the distances between the mean values, then we feel that the catalysts probably are truly different. Fig. 2 shows examples in which the groups are clearly different (a) or are clearly not different (b) or where intuition apparently fails due to the intermediate character of the data (c).

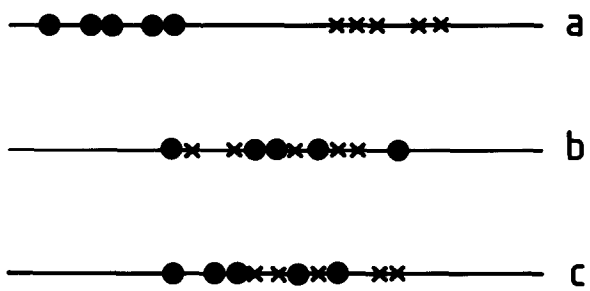

Fig. 2. Simulation data demonstrating (a) an obvious difference between two groups, (b) a lack of difference and (c) a situation requiring a statistical method (e.g. the t-test).

We will frequently return to this geometrical point of view in order to gain intuitive support for mathematical ideas.

2.3 The mathematical model

Any statistical statement is a statement of the scientist's belief that there is one mathematical model that best fits the data. We must consequently consider the models of ANOVA. Firstly, assume that there is no difference at all between the groups; that is, that the group means are really the same and any observed difference is a chance event. An accurate model would then be

$$x_{ij} = \mu + e_{ij} \tag{13}$$

where μ is the population mean and e_{ij} is the statistical error (caused by measurement error etc.). If, on the other hand, it is assumed that there is a difference between the groups then the model should be

$$x_{ij} = \mu_j + e_{ij} \tag{14}$$

Another way of expressing the same thing is to take the difference

$$\alpha_j = \mu_j - \mu \tag{15}$$

as the effect of the jth treatment and (14) can be rewritten

$$x_{ij} = \mu + \alpha_j + e_{ij} \tag{16}$$

which has a geometrical interpretation that can easily be illustrated (Fig. 3).

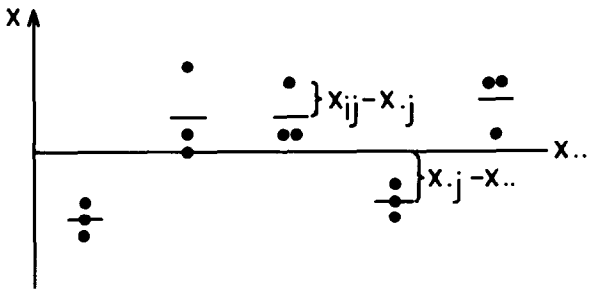

Fig. 3. Illustration of the decomposition in ANOVA of the total variation $(x_{ij} - x_{..})$ into variation within groups $(x_{ij} - x_{.j})$ and variation between groups $(x_{.j} - x_{..})$. The horizontal line is the total mean $(x_{..})$.

Because the model consists of 'effects' that are additive, one with the other, ANOVA is referred to as an additive model.

In order to choose between the models (13) and (16) we compare the residual error in (13) and (16)

$$RSS = \sum_{j=1}^{p} \sum_{i=1}^{n_j} e_{ij}^2 \tag{17}$$

To accomplish this $SSQ_b/(p-1)$ is used as an estimate of the residual variance in (13) and $SSQ_w/\Sigma(n_j - 1)$ for the residual variance of (16). The more detailed model (16) can be viewed as a decomposition of the total variance into two components, one being due to the residual error σ^2 and the other component being due to the difference between the means σ_μ^2. The decomposition is obtained as follows:

$$x_{ij} - x_{..} = (x_{ij} - x_{.j}) + (x_{.j} - x_{..}) \tag{18}$$

taking the squares of both sides and noticing that the cross product between the parentheses on the right-hand side of (18) vanish we get

$$\sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - x_{..})^2 = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2 + \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{.j} - x_{..})^2 \tag{19}$$

The component σ_μ^2 is not estimated directly. Instead $SSQ_b/(p-1)$ is an estimate of the sum $\sigma^2 + n\sigma_\mu^2$ (provided that $n = n_i = n_j$ for all i, $j = 1 \ldots p$; i this is not true the second term looks slightly different). The F-test is thus an estimate of the ratio

$$\left(\sigma^2 + n\sigma_\mu^2\right)/\sigma^2 \tag{20}$$

hence, a value of F sufficiently exceeding unity indicates that σ_μ^2 is non-zero.

