STATISTICAL MODELLING

VI. Determining the analysis of variance table

(References:

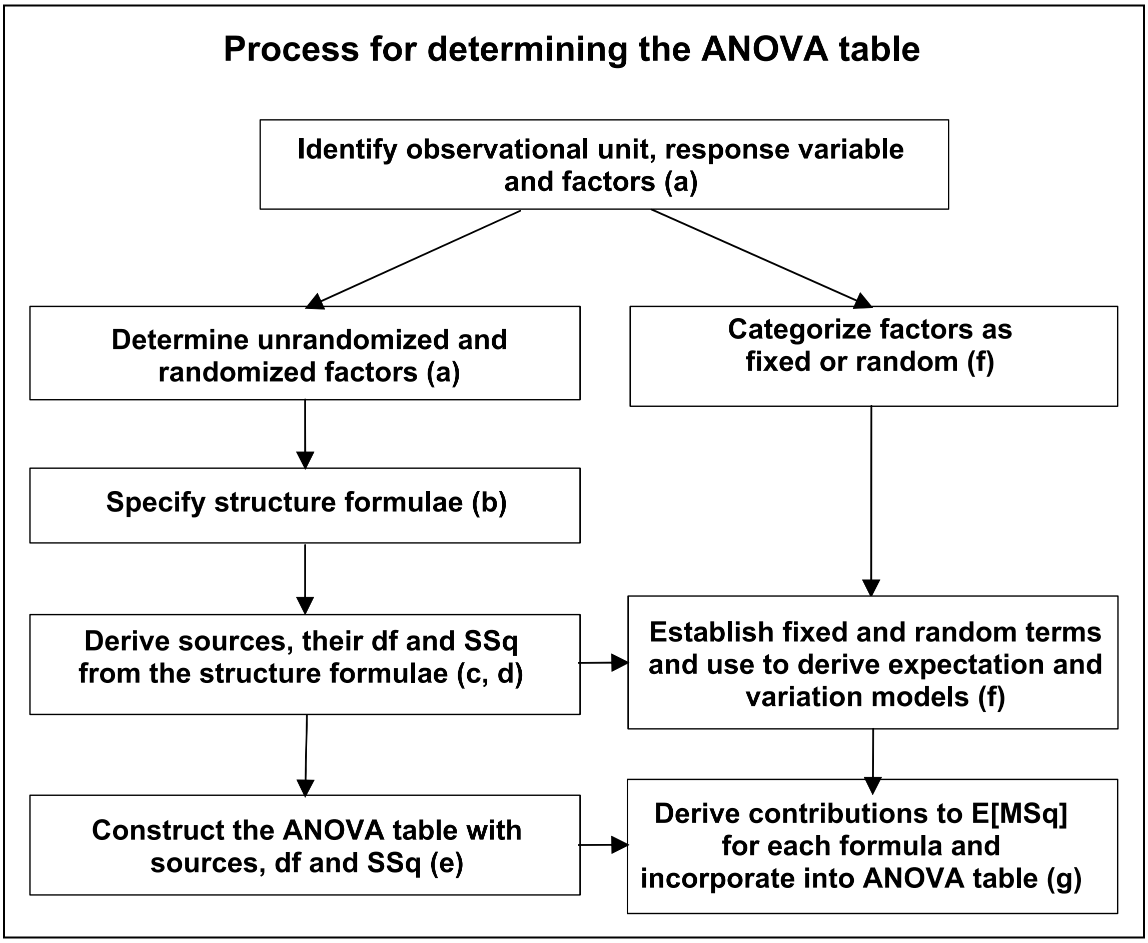
Brien, C.J. (1983) Analysis of variance tables based on experimental structure. *Biometrics*, **39**, 53-59.)

Brien (1989) A model comparison approach to linear models. *Utilitas Mathematica*, **36**, 225-254.)

Lohr, S.L. (1995) Hasse diagrams in statistical consulting and teaching. *The American Statistician*, **49**, 376–81.)

VI.A The procedure

Thus far, for each experiment we have presented an analysis of variance table that allowed us to test hypotheses about whether or not terms should be included in either the expectation or variation models. I have presented the models that might be considered in a particular situation, along with the corresponding analysis of variance table. But how did I arrive at the particular table? Are there some general principles that would allow us to determine the models and the analysis table for any experiment? Unsurprisingly, the answer to the question is "yes". In this chapter we describe a process that is summarized in the following diagram. It works for the vast majority of experimental designs used in practice.


Process for determining the ANOVA table
Identify observational unit, response variable and factors (a)
Determine unrandomized and randomized factors (a)
Categorize factors as fixed or random (f)
Specify structure formulae (b)
Derive sources, their df and SSq from the structure formulae (c, d)
Establish fixed and random terms and use to derive expectation and variation models (f)
Construct the ANOVA table with sources, df and SSq (e)
Derive contributions to E[MSq] for each formula and incorporate into ANOVA table (g)

The procedure that we will use to carry out this process will consist of the following 7 steps as indicated in the diagram:

a) Description of pertinent features of the study
b) The experimental structure
c) Sources derived from the structure formulae
d) Degrees of freedom and sums of squares
e) The analysis of variance table
f) Maximal expectation and variation models
g) The expected mean squares.

This procedure should be carried out when designing an experiment as it allows you to work out the properties of the experiment: in particular, what effects will occur in the experiment and how they will affect each other.

a) Description of pertinent features of the study

The first stage in determining the analysis of variance table is to identify the following features:

1. observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

Definition VI.1: The **observational unit** is the native physical entity which is individually measured. ■

For example, a person in a survey or a run in an experiment.

Definition VI.2: The **response variable** is the measured variable that the investigator wants to see if the factors affect its response. ■

For example, the experimenter may want to determine whether or not there are differences in yield, height, and so on for the different treatments — this is the response variable; that is, the variable of interest or for which differences might exist.

Definition VI.3: The **unrandomized factors** are those factors that would index the observational units if no randomization had been performed. ■

Definition VI.4: The **randomized factors** are those factors associated with the observational unit as a result of randomization. ■

Definition VI.5: The **type of study** is the name of the experimental design or sampling method; for example, CRD, RCBD, LS, SRS, factorial, and so on. ■

One way to decide whether a factor is unrandomized or randomized is to consider what information about the factors would be available if no randomization had been performed. Because the unrandomized factors are those that are innate to the observational units, there is no need to have performed the randomization to know which of their levels are associated with the different observational units. On the other hand, the levels of the randomized factors associated with the different observational units can only be known after the randomization has been performed. This leads to the following rule.

Rule VI.1: To determine whether a factor is unrandomized or randomized, ask the following question:

> For an observational unit, can I identify the levels of that factor associated with the unit if randomization has not been performed?
>
> If yes, then the factor is unrandomized; if no, then it is randomized. ■

Note that for some experiments, to classify factors as unrandomized and randomized is not enough; three classes of factors can be identified. However, for many, two is sufficient.

Features of experiments

Example VI.1 Calf diets

In an experiment to investigate differences between two calf diets the progeny of five dams who had twins were taken and for the two calves of each dam, one was chosen at random to receive diet A and the other diet B. The weight gained by each calf in the first 6 months was measured.

What is the observational unit? **Ans.** A calf

The observations for the experiment might be:

Observation	Dam	Calf	Diet	Weight Gain
1	1	1	A	125
2	1	2	B	.
3	2	1	B	.
4	2	2	A	.
5	3	1	B	.
6	3	2	A	.
7	4	1	A	.
8	4	2	B	.
9	5	1	A	.
10	5	2	B	.

What are the variables? **Ans.**

What is the response variable? **Ans.**

Now to determine whether the factors are unrandomized or randomized:

Is Dam randomized or unrandomized? **Ans.**

Is Calf randomized or unrandomized? **Ans.**

Is Diet randomized or unrandomized? **Ans.**

And what is the type of study? **Ans.** RCBD

In summary, the features of the study are:

1. Observational unit ______________________________
2. Response variable ______________________________
3. Unrandomized factors ______________________________
4. Randomized factors ______________________________
5. Type of study ______________________________

NOTE: Dam and Calf uniquely identify the observations in that there are no two observational units with the same combination of these two factors (for example, 2,1). ■

Example VI.2 Plant yield

Consider a CRD experiment consisting of 5 observations, each observation being the yield of a single plot which had one of three varieties applied to it.

The results of the experiment are as follows:

Plot	Variety	Yield
1	A	213
2	C	256
3	A	225
4	B	183
5	B	201

What are the features of the study?

1. Observational unit ______________________________

   Variables (including factors) are?

2. Response variable ______________________________
3. Unrandomized factors ______________________________
4. Randomized factors ______________________________
5. Type of study ______________________________

Note that two of the levels of the factor Variety are replicated twice and the third only once.

Also note that the features are reflected in the analysis of variance table, particularly the unrandomized and randomized factors.

		Source
unrandomized	➔	Plots
randomized	➔	Variety
		Residual

■

The number of unrandomized factors is a characteristic of each design.

Example VI.3 Pollution effects of petrol additives

Consider the experiment to investigate the reduction in the emission of nitrous oxides resulting from the use of four different petrol additives. Four cars and four drivers are employed in this study with additives being assigned to a particular driver-car combination according to a Latin square. The arrangement and data are as follows:

		Car			
		1	2	3	4
		B	D	C	A
	I	20	20	17	15
		A	B	D	C
	II	20	27	23	26
Driver		D	C	A	B
	III	20	25	21	26
		C	A	B	D
	IV	16	16	15	13

(Additives: A, B, C, D)

What are the features of the study?

1. Observational unit ____________________
2. Response variable ____________________
3. Unrandomized factors ____________________
4. Randomized factors ____________________
5. Type of study ____________________

■

Features of surveys

Example VI.4 Vineyard sampling

A vineyard of 125 vines is sampled at random with 15 vines being selected at random and the yields measured.

What are the features of the study?

1. Observational unit ____________________

Variables (including factors) are?

2. Response variable ____________________
3. Unrandomized factors ____________________
4. Randomized factors ____________________
5. Type of study ____________________

■

Example VI.5 Smoking effect on blood cholesterol

Consider an observational study to investigate the effect of smoking on blood cholesterol by observing 30 patients and recording whether they smoke tobacco and measuring their blood cholesterol. Suppose it happens that 11 patients smoke and 19 patients do not smoke.

What are the features of the study?

1. Observational unit ______________________________

   Variables (including factors) are?

2. Response variable ______________________________
3. Unrandomized factors ______________________________
4. Randomized factors ______________________________
5. Type of study ______________________________

■

In determining the unrandomized and randomized factors it is most important to distinguish between randomization and random sampling. They both are based on the same procedure, that is, obtaining a set of random numbers. However, their purposes are quite different:

1. randomization = random selection to assign;
2. random sampling = random selection to observe a fraction of a wholly observable population.

It is not surprising that surveys do not contain randomized factors, since they do not involve randomization.

Remember the crucial question is: If I take an observational unit, can I tell which level of this factor is associated with that unit without doing the randomization? If yes, then unrandomized, otherwise randomized.

b) The experimental structure

Having determined the unrandomized and randomized factors, one next determines the experimental structure.

Rule VI.2: Determine the experimental structure by

1. describing the nesting and crossing relationships between the unrandomized factors in the experiment,
2. describing the nesting and crossing relationships between
   i) the randomized factors, and
   ii) the randomized and the unrandomized factors, if any.

The numbers of levels of the factors are placed in front of the names of the factors. ■

Most often it will be assumed that the effects of the randomized factors are approximately the same for each observational unit so that the unrandomized and randomized factors can be treated as independent. Consequently step ii) will usually not be required.

From the above rule it is clear that the nesting (and crossing) plays a central role in determining the experimental structure. In our discussion so far it has been suggested that the nesting results from the way in which the randomization is done: randomize Units within Blocks and Units is nested within Blocks. However, this is not the whole story as two factors may be intrinsically nested.

