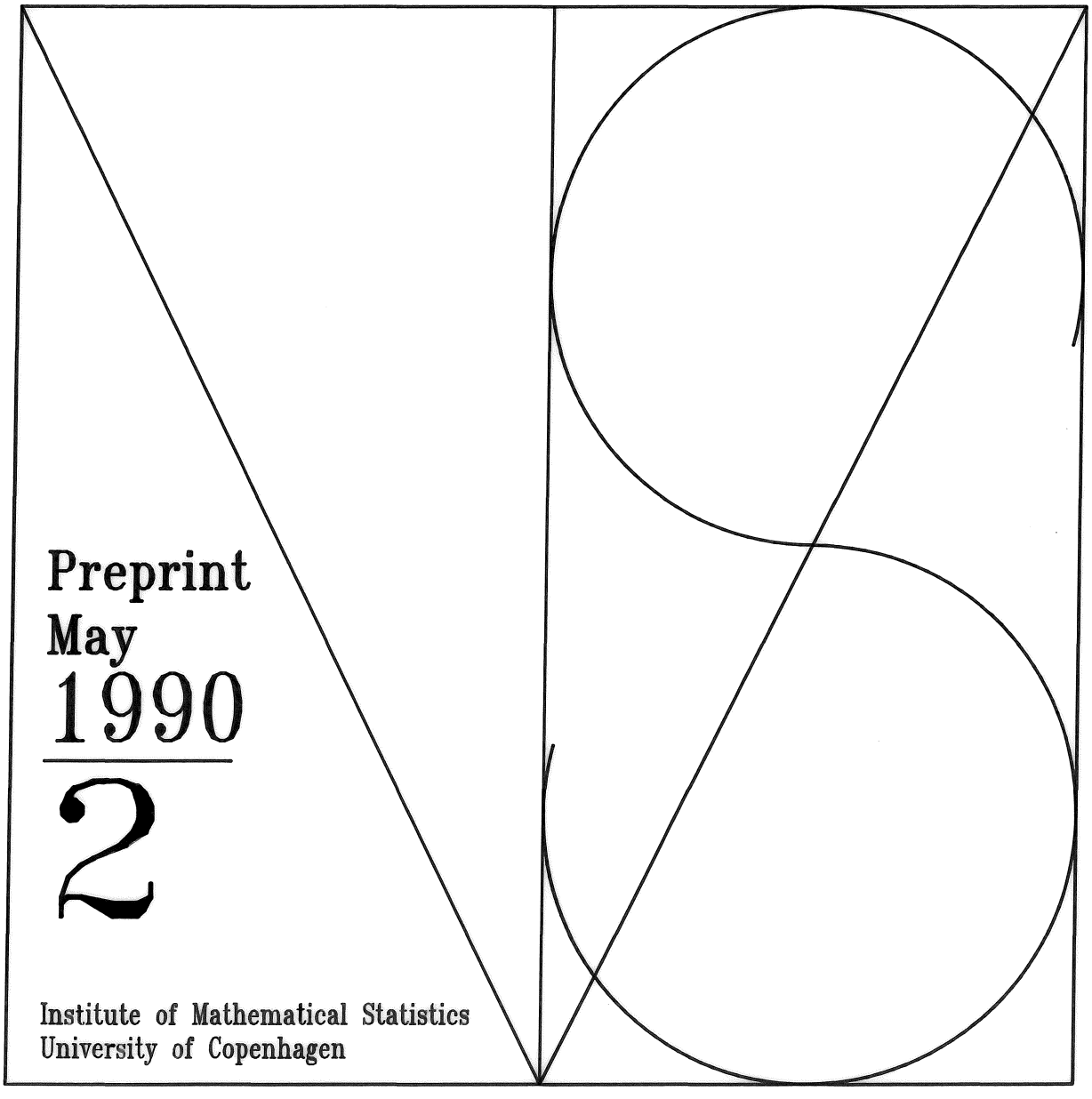
Tue Tjur

Analysis of Variance and Design of Experiments



Preprint
May
1990
2

Institute of Mathematical Statistics
University of Copenhagen

ISSN 0902-8846

Tue Tjur

ANALYSIS OF VARIANCE AND DESIGN OF EXPERIMENTS*

Preprint 1990 No. 2

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

May 1990

*) Lecture given at the 13th Nordic Conference on Statistics, June 1990.

Summary

This paper outlines a unified approach to analysis of variance and design of experiments in the orthogonal case, based on unambigouos mathematical definitions of elementary statistical concepts that are frequently left undefined or vaguely defined in the existing litterature. Particular emphasis is put on variance component models (mixed models) and the application of group theory to the construction of orthogonal designs (fractional replicates of complete factorial designs, single replicate designs with blockings).

American Mathematical Society 1980 subject classifications.

Primary 62S10; secondary 62K15.

Key words and phrases.

Analysis of variance, orthogonal design, variance component model, fractional factorial.

0. Introduction.

From a certain point of view, the contents of the phrase 'analysis of variance' (ANOVA) can be explained quite briefly as follows. The basic concept is the normal linear model, assuming that the observed data vector $\mathbf{y} = (y_i) \in \mathbf{R}^n$ is the outcome of a random variable from a multivariate normal distribution with mean vector μ in a prescribed linear subspace of $\mathbf{R}^n$ and covariance matrix of the form $\sigma^2\mathbf{I}$. The theory of estimation and hypothesis testing in the class of such models is wellknown and easy to expose in terms of concepts related to Euclidean geometry on $\mathbf{R}^n$ and linear algebra. ANOVA covers the theory of such models, mainly those for which the structure of the subspace of means can be described in terms *factors*, i.e. the partitionings of data into groups given by treatments, blocks etc. To this can be added the *variance component models* (or *mixed models*), which are models of the above mentioned type for which the effects of some of the factors are regarded as random. One may also include models with other distributional assumptions than normality, but such models are usually handled exactly as the normal models, with the reservation that distributions of test statistics and estimates are only approximate.

The phrase 'analysis of variance' has also a classical, more vague, meaning as a technique for decomposition of the total square sum $\sum(y_i - \bar{y})^2$ into components which can be ascribed to different sources of variation, like 'variation within groups', 'variation between rows', 'interaction between rows and columns' etc. This decomposition, which is usually summarized in an *analysis of variance table*, can be regarded as a purely descriptive tool.

However, both of the above attempts to define ANOVA are incomplete and not particularly exciting in themselves. What really blows life into the topic is the interplay between these two points of view. The intuitive interpretation of the square sums in the analysis of variance table finds its final justification in the way these square sums enter in expressions for F–test statistics in normal linear models and estimates for variance components in 'balanced' variance component models. Conversely, the analysis of linear models and variance component models is considerably simplified in the situations where all possible F–tests for model reductions can be computed from a single ANOVA–table. Very often, the most reasonable description of data among those in a large tree of possible models can be found almost immediately by a look at the ANOVA–table.

While the general theory of normal linear models is well explained in textbooks, ANOVA–tables and their relation to linear model theory are usually explained by examples, without any clear definition of what an ANOVA–table really is and when it can be constructed. It is the aim of the present paper to give an equally precise exposition of the latter

aspect of the topic. Obviously, this requires that seemingly trivial concepts like 'a factor', 'a two–way table' etc. are treated with the same mathematical care as usually given to concepts like 'a linear subspace', 'a matrix' etc.

What comes out of this is first of all (section 2) a characterization of the exclusive, but still very important, class of designs for which an ANOVA–table is defined, together with an unambigouos set of rules for how the ANOVA–table should be constructed and (section 3) how F–test statistics etc. in linear models should be computed from it. To this comes (section 4) a similar – but more complicated – set of rules for how to handle variance component models in these 'balanced' situations. Finally (section 5) a brief overview is given of the most important method for construction of such 'orthogonal designs', apart from the balanced k–way tables.

None of what is said in this paper is new, in any strict sense. Sections 1–4 are essentially contained in Tjur (1984), which, in turn, was a further development of other expositions, in particular Jensen (1979, a similar treatment for balanced k–way tables) and Nelder (1965). Section 5 summarizes a joint work with Christoffersen (1987), adapted to the general approach taken in Tjur (1984), but otherwise considerably overlapping with many other authors' works on group generated designs.

1. Mathematical tools: Factors and designs.

In the following, the vector of observations is denoted $\mathbf{y} = (y_i) = (y_i \mid i \in I)$, so that $\mathbf{R}^I$ is our observation space. The finite set I indexing the observations is referred to as the set of *experimental units*. A *factor* is formally defined as a mapping

$$\varphi_F \colon I \to F$$

from I to another finite set F of *factor levels* or *labels*. For convenience, we refer to a factor by the name assigned to its set of levels, e.g. 'the factor SEX = {Male, Female}' rather than 'the factor $\varphi_{\mathrm{SEX}} \colon I \to \mathrm{SEX} =$ {Male, Female}'. The representation of a factor as an assignment of levels to experimental units corresponds to the way factors are usually represented in a computer program. But for the statistical analysis, the essential property of a factor is given by the way it partitions experimental units into classes $\varphi_F^{-1}(f)$, $f \in F$. Factors that differ only by the labelling of these classes should usually not be distinguished, and levels that are not taken by at least one experimental unit can be ignored. In particular, this is subsumed when we write $|F|$ for the number of levels.

Factors are (partially) ordered in an obvious way by 'nestedness'. For example, in an experiment involving measurements on patients, the factor PATIENT (holding e.g. patient numbers) is *finer than* (or *nested in*)

the factor SEX (holding the sex of patients), or SEX is *coarser than* or *marginal to* PATIENT. We write

$$\text{SEX} \leq \text{PATIENT}$$

to indicate this relationship between the two factors. The general idea is that $F \leq G$ (F coarser than G) means that any G–class is contained in an F–class.

Two factors play a special role: The *trivial factor* 0, corresponding to a partitioning of I into a single class (or a constant mapping $\varphi_0 \colon I \to 0$, where 0 is some one–point set); and the *units factor* I, corresponding to the partitioning of I into single units (or the identity $\varphi_I \colon I \to I$). These are the two extremes in the ordering, in the sense that for any other factor F we have

$$0 \leq F \leq I.$$

For any two factors F and G, we can form their *product* $F \times G$, the cross classification by the two factors. Formally, this is again a factor, namely the mapping $\varphi_{F\times G}$ which to an experimental unit $i \in I$ assigns the level $\varphi_{F\times G}(i) = (\varphi_F(i), \varphi_G(i)) \in F \times G$. In terms of the ordering of factors, $F \times G$ can be characterized as the coarsest factor which is finer than both F and G. In this sense, we could write $F \times G = \sup\{F, G\}$ or $F \times G = F \vee G$.

An equally important – but usually less emphasized – concept comes out of the last characterization of the product $F \times G$ by reversal of the ordering. The *infimum* or *minimum* $F \wedge G$ of two factors F and G is defined as the finest factor which is coarser than both F and G. The classes in the partitioning corresponding to this factor are the minimal non–empty subsets of I which can be written as unions of classes both with respect to F and G.

EXAMPLE (the balanced k–way table). Suppose that I is given as a Cartesian product $I = F_1 \times \ldots \times F_k$, and let $\varphi_{F_j} : I \to F_j$, $j = 1, \ldots, k$, be the canonical projections. From the factors $F_1, \ldots, F_k$ (e.g. row and column factors for $k = 2$) we can form products like $F_1 \times F_2$, $F_1 \times F_3 \times F_4$ etc. There is a simple rule for the formation of infima of such factors, namely that the infimum is the product of the factors occurring in both operands, e.g.

$$(F_1 \times F_3 \times F_4) \wedge (F_2 \times F_3 \times F_4) = F_3 \times F_4.$$

In fact, this rule applies to k–way tables whether they are balanced or not, provided that all cell counts $n_{f_1 f_2 \ldots f_k}$ (= the number of observations on level $f_1 f_2 \ldots f_k$ of $F_1 \times F_2 \times \ldots \times F_k$) are positive. If some of these cell counts are zero, things may become more complicated. In a two–way table, for example, the rows factor R and the columns factor C may have the infimum $R \wedge C = 0$ in agreement with the rule above, in which case the two–way table is said to be *connected.* But in a table with many empty cells, $R \wedge C$ may be a factor with more than a single level.

The next topic to be studied is the linear structure on $\mathbf{R}^I$ imposed by one or several factors. We shall use the following notation. Vectors $\mathbf{y} = (y_i)$, $\mathbf{x} = (x_i)$, ... in $\mathbf{R}^I$ will be regarded as column vectors or $I \times 1$–matrices. The inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$(\mathbf{x}|\mathbf{y}) = \mathbf{x}^*\mathbf{y} = \sum_{i \in I} x_i y_i$$

and the (Euclidean) norm of $\mathbf{x}$ is

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}|\mathbf{x})} = \sqrt{\sum x_i^2}.$$

A factor F determines in a wellknown way an $|F|$–dimensional subspace of $\mathbf{R}^I$, consisting of the vectors which are constant on F–classes. This is the space of mean vectors in the standard one–way ANOVA model determined by F. This subspace is denoted L_F. By

$$\mathbf{X}_F \colon \mathbf{R}^F \to \mathbf{R}^I$$

we denote the standard parameterization of this subspace. As an $I \times F$–matrix, $\mathbf{X}_F$ is the *design matrix* determined by F, with elements

$$(\mathbf{X}_F)_{if} = \begin{cases} 1 \text{ if } \varphi_F(i) = f \\ 0 \text{ otherwise.} \end{cases}$$

By

$$\mathbf{P}_F \colon \mathbf{R}^I \to \mathbf{R}^I$$

we denote the orthogonal projection onto L_F. From estimation in a one–way ANOVA model it is wellknown that orthogonal projection of a vector $\mathbf{y}$ onto this subspace amounts to averaging over the F–classes. Thus, as an $I \times I$–matrix, $\mathbf{P}_F$ has elements

$$(\mathbf{P}_F)_{i_1 i_2} = \begin{cases} \frac{1}{n_f} & \text{if } \varphi_F(i_1) = \varphi_F(i_2) = f \\ 0 & \text{if } \varphi_F(i_1) \neq \varphi_F(i_2) \end{cases}$$

where, as usual, n_f denotes the number of experimental units on level f of F.