2.4 ANOVA as a regression model

An alternative way to formulate the ANOVA model is to regard each treatment as a quantitative variable that is coded as 1 if it is present and 0 if absent. Thus, for each object a treatment (design) vector (denoted z) can be constructed such that objects given the first treatment get (1,0,...,0), the objects given the second treatment get (0,1,...0) i.e. $z_i = 1$ for objects given the ith treatment and $z_j = 0$ for all $j \neq i$. With p treatments and one response variable the data may be regarded as points in a $(p+1)$-dimensional space. This reformulation of the one-factor ANOVA can be handled by multiple regression (MR) and we consequently have the following mathematical model:

$$x_{ij} = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_p z_{ip} + e_i \tag{21}$$

which, in essence, is eq. (14). For example, an object given the third treatment will have $z_{i3} = 1$ and $z_{ij} = 0$ for all $j \neq 3$ which gives $x_{i3} = b_3 + e_i$ (in (14) this is $x_{ij} = \mu_j + e_{ij}$). There are variants of (21) in which the elements of z are chosen differently [7]. The details of the computations will not be considered here; the reader is referred to more complete texts on the subject [3,7]. The importance of fact that ANOVA can be treated as a regression model lies in the insight that ANOVA is a model of the data the pros and cons of which should be considered.

2.5 The F-test

The results of the F-test (12) is a positive number which, if it exceeds 1.0, indicates that there are differences between the means (vide supra). In tables of the F-distribution one reads off the critical value of F at a given p-value (usually 0.05) for the degrees of freedom $(p-1)$

and $\Sigma(n_j - 1)$. It should be noted that the F-test is one-sided since we are only testing for the possibility that the differences between the means is disproportionately large compared to the variation within the groups. The F-test may be used to compare any pair of variances and it may therefore be useful for many situations other than straightforward ANOVA.

2.6 Interpretation and further analysis

Given a significant F-test the data are usually not left as they are unless the sole interest of the scientist is to discover only whether the groups are different. In experimental areas of research, such as chemistry, one usually has formulated some prior sub-hypotheses about group differences that should be tested, or one may wish to test all paired differences. A short description will be given of three techniques which cover most of the situations that occur in practice.

Firstly, there is a method for pairwise comparison of all groups called Scheffe's test. Although this sounds an attractive alternative, the reader is warned that the sensitivity (i.e. power) of this method decreases dramatically with an increasing number of groups. The idea behind the test is to exploit the fact that a pooled estimate of the residual variance is available from the ANOVA. Using this fact and the F-distribution any pair of means (jth and kth group) may be compared according to the formula

$$\sqrt{F} = (x_{.j} - x_{.k})/s\sqrt{(p-1)(1/n_j + 1/n_k)} \quad (22)$$

where s is the standard deviation

$$s = \sqrt{SSQ_w \sum_{j=1}^{p} (n_j - 1)} \quad (23)$$

and F has $(p-1)$ and $\Sigma(n_j - 1)$ degrees of freedom.

The second method, Dunnett's test, is designed for the situation where one of the groups constitutes a control group (the jth group) and the other groups are treatment groups (the kth group). In this situation one calculates the quantity

$$t_{p,d,\alpha/2} = (x_{.j} - x_{.k})/s\sqrt{1/n_j + 1/n_k} \quad (24)$$

which may be regarded as an adjusted t-value. A detailed table is published by Dunnett (see ref. 1) for the adjusted t-values for p groups, d degrees of freedom in the residuals at the significance level α (not to be confused with the mean effect α).

The third method is due to Tukey and uses the test statistic q which has the studentized range distribution. The test is exact only for equal sample sizes, but is approximate for small deviations. The formula for q is

$$q_{p,d,\alpha/2} = (x_{.j} - x_{.k})\sqrt{2}/s\sqrt{1/n_j + 1/n_k} \quad (25)$$

There are a number of multiple comparison procedures for the cases of unequal variances and unequal sample sizes that may also be useful [8].