Definition VI.6: Two factors are **intrinsically nested** if units with the same level of the nested factor, but different levels of the nesting factor, have no apparent characteristic in common. ■

Definition VI.7: Two factors are **intrinsically crossed** if units with the same level of one factor, but different levels of the second factor, have a common characteristic associated with the first factor. ■

A slash ('/') will be placed between two factors to indicate that they are nested, the nested factor being on the right of the slash and the nesting factor on the left. If two factors are crossed, an asterisk ('*') will be placed between them. Two other operators that sometimes occur in structure formulae are the wedge ('∧') and the plus (+). A wedge placed between two factors signifies all observed combinations of the levels of the two factors and a plus would indicate that the two factors are to be considered independently.

Note that the order of precedence of the operators in a structure formula is '∧', '/', '*' and '+', with '∧' having the highest precedence. For example, the structure formula A * B / C is the same as A * (B / C).

Our definitions evidently rely on whether there is a connection between the same levels of a factor when they involve different levels of the second factor. This is illustrated with the following examples in which the units are students.

Example VI.6 Student height — unknown age

Suppose I have three students of each of the two sexes and have measured their heights. Thus, as illustrated in the table below, the two factors indexing the six observations are Sex, with 2 levels, and Student, with 3 levels.

		Student		
		1	2	3
Sex	M	y_1	y_2	y_3
	F	y_4	y_5	y_6

Now take level 1 of Student. There are two students with this level. However, there is nothing in the table above that indicates they have anything in common, save that they happen to share the label 1. This is inconsequential; indeed, as far as we know, the labelling of students within each Sex is arbitrary. We conclude the two students with the same level of Student apparently have no common characteristic. The factor Student is intrinsically nested within Sex. ■

Example VI.7 Student height — known age

Suppose I have six students, three of each sex and that the three students of each sex consist of 1 student from each of 3 different age groups. Thus, as illustrated in the table below, the two factors indexing the six heights are Sex, with 2 levels, and Age, with 3 levels.

			Age	
		18	19	20
Sex	M	y_1	y_2	y_3
	F	y_4	y_5	y_6

Now take the level 18 of Age. There are two students with this level and they are of different Sexes. So, even though they are of different Sexes, these two students share the characteristic that they are 18. Thus the two factors Sex and Age are crossed. ■

Example VI.1 Calf diets (continued)

The factors were designated:

3. Unrandomized factors – Dam, Calf
4. Randomized factors – Diet

So are the unrandomized factors nested? That is, 'Do we have information that connects a calf from one dam with any of the calves from another dam?' **Answer** No. So they are nested. In fact, Calf is nested within Dam and Dam nests Calf. This is written symbolically as Dam/Calf.

Thus the experimental structure for this experiment is:

Structure	Formula
unrandomized	*5* Dam/*2* Calf
randomized	*2* Diet

■

Example VI.3 Pollution effects of petrol additives (continued)

The factors were designated:

3. Unrandomized factors – Driver, Car
4. Randomized factors – Additive

Are the unrandomized factors nested? That is, 'Do we have information that connects one of the drivers of a car with a driver from another car?' **Answer** Yes, one of the four from each of the other cars is the same driver. They are intrinsically crossed and this is written symbolically as Driver*Car.

Thus the experimental structure for this experiment is:

Structure	Formula
unrandomized	*4* Driver**4* Car
randomized	*4* Additive

■

Notes:

- The numbers of the levels of the factors are placed in front of the names of the factors.
- A factor will be nested within another either because they are intrinsically nested or because the randomization employed requires that they be so regarded. Hence, for two factors to be crossed requires not only that they are intrinsically crossed as in the definition, but also that the randomization employed respect this relationship. Thus, even if the same four drivers are used with each car, the relationship would be nested if the additive to be used had been randomized to the drivers within each car. Clearly, in the latter situation, we would have an RCBD not a Latin square. The following is an example of what might have been obtained if an RCBD, with Cars (columns) as blocks, had been employed.

		Car			
		1	2	3	4
	I	B	A	A	B
	II	A	D	C	A
Driver	III	D	C	D	D
	IV	C	B	B	C

(Additives: A, B, C, D)

In this design, each additive is used in a car once and only once. The same cannot be said of Drivers. The appropriate structure for this design would be Car/Driver. An RCBD would be used in preference to the Latin Square if it was thought that there was little difference between the Drivers. The RCBD has the advantage of larger Residual degrees of freedom.

■

- Some notes, including the numbers of unrandomized factors for the different designs, are as follows:

i) The set of unrandomized factors will uniquely identify the observations.
ii) Surveys have only unrandomized factors.
iii) CRD — only one unrandomized factor, the only design that does.
RCBD (& BIBD) — two unrandomized factors, one of which is nested within the other.
LS (& YS) — two unrandomized factors which are crossed.

c) Sources derived from the structure formulae

Having determined the experimental structure, the next step is to expand the formulae to obtain the sources that are to be included in the analysis of variance table.

Rule VI.3: The rules for expanding structure formulae involving two factors A and B are:

A*B = A + B + A#B

where A#B represents the **interaction of A and B** (more about interaction later)

A/B = A + B[A]

where B[A] represents the **nested effects of B within A**, that is, differences between B within A.

More generally, if L and M are two formulae, then

L*M = L + M + L#M

where L#M is the sum of all pairwise combinations of sources in L with sources in M

and L/M = L + M[*gf*(L)]

where *gf*(L) is the generalized factor (see definition VI.8 in section D, *Degrees of freedom*) formed from the combination of all factors in L. ■

Example VI.1 Calf diets (continued)

Dam/Calf = Dam + Calf[Dam]

where Calf[Dam] represents the differences between calves with the same Dam. ■

Example VI.3 Pollution effects of petrol additives (continued)

Driver*Car = Driver + Car + Driver#Car

where Driver#Car represents the interaction between Driver and Car or the extent to which Driver differences change from Car to Car. ■

d) Degrees of freedom and sums of squares

The degrees of freedom for an analysis of variance can be calculated with the aid of Hasse diagrams for Generalized-Factor Marginalities for each structure formula.

Definition VI.8: A **generalized factor** is the factor formed from several (original) factors and whose levels are the combinations that occur in the experiment of the levels of the constituent factors. The generalized factor is the "meet" of the constituent factors and it is written as the list of constituent factors separated by "wedges" or "meets" ("$\wedge$"). For convenience we include ordinary factors amongst the set of generalized factors for an experiment. ■

There is a generalized factor corresponding to each source obtained from a structure formulae — it consists of the factors in the source.

Example VI.3 Pollution effects of petrol additives (continued)

Consider the sources Car and Driver#Car from the Latin square example. The two generalized factors corresponding to these sources are Car and Driver∧Car. Driver∧Car is a factor with $4 \times 4 = 16$ levels, one for each combination of Driver and Car. ■

Now the Hasse diagram for Generalized-Factor Marginalities displays the marginality relationships between the generalized factors corresponding to the sources from a structure formula. We have previously discussed the marginality relationship as it applies to models. It is basically the same concept here, except that it applies to generalized factors which are potentially single indicator-variable terms to be included in a model.

Definition VI.9: One generalized factor, V say, is **marginal** to another, Z say, if the factors in the marginal generalized factor are a subset of those in the other and this will occur irrespective of the replication of the levels of the generalized factors. We write $V \leq Z$. ■

Note that the marginality relationship is not symmetric — it is directional, like the less-than relation that we use to symbolize it. So while $V \leq Z$, Z is not marginal to V unless $V = Z$. Of course, a generalized factor is marginal to itself.

Example VI.3 Pollution effects of petrol additives (continued)

Car is marginal to Driver∧Car (Car < Driver∧Car) as the factor in Car is a subset of those in Driver∧Car. ■

Rule VI.4: The Hasse diagrams for Generalized-Factor Marginalities for a structure formula are formed by placing generalized factors above those to which they are marginal and connecting them by an upwards arrow. Alongside each generalized factor is added its source (to save space use only 1st letter for each factor in source). The Universe factor, connecting all units that occurred in the experiment, is included at the top of the diagram. Under a generalized factor is written its number of levels. Under a source is written its degrees of freedom: it is calculated as the difference between the entry for its generalized factor and the sum of the degrees of freedom of all sources whose generalized factors are marginal to the current generalized factor.■

In constructing a Hasse diagram, having begun with the Universe factor, that involves no named factor, next consider generalized factors consisting of a single factor, followed by those with 2 factors, then those with 3 factors and so on. This strategy is appropriate because there must be less factors in a marginal generalized factor.

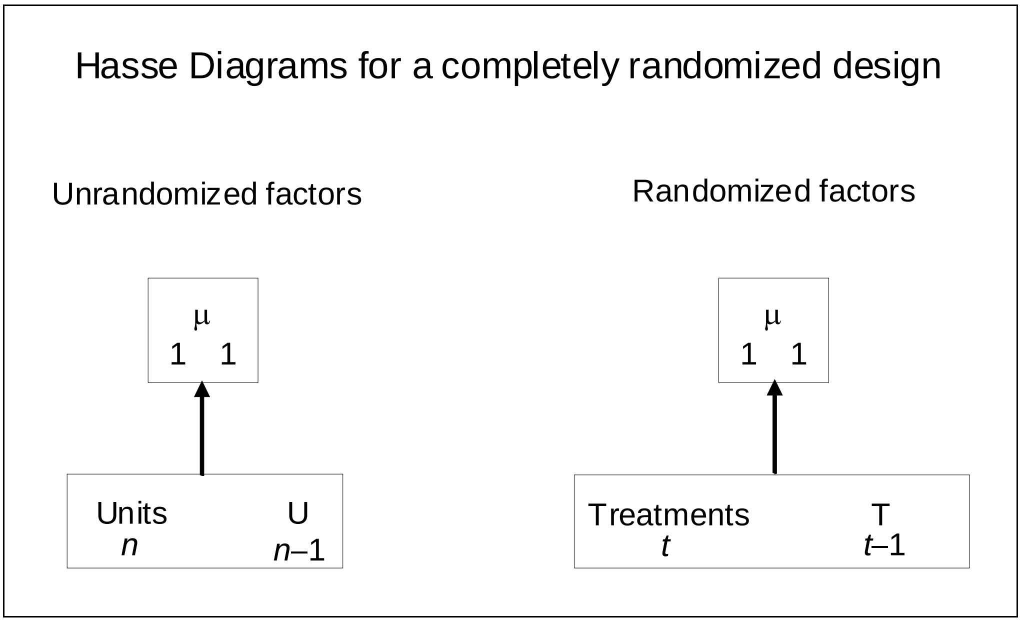
Rule VI.5: When all the factors are crossed, the degrees of freedom of any source can be calculated directly. The rule for doing this is:

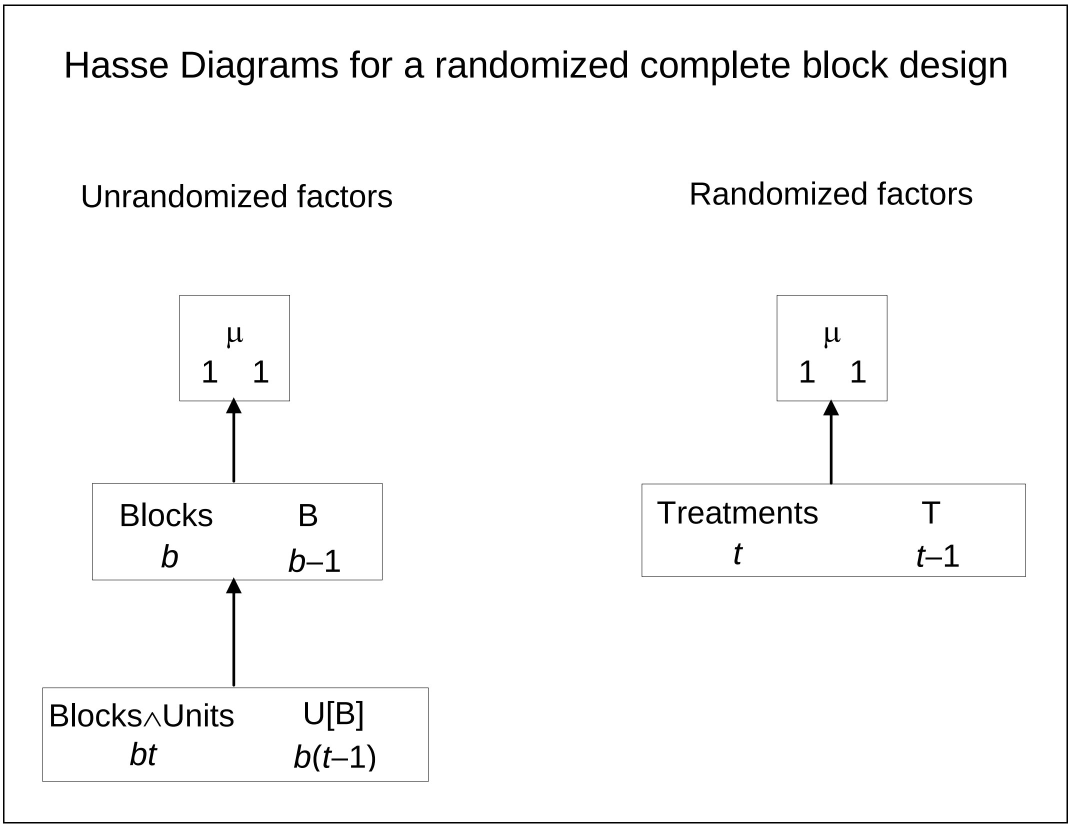
> For each factor in the source, calculate the number of levels minus one and multiply these together. ■

Example VI.3 Pollution effects of petrol additives (continued)

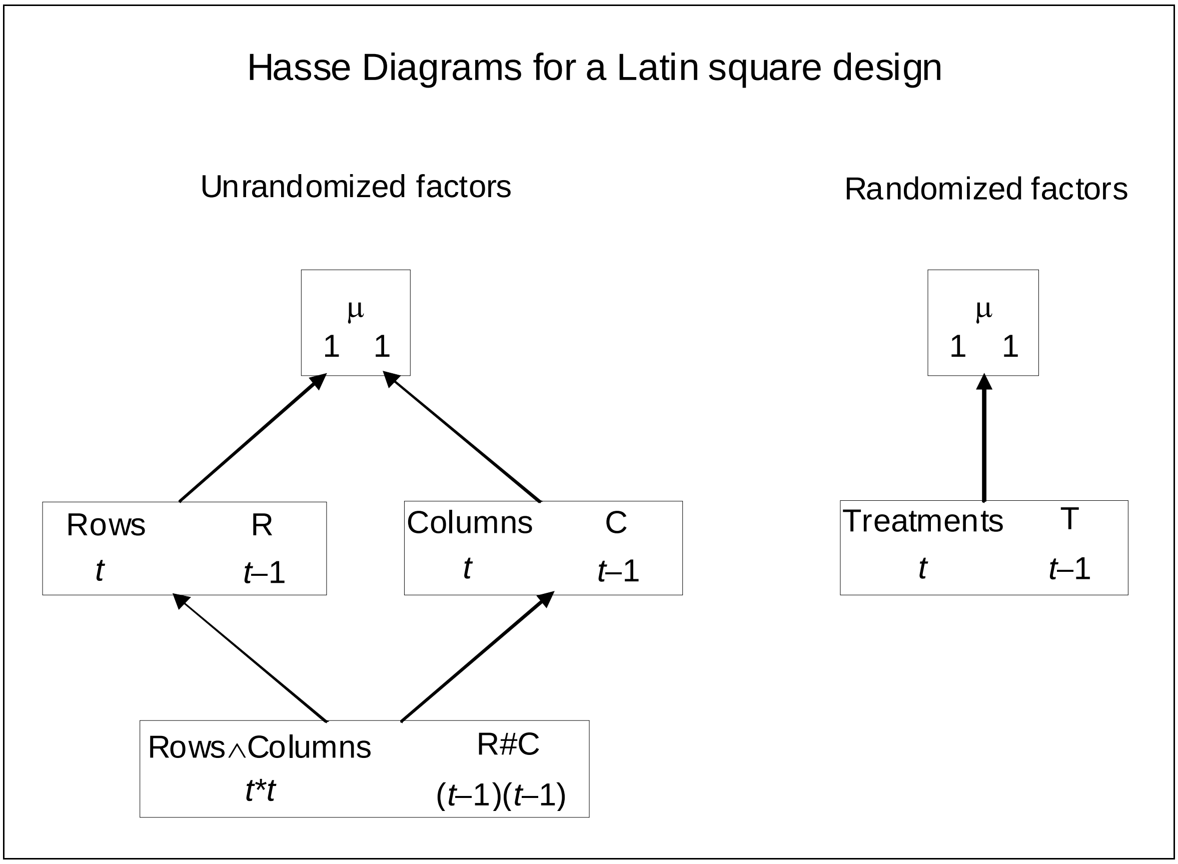
Since both Car and Driver have 4 levels, the degrees of freedom of Driver#Car is (4–1)(4–1) = 3^2 = 9. ■

The Hasse diagrams for some of the experiments you have so far encountered are as follows:





Note that the source corresponding to the generalized factor Blocks∧Units is Units[Blocks]. Blocks∧Units has *bt* levels and Units[Blocks] has $b(t-1)$ degrees of freedom.



Hasse Diagrams for a Latin square design

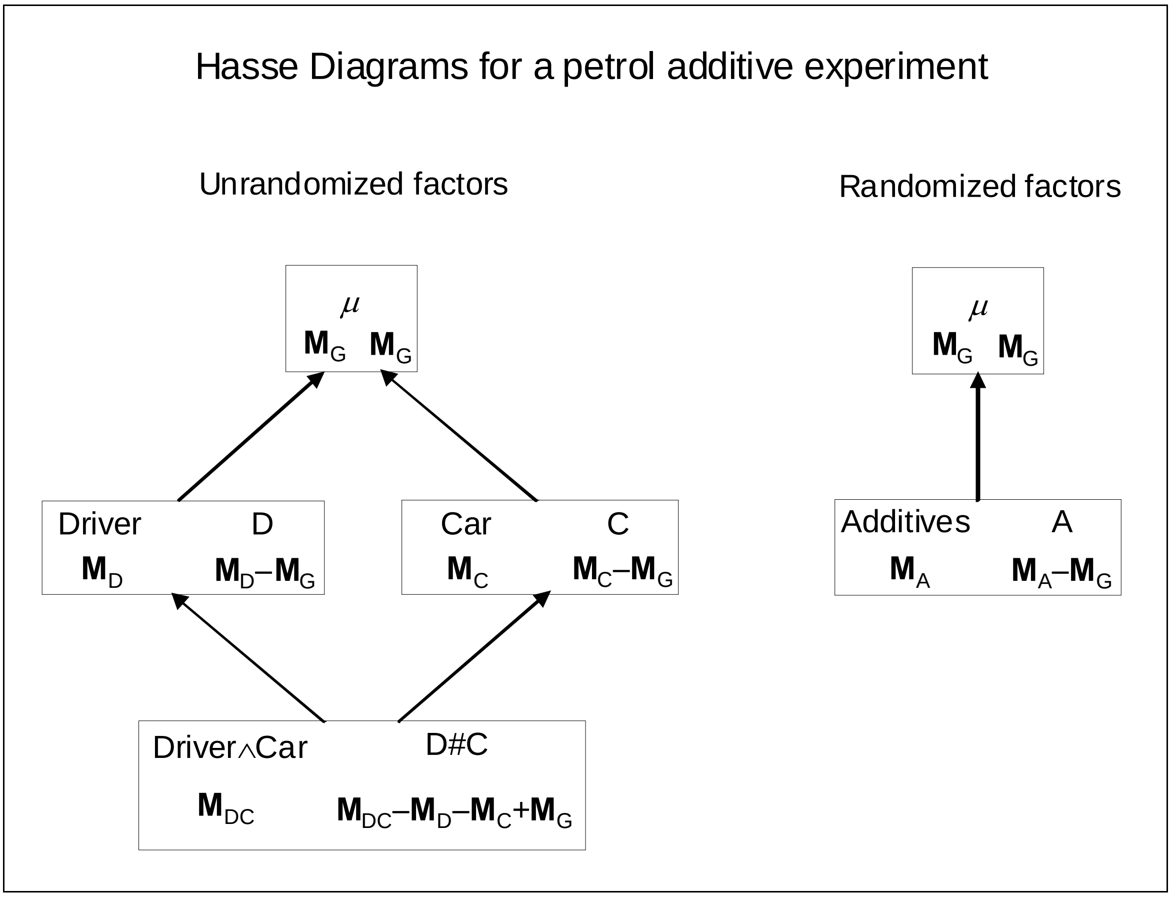
Note that the source corresponding to the generalized factor Rows∧Columns is Rows#Columns. Rows∧Columns has t^2 levels and Rows#Columns has $(t-1)^2$ degrees of freedom.

Now, we can easily write down an expression for the sums of squares in the analysis of variance table in terms of a quadratic form with the appropriate **Q** matrix as the matrix of the quadratic form. However, this still does not tell us how to compute the vector, **QY**, whose sums of squares we will obtain. To know this we need an expression for **Q** in terms of mean operator matrices, **M**. The following rule allows one to derive these easily:

Rule VI.6: There is a mean operator (**M**) and a quadratic-form (**Q**) matrix for each generalized factor and source, respectively, obtained from the structure formulae. To obtain expressions for the **Q** matrices in terms of the **M** matrices, take the Hasse diagram of generalized factors for the formula and, for each generalized factor, replace its number of levels combinations with its **M** matrix. Then, the expression for the **Q** matrix for a source replaces its degrees of freedom in the Hasse diagram. To work it out take the **M** matrix for its generalized factor and subtract all expressions for **Q** matrices of sources whose generalized factors are marginal to the generalized factor for the source whose expression you are deriving. ■

Example VI.3 Pollution effects of petrol additives (continued)