A factor F is called *balanced* if the counts n_f are equal. In this case we use the notation $n_F = n_f$ $(= |I|/|F|)$ for the common size of the F–classes. For a balanced factor F, we have the particularly simple expression

$$\mathbf{P}_F = \frac{1}{n_F} \mathbf{X}_F \mathbf{X}_F^*$$

for the orthogonal projection $\mathbf{P}_F$ in terms of the design matrix $\mathbf{X}_F$.

For two factors F and G, we have the rule

$$L_F \cap L_G = L_{F \wedge G}$$

which can be regarded as an alternative characterization of the infimum. This rule links the concept of an infimum to concepts like partial confounding, partial aliasing, overparameterization etc.

Orthogonality between two factors can be defined as follows. Two linear subpaces L_1 and L_2 of $\mathbf{R}^I$ are said to be *geometrically orthogonal* if any two vectors $\mathbf{v}_1 \in L_1 \cap (L_1 \cap L_2)^\perp$ and $\mathbf{v}_2 \in L_2 \cap (L_1 \cap L_2)^\perp$ are orthogonal (i.e. $(\mathbf{v}_1|\mathbf{v}_2) = 0$). Intuitively, this concept is wellknown from 3–dimensional geometry where two planes may be orthogonal in exactly this sense. An equivalent condition is that the corresponding orthogonal projections commute. Two factors F and G are said to be orthogonal if L_F and L_G are geometrically orthogonal, i.e.

$$\mathbf{P}_F\mathbf{P}_G = \mathbf{P}_G\mathbf{P}_F \quad (= \mathbf{P}_{F \wedge G}).$$

In terms of the cell counts of the corresponding $F \times G$ two–way table, this turns out to be equivalent to the condition that

$$n_{fg}n_h = n_f n_g$$

whenever $f \in F$ and $g \in G$ are levels such that the corresponding classes $\varphi_F^{-1}(f)$ and $\varphi_G^{-1}(g)$ are contained in the same $F \wedge G$–class $\varphi_{F \wedge G}^{-1}(h)$. Notice that for $F \wedge G = 0$, this is the usual condition of 'proportional cell counts', i.e. proportional rows (or columns) in the two–way table of cell counts. For $|F \wedge G| > 1$, the condition means that such a proportionality should hold in each of the subtables determined by the levels of $F \wedge G$.

A *design* can now be defined as a set $\mathcal{D}$ of selected factors. The idea is that $\mathcal{D}$ should consist of all factors of potential interest for the model building. For example, if data are arranged in an $R \times C$ two–way table, the factors R, C and $R \times C$ should be included in $\mathcal{D}$. To this should usually be added the factor 0 (corresponding to constant terms in models) and always the factor I (provided that $I \neq R \times C$, i.e. at least one $n_{rc} > 1$).

2. The ANOVA table.

We are now in the position to characterize a large class of designs for which an ANOVA table can be defined, i.e. a decomposition of the total square sum, holding the information required for hypothesis testing within the entire class of linear models that can be stated in terms of factors from the design. The crucial concepts turn out to be orthogonality and infima of factors.

DEFINITION. An *orthogonal design* is a set $\mathcal{D}$ of factors, satisfying the following three conditions.

($\mathcal{D}1$) $I \in \mathcal{D}$.
($\mathcal{D}2$) Any two factors in $\mathcal{D}$ are orthogonal.
($\mathcal{D}3$) $\mathcal{D}$ is closed under the formation of infima.

The justification of the name 'orthogonal' for a design satisfying these conditions is that ($\mathcal{D}2$) is the essential assumption here. If ($\mathcal{D}2$) is satisfied, the satisfaction of ($\mathcal{D}1$) and ($\mathcal{D}3$) is just a matter of adding I and the missing infima (which will not destroy the orthogonality).

The mathematical result behind the construction of the ANOVA table can now be stated as follows.

PROPOSITION 2.1. *Let $\mathcal{D}$ be an orthogonal design. Then there exists a unique decomposition*

$$\mathbf{R}^I = \bigoplus_{G \in \mathcal{D}} V_G$$

of the observation space as a direct sum of orthogonal components, one for each factor of the design, such that for each $F \in \mathcal{D}$

$$L_F = \bigoplus_{G \in \mathcal{D}, G \leq F} V_G.$$

The proof will not be given here, see Tjur (1984).

Now, let $\mathbf{Q}_G$ denote the orthogonal projection onto V_G. By the ANOVA table we mean the table

Effect	d.f.	SSD
⋮	⋮	⋮
G	$\dim V_G$	$\|\mathbf{Q}_G y\|^2$
⋮	⋮	⋮

which for each factor of the design gives the square sum of deviations $\mathrm{SSD}_G = \|\mathbf{Q}_G y\|^2$ and its degrees of freedom $d_G = \dim V_G$. The proposition gives not only the definition of the ANOVA table, but also a simple scheme for computation of SSD's and their degrees of freedom from the basic square sums

$$\mathrm{SS}_F = \|\mathbf{P}_F y\|^2 = \sum_{f \in F} n_f \bar{y}_f^2$$

(where $\bar{y}_f$ is the average of the observations on level f) and the numbers of levels

$$\dim L_F = |F|.$$

This can be done by solution of the equations

$$SS_F = \sum_{G \in \mathcal{D}, G \leq F} \mathrm{SSD}_G$$

and

$$|F| = \sum_{G \in \mathcal{D}, G \leq F} d_G$$

which are immediate consequences of proposition 2.1. Even in complicated situations where many factors are involved, it is usually easy to solve these equations recursively by means of a *factor structure diagram*, showing the ordering of factors in $\mathcal{D}$. We illustrate this by two examples.

EXAMPLE 2.1. Consider a two–way table with, say, $|R| = 4$ rows, $|C| = 5$ columns and 2 observations per cell, i.e. $|I| = 5 \times 4 \times 2 = 40$ units. As our design we take

$$\mathcal{D} = \{I, R \times C, R, C, 0\}.$$

The factor structure diagram (with coarser factors to the right, finer factors to the left, arrows indicating the canonical mappings between sets of factor levels) is

$$I_{20}^{40} \longrightarrow R \times C_{12}^{20} \begin{array}{l} \nearrow R_3^4 \\ \searrow C_4^5 \end{array} \begin{array}{l} \searrow \\ \nearrow \end{array} 0_1^1 .$$

The numbers occurring here as superscripts and subscripts indicate an easy way of performing the computations, illustrated here by computation of degrees of freedom d_G from numbers $|F|$ of factor levels, but equally applicable to computation of square sums of deviations SSD_G from 'simple square sums' SS_F. The superscripts, which should be filled in first, are numbers of levels for each factor. The subscripts are the degrees of freedom d_G, to be filled in recursively from right to left, according to the rule that each $|F|$ should be the sum of all d_G for factors G at or 'in front of' the factor F. In this case, the resulting ANOVA table is

Effect	d.f.	SSD
0	1	SS_0
R	3	$\mathrm{SS}_R - \mathrm{SS}_0$
C	4	$\mathrm{SS}_C - \mathrm{SS}_0$
$R \times C$	12	$\mathrm{SS}_{R\times C} - \mathrm{SS}_R - \mathrm{SS}_C + \mathrm{SS}_0$
I	20	$\mathrm{SS}_I - \mathrm{SS}_{R\times C}$

EXAMPLE 2.2 (design by Skovgaard and Kristensen, modified). In a field trial for comparison of three fungicides ($F = \{f_1, f_2, f_3\}$), six blocks of three plots of two subplots were treated as follows. The three fungicides were allocated to the three whole plots of each block. Two additional factors, nitrogen ($N = \{n_1, n_2\}$) and variety ($V = \{v_1, v_2\}$) on two levels were allocated to subplots in a cross–over arrangement, so that the combinations n_1v_1 and n_2v_2 occurred on half of the blocks, while the opposite combinations occurred on the other half. Thus, apart from randomization, the design might be like this:

Blocks 1, 2 and 3	$f_1n_1v_1$ $f_1n_2v_2$	$f_2n_1v_1$ $f_2n_2v_2$	$f_3n_1v_1$ $f_3n_2v_2$
Blocks 4, 5 and 6	$f_1n_1v_2$ $f_1n_2v_1$	$f_2n_1v_2$ $f_2n_2v_1$	$f_3n_1v_2$ $f_3n_2v_1$

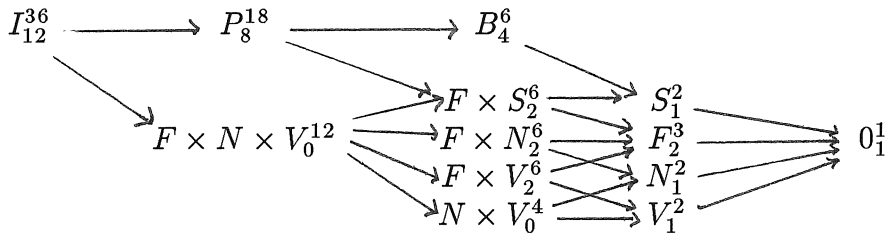
The factors to be taken into account are

0
I
$F = \{f_1, f_2, f_3\}$
$N = \{n_1, n_2\}$
$V = \{v_1, v_2\}$
$F \times N$
$F \times V$
$N \times V$
$F \times N \times V$
$B = \{1, \ldots, 6\}$ (block)
$P = \{1, \ldots, 18\}$ (plot)

It is a matter of straightforward hard work to check that these factors are orthogonal and to identify their infima. Orthogonality is present in all cases (in fact, this is a triviality, since the design is group generated, cfr. section 5), but some of the infima are related to an additional factor, which we may call S for 'superblock'. It is the factor on two levels, classifying observations according to the 'kind of block' they occur on, cfr. the cross–over arrangement. This factor must be included since it is the infimum of $N \times V$ and B. Also the factor $F \times S$ on six levels must be included, since it is the infimum of $F \times N \times V$ and P. After the addition of these two factors, the design satisfies ($\mathcal{D}1$)–($\mathcal{D}3$). Notice that the role of such 'pseudofactors' is purely formal. A main effect of S makes no sense in itself, but the structure of the design is such that a main effect of S can not be distinguished from an interaction between N and V. Similarly, an interaction between F and S turns out to be an alias for three factor interaction between F, N and V. This is easily

seen by comparison of the ANOVA table with the table one gets in the simpler design where the 'blocking factors' B and P are not included (which is merely a complete $F \times N \times V$–table with three observations per cell).

The factor structure diagram (with degrees of freedom) looks like this:



From this, the ANOVA table is eaily constructed. However, we postpone the construction to section 4, to add some features related to the analysis of a variance component model with random effects of the factors B and P.

Notice that two degrees of freedom ($d_{N\times V}$ and $d_{F\times N\times V}$) are zero. Of course, the corresponding square sums will also be zero, and these two lines of the ANOVA table can be omitted, if desired. The square sums that really measure $N \times V$–interaction and three factor interaction between F, N and V occur elsewhere in the table, as indicated above.

3. The linear model.

Let $\mathcal{D}$ be an orthogonal design. By a *model formula* we mean, simply, a subset $\mathcal{T}$ of $\mathcal{D}$. Following (to some extent) GENSTAT's conventions, we write the elements of $\mathcal{T}$ separated by plusses, so that, for example, $R + C$ means the subset $\{R, C\}$ of (say) $\mathcal{D} = \{I, R \times C, R, C, 0\}$.

A model formula determines a normal linear model in an obvious way. For example, the formula $R + C$ represents the usual additive model in a two–way table

$$y_i = \alpha_r + \beta_c + \sigma u_i,$$

where the u_i, $i \in I$, are i.i.d. N(0,1). The general idea is that $\mathcal{T} \subseteq \mathcal{D}$ represents a model of the form

$$y_i = \sum_{T\in\mathcal{T}} \alpha_t^T + \sigma u_i.$$

In such expressions we subsume, of course, that the factor levels on the right correspond to the unit i on the left, i.e. $t = \varphi_T(i)$.