Finally a word on non-significant ANOVA. Firstly, this constitutes a lack of evidence for rejecting the null hypothesis that there is 'no difference between the groups'; it does not provide any evidence that the null hypothesis is true. The nonsignificance may well be due to the small number of objects used or that the assumptions underlying ANOVA have not been fulfilled (see Section 5). The best way to avoid misinterpreting an ANOVA is probably to plot the data in all possible useful ways.

3 CROSSED ANOVA

3.1 Extending one-factor ANOVA

As has been pointed out in Section 2.3, a one-factor ANOVA may be regarded as a decomposition of the data into various sources of variance (i.e. variance due to the factor and the residual variance). But let us introduce a second source of variance due to a second factor. The model is then a two-factor crossed ANOVA. Denote by p_a the number of different treatments for the first factor and p_b the number of treatments for the second factor. In the crossed arrangement all possible combinations of treatments from the two factors are tested giving $p_a p_b$ factor combinations in all. This is called a $p_a \times p_b$ factorial design (see Section 7). An example from quantitative struc-

ture–activity studies (QSAR) is the synthesis of drug molecules having two substituent sites and substituents H, CH_3 and CH_2CH_3. In this case there will be 3×3 molecules to synthesize and the pharmacological activity of each of must be measured. It should be noted that the possibility exists that the effect of a substituent on one site is dependent on the substituent at the other site; i.e. the sites interact.

A second example is a comparison between the pain relief following administration of three different pharmaceutical preparations of aspirin where both males and females have participated as experimental subjects. It may well be that that sex is a source of variance that tends to blur the result. If, however, we could subtract the variance due to sex we may be able to improve our estimate of the effect of pharmaceutical preparation. The effect of sex is then referred to as a blocking effect. To be able to do this it it necessary that neither of the sexes reacts particularly to one of the preparations.

Before developing the mathematical model and specific uses of two-factor ANOVA the concept of interaction must be developed further.

3.2 The interaction concept

By an interaction we mean that the effect of one factor is dependent upon the another factor. A very simple example is that treatment with the milk production-stimulating hormone prolactin is likely to result in milk production in females but not in men. Thus the factor (prolactin treatment) interacts with the sex of the subject. Another example is the effect of a catalyst on a chemical reaction which may be dependent on the temperature; i.e. the effectiveness of the catalyst has a relationship with the reaction temperature. These are the so-called second order interactions. The last example is easily extended to higher order interactions; e.g. by adding the factors pH and reaction time we may have four-factor interactions by which we mean an effect which is the unique result of combining all four factors. An illustration of a two-factor interaction is given in Fig. 4.

3.3 The mathematical model

In the same way as with the one-factor ANOVA we write the crossed two-factor ANOVA model without interactions with one additive component for each effect

$$x_{ijk} = \mu + \alpha_j + \beta_k + e_{ijk} \tag{26}$$

In this model we assume that the effects α and β are additive: i.e. the effect of the jth treatment of the first factor is the same for all k treatments on the second factor. In the presence of interactions we introduce the alternative model

$$x_{ijk} = \mu + \alpha_j + \beta_k + \delta_{jk} + e_{ijk} \tag{27}$$

An important use of crossed two-factor ANOVA is the discovery of interactions. By fitting the two models (26) and (27) their residual errors can be compared by an F-test. In the following we consider the case where all groups are of the same size since the formulae for unequal sample sizes are complicated and there is an abundance of com-

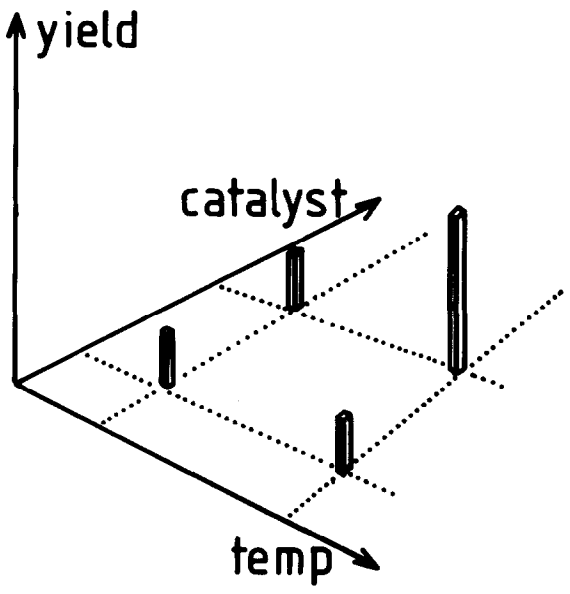

Fig. 4. Example of a two-factor interaction. The yield cannot be improved by increasing only the temperature or only the amount of catalyst. Both must be increased to improve the yield. Hence, the effects of the amount of catalyst and the temperature are dependent on one another.