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

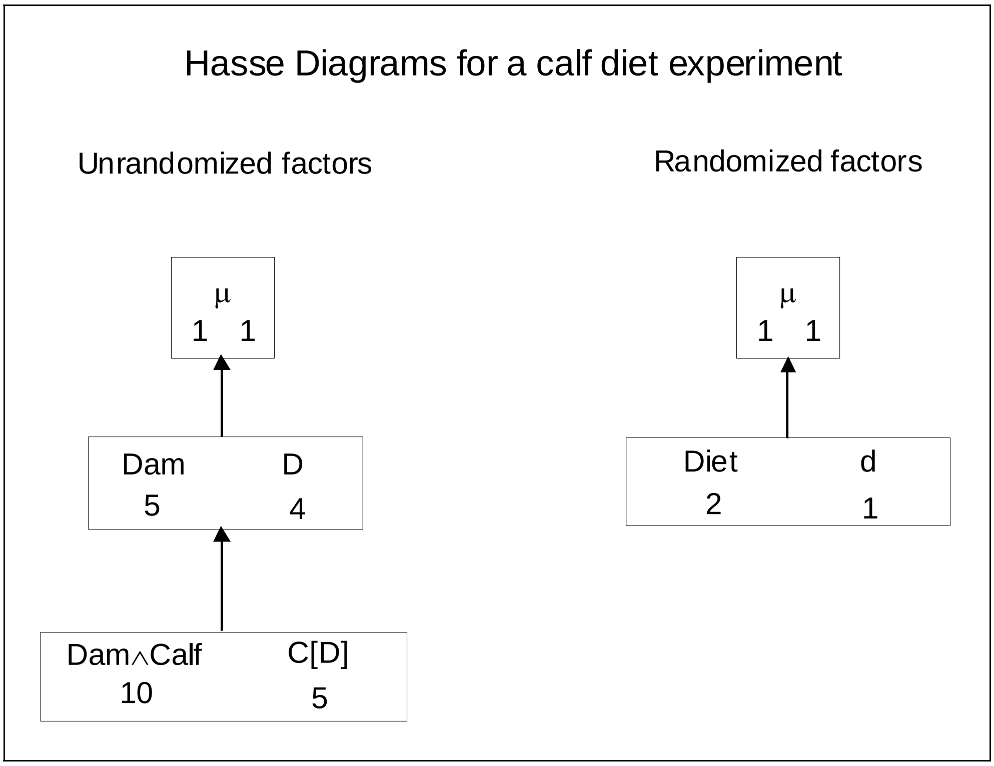


Thus the estimators of the quadratic forms or sums of squares on which the analysis of variance will be based are $\mathbf{Y}'\mathbf{Q}_{\mathrm{D}}\mathbf{Y}$, $\mathbf{Y}'\mathbf{Q}_{\mathrm{C}}\mathbf{Y}$, $\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Q}_{\mathrm{A}}\mathbf{Y}$ where $\mathbf{Q}_{\mathrm{D}}=\mathbf{M}_{\mathrm{D}}-\mathbf{M}_{\mathrm{G}}$, $\mathbf{Q}_{\mathrm{C}}=\mathbf{M}_{\mathrm{C}}-\mathbf{M}_{\mathrm{G}}$, $\mathbf{Q}_{\mathrm{DC}}=\mathbf{M}_{\mathrm{DC}}-\mathbf{M}_{\mathrm{D}}-\mathbf{M}_{\mathrm{C}}+\mathbf{M}_{\mathrm{G}}$ and $\mathbf{Q}_{\mathrm{A}}=\mathbf{M}_{\mathrm{A}}-\mathbf{M}_{\mathrm{G}}$. Note that we do not include a source for the intercept or grand-mean term in the analysis of variance table and so the corresponding quadratic-form estimator $\mathbf{Y}'\mathbf{Q}_{\mathrm{G}}\mathbf{Y}$ is not required.

Now we have these expressions it is possible to find an expression for a sum of squares as a summation formula of elements of a vector. For example, consider $\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y}$. It is the sum of squares of $\mathbf{Q}_{\mathrm{DC}}\mathbf{Y}=\left(\mathbf{M}_{\mathrm{DC}}-\mathbf{M}_{\mathrm{D}}-\mathbf{M}_{\mathrm{C}}+\mathbf{M}_{\mathrm{G}}\right)\mathbf{Y}=\mathbf{Y}-\bar{\mathbf{D}}-\bar{\mathbf{C}}+\bar{\mathbf{G}}$ and so the element for the *i*th Driver and the *j*th Car is $Y_{ij}-\bar{D}_i-\bar{C}_j+\bar{G}$ and the Driver#Car sums of squares is $\sum_{i=1}^{4}\sum_{j=1}^{4}\left(Y_{ij}-\bar{D}_i-\bar{C}_j+\bar{G}\right)^2$ ■

Example VI.1 Calf diets (continued)

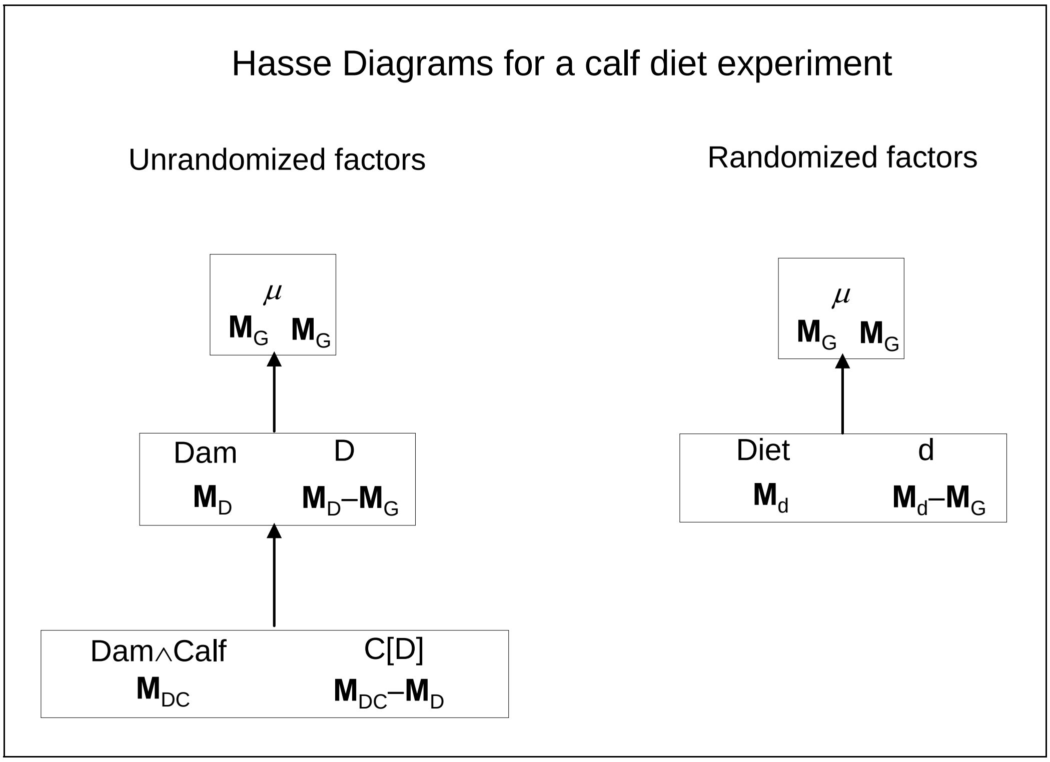
The Hasse diagrams, with degrees of freedom, for this study are:



Hasse Diagrams for a calf diet experiment

Note the use of lower case "d" for Diet to distinguish it from "D" for Dam.

The Hasse diagrams, with **M** and **Q** matrices, for this study are:



Hasse Diagrams for a calf diet experiment

Unrandomized factors
Randomized factors
μ
$\mathbf{M}_G$ $\mathbf{M}_G$
Dam
D
$\mathbf{M}_D$
$\mathbf{M}_D - \mathbf{M}_G$
Dam∧Calf
C[D]
$\mathbf{M}_{DC}$
$\mathbf{M}_{DC} - \mathbf{M}_D$
μ
$\mathbf{M}_G$ $\mathbf{M}_G$
Diet
d
$\mathbf{M}_d$
$\mathbf{M}_d - \mathbf{M}_G$

Thus the estimators of the quadratic forms or sums of squares on which the analysis of variance will be based are $\mathbf{Y}'\mathbf{Q}_D\mathbf{Y}$, $\mathbf{Y}'\mathbf{Q}_{DC}\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Q}_d\mathbf{Y}$ where $\mathbf{Q}_D = \mathbf{M}_D - \mathbf{M}_G$, $\mathbf{Q}_{DC} = \mathbf{M}_{DC} - \mathbf{M}_D$ and $\mathbf{Q}_d = \mathbf{M}_d - \mathbf{M}_G$. ■

e) The analysis of variance table

At this step an analysis of variance table with the sources, their degrees of freedom and quadratic-form estimators is formulated. Use following rule, although it only specifies the quadratic-form estimators for orthogonal experiments.

Rule VI.7: The analysis of variance table is formed by:

1. Listing down all the unrandomized sources in the Source column, and their degrees of freedom in the df column and the quadratic forms in the SSq column.
2. Then the randomized sources are placed indented under the unrandomized sources with which they are confounded, along with their degrees of freedom and, if the design is orthogonal, their quadratic forms.
3. Residual sources are added to account for the left-over portions of unrandomized sources and their degrees of freedom and quadratic forms are computed by difference. For orthogonal experiments, the matrix of the Residual quadratic form is the difference of the matrices of the quadratic forms from which it is computed. ■

Example VI.1 Calf diets (continued)

	Source	df	SSq
unrandomized ➔	Dam	4	$\mathbf{Y}'\mathbf{Q}_{\mathrm{D}}\mathbf{Y}$
	Calf[Dam]	5	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y}$
randomized ➔	Diet	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{d}}\mathbf{Y}$
	Residual	4	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}}\mathbf{Y}$

Now $\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}}\mathbf{Y} = \mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y} - \mathbf{Y}'\mathbf{Q}_{\mathrm{d}}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{Q}_{\mathrm{DC}} - \mathbf{Q}_{\mathrm{d}}\right)\mathbf{Y}$. That is, $\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}} = \mathbf{Q}_{\mathrm{DC}} - \mathbf{Q}_{\mathrm{d}}$ and it can be proven that this matrix is symmetric and idempotent. ■

Example VI.3 Pollution effects of petrol additives (continued)

	Source	df	SSq
	Driver	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{D}}\mathbf{Y}$
unrandomized ➔	Car	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{C}}\mathbf{Y}$
	Driver#Car	9	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y}$
randomized ➔	Additive	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{A}}\mathbf{Y}$
	Residual	6	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}}\mathbf{Y}$

Now $\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}}\mathbf{Y} = \mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y} - \mathbf{Y}'\mathbf{Q}_{\mathrm{A}}\mathbf{Y} = \mathbf{Y}'\left(\mathbf{Q}_{\mathrm{DC}} - \mathbf{Q}_{\mathrm{A}}\right)\mathbf{Y}$. That is, $\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}} = \mathbf{Q}_{\mathrm{DC}} - \mathbf{Q}_{\mathrm{A}}$ and it can be proven that this matrix is symmetric and idempotent. ■

f) Maximal expectation and variation models

In the discussion of the analysis of experiments I have been writing down expectation and variation models as the sum of a set of indicator-variable terms, these terms being derived from the generalized factors. The rules for obtaining these terms are as follows:

Rule VI.8: To obtain the terms in the expectation and variation model:

1. Designate each original factor in the experiment as either fixed or random.

2. Determine whether a generalized factor is a potential expectation or variation term as follows: generalized factors that involve *only fixed* (original) factors are potential expectation terms and generalized factors that contain *at least one random* (original) factor will become variation terms. If there is no unrandomized factor that has been classified as random, the term consisting of all unrandomized factors will be designated as random.

3. The maximal expectation model is then the sum of all the potential expectation terms except those that are marginal to another expectation term; if there are no expectation terms, the model consists of a single term for the grand mean.

4. The maximal variation model is the sum of all the variation terms. ■

So the first step in determining the model for an experiment is to classify **all** the factors in the experiment as fixed or random.

Definition VI.10: A factor will be designated as **random** if it is anticipated that the distribution of effects associated with the population set of levels for the factor can be described using a probability distribution function. ■

Definition VI.11: A factor will be designated as **fixed** if it is anticipated that a probability distribution function will not provide a satisfactory description the set of effects associated with the population set of levels for the factor. ■

So when we are deciding whether a factor is random or fixed, we are choosing which mathematical model best describes the population distribution for the response variable. The above definitions provide us with a basis for making the choice. One needs to consider the population set of levels and how the set of response variable effects corresponding to this set of levels might behave.

Of the two designations, that of random factors is the more restrictive in that, while the associated effects should be able to be viewed as being generated randomly, they must be "well behaved" to the extent that they display a pattern that conforms to a regularly-shaped probability distribution function. This also implies that, for a random factor, the number of levels in the population must be large. On the other hand, a fixed factor might have a small or a large number of levels in the population and the effects associated with fixed factors may vary arbitrarily — that is they may or may not display some pattern such as i) conforming to a distribution or ii) showing a systematic trend across the levels of the corresponding factor.