In vector notation, we can say the same thing as follows. $\mathbf{y}$ is (the outcome of) an $|I|$–dimensional multivariate normal random variable with covariance matrix $\sigma^2\mathbf{I}$ and mean

$$\mu = \sum_{T\in\mathcal{T}} \mathbf{X}_T \alpha^T$$

in the subspace

$$L = \sum_{T \in \mathcal{T}} L_T.$$

Notice that different model formulas may correspond to different parameterizations of the same model. For example, the above mentioned additive model in a two–way table $\mathcal{D} = \{I, R \times C, R, C, 0\}$ can be written either as $R + C$ or as $0 + R + C$, since the addition of a constant term γ to the expression $\alpha_r + \beta_c$ merely creates some (further) overparameterization. Minimization of the number of such redundant terms and ties on specific parameters to obtain one–to–one parameterizations are of interest when the parameters are to be estimated. But since we are only going to discuss hypothesis testing (and estimation of the variance), no restrictions will be imposed on the degree of overparameterization or partial aliasing that may occur in a model formula. On the contrary, it turns out to be convenient to represent the model given by $\mathcal{T} \subseteq \mathcal{D}$ by its *maximal model formula*

$$\mathcal{T}^* = \{F \in \mathcal{D} \mid F \leq T \text{ for some } T \in \mathcal{T}\}.$$

This can be regarded as the model formula with the maximal amount of overparameterization. The advantage of this representation is that it gives the following simple rules for how to form residual sums of squares and F–test statistics from the ANOVA–table:

For a given model $\mathcal{T}$, the residual sum of squares is

$$\mathrm{SSD}_{\mathrm{res}}(\mathcal{T}) = \sum_{G \notin \mathcal{T}^*} \mathrm{SSD}_G,$$

and the degrees of freedom for the residual square sum is

$$d_{\mathrm{res}}(\mathcal{T}) = \sum_{G \notin \mathcal{T}^*} d_G.$$

The usual (restricted maximum likelihood, or maximum likelihood corrected for bias) estimate of the variance is

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_{\mathrm{res}}(\mathcal{T})}{d_{\mathrm{res}}(\mathcal{T})}.$$

In words, the residual sum of squares is computed by summation of all SSD's corresponding to factors that are not in the maximal model formula.

For a submodel, given by a model formula $\mathcal{T}_0$ such that $\mathcal{T}_0^* \subseteq \mathcal{T}^*$, The F–test for this simplification of mean structure can be written as

$$F(d_{\mathrm{dif}}, d_{\mathrm{res}}(\mathcal{T})) = \frac{\mathrm{SSD}_{\mathrm{dif}}/d_{\mathrm{dif}}}{\hat{\sigma}^2}$$

where

$$\mathrm{SSD}_{\mathrm{dif}} = \sum_{G \in \mathcal{T}^* \setminus \mathcal{T}_0^*} \mathrm{SSD}_G$$

and, similarly,

$$d_{\mathrm{dif}} = \sum_{G \in \mathcal{T}^* \setminus \mathcal{T}_0^*} d_G.$$

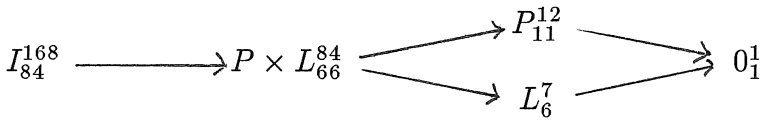
Again, this is a wellknown rule, in particular for situations where $\mathcal{T}^* \setminus \mathcal{T}_0^*$ consists of a single factor. But these general rules are usually not stated explicitly in textbooks — and obviously they cannot be, without an unambigous definition of the ANOVA table.

Rules for estimability of simple contrasts of the form $\alpha_{t'}^T - \alpha_{t''}^T$ will not be discussed in details here (see Tjur 1984). We shall confine ourselves to the remark that if such a contrast is estimable, the maximum–likelihood or least squares estimate of it equals the corresponding difference $\bar{y}_{t'} - \bar{y}_{t''}$ between averages over the classes determined by the two levels of T.

4. The variance component model.

Variance component models, or mixed models, can be regarded as ordinary linear models like those considered in section 3, in which some of the terms in the expression for the mean are regarded as random, in the sense that the set of level parameters for the corresponding factor is thought of as a random sample from a normal population.

EXAMPLE 4.1 (Paulev et al. 1981, slightly simplyfied). Blood samples from 12 patients are taken. Each sample is divided into 14 portions. These portions are sent to 7 laboratories, each laboratory recieving 2 portions from each patient. The laboratories perform (among other things) a measurement of the concentration of CO_2 in the blood, and these concentrations (or rather their logarithms) constitute our data vector $\mathbf{y} = (y_i)$. The structure of the design is shown here (with $P = \{1, \ldots, 12\}$ (patient), $L = \{a, b, \ldots, g\}$ (laboratory)):



This is obviously an orthogonal design (a balanced 12×7 table with two observations per cell), but the relevant model is not a linear model. In fact, the application of standard linear model theory in situations like this will typically end almost before it started, at the model $P \times L$ specifying a separate mean for each combination of patient and laboratory, which is not of much help to anyone. The purpose of the experiment was to detect and quantify the sources of variation, and a model for this purpose can be stated as follows:

$$y_i = \alpha_p + \delta v_l + \omega w_{pl} + \sigma u_i,$$

where v_l ($l \in L$), w_{pl} ($(p,l) \in P \times L$) and u_i ($i \in I$) are independent normalized normal variables. The intiutive idea is that α_p is the true concentration of CO_2 in the blood sample from person p; δv_l is the 'baseline error', present at laboratory l (on that day); the term σu_i represents the measurement–to–measurement variation present on any laboratory; and the 'random interaction' ωw_{pl} represents any sort of error which is specific to the laboratory and the given blood sample (e.g. confusion of CO_2 with other components) but common to the two measurements. In this model the variance on a single observation is

$$\text{var}(y_i) = \delta^2 + \omega^2 + \sigma^2$$

and the parameters δ^2, ω^2 and σ^2 are, accordingly, called the *variance components.*

More generally, a variance component model in a design $\mathcal{D}$ can be specified on the form

$$y_i = \sum_{T \in \mathcal{T}} \alpha_t^T + \sum_{B \in \mathcal{B}} \sigma_B u_b^B$$

where $\mathcal{T}$ (the fixed effects, or the 'treatment stucture') and $\mathcal{B}$ (the random effects, or the 'block structure') are selected subsets of $\mathcal{D}$. In the example, $\mathcal{T} = \{P\}$ and $\mathcal{B} = \{L, P \times L, I\}$. For convenience, we shall represent variance component models by single model formulas, with the elements of $\mathcal{B}$ in brackets, so that e.g. the model of example 4.1 is written

$$P + [L + P \times L + I].$$

This general definition of a variance component model does not require any further conditions, like orthogonality of factors etc. Conceptually, variance component models are not hard to understand, provided that certain obvious conditions are satisfied. One of these conditions is that I should be an element of $\mathcal{B}$, i.e. a 'unit–to–unit variation' should be present in the model. Another condition, which is required for the estimation of the variance components, is that none of the random factors $B \in \mathcal{B}$ should be such that L_B is contained in the mean space $\sum L_T$, since this would obviously imply that random variation due to this factor would be indistinguishable from the fixed effects. However, even under such assumptions, the solution of the general 'unbalanced' variance component model is a complicated affair, involving either numerical maximization of likelihood functions or approximate solutions, approximations to distributions of estimators and test statistics, etc.

There is, however, an exclusive class of variance component models, usually referred to as the 'balanced models', for which an algebraically simple treatment can be given. For these models, the analysis can be broken into components called *error strata.* There is one stratum for

each variance component, and each stratum corresponds to a component of the data vector according to a decomposition of $\mathbf{R}^I$ as a direct sum of orthogonal subspaces. These components are stochastically independent and described by models which are, essentially, linear models of the kind discussed in section 3. This solution relies on an alternative parameterization of the covariance, which is only available in the 'balanced' case. We shall refer to this parameterization as the *eigenvalue parameterization*, because the essential parameters are simply the eigenvalues of the covariance matrix.

The presence of this solution has, more than anything else, contributed to the present confusion around the topic. Some textbooks present the variance component models directly in terms of the eigenvalues (or closely related parameters) in order to simplify the mathematics. Consequently, it is a widespread misunderstanding that there are two 'kinds' of variance component models, those with and those without 'constraints' on the random effects. It is hoped that the following pages will contribute to the clarification of these matters.

First of all, notice that the variance component model can be regarded as a multivariate normal model specified by the mean

$$\mu = \mathrm{E}(\mathbf{y}) = \sum_{T \in \mathcal{T}} \mathbf{X}_T \alpha^T$$

and the covariance matrix

$$\Sigma = \mathrm{cov}(\mathbf{y}) = \sum_{B \in \mathcal{B}} \sigma_B^2 \mathbf{X}_B \mathbf{X}_B^*.$$

The covariance matrix is linear in the parameters σ_B^2, and the essential property of the class of models considered in the following is that the spectral decomposition of $\mathbf{R}^I$ as the direct sum of eigenspaces for the covariance matrix is independent of the variance components σ_B^2.

We shall make the following assumptions. As in the case of linear models, we assume that $\mathcal{D}$ is an orthogonal design, and no further conditions are put on the set $\mathcal{T} \subseteq \mathcal{D}$ of fixed effects. But the set $\mathcal{B} \subseteq \mathcal{D}$ of random effects is assumed to satisfy the following four conditions:

- ($\mathcal{B}1$) $I \in \mathcal{B}$.
- ($\mathcal{B}2$) Any factor $B \in \mathcal{B}$ is balanced.
- ($\mathcal{B}3$) $\mathcal{B}$ is closed under the formation of infima.
- ($\mathcal{B}4$) The matrices $\mathbf{X}_B \mathbf{X}_B^*$ are linearly independent.

Comments.

Condition ($\mathcal{B}1$) has already been commented on.

Condition ($\mathcal{B}2$) is restrictive, of course, but necessary for a simple solution. It is wellknown that even the simple model $0 + [G + I]$ (a one–way

ANOVA model with random variation between groups) becomes a lot more complicated when the groups are of unequal sizes. The mathematical reason for this is that the simple solution relies heavily on the equation $\mathbf{X}_B\mathbf{X}_B^* = n_B\mathbf{P}_B$, which is valid only for balanced factors, and which enables us to write the covariance matrix as a linear combination of the projections $\mathbf{P}_B$.

Condition ($\mathcal{B}3$) is restrictive too, but also essential for the simple solution. This condition implies that the set $\mathcal{B}$ of 'random factors' is an orthogonal design in the sense of section 2. Accordingly, it induces a decomposition of $\mathbf{R}^I$ as a direct sum of orthogonal subspaces, which, as we shall see, are the eigenspaces for the covariance matrix. Situations where ($\mathcal{B}3$) is not satisfied can, to some extent, be handled by formal extension of $\mathcal{B}$ by the missing minima.

Condition ($\mathcal{B}4$) is obviously necessary for estimability of the variance components. Situations where this condition is not satisfied are rarely met in practice.

Now, let

$$\mathbf{R}^I = \bigoplus_{B\in\mathcal{B}} V_B^o$$

be the decomposition induced by $\mathcal{B}$ according to proposition 2.1. Not surprisingly, this decomposition is coarser than the decomposition induced by the whole design $\mathcal{D}$, in the sense that any of the subspaces V_B^o can be written as the direct sum of some of the subspaces V_G in the decomposition induced by $\mathcal{D}$. Hence, the 'block structure' $\mathcal{B}$ induces a classification of the original factors $G \in \mathcal{D}$, reflecting the way subspaces V_G should be collapsed to form the subspaces V_B^o. These classes $\mathcal{D}_B$, $B \in \mathcal{B}$ — or the subspaces V_B^o, or the corresponding orthogonal components of the data vector — are what is vaguely referred to as the *error strata*. The exact rule for allocation of factors to strata is as follows.

PROPOSITION 4.1. *Let $\mathcal{D}_B$ be the set of $G \in \mathcal{D}$ such that B is the coarsest factor in $\mathcal{B}$ which is finer than (or equal to) G. Then*

$$V_B^o = \bigoplus_{G\in\mathcal{D}_B} V_G.$$

Accordingly, if $\mathbf{Q}_B^o$ denotes the orthogonal projection onto V_B^o, then

$$\mathbf{Q}_B^o = \sum_{G\in\mathcal{D}_B} \mathbf{Q}_G.$$

The mathematical result that defines the eigenvalue parameterization can now be stated as follows.

PROPOSITION 4.2. *The linear space of $I \times I$–matrices spanned by the matrices $\mathbf{X}_B\mathbf{X}_B^*$, $B \in \mathcal{B}$, equals the linear space spanned by the matrices $\mathbf{Q}_B^o$, $B \in \mathcal{B}$.*

This follows from proposition 2.1 and the relation $\mathbf{X}_B\mathbf{X}_B^* = n_B\mathbf{P}_B$ for balanced factors.