puter programs available. For the interaction test we do, however, provide an approximate formula. See ref. 2 for details on the exact procedures.

When the sample sizes are equal, the residual variance is

$$\sum_{j=1}^{p_a}\sum_{k=1}^{p_b}\sum_{i=1}^{n}(x_{ijk}-x_{\cdot jk})^2/p_a p_b(n-1) \qquad (28)$$

and an independent estimate of the same variance under the null hypothesis (no interaction) is

$$n\sum_{j=1}^{p_a}\sum_{k=1}^{p_b}(x_{\cdot jk}-x_{\cdot j\cdot}-x_{\cdot\cdot k}+x_{\cdots})^2 /(p_a-1)(p_b-1) \qquad (29)$$

In the case of slightly unequal sample sizes the within group variance is estimated as

$$\sum_{j=1}^{p_a}\sum_{k=1}^{p_b}\sum_{i=1}^{n_{ab}}(x_{ijk}-x_{\cdot jk})^2 \Bigg/ \sum_{j=1}^{p_a}\sum_{k=1}^{p_b}(n_{jk}-1) \qquad (30)$$

while the variance due to interaction plus error uses the harmonic mean

$$1/n_h = p_a p_b(1/n_{11}+1/n_{12}+\cdots+1/n_{jk}+\cdots) \qquad (31)$$

in place of the group size n in (29).

3.4 F-tests in two-factor ANOVA

The ratio of (29) to (28) is an F-test with $(p_a-1)(p_b-1)$ and $p_a p_b(n-1)$ degrees of freedom of the presence of an interaction. This is a necessary test in crossed ANOVA since its significance complicates the continuation of the ANOVA.

With no significant interaction, the model (26) is sufficient and the hypotheses of significant factor effects may be investigated. In much the same way as in one-factor ANOVA we may take

$$p_b n\sum_{j=1}^{p_a}(x_{\cdot j\cdot}-x_{\cdots})^2/(p_a-1) \qquad (32)$$

as an independent estimate of the residual variance under the null hypothesis of no treatment effect of the first factor and

$$p_a n\sum_{k=1}^{p_b}(x_{\cdot\cdot k}-x_{\cdots})^2/(p_b-1) \qquad (33)$$

analogously for the second factor. An F-test with (p_a-1) and $p_a p_b(n-1)$ degrees of freedom for the first factor is (32) divided by (28) and similarly (33) divided by (28) is used to test for the second factor. Testing differences between individual means can be accomplished by the methods outlined in Section 2.6.

In the presence of a significant interaction, tests for treatment effects are difficult. Some authors hold that the interaction makes any interpretation of a factor effect in isolation meaningless because it is dependent on other factors. On the other hand, it may be argued that a significant factor effect can be interpreted locally. Our view is that isolated interpretation of factor effects are better avoided; the F-test does provide information, however, and may be computed. In any case

TABLE 3

Yield of a product using three catalysts and four reagents

Two independent experiments are made for each combination of catalyst and reagent. The crossed two-factor ANOVA table contains only one significant F-test which revealed an interaction. * denotes $p < 0.05$.

	Catalyst		
	1	2	3
Reagent A	4	11	5
	6	7	9
Reagent B	6	13	9
	4	15	7
Reagent C	13	15	13
	15	9	13
Reagent D	12	12	7
	12	14	9

	ANOVA table		
	SSQ	*df*	*F*
Catalyst	48	2	1.7
Reagent	120	3	2.9
Interaction	84	6	3.5 *
Residual	48	12	

it is best to plot the data to get an impression of what they really look like. It should be noted that (29) rather than (30) must be used as the denominator in the F-test for factor effects in the presence of an interaction since (29) measures both residual and interaction variance and the same is true for (32) and (33) under the null hypothesis of no factor effect.