It is clear that, if it is anticipated that the effects of a factor will display a systematic trend, then this must be modelled using an expectation model, perhaps involving polynomial submodels. Also, the factor for a small set of treatments that are to be compared would be modelled using a term in the expectation model. In both cases, it seems inappropriate to model the effects as being, say normally distributed — the pattern in the treatment means may well be quite irregular and there is no interest in the form of this distribution.

However, the effects from individual units treated alike (for example, animals, plots of land, runs of a chemical reactor) are anticipated to arise randomly and the effects could well follow a probability distribution, say a normal distribution. Hence it is appropriate to model them via a term in the variation model.

Notwithstanding any of this, you must always model terms to which other terms have been randomized as random effects. For example, because Treatments are randomized to Units (within Blocks) in an RCBD, Units must be a random factor.

The definitions of fixed and random factors given above are not universally used. Take for example Montgomery (DOE, 2000, p.511). He suggests that fixed-factors apply when "the levels of the factors used by the experimenter were the specific levels of interest", although he does recognize that for a quantitative factor (such as temperature) we might be interest in the "response over the region spanned by the factor levels used in the experimental design". On the other hand, random factors apply when "the factor levels are chosen at random from a larger population of possible levels, and the experimenter wishes to draw conclusions about the entire population of levels, not just those used in the experimental design."

The following argument demonstrates the unsatisfactory nature of this basis for making the distinction between fixed and random factors. If I randomly sample a set of temperatures, perhaps because they are not under control, does this mean that the factor Temperature is to be classified as random? The definition given above would lead you to say no, because the temperature effects are likely to display some systematic trend (low order polynomial?) and therefore are not going to conform to a distribution. It seems that the definition given above is based on the essence of the difference between them — the property that decides whether they should contribute to the expectation or variation component of the model.

However, this is not to say that the fact, as to whether or not a random sample of a set of levels has been obtained, is not useful in determining the type of a factor. As mentioned above, any random factor needs to involve, effectively, a random sample from a large population of levels. So if it is not, it cannot be a random factor. But, it is clearly not true that any random sample of levels leads to a random factor.

In practice

- Random if
  1. large number of population levels and
  2. random behaviour
- Fixed if
  i. small or large number of population levels and
  ii. systematic or other non-random behaviour

It often happens, but not always, that all unrandomized factors are designated as random and so all terms involving them occur in the variation model — all the randomized factors are designated as fixed and all terms involving them, minus marginal terms, occur in expectation models.

Example VI.1 Calf diets (continued)

One maximal model for this experiment, based on the usual maximal model for an RCBD, is

$$E\left[Y_{ijk}\right] = \beta_i + \tau_k \text{, var}\left[Y_{ijk}\right] = \sigma^2_{\text{DC}} \text{ and cov}\left[Y_{ijk}, Y_{i'j'k'}\right] = 0, \quad i \neq i', j \neq j' \text{ or } k \neq k'$$

or, in matrix form,

$$\boldsymbol{\psi} = E\left[\mathbf{Y}\right] = \mathbf{X}_{\text{D}}\boldsymbol{\beta} + \mathbf{X}_{\text{d}}\boldsymbol{\tau} \text{ and var}\left[\mathbf{Y}\right] = \sigma^2_{\text{DC}}\mathbf{I}_{10} = \sigma^2_{\text{DC}}\mathbf{M}_{\text{DC}}\,.$$

where σ^2_{DC} is equivalent to what has been previously designated σ^2.

This will be expressed symbolically as

ψ = E[Y] = Diet + Dam

and V[Y] = Dam∧Calf.

However, this is not the only possible maximal model for the experiment. An alternative model would be:

$$E\left[Y_{ijk}\right] = \tau_k \text{, var}\left[Y_{ijk}\right] = \sigma^2_{\text{D}} + \sigma^2_{\text{DC}}\text{,}$$

$$\text{cov}\left[Y_{ijk}, Y_{ij'k'}\right] = \sigma^2_{\text{D}}, \quad j \neq j' \;\left(\text{so } k \neq k'\right) \text{ and cov}\left[Y_{ijk}, Y_{i'j'k'}\right] = 0, \quad i \neq i'$$

or, in matrix form,

$$E\left[\mathbf{Y}\right] = \mathbf{X}_{\text{d}}\boldsymbol{\tau} \text{ and } \mathbf{V} = \sigma^2_{\text{DC}}\mathbf{I}_{10} + \sigma^2_{\text{D}}\left(\mathbf{I}_5 \otimes \mathbf{J}_2\right) = \sigma^2_{\text{DC}}\mathbf{M}_{\text{DC}} + 2\sigma^2_{\text{D}}\mathbf{M}_{\text{D}}\,.$$

This will be written symbolically as

ψ = E[Y] = Diet

and V[Y] = Dam + Dam∧Calf.

Note that this model is the one for which all the unrandomized factors are regarded as random and the randomized factors as fixed.

The model for the variance is that the variability of a particular observation is due to the individual itself and to its dam. Further, there is covariance between individuals from the same dam. The variance matrix is as follows:

	Dam									
	I		II		III		IV		V	
Calf	1	2	1	2	1	2	1	2	1	2
1	$\sigma^2_{DC} + \sigma^2_D$	σ^2_D	0	0	0	0	0	0	0	0
2	σ^2_D	$\sigma^2_{DC} + \sigma^2_D$	0	0	0	0	0	0	0	0
1	0	0	$\sigma^2_{DC} + \sigma^2_D$	σ^2_D	0	0	0	0	0	0
2	0	0	σ^2_D	$\sigma^2_{DC} + \sigma^2_D$	0	0	0	0	0	0
1	0	0	0	0	$\sigma^2_{DC} + \sigma^2_D$	σ^2_D	0	0	0	0
2	0	0	0	0	σ^2_D	$\sigma^2_{DC} + \sigma^2_D$	0	0	0	0
1	0	0	0	0	0	0	$\sigma^2_{DC} + \sigma^2_D$	σ^2_D	0	0
2	0	0	0	0	0	0	σ^2_D	$\sigma^2_{DC} + \sigma^2_D$	0	0
1	0	0	0	0	0	0	0	0	$\sigma^2_{DC} + \sigma^2_D$	σ^2_D
2	0	0	0	0	0	0	0	0	σ^2_D	$\sigma^2_{DC} + \sigma^2_D$

The first model corresponds to Dam being fixed, whereas the second model corresponds to Dam random. Other models would also have Diets as random. Which model will we use here?

It is most unlikely that the two observed levels will be thought of as being representative of a very large group of Diets, and as well we anticipate arbitrary differences between them, so it is designated to be a fixed factor. However, we can envisage a very large population of dams that the five observed dams represent and it is likely that dam differences will be random and so we believe that the population dam effects could be modelled using a probability distribution. It seems appropriate to make Dam a random factor. Similarly, there are many calves associated with our population of dams and we believe that calf effects are also likely to be random and so we believe that calf effects could be described using a probability distribution. Anyway, Diets were randomized to Calf within Dams. Thus, it is also appropriate to designate Calf to be a random factor. The generalized factors in the experiment are Dam, Dam∧Calf and Diet. As the first two generalized factors contain at least one random factor, they are variation terms. The generalized factor Diet consists of only a fixed factor and so it is a potential expectation term. Hence, the maximal model to be used for this experiment is the alternative model given above:

ψ = E[Y] = Diet
and V[Y] = Dam + Dam$\wedge$Calf. ■

While in this RCBD experiment the blocks factor (Dam) has been designated as random, it will sometimes be appropriate to designate the Blocks as fixed and other times as random — it varies from experiment to experiment.

g) The expected mean squares

Rule VI.9: The steps for constructing the expected mean squares for an orthogonal experiment are:

1. For each structure formula, take the Hasse diagram of generalized factors for the formula and, for each generalized factor *F*, replace its number of levels combinations, *f*, with a) $(n/f)\sigma_F^2$ if *F* is a term in the variation model, or b) $q_F(\boldsymbol{\psi})$ if *F* is a potential expectation model term. Under the source corresponding to a generalized factor *F*, form the source's contribution to the expected mean squares by including
   a) the expression on the left for every variation generalized factor *V* to which *F* is marginal $(F < V)$, starting with the bottommost V and working up to *F* — that is the left hand expression for every *V* directly or indirectly connected to *F* from below — and
   b) the expression under *F*.

2. Add contributions of unrandomized factors to expected mean squares, computed in the Hasse diagram, to the analysis of variance table by putting each contribution against its source in the table, unless the source has been partitioned, in which case put the contribution against the sources into which it has been partitioned.

3. Repeat rule 2 for the other structure formula(e), adding the contributions to those already in the table. However, if a generalized factor occurs in more than one Hasse diagram, its $(n/f)\sigma_F^2$ or $q_F(\boldsymbol{\psi})$ is added only once. ■

Note that if all the factors in a structure formula are fixed, then it is unnecessary to use the Hasse diagram to work out the contributions of their generalized factors to the expected mean squares. For each generalized factor, *F*, the contribution will be simply $q_F(\boldsymbol{\psi})$, as there are no variation generalized factors to take into account in step 1.a) of rule VI.9. For example, whenever all the randomized factors are fixed, work out the contributions of just the unrandomized factors using a Hasse diagram.

VI.B The Latin square example

Example VI.3 Pollution effects of petrol additives (continued)

We shall determine the expected mean squares for the Latin square example. However, to recap what we have done so far with this example.

a) Description of pertinent features of the study

1.	Observational unit	– a car with a driver
2.	Response variable	– Reduction
3.	Unrandomized factors	– Driver, Car
4.	Randomized factors	– Additive
5.	Type of study	– Latin square

b) The experimental structure

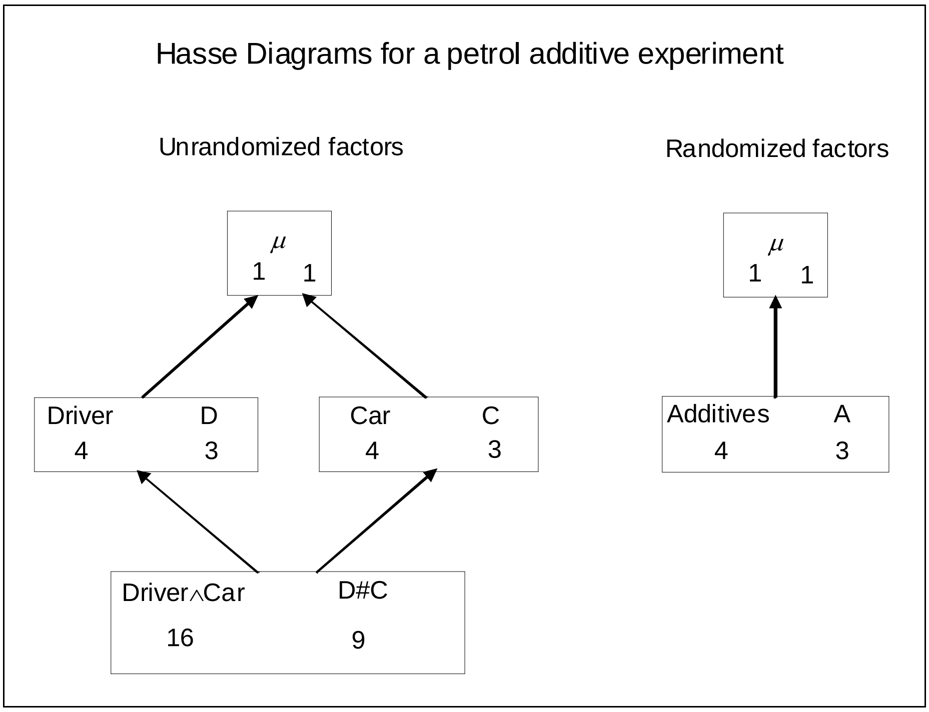
Structure	Formula
unrandomized	*4* Driver**4* Car
randomized	*4* Additive