An immediate consequence of this is that the covariance matrix Σ has a unique representation on the form

$$\Sigma = \sum_{B\in\mathcal{B}} \lambda_B \mathbf{Q}_B^o.$$

Since the matrices $\mathbf{Q}_B^0$ are the projections on components of an orthogonal decomposition, this is the spectral decomposition of the covariance matrix and the parameters λ_B are its eigenvalues. Their relation to the variance components (which can be derived from the relation $\mathbf{P}_B = \sum_{B'\in\mathcal{B},B'\leq B} \mathbf{Q}_{B'}^o$, cfr. proposition 2.1) is

$$\lambda_B = \sum_{B'\in\mathcal{B},B'\geq B} n_{B'}\sigma_{B'}^2.$$

Notice that we always have $\lambda_I = \sigma_I^2$, i.e. the eigenvalue parameter for I–stratum or 'units–stratum' equals the measurement–to–measurement variance. But for other factors B there is no such simple relation between σ_B^2 and λ_B, and usually no easy interpretation of the eigenvalue parameter λ_B. An essential assumption about the variance components is that they are independent of the design, so that for example the addition of more laboratories to the design of example 4.1 would not change the variance components. The eigenvalues are not design–independent in this sense. But from a mathematical point of view, the eigenvalue parameterization is the nice one, which enables us to reduce the analysis to seperate analyses of linear models in strata.

EXAMPLE 4.2. In example 4.1, the covariance matrix can be rewritten as follows:

$$\begin{aligned}\Sigma &= \delta^2\mathbf{X}_L\mathbf{X}_L^* + \omega^2\mathbf{X}_{P\times L}\mathbf{X}_{P\times L}^* + \sigma^2\mathbf{I}\\ &= 24\delta^2\mathbf{P}_L + 2\omega^2\mathbf{P}_{P\times L} + \sigma^2\mathbf{I}\\ &= (24\delta^2 + 2\omega^2 + \sigma^2)\mathbf{P}_L + (2\omega^2+\sigma^2)(\mathbf{P}_{P\times L} - \mathbf{P}_L) + \sigma^2(\mathbf{I} - \mathbf{P}_{P\times L})\end{aligned}$$

where, obviously, $\mathbf{P}_L = \mathbf{Q}_L^0$, $\mathbf{P}_{P\times L} - \mathbf{P}_L = \mathbf{Q}_{P\times L}^0$ and $\mathbf{I} - \mathbf{P}_{P\times L} = \mathbf{Q}_I^0$ are the orthogonal projections in the decomposition induced by $\mathcal{B}$. Accordingly, the relation between the two parameterizations is given by

$$\begin{aligned}\lambda_L &= 24\delta^2 + 2\omega^2 + \sigma^2,\\ \lambda_{P\times L} &= 2\omega^2 + \sigma^2,\\ \lambda_I &= \sigma^2,\end{aligned}$$

in agreement with the general formula above.

Negative variance components. A problem with the eigenvalue parameterization, clearly illustrated by the above example, is that the canonical domain of variation for the eigenvalue parameters $\{\lambda_B > 0\}$ (given by the condition that the covariance matrix should be positive definite) represents an extension of the domain of variation subsumed in the original parameterization, given by the obvious condition that the variance components should be positive. The easy way of solving this problem is to ignore it, just estimating the eigenvalue parameters in their canonical domain of variation. This implies that the backwards calculation of estimates for the variance components from estimated eigenvalues may result in negative estimates for some of the variance components. But this is not as bad as it sounds, for two reasons. First of all, in a proper variance component model, like that of example 4.1 or any other case where one should really expect the variance components to be positive, a negative value of an estimate can very often be taken as a wellcome opportunity to set that component to zero, thus removing this source of variation from the model. Of course, a *significantly* negative estimate of a parameter, which is only meaningful as a non–negative parameter, will always be a problem; but such a lack of fit to the assumptions can obviously not be prevented by formal manipulations with the domain. Secondly, there are situations where 'negative variance components' are meaningful. A classical example is the field trial, where a negative correlation between yields from plots on the same block (i.e. a 'negative variance component' for the random effect of the blocks factors) may be due to competition between plants on neighbouring plots or vaguely defined borders between plots. Obviously, the 'variance components' can not be interpreted as proper population variances here (most certainly not when they are negative), but it is nice to have this possibility included in our models. And since the mathematics becomes a lot easier in this way, we shall consider the extended model given by $\lambda_B > 0$ throughout the paper.

The formal solution of the variance component model now goes as follows. Consider the 'data components'

$$\mathbf{y}_B = \mathbf{Q}_B^0 \mathbf{y}, \quad B \in \mathcal{B},$$

i.e. the orthogonal projections of the data vector on the strata subspaces. It is easy to show that these components are stochastically independent, and that the (normal) distribution of $\mathbf{y}_B$ on the subspace V_B^0 is given by the mean $\mathrm{E}\mathbf{y}_B = \mu_B = \mathbf{Q}_B^0 \mu$ and the covariance matrix $\mathrm{cov}(\mathbf{y}_B) = \lambda_B \mathbf{Q}_B^0$. Estimation in one of these 'strata models' is straightforward, because the model is essentially a linear model of the kind considered in section 3, with V_B^0 taking the role of $\mathbf{R}^I$, the subspace

$$V_B^0 \cap \left(\sum_{T \in \mathcal{T}} L_T \right) = V_B^0 \cap \left(\bigoplus_{G \in \mathcal{T}^*} V_G \right) = \bigoplus_{G \in \mathcal{T}^* \cap \mathcal{D}_B} V_G$$

taking the role of the mean space $L = \sum_{T \in \mathcal{T}} L_T$, $\mathbf{Q}_B^0$ taking the role of the unit matrix $\mathbf{I}$ and λ_B taking the role of the variance σ^2. Thus, estimating as usual in a linear model, the mean vector μ_B is estimated by the projection

$$\hat{\mu}_B = \sum_{G \in \mathcal{T}^* \cap \mathcal{D}_B} \mathbf{Q}_G \mathbf{y} = \mathbf{Q}_B^0 \left(\sum_{G \in \mathcal{T}^*} \mathbf{Q}_G \right) \mathbf{y}.$$

Combining these estimates of mean components from the strata, we obtain the estimate

$$\hat{\mu} = \sum_{B \in \mathcal{B}} \hat{\mu}_B = \left(\sum_{B \in \mathcal{B}} \mathbf{Q}_B^0 \right) \left(\sum_{G \in \mathcal{T}^*} \mathbf{Q}_G \right) \mathbf{y} = \left(\sum_{G \in \mathcal{T}^*} \mathbf{Q}_G \right) \mathbf{y},$$

which coincides with the estimate for the mean in a linear model with the same mean structure $\mathcal{T}$. Thus, treatment contrasts should be estimated exactly as in a linear model, i.e. by the corresponding differences between averages. The expressions for variances on these contrast estimates are, of course, more complicated, but this will not be discussed in details here (see Tjur 1984).

Notice that an essential condition for the above argument to hold is that the mapping $\mu \to (\mu_B \mid B \in \mathcal{B})$, which splits the mean into its strata components, is one–to–one, due to the geometric orhogonality of the mean space $L = \sum_{T \in \mathcal{T}} L_T$ to the strata subspaces V_B^0. Models that do not satisfy this condition, i.e. models with a similarly simple covariance structure but with an arbitrary mean structure (involving e.g. covariates in more or less arbitrary directions or non–orthogonal factors), can be handled in essentially the same way stratum by stratum. But the recombination of strata estimates is more complicated in this case, because one obtains estimates of the same parameters from different strata. 'Recovery of interblock information' in an incomplete block design is the classical example of this.

Estimation of the covariance structure is performed similarly, by recombination of the strata estimates. From B–stratum, we have the estimate

$$\hat{\lambda}_B = \frac{\mathrm{SSD}_{\mathrm{res}}^B}{d_{\mathrm{res}}^B},$$

where

$$\mathrm{SSD}_{\mathrm{res}}^B = \|\mathbf{y}_B - \hat{\mu}_B\|^2 = \|\mathbf{Q}_B^0 \mathbf{y} - \sum_{G \in \mathcal{T}^* \cap \mathcal{D}_B} \mathbf{Q}_G \mathbf{y}\|^2$$

$$= \| \sum_{G \in \mathcal{D}_B \setminus \mathcal{T}^*} \mathbf{Q}_G \mathbf{y} \|^2 = \sum_{G \in \mathcal{D}_B \setminus \mathcal{T}^*} \mathrm{SSD}_G$$

and, similarly,

$$d^B_{\mathrm{res}} = \sum_{G \in \mathcal{D}_B \setminus \mathcal{T}^*} d_G.$$

Notice that the computation of the residual sum of squares and its degrees of freedom from the ANOVA table goes exactly as for the linear model, except that only factors in the stratum should be taken into account. From the estimates $\hat{\lambda}_B$ of the eigenvalue parameters, estimates $\hat{\sigma}^2_B$ of the variance components can be computed by solution of the equations connecting the two parameterizations.

Hypothesis testing: mean structure. Removal of one or more terms from $\mathcal{T}$ can be handled in a similar 'stratum–wise' manner. Let $\mathcal{T}_0$ be a model formula such that $\mathcal{T}_0^* \subset \mathcal{T}^*$. For the linear model in B–stratum, the consequence of this reduction is that the mean space

$$V^0_B \cap \left(\sum_{T \in \mathcal{T}} L_T \right) = \bigoplus_{G \in \mathcal{T}^* \cap \mathcal{D}_B} V_G$$

is replaced with

$$V^0_B \cap \left(\sum_{T \in \mathcal{T}_0} L_T \right) = \bigoplus_{G \in \mathcal{T}_0^* \cap \mathcal{D}_B} V_G.$$

Accordingly, we obtain an F–test for this reduction of the form

$$F\left(d^B_{\mathrm{dif}}, d^B_{\mathrm{res}}(\mathcal{T})\right) = \frac{\mathrm{SSD}^B_{\mathrm{dif}} / d^B_{\mathrm{dif}}}{\hat{\lambda}_B}$$

where

$$\mathrm{SSD}^B_{\mathrm{dif}} = \sum_{G \in \mathcal{D}_B \cap (\mathcal{T}^* \setminus \mathcal{T}_0^*)} \mathrm{SSD}_G$$

and

$$d^B_{\mathrm{dif}} = \sum_{G \in \mathcal{D}_B \cap (\mathcal{T}^* \setminus \mathcal{T}_0^*)} d_G.$$

Again, the rules for which SSD's to include in nominator and denominator are exactly as for the linear model, with omission of factors that are not in B–stratum. In this way we obtain for each stratum an F–test for the hypothesis considered. In most cases, all but one of these F–tests are trivial, in the sence that no reduction of mean structure takes place ($\mathcal{D}_B \cap \mathcal{T}^* = \mathcal{D}_B \cap \mathcal{T}_0^*$). In particular, this is so if $\mathcal{T}_0^*$ is obtained from $\mathcal{T}^*$ by removal of a single factor. In more complex designs, the F–test

may be non–trivial in two or more strata, and in such cases some kind of weighted test statistic, summarizing the results of these F–tests, may be considered. This typically happens when the initial construction of the orthogonal design required addition of one or more 'pseudofactors', so that removal of a single factor T from $\mathcal{T}$ may imply the removal of T and one or more pseudofactors from $\mathcal{T}^*$.

Hypothesis testing: Covariance structure. The hypotheses of interest here are those of the form $\sigma_B^2 = 0$, stating that a given variance component is zero. In the present framework, a hypothesis like this can only be tested if the resulting covariance structure $\mathcal{B}_0 = \mathcal{B} \setminus \{B\}$ satisfies our conditions $(\mathcal{B}1)$–$(\mathcal{B}4)$. For example, the hypothesis $\sigma_I^2 = 0$ is strictly forbidden. Apart from this, the only condition that does not carry over automatically from $\mathcal{B}$ to $\mathcal{B}_0$ is the requirement that the new set of random factors should be closed under the formation of infima. This is easily seen to be the case for $\mathcal{B}_0$ if and only if the factor

$$B_0 = \inf_{B' \in \mathcal{B}, B \leq B'} B'$$

(= the infimum of all factors in $\mathcal{B}$ which are strictly finer than the one to be removed) is distinct from the factor B to be removed. Under this condition, it follows from the relation between the two parameterizations that

$$\lambda_B = \sum_{B' \geq B} n_{B'} \sigma_{B'}^2 = n_B \sigma_B^2 + \sum_{B' \geq B_0} n_{B'} \sigma_{B'}^2 = n_B \sigma_B^2 + \lambda_{B_0},$$

which means that our hypothesis $\sigma_B^2 = 0$ is equivalent to the hypothesis that the two eigenvalues λ_B and λ_{B_0} are equal. Since we have independent, χ^2–distributed estimates $\hat{\lambda}_B$ and $\hat{\lambda}_{B_0}$ of these two parameters, the obvious test for this is a two–sided evaluation of the F–statistic

$$F\left(d_{\text{res}}^B, d_{\text{res}}^{B_0}\right) = \hat{\lambda}_B / \hat{\lambda}_{B_0}.$$

Large values of this test statistic indicate that $\sigma_B^2 > 0$, small values that $\sigma_B^2 < 0$. If the hypothesis is accepted, B_0–stratum of the new model is formed by collapse of the two strata in the old model.