Another, more conservative, way to handle a significant interaction is to proceed by analysing the data treatment-wise or to reformulate the problem to a one-factor ANOVA and use multiple comparisons to study differences between individual groups.

A numerical example of a crossed two-factor ANOVA is given in Table 3. This example illustrates the presence of an interaction.

3.5 Multifactor ANOVA

More complicated data in which combinations of three or more factors are studied can also be analyzed with ANOVA. For such data there will be one two-factor interaction for each pair of factor (3 in a three-factor ANOVA, 6 in a four-factor ANOVA etc.), one three-factor interaction for each triplet of factors etc. The computation of these can be found in ref. 3. Since already the crossed two-factor ANOVA was somewhat difficult in the presence of an interaction, it is easy to see that the risk of complications rise rapidly with the number of factor. Another problem is the many computer programs take only data with n equal in all treatment combinations, a demand which is not easy to fulfill in practice. For data of this kind it may be better to use a regression method (see Section 7.2).

4 REPEATED MEASUREMENTS

A not uncommon form of experimental design is one in which one subject is given a number of different treatments, thus serving the role as its own control. In this case we have treatments as usual but regard the individual as a source of variance to be removed (e.g. by blocking, see Section 3.1). An example is given where a number of blood samples were used to compare a number of laboratory assays for prothrombin activity and each sample was analyzed by all tests. In order to compare the assays a repeated measure ANOVA was set up with the assays as one factor and the samples as the other (blocking) factor (Table 4). Notice that this design does not allow for an estimation of the interaction term which, hence, is confounded with the residual variance. Thus, the sum of the squares due to treatment is (with $p-1$ degrees of freedom)

TABLE 4

Prothrombin activity measured by three methods on eight samples

In this two-factor ANOVA of repeated measures it is not possible to independently estimate interaction and residual variance.

Method		
simpl. A	nycotest	SPA
109	89	93
19	15	19
59	46	55
110	81	93
95	82	103
70	64	75
90	71	83
76	67	73

ANOVA table

	SSQ	*df*	*F*
Method	840.3	2	13.9
Residual	422.2	14	

$$SSQ_T = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{.j} - x_{..})^2 \qquad (34)$$

The sum of squares due to the samples is (with $n-1$ degrees of freedom)

$$SSQ_S = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{i.} - x_{..})^2 \qquad (35)$$

The residual sum of squares is (with $(p-1)(n-1)$ degrees of freedom)

$$SSQ_R = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - x_{i.} - x_{.j} + x_{..})^2 \qquad (36)$$

The F-test of interest is the ratio (34) to (36), each divided by its degrees of freedom. Although this data set can be analyzed in this way, very similar data should not be analyzed by repeated measures ANOVA. This includes measurements taken over time or over different doses (concentrations, temperatures, etc.) for which regression methods combined with analysis of summary statistics may be a more efficient approach. In this connection it should also be pointed out that alternative methods to the additive ANOVA model are available. All these methods use a multiplicative decomposition of the two-way table in a way related to principal component analysis (PCA) [9–11]. It is interesting to notice that Fisher and MacKenzie recommend PCA before ANOVA, writing in their 1923 paper [9] that "the summation formula for combining the effects of variety (or potatoes) and manurial treatment is evidently quite unsuitable for the purpose. ... A far more natural assumption is that the yield should be the product of two factors, one depending on the variety and one on the manure."

5 VARIANCE COMPONENTS

Another application of ANOVA is the assessment of the size of independent sources of variance in a chemical system. It is very useful in the analysis of analytical procedures to determine where precision can be improved. We use the data given in ref. 12 to illustrate the procedure. An automated atomic absorption method for lead determination using the Delves Cup technique was analyzed by using 10 different cups and 9 replications (Table 5; a plot is given in Fig. 5) were run. Since the cups were selected at random it is a model II ANOVA and we have a nested arrangement. Now the total variance is

$$\text{var(tot)} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - x_{..})^2 \Big/ \left(\sum_{j=1}^{p} n_j - 1 \right) \quad (37)$$

and the residual variance is

$$\text{var(res)} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2 \Big/ \sum_{j=1}^{p} (n_j - 1) \quad (38)$$

implying, since the two components are assumed

TABLE 5

Determination of lead from a standard solution made in nine replications for each of ten randomly selected Delves Cups

The latter contribute 9.4% to the total variance while the residual variance was 90.6%. This is also reflected in a non-significant F-test.