c) Sources derived from the structure formulae

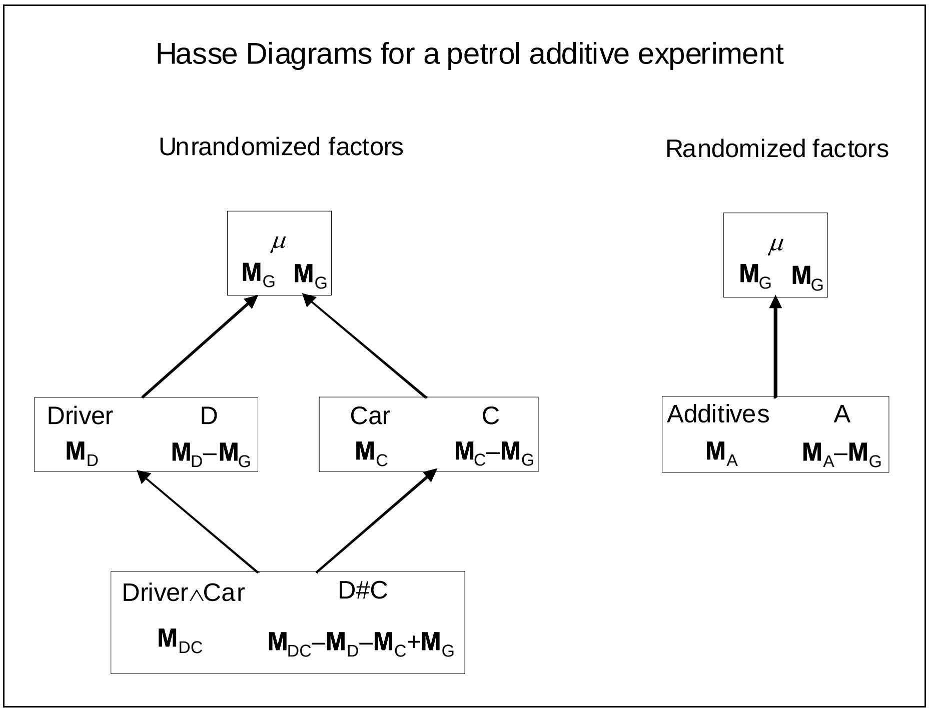
Driver*Car = Driver + Car + Driver#Car
Additive = Additive

d) Degrees of freedom and sums of squares

The Hasse diagrams, with degrees of freedom, for this study are:



The Hasse diagrams, with **M** and **Q** matrices, for this study are:



e) **The analysis of variance table**

Source	df	SSq
Drivers	3	$\mathbf{Y'Q_DY}$
Cars	3	$\mathbf{Y'Q_CY}$
Drivers#Cars	9	$\mathbf{Y'Q_{DC}Y}$
Additive	3	$\mathbf{Y'Q_AY}$
Residual	6	$\mathbf{Y'Q_{DC_{Res}}Y}$

f) **Maximal expectation and variation models**

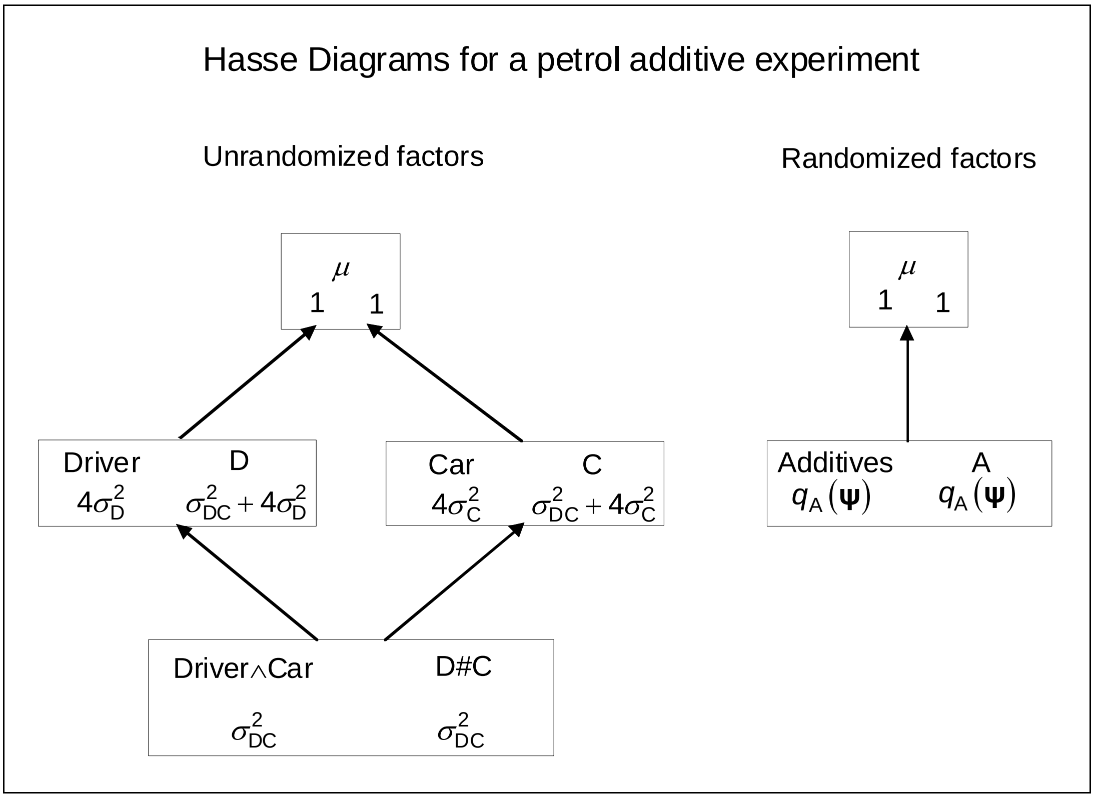
Take the random factors to be Drivers and Cars and the fixed factor to be Additive. Then, referring to the Hasse diagram of Generalized-factor Marginalities, the maximal expectation and variation models are

$$\psi = E[Y] = \text{Additive and}$$
$$\text{var}[Y] = \text{Driver} + \text{Car} + \text{Driver}\wedge\text{Car}$$

g) **The expected mean squares.**

Additive is an expectation term and Driver, Car, and Driver∧Car are the variation terms. Hence, the variation components will be σ_D^2, σ_C^2 and σ_{DC}^2, respectively. In this experiment there are 16 observations and the values of *f* for Driver, Car, and Driver∧Car are 4, 4, and 16, respectively — see Hasse diagram with degrees of freedom. The multipliers of the components are $16/4 = 4$, $16/4 = 4$ and $16/16 = 1$, respectively.

The Hasse diagrams, with contributions to expected mean squares, for this study are:


Hasse Diagrams for a petrol additive experiment
Unrandomized factors
Randomized factors
μ
1 1
μ
1 1
Driver D
$4\sigma^2_{\mathrm{D}}$ $\sigma^2_{\mathrm{DC}} + 4\sigma^2_{\mathrm{D}}$
Car C
$4\sigma^2_{\mathrm{C}}$ $\sigma^2_{\mathrm{DC}} + 4\sigma^2_{\mathrm{C}}$
Additives A
$q_{\mathrm{A}}(\boldsymbol{\psi})$ $q_{\mathrm{A}}(\boldsymbol{\psi})$
Driver∧Car D#C
σ^2_{DC} σ^2_{DC}

Source	Df	SSq	E[MSq]
Driver	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{D}}\mathbf{Y}$	$\sigma^2_{DC} + 4\sigma^2_{D}$
Car	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{C}}\mathbf{Y}$	$\sigma^2_{DC} + 4\sigma^2_{C}$
Driver#Car	9	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}}\mathbf{Y}$	
Additive	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{A}}\mathbf{Y}$	$\sigma^2_{\mathrm{DC}} + q_{\mathrm{A}}(\boldsymbol{\psi})$
Residual	6	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DC}_{\mathrm{Res}}}\mathbf{Y}$	σ^2_{DC}
Total	15		

■

VI.C Usage of the procedure

As mentioned at the beginning of this chapter, the procedure should be carried at the design stage. That way you can check the properties of the design that you propose to use. It is a good idea to obtain a layout for the design in R and carry out an analysis on randomly-generated data. The experimental structure that you have determined will be central to this as it contains all the factors to be generated and the model formula will be of the form:

```
Y ~ randomized structure + fixed unrandomized sources
          + Error(unrandomized structure)
```

where `Y` is some randomly generated data needed in advance of doing the experiment so that R has something to analyze.

The advantage of doing this analysis is that it will provide a check of what you have done. In particular, are the sources, and where they occur in the analysis, what you expected? Are the degrees of freedom what you have calculated? Are the correct F-tests being performed?

VI.D Rules for determining the analysis of variance table — summary

a) Description of pertinent features of the study

The first stage in determining the analysis of variance table is to identify the following features:

Definition VI.1: The **observational unit** is the native physical entity which is individually measured. For example, a person in a survey or a run in an experiment. ■

Definition VI.2: The **response variable** is the unrestricted variable that the investigator wants to see if the factors affect its response. For example, the experimenter may want to determine whether or not there are differences in yield, height, and so on for the different treatments — this is the response variable; that is, the variable of interest or for which differences might exist. ■

Definition VI.3: The **unrandomized factors** are those factors that would index the observational units if no randomization had been performed. ■

Definition VI.4: The **randomized factors** are those factors associated with the observational unit as a result of randomization. ■

Definition VI.5: The **type of study** is the name of the experimental design or sampling method; for example, CRD, RCBD, LS, YS, SRS, factorial, and so on. ■

Rule VI.1: To determine whether a factor is unrandomized or randomized, ask the following question:

> For an observational unit, can I identify the levels of that factor associated with the unit if randomization has not been performed?
>
> If yes then the factor is unrandomized, if no then it is randomized. ■

b) The experimental structure

Having determined the unrandomized and randomized factors, one next determines the experimental structure.

Rule VI.2: Determine the experimental structure by

3. describing the nesting and crossing relationships between the unrandomized factors in the experiment,
4. describing the nesting and crossing relationships between
   iii) the randomized factors, and
   iv) the randomized and the unrandomized factors, if any.

The numbers of levels of the factors are placed in front of the names of the factors. ■

Definition VI.6: Two factors are **nested** if units with the same level of the nested factor, but different levels of the nesting factor, have no apparent characteristic in common. ■

Definition VI.7: Two factors are **crossed** if units with the same level of one factor, but different levels of the second factor, have a common characteristic associated with the first factor. ■

c) Sources derived from the structure formulae

Having determined the experimental structure, the next step is to expand the formulae to obtain the terms that are to be included in the model and in the analysis of variance table.

Rule VI.3: The rules for expanding structure formulae involving two factors A and B are:

A*B = A + B + A#B
A/B = A + B[A].

More generally, if L and M are two formulae
L*M = L + M + L#M
where L#M is the sum of all pairwise combinations of terms in L with terms in M
and L/M = L + M[*gf*(L)]
where *gf*(L) is the generalized factor (see definition VI.8 in section D, *Degrees of freedom*) formed from the combination of all factors in L. ■

d) Degrees of freedom and sums of squares

The degrees of freedom for an analysis of variance can be calculated with the aid of Hasse diagrams for Generalized-Factor Marginalities for each structure formula.