EXAMPLE 4.3. Consider the design of example 2.2. Since the treatment factors here are confounded in a non–trivial manner with the factors B (block) and P (plot), all treatment effects can not be estimated in a simple linear model unless the factors B and P are ignored. This is obviously not satisfactory, and a more reasonable idea is to consider the model with random effects of B and P,

$$F \times N \times V + [B + P + I],$$

i.e.

$$y_i = \alpha_{fnv} + \omega w_b + \delta v_P + \sigma u_i.$$

The partitioning of $\mathcal{D}$ into strata is seen to be as follows (cfr. the factor structure diagram and proposition 4.1):

$$\begin{aligned}
\mathcal{D}_B &= \{B, S, 0\},\\
\mathcal{D}_P &= \{P, F \times S, F\},\\
\mathcal{D}_I &= \{I, F \times N \times V, F \times N, F \times V, N \times V, N, V\}.
\end{aligned}$$

Accordingly, the ANOVA–table, with its lines arranged according to the classification by strata (as one should always do, for convenience) becomes

Stratum	G	d_G	SSD_G
B	B	4	$\mathrm{SS}_B - \mathrm{SS}_S$
	S	1	$\mathrm{SS}_S - \mathrm{SS}_0$
	0	1	SS_0
P	P	8	$\mathrm{SS}_P - \mathrm{SS}_B - \mathrm{SS}_{F\times S} + \mathrm{SS}_S$
	$F \times S$	2	$\mathrm{SS}_{F\times S} - \mathrm{SS}_S - \mathrm{SS}_F + \mathrm{SS}_0$
	F	2	$\mathrm{SS}_F - \mathrm{SS}_0$
I	I	12	
	$F \times N \times V$	0	
	$F \times N$	2	
	$F \times V$	2	
	$N \times V$	0	
	N	1	
	V	1	

(we have omitted formulas for SSD's in I–stratum because some of them are too long for the table, and too complicated to be informative).

The first thing to be tried here would probably be removal of the three–factor interaction from the model, i.e. reduction to

$$F \times N + F \times V + N \times V + [B + P + I].$$

The maximal model formula for the mean structure of this model is

$$\mathcal{T}_0^* = \{F \times N, F \times V, N \times V, S, F, N, V, 0\}.$$

Thus, $\mathcal{T}_0^*$ is obtained from $\mathcal{T}^*$ by removal of the two factors $F \times S$ and $F \times N \times V$. But since $\mathrm{SSD}_{F\times N\times V} = 0$, there is only one F–test to be performed, namely a test in P–stratum

$$F(2,8) = \frac{\mathrm{SSD}_{F\times S}/2}{\mathrm{SSD}_P/8}.$$

If this hypothesis is accepted we may proceed with removal of two–factor interactions. Removal of $F \times N$ or $F \times V$ from the model corresponds to removal of the same factors from the maximal model formula, and are thus tested for in I–stratum. Removal of $N \times V$ corresponds to removal of both $N \times V$ and S from the maximal model formula, and since $\mathrm{SSD}_{N\times V} = 0$ we obtain the test in B–stratum given by

$$F(1,4) = \frac{(\mathrm{SS}_S - \mathrm{SS}_0)/1}{(\mathrm{SS}_B - \mathrm{SS}_S)/4}.$$

This test is seen to be equivalent to a test for total homogeniety in a one–way analysis of a data set consisting of the six block averages, grouped according to the factor S ('superblock', or cross–over pattern). This is in accordance with the fact noticed earlier, that a main effect of the pseudo–factor S can not be distinguished from an $N \times V$–interaction. In more conventional terms, $N \times V$–interaction is fully confounded with blocks. A similar, more complicated, interpretation can be given to the test for three–factor interaction. There, the same F–statistic occurs in the test for 'group–independence of the column effect' in an additive model for the 6×3 table of plot averages, with rows classified in two groups according to the factor S.

Depending on our success in removal of two–factor interactions, we may finally test for main effects. The test for F–effect takes place in P–stratum, the two other tests in I–stratum. This is intuitively obvious from the beginning, since fungicides have been allocated to whole plots while main effects of the two other factors obviously have to do with differences between yields on subplots of the same plot. This is also reflected by the expressions for variances on estimated contrasts, which we have not given here. The variance on an estimated F–contrast is a linear combination of σ_P^2 and σ_I^2, while the variances on estimated N– and V–contrasts are just proportional to σ_I^2.

Simplification of the covariance structure is usually of limited interest in situations like this. However, for completeness, the 'allowed' hypotheses are

$$\sigma_B^2 = 0 \text{ or } \lambda_B = \lambda_P$$

and

$$\sigma_P^2 = 0 \text{ or } \lambda_P = \lambda_I.$$

Since the formulas for the F–statistics depend on the mean structure assumed, there is nothing to be added here to the general theory.

To illustrate the way pseudo–factors take care of the correct allocation to strata in more complex designs, we give a last example.

EXAMPLE 4.4 ('Twin cross–over' design, Aa. Vølund and E. Sørensen, Novo Nordisk). 40 mice were divided at random into 4 groups of 10 mice. The experiment took place over two days, each mouse occuring twice in the experiment in a test on each day. A 'test' here means injection of a certain dose of a certain type of insulin, followed by a measurement (after a given time) of glucose concentration in the blood. Two types of insulin in two different doses were used, and these treatments were given according to the following table, where t (test) and s (standard) denote the two types of insulin while l (low) and h (high) denote the two doses.

Group	day 1	day 2
1	tl	sh
2	th	sl
3	sl	th
4	sh	tl

The treatment factors of immediate interest are

$$D = \{l, h\} \text{ (dose)}$$
$$T = \{s, t\} \text{ (type of insulin)}$$

To this must be added

$$N = \{1, 2\} \text{ (number of day)}$$

because there may be a difference between the two days, e.g. an after-effect on day 2 of the test given on day 1. To this we add all possible products of these three factors, the trivial factor 0 and the 'units' factor I. This design is just an ordinary $2 \times 2 \times 2$ table with 10 observations per cell. To this comes now the 'blocking' factor

$$M = \{1, \ldots, 40\} \text{ (mouse)}$$

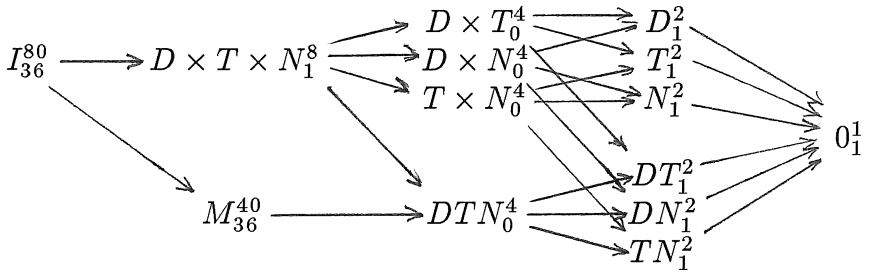
which groups the 80 observations of the table in pairs in such a way that the two observations of a pair always occur in opposite corners of the cubic table. Obviously, we shall consider a model with random effect of this factor M, and the advantage of this complicated blocking is that the main effects of D, T (and N) are estimated with maximal accuracy, in the sense that these factors end up in I–stratum.

All factors included until now are easily seen to be orthogonal, but the following pseudo factors must be added to obtain closedness under infima:

$$
\begin{aligned}
DTN &= M \wedge (D \times T \times N) \text{ on 4 levels} \\
DT &= M \wedge (D \times T) \text{ on 2 levels} \\
DN &= M \wedge (D \times N) \text{ on 2 levels} \\
TN &= M \wedge (T \times N) \text{ on 2 levels}
\end{aligned}
$$

Notice that DTN is just the classification according to 'group of mice'. The three other pseudofactors are formed from this by pairwise collapse of these groups, in a structure similar to the 2×2 latin square.

The factor structure diagram follows here:



The allocation of factors to strata in the model with $\mathcal{B} = \{M, I\}$ is easily seen to be as indicated by the following ANOVA–table (where zero SSD's and formulas for SSD's are omitted):

Stratum	Factor	d.f.	SSD
I	I	36	
	$D \times T \times N$	1	
	D	1	
	T	1	
	N	1	
M	M	36	
	DT	1	
	DN	1	
	TN	1	
	0	1	

Notice that the SSD's in this table are exactly the same as one would get for the $2 \times 2 \times 2$ complete factorial design without the blocking factor

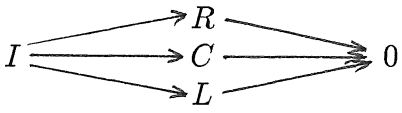
M, except that the residual SSD is split in two (SSD_I and SSD_M). For example, SSD_{DT} is the usual square sum for interaction between D and T. Accordingly, the tests for two–factor interactions take place in M–stratum, while the tests for three–factor interaction and — most importantly — the tests for main effects take place in I–stratum.

5. Group generated designs.

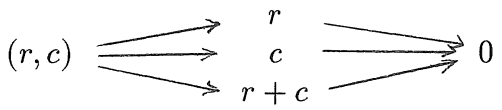
Apart from the designs derived from complete balanced k–way tables, the most important class of orthogonal designs is the class of *group generated designs.* This class includes the designs known as fractional factorials, and (more or less as a byproduct) complete single–replicate designs with blockings that confound some of the main effects and/or interactions.

Such designs have been known for a long time (Yates 1935, Fisher 1942, Finney 1945, Bose 1947), and so has their relation to the theory of Abelian groups (see e.g. Fisher 1942). Recently, fractional factorials have received new attention due to their role in Taguchi's methods for industrial statistics and quality control. Also quite recently, several authors have studied the possibility of a more efficient use of group theory to the construction of such designs. The present section summarizes a joint work with Christoffersen (1987), which can be regarded as an attempt to put the classical constructions of fractional factorials etc. into a more modern framework. Almost all results presented here can be found in earlier papers by El Mossadeq et al. (1985), Kobilinsky (1985) and, in particular, Bailey (1985).

The simplest example of a group generated design is the cyclic latin square. In the present context, a latin square of order k can be defined as a design of the form



with its three factors R (row), C (column) and L (latin letter) balanced, $|R| = |C| = |L| = k$ and $|I| = k^2$. It is wellknown how a latin square can be constructed for any k by a simple cyclic arrangement of the first k letters of the alphabet in a $k \times k$ square such that each letter occurs exactly once in each row and each column. Mathematically, this construction can also be explained as follows: Take $I = \mathbf{Z}_k \times \mathbf{Z}_k$, where $\mathbf{Z}_k$ denotes the cyclic group of order k (= the integers modulo k). Put $R = C = L = \mathbf{Z}_k$, and define the allocation of factor levels to units according to the scheme



A more complicated example is the split–plot design of example 2.2. Here, we can take

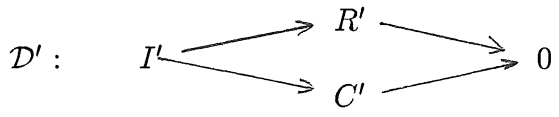
$$
\begin{aligned}
F &= \mathbf{Z}_3 \\
N &= V = \mathbf{Z}_2 \\
S &= \mathbf{Z}_2 \\
B &= S \times \mathbf{Z}_3 = \mathbf{Z}_2 \times \mathbf{Z}_3 \\
P &= B \times F = \mathbf{Z}_2 \times \mathbf{Z}_3 \times \mathbf{Z}_3 \\
I &= B \times F \times N = \mathbf{Z}_2 \times \mathbf{Z}_3 \times \mathbf{Z}_3 \times \mathbf{Z}_2
\end{aligned}
$$

and define the factor levels for unit $i = (s, j, f, n) \in I$ by

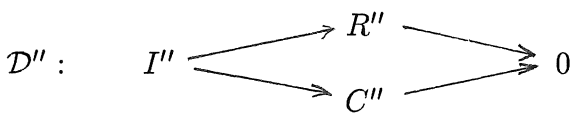
$$
\begin{aligned}
\varphi_F(s, j, f, n) &= f \\
\varphi_N(s, j, f, n) &= n \\
\varphi_V(s, j, f, n) &= s + n \\
\varphi_B(s, j, f, n) &= (s, j) \\
\varphi_P(s, j, f, n) &= (s, j, f)
\end{aligned}
$$

The design of example 4.4 can be constructed in a similar way.

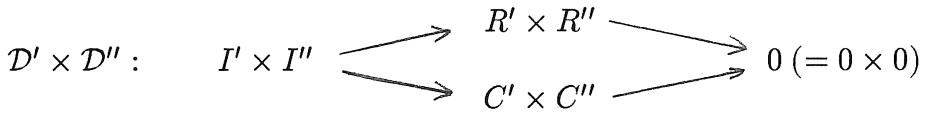
Products of designs. Before we proceed with group generated designs, a few general concepts are required. Suppose that we have two designs of the same 'shape', e.g.



and



We can then form a new design $\mathcal{D}' \times \mathcal{D}''$, again of the same 'shape', as follows. As the units set take $I = I' \times I''$, and define the sets of factor levels similarly as Cartesian products of the corresponding sets of factor levels in the two original designs:



The mappings are defined in the obvious canonical way, e.g.

$$\varphi_{R' \times R''}(i', i'') = (\varphi_{R'}(i'), \varphi_{R''}(i'')) .$$

Notice that the Cartesian products occurring here are 'exterior' — not to be confused with cross–classifications which are products of factors in the same design. Notice also that the design $\mathcal{D}' \times \mathcal{D}''$ is, strictly speaking, not unique when $\mathcal{D}'$ and $\mathcal{D}''$ are given, since it depends on the way we draw the factor structure diagram (in the example, we might turn one of the original diagrams upside down, and probably obtain something different). Apart from this, the construction is a triviality, which deserves attention here only because we need it for the formulation of a fundamental result for group generated designs, stating that such a design can be decomposed as a product $\mathcal{D} = \mathcal{D}_2 \times \mathcal{D}_3 \times \mathcal{D}_5 \times \mathcal{D}_7 \times \ldots$ of 'prime components'.