	Cup No.									
	1	2	3	4	5	6	7	8	9	10
	3.104	3.126	3.084	3.060	3.196	3.120	2.886	2.982	3.252	3.099
	3.055	2.823	2.953	2.983	2.785	3.077	2.794	3.110	2.937	3.016
	2.908	2.758	2.896	2.940	2.902	2.926	2.719	2.933	2.933	3.020
	3.053	2.809	2.811	2.782	2.958	2.944	2.677	2.909	2.944	2.972
	2.893	2.667	2.915	2.844	2.935	3.031	2.752	2.984	2.781	3.036
	2.864	2.888	2.896	3.010	2.825	2.899	2.889	2.943	2.073	2.880
	2.919	2.831	2.823	2.857	2.863	2.928	3.034	2.736	2.944	3.013
	2.760	2.843	2.929	2.974	2.893	2.770	2.846	2.836	2.859	2.901
	2.992	3.019	3.031	2.952	2.890	2.880	2.850	2.855	2.975	2.971
Mean	2.950	2.863	2.925	2.934	2.916	2.953	2.827	2.921	2.966	2.990

ANOVA table

	SSQ	*df*	*F*
Cups	0.209	9	2.0
Residual	0.914	80	

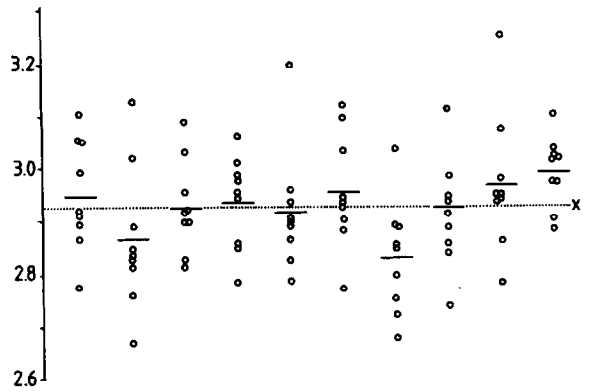

Fig. 5. Plot of the Delves Cup data used to illustrate variance components. The plot is in the same style as Fig. 1b and Fig. 3. The dotted line is the overall mean ($x_{..}$).

to be independent, that the variance due to the Delves Cups is

$$\text{var(Delves Cups)} = \text{var(tot)} - \text{var(res)} \qquad (39)$$

The actual figures are given in Table 5. Notice that the variance due to the Delves Cups cannot be calculated directly using

$$\sum_{j=1}^{p} (x_{.j} - x_{..})^2/(p-1)$$

because this expression contains both the variance due to the Delves Cups and the residual variance (see Section 2.3). The Delves Cup data is an example of a two-level variance component problem. An analysis at more levels is possible given that replications are made at each level and that there are good reasons to assume that the levels are independent. It should be noted that each level, except the highest must have model II ANOVA properties. Detailed computational procedures are given in ref. 3 for the more complicated designs.

The *F*-tests in variance component analysis is different due to the model II nature of the nested arrangement. The kind of *F*-tests used in crossed ANOVA cannot be used. In the nested ANOVA, each level has a total variance which is the sum of the variance due to this level and the variance from all subordinate levels. Thus, for a three (or more) level variance component problem the *F*-ratio is formed between adjacent levels.

6 ASSUMPTIONS

6.1 Normality

Hypothesis testing in ANOVA by the *F*-test is based on the assumption that the data are drawn from a population with a normal distribution, by which we mean that the term e_{ij} (or e_{ijk}) is a normal random variable with zero mean and some finite variance. Strictly speaking, any data not normally distributed cannot be analyzed by ANOVA. This is, however, too narrow-minded an approach. The fact that, by the central limit theorem, the mean is always more normally distributed than the individual objects, can be used as a motive for accepting moderate departures from normality. Nevertheless, non-normality may also result in loss of power and some transformation may therefore be warranted (see Section 5.3). A good rule of thumb is not to analyze data which are severely skewed in their distributions.