Definition VI.8: A **generalized factor** is the factor formed from several (original) factors and whose levels are the combinations that occur in the experiment of the levels of the constituent factors. The generalized factor is the "meet" of the constituent factors and it is written as the list of constituent factors separated by "wedges" or "meets" ("$\wedge$"). For convenience we include ordinary factors amongst the set of generalized factors for an experiment. ■

Definition VI.9: One generalized factor, V say, is **marginal** to another, Z say, if the factors in the marginal generalized factor are a subset of those in the other and this will occur irrespective of the replication of the levels of the generalized factors. We write $V \leq Z$. ■

Rule VI.4: The Hasse diagrams for Generalized-Factor Marginalities for a structure formula is formed by placing generalized factors above those to which they are marginal and connecting them by an upwards arrow. Alongside each generalized factor is added its source. The Universe factor, connecting all units that occurred in the experiment, is included at the top of the diagram. Under a generalized factor is written its number of levels. Under a source is written its degrees of freedom: it is calculated as the difference between the entry for its generalized factor and the sum of the degrees of freedom of all sources whose generalized factors are marginal to the current generalized factor. ■

Rule VI.5: When all the factors are crossed, the degrees of freedom of any source can be calculated directly. The rule for doing this is:

> For each factor in the source, calculate the number of levels minus one and multiply these together. ■

In addition the quadratic forms that form the basis of the expressions for sums of squares in the analysis of variance table are obtained using the following rule:

Rule VI.6: There is a mean operator (**M**) and a quadratic-form (**Q**) matrix for each generalized factor and source, respectively, obtained from the structure formulae. To obtain expressions for the **Q** matrices in terms of the **M** matrices, take the Hasse diagram of generalized factors for the formula and, for each generalized factor, replace its number of levels combinations with its **M** matrix. Then, the expression for the **Q** matrix for a source replaces its degrees of freedom in the Hasse diagram. To work it out take the **M** matrix for its generalized factor and subtract all expressions for **Q** matrices of sources whose generalized factors are marginal to the generalized factor for the source whose expression you are deriving. ■

e) The analysis of variance table

Rule VI.7: The analysis of variance table is formed by:

1. Listing down all the unrandomized sources in the Source column, and their degrees of freedom in the df column and the quadratic forms in the SSq column.
2. Then the randomized sources are placed indented under the unrandomized sources with which they are confounded, along with their degrees of freedom and, if the design is orthogonal, their quadratic forms.
3. Residual sources are added to account for the left-over portions of unrandomized sources and their degrees of freedom and quadratic forms are computed by difference. For orthogonal experiments, the matrix of the Residual quadratic form is the difference of the matrices of the quadratic forms from which it is computed. ■

f) Maximal expectation and variation models

Rule VI.8: To obtain the terms in the expectation and variation model:

1. Designate each factor in the experiment as either fixed or random.
2. Determine whether a generalized factor is a potential expectation or variation term as follows: generalized factors that involve *only fixed* (original) factors are potential expectation terms and generalized factors that contain *at least one random* (original) factor will become variation terms. If there is no unrandomized factor that has been classified as random, the term consisting of all unrandomized factors will be designated as random.

3. The maximal expectation model is then the sum of all the expectation terms except those that are marginal to a term in the model; if there is no expectation terms, the model consists of a single term for the grand mean.
4. The maximal variation model the sum of all the variation terms. ■

Definition VI.10: A factor will be designated as **random** if it is anticipated that the distribution of effects associated with the population set of levels for the factor can be described using a probability distribution function. ■

Definition VI.11: A factor will be designated as **fixed** if it is anticipated that a probability distribution function will not provide a satisfactory description the set of effects associated with the population set of levels for the factor. ■

It often happens, but not always, that all unrandomized factors are designated as random and so all terms involving them occur in the variation model — all the randomized factors are designated as fixed and all terms involving them, minus marginal terms, occur in expectation models.

g) The expected mean squares

Rule VI.9: The steps for constructing the expected mean squares for an orthogonal experiment are:

1. For each structure formula, take the Hasse diagram of generalized factors for the formula and, for each generalized factor *F*, replace its number of levels combinations, *f*, with a) $(n/f)\sigma_F^2$ if *F* is a term in the variation model, or b) $q_F(\boldsymbol{\psi})$ if *F* is a potential expectation model term. Under the source corresponding to a generalized factor *F*, form the source's contribution to the expected mean squares by including
   a) the expression on the left for every variation generalized factor *V* to which *F* is marginal $(F < V)$, starting with the bottommost V and working up to *F* — that is the left hand expression for every *V* directly or indirectly connected to *F* from below — and
   b) the expression under *F*.
2. Add contributions of unrandomized factors to expected mean squares, computed in the Hasse diagram, to the analysis of variance table by putting each contribution against its source in the table, unless the source has been partitioned, in which case put the contribution against the sources into which it has been partitioned.
3. Repeat rule 2 for the other structure formula(e), adding the contributions to those already in the table. However, if a generalized factor occurs in more than one Hasse diagram, its $(n/f)\sigma_F^2$ or $q_F(\boldsymbol{\psi})$ is added only once. ■

Note that if all the factors in a structure formula are fixed, then for each generalized factor, *F*, the contribution will be simply $q_F(\boldsymbol{\psi})$.

VI.E Determining the analysis of variance table – further examples

For each of the following studies determine the analysis of variance table (Source, df, SSq and E[MSq]) using the following seven steps:

a) Description of pertinent features of the study
b) The experimental structure
c) Sources derived from the structure formulae
d) Degrees of freedom and sums of squares
e) The analysis of variance table
f) Maximal expectation and variation models
g) The expected mean squares.

Example VI.8 Mathematics teaching methods

An educational psychologist wants to determine the effect of three different methods of teaching mathematics to year 10 students. Five metropolitan schools with three mathematics classes in year 10 are selected and the methods of teaching randomized to the classes in each school. After being taught by one of the methods for a semester, the students sit a test and their average score is recorded.

a) *Description of pertinent features of the study*

1.	the observational unit	–	a class
2.	response variable	–	Test score
3.	unrandomized factors	–	Schools, Classes
4.	randomized factors	–	Methods
5.	type of study	–	an RCBD

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) *Sources derived from the structure formulae*

Schools/Classes = Schools + Classes[Schools]

Methods = Methods

d) *Degrees of freedom and sums of squares*

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that randomized factors should be fixed and unrandomized factors should be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

■

Example VI.9 Smoke emission from combustion engines

A manufacturer of combustion engines is concerned about the percentage of smoke emitted by engines of a particular design in order to meet air pollution standards. An experiment was conducted involving engines with three timing levels, three throat diameters, two volume ratios in the combustion chamber, and two injection systems. Thirty-six engines were designed to represent all possible combinations of these four factors and the engines were then operated in a completely random order so that each engine was tested twice. The percent smoke emitted was recorded at each test.

a) *Description of pertinent features of the study*

1.	Observational unit	– a test (involving some engine)
2.	Response variable	– Percent Smoke Emitted
3.	Unrandomized factors	– Tests
4.	Randomized factors	– Timing, Diameter, Ratios, Systems
5.	Type of study	– Complete Four-factor CRD

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) Sources derived from the structure formulae

Tests $=$ Tests

Timing*Diameter*Ratios*Systems

$=$ Timing + Diameter + Ratios + Systems
+ Timing#Diameter + Timing#Ratios + Timing#Systems
+ Diameter#Ratios + Diameter#Systems + RatiosSystems
+ Timing#Diameter#Ratios + Timing#Diameter#Systems
+ Timing#Ratios#Systems + Diameter#Ratios#Systems
+ Timing#Diameter#Ratios#Systems

d) Degrees of freedom and sums of squares

Note that factors in the randomized structure are completely crossed so that the degrees of freedom of a source from that structure can be obtained by computing the number of levels minus one for each factor in the source and forming their product.

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that randomized factors should be fixed and unrandomized factor should be random. Hence, the maximal variation and expectation models are:

Var[Y] $=$

$\psi =$ E[Y] $=$

g) The expected mean squares.

In this case, using the Hasse diagrams to compute the E[MSq]s would be overkill. There is one variation component for Tests: $\left(72/72\right)\sigma_{\mathrm{t}}^{2} = \sigma_{\mathrm{t}}^{2}$. Since Tests is partitioned into the other sources in the table, this component occurs against all lines except Tests. All the randomized generalized factors are potential contributors to the expectation model and so their contributions to the E[MSq]s for the corresponding sources are of the form $q_{F}\left(\boldsymbol{\psi}\right)$.

The analysis of variance table is:

Source	df	SSq	E[MSq]
Tests	71	$\mathbf{Y}'\mathbf{Q}_{\mathrm{t}}\mathbf{Y}$	
Timing	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{T}}(\boldsymbol{\psi})$
Diameter	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{D}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{D}}(\boldsymbol{\psi})$
Ratios	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{R}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{R}}(\boldsymbol{\psi})$
Systems	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{S}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{S}}(\boldsymbol{\psi})$
Timing#Diameter	4	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TD}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TD}}(\boldsymbol{\psi})$
Timing#Ratios	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TR}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TR}}(\boldsymbol{\psi})$
Timing#Systems	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TS}}(\boldsymbol{\psi})$
Diameter#Ratios	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DR}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{DR}}(\boldsymbol{\psi})$
Diameter#Systems	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{DS}}(\boldsymbol{\psi})$
Ratios#Systems	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{RS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{RS}}(\boldsymbol{\psi})$
Timing#Diameter#Ratios	4	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TDR}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TDR}}(\boldsymbol{\psi})$
Timing#Diameter#Systems	4	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TDS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TDS}}(\boldsymbol{\psi})$
Timing#Ratios#Systems	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TRS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TRS}}(\boldsymbol{\psi})$
Diameter#Ratios#Systems	2	$\mathbf{Y}'\mathbf{Q}_{\mathrm{DRS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{DRS}}(\boldsymbol{\psi})$
Timing#Diameter#Ratios#System	4	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TDRS}}\mathbf{Y}$	$\sigma_{\mathrm{t}}^2 + q_{\mathrm{TDRS}}(\boldsymbol{\psi})$
Residual	36	$\mathbf{Y}'\mathbf{Q}_{\mathrm{t}_{\mathrm{Res}}}\mathbf{Y}$	σ_{t}^2

■

Example VI.10 Lead concentration in hair

An investigation was performed to discover trace metal concentrations in humans in the five major cities in Australia. The concentration of lead in the hair of fourth grade school boys was determined. In each city, 10 primary schools were randomly selected and from each school ten students selected. Hair samples were taken from the selected boys and the concentration of lead in the hair determined.

a) *Description of pertinent features of the study*

1. the observational unit – hair of a boy
2. response variable – Lead concentration
3. unrandomized factors – Cities, Schools, Boys
4. randomized factors – not applicable
5. type of study – a multistage survey

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) Sources derived from the structure formulae

d) Degrees of freedom and sums of squares

The Hasse diagram, with degrees of freedom, for the unrandomized generalized factors in this study is:

The Hasse diagram, with **M** and **Q** matrices, for the unrandomized generalized factors in this study is:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that Cities should be fixed and Schools and Boys should be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is

Source	df	SSq	E[MSq]

■

Example VI.11 Penicillin pain

An experiment was conducted to investigate the degree of pain experienced by patients when injected with penicillin of different potencies. In the experiment there were 10 groups of three patients, all three patients being simultaneously injected with the same dose. Altogether there were six different potencies, all of which were administered over six consecutive times to the groups of patients; the order in which they were administered is randomized for each group. The total of the degree of pain experienced by the three patients was obtained, the pain for each patient having been measured on a five-point scale 0–4.

a) Description of pertinent features of the study

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) The experimental structure

Structure	Formula
unrandomized	
randomized	

c) Sources derived from the structure formulae

d) Degrees of freedom and sums of squares

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that randomized factors should be fixed and unrandomized factors should be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

■

Example VI.12 Plant rehabilitation study

In a plant rehabilitation study, the increase in height of plants of a certain species during a 12-month period was to be determined at three sites differing in soil salinity. Each site was divided into five parcels of land containing 4 plots and four different management regimes applied to the plots, the regimes being randomized to the plots within a parcel. In each plot, six plants of the species were selected and marked and the total increase in height of all six plants measured.