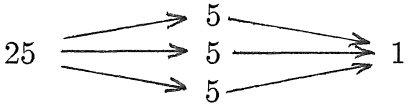
It is relatively easy to show that the formation of products of two or more designs preserves the most important properties of these designs, in the sense that

(1) If the original designs $\mathcal{D}', \mathcal{D}'', \ldots$ are closed under formation of infima, so is the product $\mathcal{D}' \times \mathcal{D}'' \times \ldots$ (infima in a product design are given by the rule $(F' \times F'' \times \ldots) \wedge (G' \times G'' \times \ldots) = (F' \wedge G') \times (F' \wedge G') \times \ldots$).

(2) A factor $F' \times F'' \times \ldots$ in the product design is balanced if all corresponding factors $F', F'', \ldots$ in the original designs are balanced.

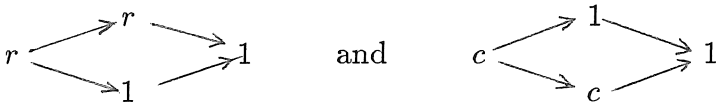
(3) A product of orthogonal designs is again an orthogonal design.

By the *dimensions* of a design we mean a factor structure diagram which displays the numbers of levels of factors, rather than their names, like

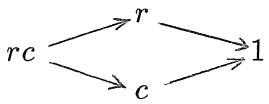


which shows the dimensions of a 5×5 latin square. The purpose of this definition is just to avoid names of factors when they are not needed. Most of what is said in the following has to do with existence of designs of given dimensions, without regard to what the factors stand for.

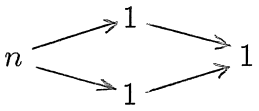
Just to settle these concepts, consider the a simple example. For any non–negative integers r and c we have orthogonal designs of dimensions



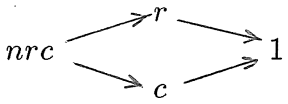
These designs are trivial in the sense that they contain only the units factor and the trivial factor (more than once, but this is allowed in the present context). The product of these two designs is obviously a two–way table with one observation per cell, of dimensions



If we form the product of this design with another trivial design of dimensions



we obtain a design of dimensions



which is nothing but an $r \times c$ table with n observations per cell.

In the following, all factors in our designs will be balanced. Under this condition, it is roughly correct to say that the properties of an orthogonal design are determined by its dimensions. This is true as far as its formal statistical properties are concerned. It is *not* correct that designs of the same dimensions are isomorphic, in the sense that there is a one–to–one mapping between their units sets which preserves the entire structure up to relabelling of levels. For example, the two 4×4 latin squares

$$\begin{array}{cccc} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{array} \qquad \text{and} \qquad \begin{array}{cccc} A & B & C & D \\ B & A & D & C \\ C & D & A & B \\ D & C & B & A \end{array}$$

are not isomorphic in this sense. But their statistical properties are the same, formally.

Group generated designs. Recall that an *Abelian group* is a set G with operations $+$ and $-$, behaving according to the usual rules for addition and subtraction, and a selected element 0, which has the usual role of a zero. The simplest example is the cyclic group $\mathbf{Z}_n$, usually represented

as the integers $0, 1, 2, \ldots, n-1$ with addition modulo n. According to a fundamental theorem in group theory, any Abelian group can be represented as a Cartesian product of such cyclic groups, with coordinatewise addition.

For a given design $\mathcal{D}$, suppose that the set I and all other sets F of factor levels are equipped with structures as Abelian groups, in such a way that the mappings φ_F assigning factor levels to units are group homomorphisms $(\varphi_F(i_1 \pm i_2) = \varphi_F(i_1) \pm \varphi_F(i_2))$. The examples of designs given earlier in this section are all of this kind. In this case we say that the design is *group generated.*

For a factor F in a group generated design, the partitioning of I induced by φ_F consists of the *kernel*

$$K_F = \mathrm{Ker}(\varphi_F) = \varphi_F^{-1}(0)$$

of the homomorphism φ_F and its *cosets* $i + K_F$, $i \in I$. From elementary group theory, it is wellknown that these cosets are all of the same size, which means that the factor is balanced, provided that φ_F is surjective. Moreover, a factor is characterized (up to relabelling of classes) by its kernel, because $i_1, i_2 \in I$ are on the same level of F if and only if $i_1 - i_2 \in K_F$. This correpondence between factors $\varphi_F \colon I \to F$ and subgroups K_F of I is order–reversing, in the sense that

$$F \leq G \Longleftrightarrow K_F \supseteq K_G,$$

and the formation of minima of factors corresponds to the operation 'addition of subgroups',

$$K_{F \wedge G} = K_F + K_G = \{k' + k'' \mid k' \in K_F, k'' \in K_G\}.$$

Similarly, the formation of products of factors corresponds to intersection of their kernels,

$$K_{F \times G} = K_F \cap K_G.$$

Most important, perhaps, is the fact that *any two factors in a group generated design are orthogonal.* This is a trivial consequence of the fact that the classes $\{i \mid \varphi_F(i) = f, \varphi_G(i) = g\}$ for f and g on the same level of $F \wedge G$ are all of the same size, since they are cosets of the subgroup $K_F \cap K_G$ $(= K_{F \times G})$.

Let $\mathcal{K}$ be the set $\{K_F \mid F \in \mathcal{D}\}$ of kernels for factors in a given group generated design $\mathcal{D}$. From what has been said above, it follows that this system $\mathcal{K}$ of subgroups of I determines the design uniquely, up to relabelling of factor levels. Moreover, it follows that an *orthogonal* design, in the sense of section 2, can be represented by such a set $\mathcal{K}$ of subgroups which is closed under addition of subgroups and contains

the trivial group $\{0\}$. Notice that the orders of these subgroups are complementary to numbers of levels of corresponding factors in the sense that $|K_F| = |I|/|F|$.

By duality, this representation of an orthogonal design can be translated to a representation by a set $\mathcal{R}$ of subgroups of the *dual group* I^*, which is closed under intersections and has I^* as an element. This representation has the advantage that the size of a subgroup here is simply equal to the number of levels of the corresponding factor, and infima of factors correspond to intersections of subgroups. But since the dual group is a slightly abstract concept, we shall postpone this to the discussion of the special case $I = \mathbf{Z}_p^n$, p prime, where the dual group can be given a more concrete interpretation.

The prime components of a group generated design. For a given (finite) Abelian group $(G, +)$ and a given prime integer $p \in \mathbf{P} = \{2, 3, 5, 7, \dots\}$, define

$$G[p] = \{g \in G \mid p^n g = 0 \text{ for some power } p^n \text{ of } p\},$$

where multiplication of a group element g by an integer m is defined in the obvious way by $mg = g + \cdots + g$ (m terms). Then, $G[p]$ is a subgroup of G, and it is not difficult to show that

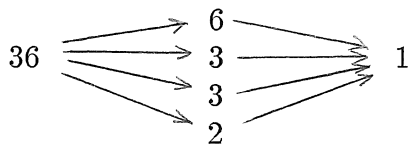
$$G = \bigoplus_{p \in \mathbf{P}} G[p]$$

i.e. any group element has a unique representation as a (finite) sum $g = g_2 + g_3 + g_5 + g_7 + \dots$, $g_p \in G[p]$. The order of (number of elements in) the subgroup $G[p]$ is a power of p, namely the greatest such power that divides $|G|$. Moreover, any group homomorphism $\varphi\colon G \to H$ into som other finite Abelian group H has the property that it maps the subgroup $G[p]$ into the subgroup $H[p]$. For our group generated designs, the consequence of this is that *any such design $\mathcal{D}$ can be decomposed as a product design*

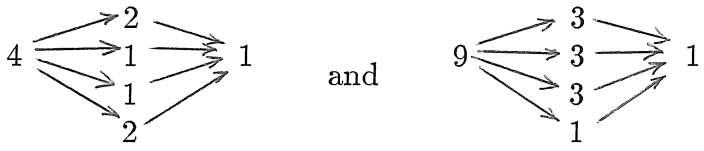
$$\mathcal{D} = \mathcal{D}_2 \times \mathcal{D}_3 \times \mathcal{D}_5 \times \dots,$$

where the numbers of levels for factors in $\mathcal{D}_p$ are powers of p.

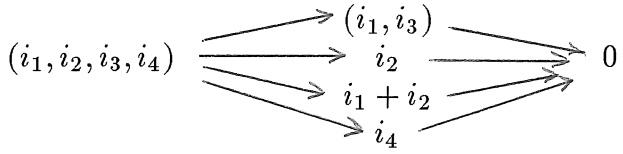
EXAMPLE 5.1. Suppose that we want to construct a design of dimensions



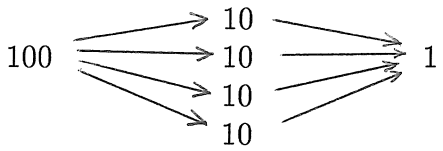
(a 1/3 replicate of a $6 \times 3 \times 3 \times 2$ factorial). Existence of a group generated design of these dimensions is equivalent to existence of its prime components for $p = 2, 3$, of dimensions



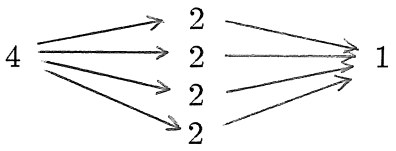
respectively. The 2–component trivially exists, and so does the 3–component which is just a (trivially extended) latin square of order 3. Hence, we can construct a design of the given dimensions as follows: Take $I = \mathbf{Z}_3 \times \mathbf{Z}_3 \times \mathbf{Z}_2 \times \mathbf{Z}_2$, and define mappings (and sets of factor levels) as indicated by the scheme



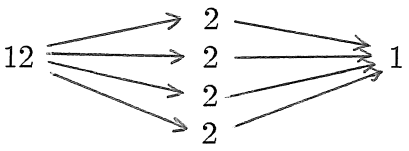
EXAMPLE 5.2. A group generated design of dimensions



does not exist, since the 2–componont of this would be of dimensions



which is obviously impossible, for the simple reason that calculation of degrees of freedom results in a negative value of d_I. However, a graeco-latin square of order 10 is known to exist. What we have shown is only that it can not be group generated. A smaller (probably the smallest, in some sense) set of design constants for which a design exists, but not a group generated one, is



$\mathbf{Z}_p$–vectorspace designs. For group generated designs, the decomposition into prime components gives a considerable reduction of the amount of work required from the experimental designer. Essentially, we can now restrict our attention to designs for which the size of the units set and all sets of factor levels are powers of the same prime p. In practice, it is

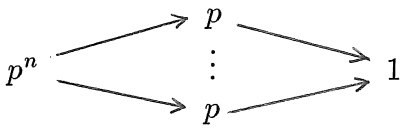
possible to reduce the efforts a little further, by restriction to groups of the form $\mathbf{Z}_p^n$. This is so because the system of subgroups of $\mathbf{Z}_p^n$ is richer than the system of subgroups in any other group of the same order. Unfortunately, this is not a result that we have been able to prove (or even state) in general, it is mainly a matter of experience. Nevertheless, in the following we shall restrict our attention to groups of this form.

It is wellknown that $\mathbf{Z}_p$, the integers modulo a prime p, has the structure of a *field* under addition and multiplication modulo p (the usual rules for addition and multiplication hold, and division by non–zero elements is also allowed). The Abelian group $\mathbf{Z}_p^n$ (and any of its subgroups) has the structure of a *vectorspace* over the field $\mathbf{Z}_p$, and any group homomorphism between such groups is also a linear mapping. In particular, any homomorphism $\varphi\colon \mathbf{Z}_p^n \to \mathbf{Z}_p^k$ can be represented by a $k \times n$–matrix in the usual way,

$$\varphi(z) = \varphi(z_1, \ldots, z_n) = \mathbf{A}z = \begin{bmatrix} a_{11}z_1 + \cdots + a_{1n}z_n \\ \vdots \\ a_{k1}z_1 + \cdots + a_{kn}z_n \end{bmatrix}$$

(subsuming here and in the following that elements $z = (z_1, \ldots, z_n) \in \mathbf{Z}_p^n$ are regarded as $n \times 1$ columns whenever they occur in equations involving matrix operations).

EXAMPLE 5.3. Consider a design of dimensions



How many factors on p levels can be put into this? If each factor is a linear mapping

$$\varphi_j(z_1, \ldots, z_n) = a_{j1}z_1 + \ldots a_{jn}z_n$$

from $\mathbf{Z}_p^n$ to $\mathbf{Z}_p$, the question can be rephrased as follows:

How many linear mappings $\varphi_1, \ldots, \varphi_m$ can we find such that any two $\varphi_{j'}$ and $\varphi_{j''}$ among them have the property that

$$(\varphi_{j'}, \varphi_{j''})\colon \mathbf{Z}_p^n \to \mathbf{Z}_p^2$$

is surjective?