6.2 Homoscedasticity

An implicit assumption of ANOVA is that the residual variance is the same in all treatment groups. As a rule of thumb the largest and the smallest variance within groups should not differ by more than one order of magnitude. The reason for assuming equal residual variance (homoscedasticity) is that the variances are pooled to give an estimate of the variance that can be used for all groups simultaneously. If this assumption is violated a serious loss of power may result. It is the authors' experience that slavish adherence to the formulae in the face of a violation homoscedasticity assumption is one of the most common mistakes in the use of ANOVA which often result in a tragic abuse of really good data. The cure may be a well-chosen transformation.

6.3 Transformations

Transformation of data is sometimes regarded with suspicion as a form of manipulation to improve the data. The contrary is often the truth and is far worse; the wrong statistics are used on the right data. In fact, a statistical method is

nothing but a mathematical model associated with an element of chance and therefore the best model available should be used to analyse the data. The most common model is when the standard deviation is proportional to the mean in which case the data are often approximately log-normally distributed. Such data are readily identified in plots like that shown in Fig. 6 in which the mean is plotted against the residuals. A funnel-like shape of the data indicates a transformable data set. This is so common that it has been suggested that this may constitute a distribution which occurs naturally just as often as the famous normal distribution [13]. In this situation it is assumed that the data follow a model like

$$x_{ij} = \mu\alpha_j e_{ij} \tag{40}$$

where e_{ij} is log-normally distributed. It is easily seen that this can be transformed to the usual additive model by taking logarithms

$$\ln(x_{ij}) = \ln(\mu) + \ln(\alpha_j) + \ln(e_{ij}) \tag{41}$$

which can be properly analysed by ANOVA. This is however NOT the case for data modelled by eq. (40).

Another, common, type of data are proportions (sometimes given as percentages; notice that percentages are not always proportions). Since the range is limited (0–1), the normal approximation may be misleading. The arcsine function can be used to solve the problem

$$\arcsin\sqrt{x_{ij}} \tag{42}$$

This transformation is not necessary for proportions in the range (0.3–0.7).

The Box–Cox transformations and other methods of wide applicability are discussed in refs. 1 and 3.

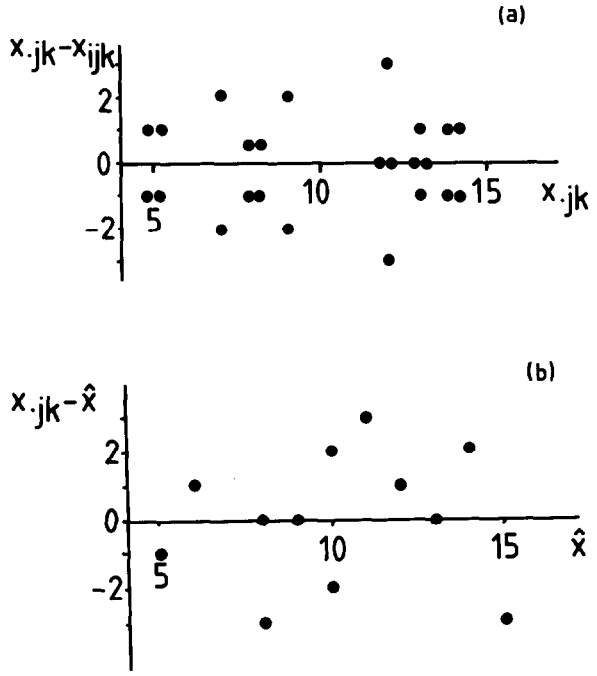

Fig. 6. Data from Table 3. (a) Plot of the mean in each group against the residual. There is no evidence that the residuals increase with the mean. (b) Plot of predicted group mean using the catalyst effect and the temperature effect, against the difference between observed and predicted means. Although the interaction is significant there are no indications of a curved relation.