But since $(\varphi_{j'}, \varphi_{j''})$ is surjective if and only if the two rows of its matrix

$$\begin{bmatrix} a_{j'1} & \ldots & a_{j'n} \\ a_{j''1} & \ldots & a_{j''n} \end{bmatrix}$$

are linearly independent, we can also state the question as follows:

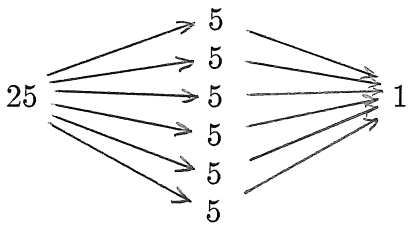
How many non–zero n–vectors $(a_{j1}, \ldots, a_{jn})$, $j = 1, \ldots, m$, can we find such that no two of them are proportional?

The answer to the last question is obviously

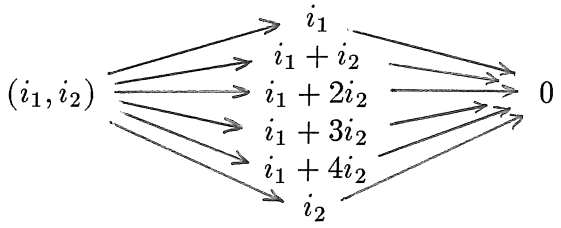
$$m = (p^n - 1)/(p - 1) = p^{n-1} + p^{n-2} + \cdots + p + 1,$$

since the equivalence relation 'proportionality' divides the $p^n - 1$ non–zero elements of $\mathbf{Z}_p^n$ into classes of size $p - 1$ (lines through 0, excluding the point 0), and exactly one vector can be chosen from each.

For example, a design of dimensions



with 25 units, $(25 - 5)/(5 - 1) = 6$ factors on 5 levels, exists. In order to construct it, we must define the six factors by linear expressions with pairwise linearly independent coefficient sets, for example



Dualization. Example 5.3 above illustrates the following more general principle. Suppose that we want to investigate existence of a design of given dimensions. Assuming that all dimensions are powers of the same prime p, and restricting attention to $\mathbf{Z}_p$–vectorspace designs, the problem is then to construct a system of linear mappings between $\mathbf{Z}_p$–vectorspaces of the given sizes, such that the formation of infima of factors (addition of the kernels) matches the structure of the given diagram. Let I^* denote the *dual* of $I = \mathbf{Z}_p^n$, i.e. the vectorspace of linear mappings $I \to \mathbf{Z}_p$. It is natural to think of I^* as the n–dimensional space of $1 \times n$ row matrices, so that the value of $a \in I^*$ at $i \in I$ is simply the matrix product ai. For any factor $\varphi_F \colon \mathbf{Z}_p^n \to \mathbf{Z}_p^k$, let $R_F \subseteq I^*$ denote the subspace of linear mappings that can be written as functions of $\varphi_F(i)$; or, in matrix terms, the subspace of I^* spanned by the rows of the matrix for φ_F. This space R_F determines the factor uniquely, up to

relabelling of classes. The relation to the representation of a factor by its kernel is given by

$$R_F = \{a \in I \mid ai = 0 \text{ for } i \in K_F\}$$

and the relation between the order of R_F and the number of levels for F is simply that $|F| = |R_F|$. The assignment of a subspaces R_F of I^* to factors F is order preserving, in the sense that

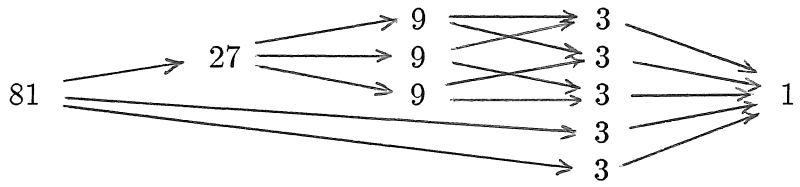
$$F \leq G \Longleftrightarrow R_F \subseteq R_G,$$

and the formation of minima of factors corresponds to intersection of these subspaces,

$$R_{F \wedge G} = R_F \cap R_G.$$

Hence, we have an alternative representation of orthogonal $\mathbf{Z}_p$–vector-space designs by systems $\mathcal{R} = \{R_F \mid F \in \mathcal{D}\}$ of subspaces of I^* ($\approx \mathbf{Z}_p^n$), which are closed under intersections and include I^*. The problem to find a design of given dimensions reduces to the problem of finding such a system of subspaces of the desired sizes such that the formation of intersections matches with the structure of the diagram.

EXAMPLE 5.4. Suppose that we want to construct a design of dimensions



i.e. a 1/3 replicate of a 3^5 factorial which fully allows for estimation and test of hypotheses involving interactions between the first three factors, provided that no other interactions are present. This is solved if we can find 5 subspaces $R_A, \ldots, R_E \subseteq \mathbf{Z}_3^4$ of dimension 1 such that $R_{A \times B \times C} = R_A + R_B + R_C$ is of dimension 3 and

$$R_D \cap R_{A \times B \times C} = R_E \cap R_{A \times B \times C} = R_D \cap R_E = \{0\}.$$

But this is easy, since we can just take

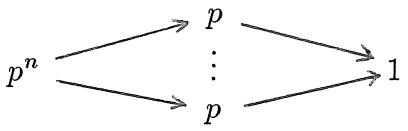
$$\begin{aligned} R_A &= \text{span}\{(1,0,0,0)\} \\ R_B &= \text{span}\{(0,1,0,0)\} \\ R_C &= \text{span}\{(0,0,1,0)\} \\ R_D &= \text{span}\{(0,0,0,1)\} \\ \text{and, say,} \quad R_E &= \text{span}\{(1,1,1,1)\} \end{aligned}$$

Accordingly, we can define the allocation of factor levels to units by

$$\begin{aligned}
\varphi_A(i_1,i_2,i_3,i_4) &= i_1,\\
\varphi_B(i_1,i_2,i_3,i_4) &= i_2,\\
\varphi_C(i_1,i_2,i_3,i_4) &= i_3,\\
\varphi_D(i_1,i_2,i_3,i_4) &= i_4,\\
\varphi_E(i_1,i_2,i_3,i_4) &= i_1+i_2+i_3+i_4.
\end{aligned}$$

In fact, there is room for much more than two 'non–interacting' factors D and E in this design. Since any one–dimensional subspace not contained in $R_{A\times B\times C}$ can be used for the construction of such a factor, it is possible to include up to $(81-27)/(3-1)=27$.

Maximal packings. The design of dimensions



with $(p^n-1)/(p-1)$ factors on p levels is an example of a design for which calculation of degrees of freedom results in d_I=0, i.e. 'zero degrees of freedom for the residual in the maximal model'. For the corresponding system $\mathcal{R}$ of row spaces $R_F \subseteq I^*$, this means that it constitutes a *maximal packing* of $I^* \approx \mathbf{Z}_p^n$, in the sense that $I^* \setminus \{0\}$ is the disjoint union of the sets $R_F \setminus \{0\}$, $F \in \mathcal{D}$, $F \neq I$. Maximal packings are interesting because they represent designs with the maximal number of factors included, such that all main effects can be estimated (but the introduction of any interaction will introduce non–trivial confoundedness–relations). Another example (a straightforward generalization of the last remarks of example 5.4) is a design with p^n units, one factor on p^k levels ($k<n$) and $(p^n-p^k)/(p-1)$ factors on p levels.

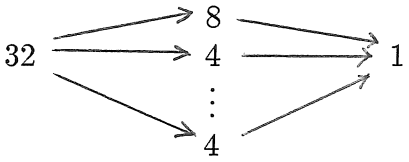
A less obvious result on existence of maximal packings is the following.

PROPOSITION 5.1. *For any prime p and non–negative integers n_1 and n_2, $n_1 \geq n_2$, there exist subspaces R_0 of dimension n_1 and $R_1,\ldots,R_{p^{n_1}}$ of dimension n_2 of $\mathbf{Z}_p^{n_1+n_2}$ such that $R_{j'} \cap R_{j''} = \{0\}$ for $j' \neq j''$.*

This is a maximal packing, since

$$p^{n_1+n_2}-1 = (p^{n_1}-1)+p^{n_1}(p^{n_2}-1).$$

EXAMPLE 5.5. For $p=2$, $n_1=3$ and $n_2=2$, the proposition states that a design of dimensions



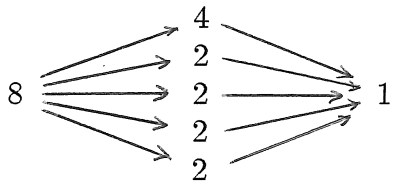
with one factor on 8 levels and 8 factors on 4 levels exists.

Outline of proof. It can be shown that there exists a one–to–one linear mapping $C\colon \mathbf{Z}_p^{n_1} \to \mathbf{Z}_p^{n_1}$ with the property that any of the linear mappings $C^{k_1} - C^{k_2}$ is also one–to–one when $k_1 - k_2$ is non–zero modulo $p^{n_1} - 1$. This (an immediate consequence of the fact that the multiplicative group in the Galois–field of order p^{n_1} is cyclic) will not be explained in details here. For an arbitrary injective linear mapping $A\colon \mathbf{Z}_p^{n_2} \to \mathbf{Z}_p^{n_1}$, define

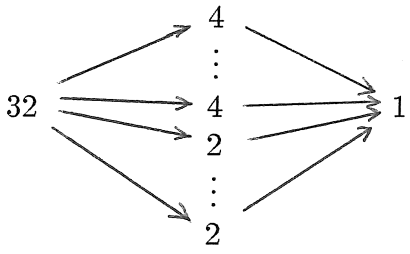
$$
\begin{aligned}
R_0 &= \mathbf{Z}_p^{n_1} \times \{0\} \\
R_1 &= \{(CAz, z) \mid z \in \mathbf{Z}_p^{n_2}\} \\
R_2 &= \{(C^2Az, z) \mid z \in \mathbf{Z}_p^{n_2}\} \\
&\vdots \\
R_{p^{n_1}-1} &= \{(C^{p^{n_1}-1}Az, z) \mid z \in \mathbf{Z}_p^{n_2}\} \\
R_{p^{n_1}} &= \{0\} \times \mathbf{Z}_p^{n_2}
\end{aligned}
$$

This system of subspaces is seen to satisfy the conditions.

From maximal packings, other maximal packings can be constructed by further splitting of the factors. For example, if we replace the factor on 8 levels in the design of example 5.5 with the factors in the maximal packing of dimensions

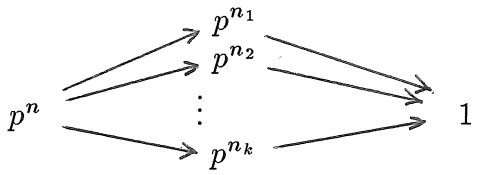


we obtain a design of dimensions



with 9 factors on 4 levels and 4 factors on 2 levels. By this technique, possibly followed by removal of some of the factors, it is possible to construct a large number of designs.

Existence of 'main–effect' designs. We may now pose the following general question. For a given prime p, for which sets $n, n_1, n_2, \ldots, n_k$ of integers does a $\mathbf{Z}_p$–vectorspace design of dimensions



exist? We can not give a definitive answer to this. However, it is easy to give two necessary conditions for such a design to exist, and these turn out to be sufficient also, for most practical purposes. The conditions are

(1) The degrees–of–freedom book–keeping must not result in a negative value of d_I; i.e. we should have

$$p^n - 1 \geq (p^{n_1} - 1) + \cdots + (p^{n_k} - 1).$$

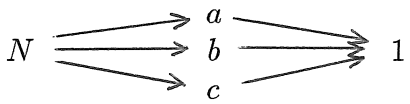
(2) No two dimensions must sum to more than n; i.e. we should have

$$n \geq n_{j'} + n_{j''} \text{ for } j' \neq j''.$$

For $n \leq 4$, it is not hard to show that these conditions will ensure that a design can actually be constructed by (possibly repeated) use of proposition 5.1. Christoffersen (1987) showed, by a detailed examination of cases, that the conditions are sufficient for existence of the design when $k \leq 9$. For $k = 10$ this is not true. In the case $k = 10$, $n = 5$, $n_1 = n_2 = \cdots = n_{10} = 2$, it can be shown that no $\mathbf{Z}_2$–vectorspace design exists.

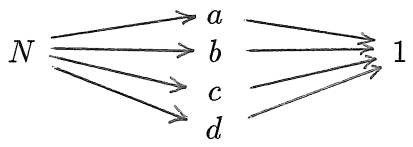
Just to illustrate the consequences for the general problem of designing 'main–effect fractional factorials', consider the following two propositions (which are immediate consequences of the above results).

PROPOSITION 5.2. *A group generated design of dimensions*



exists if and only if N is an integer multiple of any of the three numbers ab, ac and bc.

PROPOSITION 5.3. *A group generated design of dimensions*

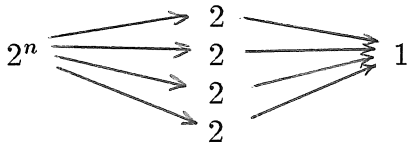


exists if and only if the following two conditions are satisfied.

(i) N is an integer multiple of any of the six numbers ab, ac, ad, bc, bd and cd.

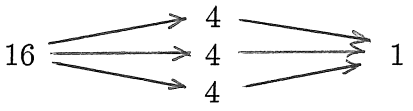
(ii) If the four numbers a, b, c and d are all even, N must be a multiple of 8.