7 EXPERIMENTAL DESIGN

7.1 Factorial designs

Experimental design is intimately connected with ANOVA (see e.g. ref. 1). ANOVA is in no way a cure for bad data and the 'garbage in–garbage out law' is always applicable. The one-factor ANOVA is the result of an uncomplicated design which, given the assumptions are fulfilled, is almost always valid. The two-factor ANOVA is more complicated. In particular when there is no possibility to discriminate interaction and residual variance it may easily happen that the analysis fails. Factorial designs can be used to design experiments so that interactions can be detected and estimated. An illustration of a three-factor complete factorial design is given in Fig. 7.

7.2 Response surfaces

When a combination of factorial designs and perhaps some additional experiments have been performed it may be useful not to use ANOVA but instead exploit the fact that the factors are

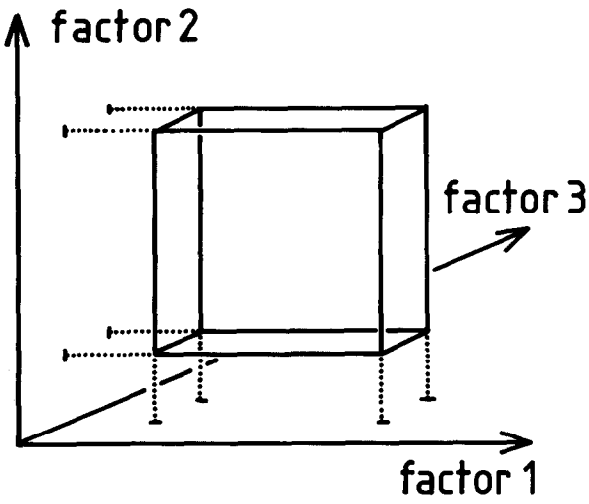

Fig. 7. Illustration of an experimental design with three factors. The design allows estimation of both 2- and 3-factor interactions.

quantitative variables. One well-known possibility is to calculate the formula for a polynomial surface by means of some stepwise multiple regression procedure [1]; alternatively, some of the newly developed methods such as partial least squares (PLS) analysis [14] may be used.

8 CONCLUSIONS

ANOVA is useful for analysing an array of experimental data in chemistry as long as one remembers the underlying assumptions and the linear model. Caution should be taken to scrutinize the data for the need for transformations. Designs where the interaction term can be estimated is preferable. Well designed experiments are the key to a successful use of ANOVA, as well as any other statistical analysis.

9 ACKNOWLEDGEMENTS

This study was supported by Svenska Läkarsällskapet, Karolinska Institutet, the Swedish Medical Research Council and the Swedish Natural Science Research Council. Dr Nils Egberg has generously allowed us to cite experimental data.

REFERENCES

1 G.E.P. Box, W.G. Hunter and J.S. Hunter, *Statistics for Experimenters,* Wiley, New York, 1978.
2 G.W. Snedecor and W.G. Cochran, *Statistical Methods*, Iowa State University Press, Ames, IA, 1967.
3 R.R. Sokal and F.J. Rohlf, *Biometry*, Freeman, New York, 1981.
4 T.H. Wonnacott and R.J. Wonnacott, *Introductory Statistics,* Wiley, New York, 1978.
5 R.J. Larsen and M.L. Marx, *An Introduction to Mathematical Statistics and Its Applications,* Prentice Hall, New York, 1986.
6 B.W. Lindgren, *Statistical Theory*, MacMillan, New York, 1976.
7 N. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 1981.
8 P.A. Games, H.J. Keselman and J.C. Rogan, Simultaneous multiple comparison procedures for means when sample sizes are unequal, *Psychology Bulletin,* 3 (1981) 594–598.
9 R.A. Fisher and W.A. MacKenzie, Studies in crop variation. II. The manurial response of different potato varieties, *Journal of Agricultural Science*, 13 (1923) 311–320.
10 V. Hegemann and D.E. Johnson, On analyzing two-way AoV data with interaction, *Technometrics*, 18 (1976) 273–281.
11 J. Mandel, A new analysis of variance model for non-additive data, *Technometrics*, 13 (1971) 1–18.
12 R.F. Hirsch, Analysis of variance in analytical chemistry, *Analytical Chemistry,* 49 (1977) 691A–700A.
13 S. Newcomb, Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics*, 4 (1881) 39–40.
14 S. Wold, A. Ruhe, H. Wold and W.J. Dunn III, The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses, *SIAM Journal of Statistics and Computation,* 5 (1984) 735–743.