The condition (ii) is obviously there to ensure existence of a 2–component of the form

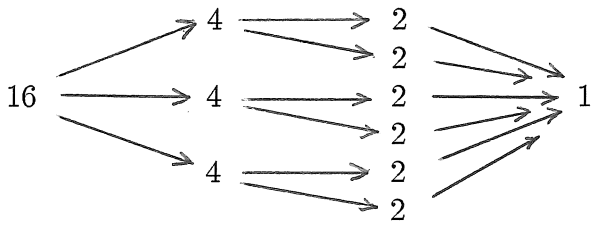


which requires $n \geq 3$. The sequence of propositions is easy to continue. For designs with five factors, a new condition for existence of the 3–component should be introduced, etc. For concrete purposes, however, it is usually easier to refer directly to the results for existence of the prime components.

Fractional factorials with estimable interactions. The results for 'main–effect designs' are trivially applicable also in situations where one or more interactions are to be included in the initial model, provided that these interactions involve disjoint sets of main effects. For example, existence of a design of dimensions



trivially implies existence of a design of dimensions



which we may think of as a 1/4 replicate of a 2^6 design with the property that the two–factor interactions $A \times B$, $C \times D$ and $E \times F$ can be included

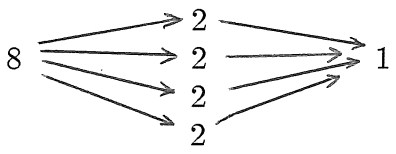
in the initial model, with no further aliasing than already present in a complete 2^6 design (provided, of course, that all other interactions can be assumed vanishing).

Designs for estimation of a more or less arbitrary pattern of main effects and interactions are more difficult to classify. At least, we can say nothing about them, beyond what has already been said about group generated designs in general.

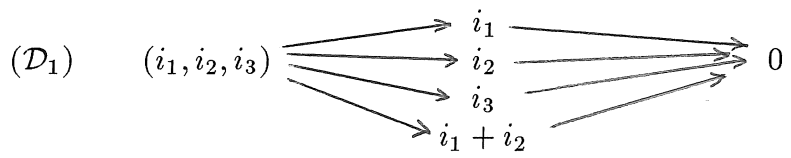
However, in the case of k factors on p levels (p prime), it is possible to say a little more, provided that the condition for estimability of interactions is symmetric in the k factors. By convention, a design of this form is said to be of *resolution* w if the design is a complete factorial in any $w-1$ of the factors, when the remaining $k-w+1$ factors are ignored. In this terminology, a 'main effect design' is a design of resolution 3. More generally, for w odd, it is easy to see that designs of resolution w are designs with the property that any set of interaction terms up to and including interactions between $\frac{w-1}{2}$ factors can be included in a model without introducing any further aliasing than present in the complete p^k design. For w even, the interpretation is slightly more complicated. A design of resolution 4 is essentially a main effect design, since the condition only ensures that one two–factor interaction may be present. But since this interaction term may be chosen arbitrarily, the design has also the 'robustness property' that if any of the two–factor–interactions, initially assumed vanishing, happens to be present, then the damage is restricted to meaningless identification of the corresponding main effects. Estimates of other (welldefined) main effects do not become biased.

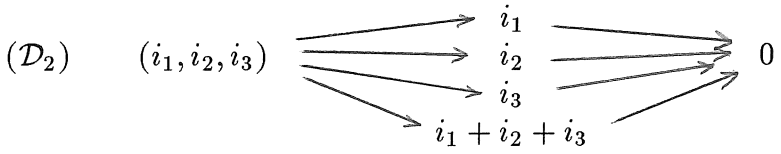
From the experimental designers point of view, the important aspect of the concept of resolution is that once the dimensions of a design are settled, the design should usually be chosen with the maximal possible resolution.

EXAMPLE 5.6. A design of dimensions



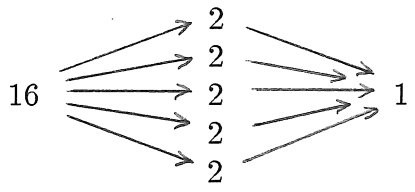
can be constructed in many ways. Two of them are given here (subsuming $I = \mathbf{Z}_2^3$ etc.):



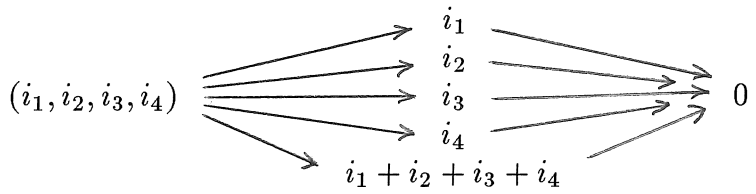


For $\mathcal{D}_1$, an (unexpected) interaction between the first two factors would result in a false main effect of the fourth factor. This (or any similar phenomena) does not occur for $\mathcal{D}_2$, since this design is of resolution 4.

EXAMPLE 5.7. Consider the dimensions



(a 1/2 replicate of a 2^5 design). Here it is possible to select a design of resolution 5, e.g. by



This design has the obvious advantage that all two–factor interactions can be included (and tested for) in the initial model.

As should be obvious from the definition (and also from the examples), the property that a p^k fractional factorial is of resolution w is equivalent to the property of the k sets of coefficients $(a_{j1}, \ldots, a_{jn}) \in \mathbf{Z}_p^n$, $j = 1, \ldots, k$, defining the k factors, that any $w-1$ among them are linearly independent. Pairwise linear independent coefficient sets give resolution 3 (main effect design), triplewise linear independent coefficient sets give a design of resolution 4, etc. Hence, we are left with the following problem:

How many vectors $a_1, \ldots a_k \in \mathbf{Z}_p^n$ can be selected such that any $w-1$ among them are linearly independent?

The answer, $k_{\max}$ say, is the maximal number of factors on p levels that can be present in a design of resolution w with p^n units.

Only partial answers to this question are available. For $w = 3$, we already know that $k_{\max} = (p^n - 1)/(p-1)$. For $w = 4$ and $p = 2$, Bose (1947) showed that $k_{\max} = 2^{n-1}$. This is obtained by selecting $a_1, \ldots, a_k$ as the 2^{n-1} elements of $\mathbf{Z}_2^n$ with coordinate sum 1, which are easily seen to constitute a set of triple–wise linearly independent vectors. For fractional replicates of 2^k designs, this has the important

consequence that whenever the number of factors is less than or equal to half the number of experimental units, a resolution 4 design can (and should) be selected, cfr. example 5.6. For $w > 3$ and/or $p > 2$, things are more complicated, but in some cases the following result and its proof may help to bring the problem on a more concrete form.

Call a subgroup G of $\mathbf{Z}_p^k$ *w–heavy*, if it has the property that any of its non–zero elements $g = (g_1, \ldots, g_k) \in G$ has at least w non–zero coordinates.

PROPOSITION 5.4 (Bailey 1985). *A resolution w $\mathbf{Z}_p$–vectorspace design with k factors on p levels, p^n units (p prime, $k > n$), exists if and only if $\mathbf{Z}_p^k$ has a w–heavy subgroup with p^{k-n} elements.*

Proof. Suppose that a design with the given properties exists, and let $a_1, \ldots, a_k \in I^* \approx \mathbf{Z}_p^n$ be the sets of coefficients defining the factors. Let $G \subseteq \mathbf{Z}_p^k$ be the set of $(g_1, \ldots, g_k)$ such that $g_1 a_1 + \cdots + g_k a_k = 0$. The condition that any $w-1$ among $a_1, \ldots, a_k$ are linearly independent obviously implies that any such k–tuple $(g_1, \ldots, g_k)$ must either have all coordinates equal to zero or at least w coordinates non–zero. Thus, G is a w–heavy subgroup, and the dimension of G as a subspace of $\mathbf{Z}_p^k$ is obviously $k - n$ (or more, if $a_1, \ldots, a_k$ do not span I^*, but this is even stronger). Conversely, if G is a w–heavy subspace of $\mathbf{Z}_p^k$ of dimension $k-n$, a design with the desired properties can be constructed as follows. Take

$$I = \{i \in \mathbf{Z}_p^k \mid g^* i = 0 \text{ for any } g \in G\}$$

(g^* denoting transpose to g), and define the factors by (restriction of) the k coordinate projections. This is easily shown to give a design with the desired properties. Intuitively, this is so because the constraints $g^* i = 0$ involve at least w coordinates of i and therefore can never introduce a tie on $w - 1$ of the coordinates.

Single replicate designs with blocking. We shall say very little about this, because a few basic observations show that these designs and their properties are essentially special cases of what has been said about fractional factorials. The only difference is that the design is now assumed to be a complete balanced k–way table (usually with one observation per cell) in some of the factors, namely those that are determined by the treatments. In addition to these treatment factors we have one or more factors determining the partitioning into blocks. The treatment effects that can be estimated in I–stratum in a model with random effect(s) of the block factor(s) are exactly those that can be estimated in a linear model with fixed effect(s) of the block factor(s). This applies also to interactions, with the reservation that a higher order interaction may appear in I–stratum, even when lower order interactions marginal to it do not (see example 4.3 and 4.4). However, this is usually not important for the design of an experiment, since the conditions of interest

there require that a given set of main effects and interaction terms in a *maximal* model formula should be in I–stratum. Thus, any treatment factor or product of treatment factors, that has a non–trivial infimum with a block factor, is essentially 'sacrificed' in the sense that the factor itself or some factor marginal to it will be put in some other stratum than I–stratum.

For designs with more than one block factor, one may of course be interested in a more detailed analysis of the allocation of the factors that are not in I–stratum. This is obvious in example 4.3, where the allocation of the main effect of F to P–stratum may be acceptable, whereas a confounding of F with B would probably not be so. However, for hierarchical block structures this problem is easy to solve (in example 4.3, just ignore the block factor P to analyse the allocation of factors to either B–stratum or the two more accurate strata). Crossed block structures may require further considerations, but we have no general remarks here that lead to more useful methods than the obvious 'trial and error' technique.

References.

Bailey, R.A. (1985).
Factorial designs and Abelian groups.
Linear Algebra and its Applications **70**, 349–368.

Bjældager, P.A.L, Jensen, H.A., Larsen, E., Lauritsen, O.S., Paulev, P.E., Tjur, T. and Uldall, A. (1981).
Inter–laboratory comparison of acid–base variables in human blood and in quality control materials.
Clinica Chimica Acta **116**, 289–300.

Bose, R.C. (1947).
Mathematical theory of the symmetrical factorial design.
Sankhya **8**, 107–166.

Christoffersen, K. (1987).
Gruppefrembragte Designs.
Candidates thesis, University of Copenhagen (manuscript).

Cochran, W.G. and Cox, G.M. (1957).
Experimental Designs.
Wiley.

El Mossadeq, A., Kobilinsky, A. and Collombier, D. (1985).
Construction d'orthogonaux dans les groupes Abéliens finis et confusions d'effets dans les plans factoriels.
Linear Algebra and its Applications **70**, 303–320.

Finney, D.J. (1945).
The fractional replication of factorial arrangements.
Annals of Eugenics **12**, 291–301.

Fisher, R.A. (1942).
The theory of confounding in factorial experiments in relation to the theory of groups.
Annals of Eugenics **11**, 341–353.

Jensen, S.T. (1979). *Varianskomponentmodeller i fuldstændigt balancerede forsøg.* Inst. Math. Stat., University of Copenhagen (manuscript).

Kobilinsky, A. (1985).
Confounding in relation to duality of finite Abelian groups.
Linear Algebra and its Applications **70**, 321–347.

Nelder, J.A. (1965).
The analysis of randomized experiments with orthogonal block structure, I and II.
Proc. R. Soc. Lond. A **283**, 147–178.

Speed, T.P. (1987).
What is analysis of variance?
The Annals of Statistics **15**, 885–910.

Tjur, T. (1984).
Analysis of variance models in orthogonal designs.
International Statistical Review **52**, 33–81.

Yates, F. (1935).
Complex experiments.
J. Roy. Statist. Soc. Suppl. **2**, 181–247.

PREPRINTS 1989

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33 .

No. 1 Bertelsen, Aksel: Asymptotic Expansion of a Complex Hypergeometric Function.

No. 2 Davidsen, Michael and Jacobsen, Martin: Weak Convergence of Twosided Stochastic Integrals, with an Application to Models for Left Truncated Survival Data.

No. 3 Johansen, Søren: Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models.

No. 4 Johansen, Søren and Juselius, Katarina: The Full Information Maximum Likelihood Procedure for Inference on Cointegration - with Applications.

No. 5 Thompson, Steven K.: Adaptive Cluster Sampling.

No. 6 Thompson, Steven K.: Adaptive Cluster Sampling: Designs with Primary and Secondary Units.

No. 7 Thompson, Steven K.: Stratified Adaptive Cluster Sampling.

No. 8 Johansen, Søren: The Power Function of the Likelihood Ratio Test for Cointegration.

No. 9 Andersson, S.A. and Perlman, M.D.: Lattice-Ordered Conditional Independence Models for Missing Data.

PREPRINTS 1990

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 31 35 31 33.

No. 1 Johansen, Søren and Juselius, Katarina: Some Structural Hypotheses in a Multivariate Cointegration Analysis of the Purchasing Power Parity and the Uncovered Interest Parity for UK.

No. 2 Tjur, Tue: Analysis of Variance and Design of Experiments.