Note that the researcher is interested in seeing if any regime differences vary from site to site.

a) *Description of pertinent features of the study*

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) *Sources derived from the structure formulae* *d) Degrees of freedom and sums of squares*

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that Sites and Regimes should be fixed and Parcels and Plots should be random. Hence, the maximal variation and expectation models are:

g) *The expected mean squares.*

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is

■

Example VI.13 Pollution effects of petrol additives

Suppose a study is to be conducted to investigate whether four petrol additives differ in the amount by which they reduce the emission of oxides of nitrogen. Four cars and four drivers are to be employed in the study and the following Latin square arrangement is to be used to assign the additives to the driver-car combinations:

		Car			
		1	2	3	4
	I	B	D	C	A
	II	A	B	D	C
Drivers	III	D	C	A	B
	IV	C	A	B	D

(Additives A, B, C, D)

a) *Description of pertinent features of the study*

1. the observational unit – a driver in a car
2. response variable – Reduction
3. unrandomized factors – Drivers, Cars

4. randomized factors – Additives
5. type of study – a Latin square

b) The experimental structure

Structure	Formula
unrandomized	*4* Drivers**4* Cars
randomized	*4* Additives

c) Sources derived from the structure formulae

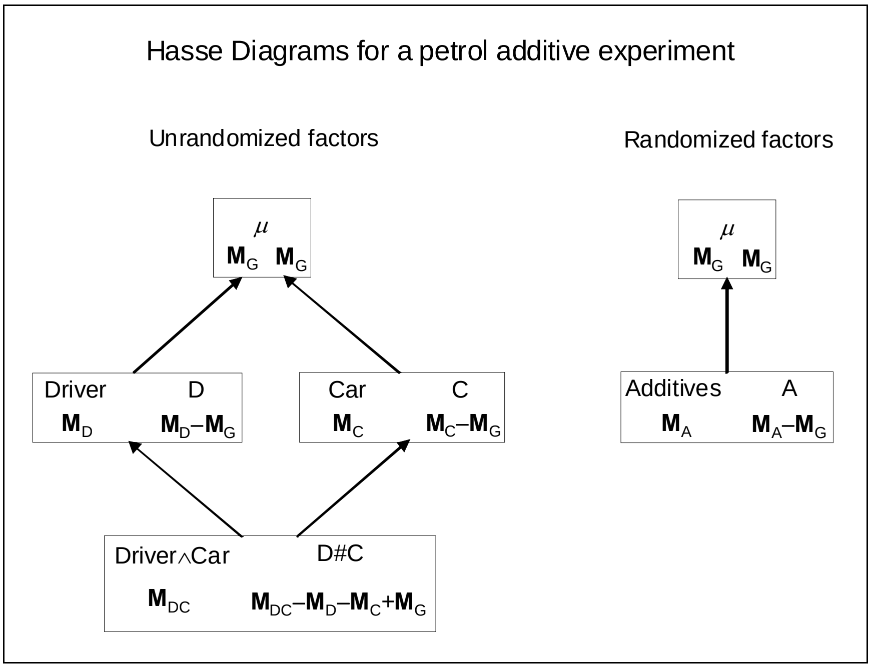
Drivers*Cars = Drivers + Cars + Drivers#Cars

Additives = Additives

d) Degrees of freedom and sums of squares

The degrees of freedom for this study can be worked out using the rule for completely crossed structures.

The Hasse diagrams, with **M** and **Q** matrices, for this study are:


Hasse Diagrams for a petrol additive experiment
Unrandomized factors
Randomized factors
μ
$\mathbf{M}_G$ $\mathbf{M}_G$
Driver D
$\mathbf{M}_D$ $\mathbf{M}_D-\mathbf{M}_G$
Car C
$\mathbf{M}_C$ $\mathbf{M}_C-\mathbf{M}_G$
Driver∧Car D#C
$\mathbf{M}_{DC}$ $\mathbf{M}_{DC}-\mathbf{M}_D-\mathbf{M}_C+\mathbf{M}_G$
μ
$\mathbf{M}_G$ $\mathbf{M}_G$
Additives A
$\mathbf{M}_A$ $\mathbf{M}_A-\mathbf{M}_G$

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

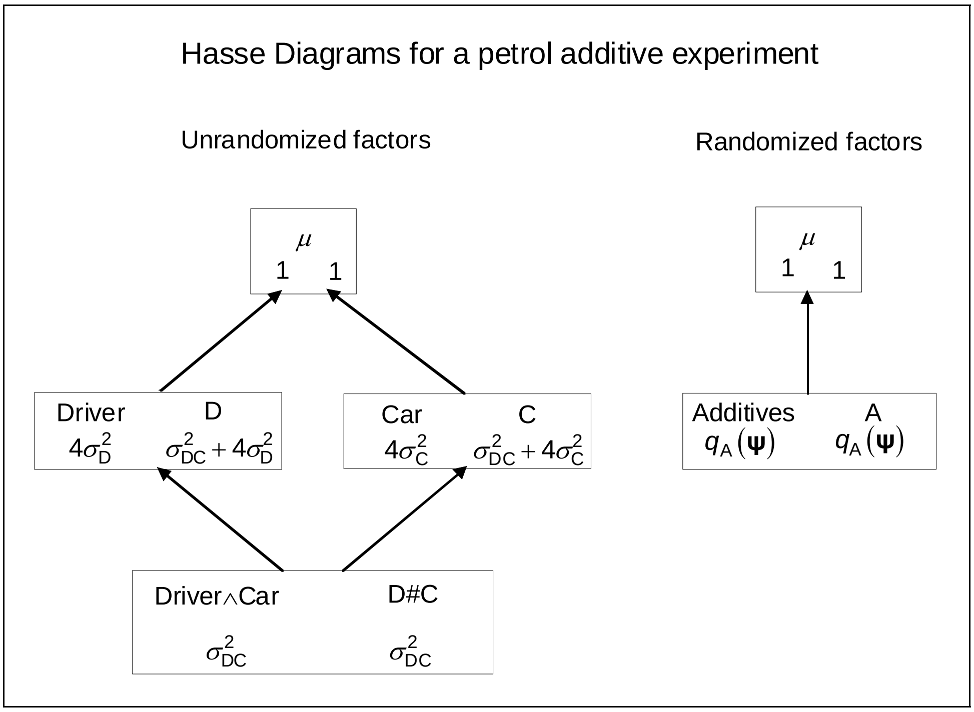
Will take it that randomized factors should be fixed and unrandomized factors should be random. Hence, the maximal variation and expectation models are:

$$\text{Var[Y]} = \text{Drivers} + \text{Cars} + \text{Drivers}\wedge\text{Cars}$$

$$\psi = \text{E[Y]} = \text{Additives}$$

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:


Hasse Diagrams for a petrol additive experiment
Unrandomized factors
Randomized factors
μ
1 1
Driver D
$4\sigma^2_\text{D}$ $\sigma^2_\text{DC} + 4\sigma^2_\text{D}$
Car C
$4\sigma^2_\text{C}$ $\sigma^2_\text{DC} + 4\sigma^2_\text{C}$
Driver∧Car D#C
σ^2_DC σ^2_DC
μ
1 1
Additives A
$q_\text{A}(\boldsymbol{\psi})$ $q_\text{A}(\boldsymbol{\psi})$

The analysis of variance table is:

Source	Df	SSq	E[MSq]
Driver	3	$\mathbf{Y}'\mathbf{Q}_\text{C}\mathbf{Y}$	$\sigma^2_{DC} + 4\sigma^2_D$
Car	3	$\mathbf{Y}'\mathbf{Q}_\text{D}\mathbf{Y}$	$\sigma^2_{DC} + 4\sigma^2_C$
Driver#Car	9	$\mathbf{Y}'\mathbf{Q}_\text{DC}\mathbf{Y}$	
Additive	3	$\mathbf{Y}'\mathbf{Q}_\text{A}\mathbf{Y}$	$\sigma^2_\text{DC} + q_\text{A}(\boldsymbol{\psi})$
Residual	6	$\mathbf{Y}'\mathbf{Q}_{\text{DC}_\text{Res}}\mathbf{Y}$	σ^2_{DC}
Total	15		

Suppose that the experiment is to be repeated by replicating the Latin square twice using the same cars but new drivers on a second occasion. What are the features of the study?

a) *Description of pertinent features of the study*

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) *Sources derived from the structure formulae*

d) *Degrees of freedom and sums of squares*

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Again, will take it that randomized factors should be fixed and unrandomized factors should be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

■

Example VI.14 Controlled burning

Suppose an environmental scientist wants to investigate the effect on the biomass of burning Areas of natural vegetation. There are available two Areas separated by several kilometres for use in the investigation. It is only possible to either burn or not burn an entire Area. The scientist randomly selects to burn one Area and the other Area is left unburnt as a control. She randomly samples 30 locations in each Area and measures the biomass at each location.

a) *Description of pertinent features of the study*

3. the observational unit
4. response variable
5. unrandomized factors
6. randomized factors
7. type of study

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) *Sources derived from the structure formulae*

d) *Degrees of freedom and sums of squares*

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that randomized factor should be fixed and unrandomized factors should be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

■

Example VI.15 Generalized randomized complete block design

A generalized randomized complete block design is the same as the ordinary randomized complete block design, except that each treatment occurs more than once in a block — see section IV.G, *Generalized randomized complete block design*. For example, suppose four treatments are to be compared when applied to a new variety of wheat. I employed a generalized randomized complete block design with 12 plots in each of 2 blocks so that each treatment is replicated 3 times in each block. The yield of wheat from each plot was measured. A possible layout for this experiment is shown in the table given below.

Layout for a generalized randomized complete block experiment

		Plots											
		1	2	3	4	5	6	7	8	9	10	11	12
Blocks	I	C	B	B	A	C	D	A	A	D	B	D	C
	II	C	B	A	D	D	D	A	A	B	C	B	C

Here work out the analysis for this experiment that includes a source for Block#Treatment interaction, and assumes that the unrandomized factors are random and the randomized factors are fixed. Having done this, derive the analysis for only Plots random and the rest of the factors fixed.

a) *Description of pertinent features of the study*

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) Sources derived from the structure formulae

d) Degrees of freedom and sums of squares

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Take the randomized factors to be fixed and unrandomized factors to be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

For only Plots random, the maximal variation and expectation models are:

$$\mathrm{Var}[\mathrm{Y}] = \mathrm{Blocks} \wedge \mathrm{Plots}$$

$$\psi = \mathrm{E}[\mathrm{Y}] = \mathrm{Treatments} \wedge \mathrm{Blocks}$$

In this case there will be the one variation component, σ^2_{BP}. The remaining contributions will be of the form $q_{\mathrm{F}}\left(\boldsymbol{\psi}\right)$.

The analysis of variance table is:

Source	df	SSq	E[MSq]
Blocks	1	$\mathbf{Y}'\mathbf{Q}_{\mathrm{B}}\mathbf{Y}$	$\sigma^2_{\mathrm{BP}} \quad +q_{\mathrm{B}}\left(\boldsymbol{\psi}\right)$
Plots[Blocks]	22	$\mathbf{Y}'\mathbf{Q}_{\mathrm{BP}}\mathbf{Y}$	
Treatments	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{T}}\mathbf{Y}$	$\sigma^2_{\mathrm{BP}} \quad +q_{\mathrm{T}}\left(\boldsymbol{\psi}\right)$
Treatments#Blocks	3	$\mathbf{Y}'\mathbf{Q}_{\mathrm{TB}}\mathbf{Y}$	$\sigma^2_{\mathrm{BP}} \quad +q_{\mathrm{BT}}\left(\boldsymbol{\psi}\right)$
Residual	16	$\mathbf{Y}'\mathbf{Q}_{\mathrm{BP}_{\mathrm{Res}}}\mathbf{Y}$	σ^2_{BP}
Total	23		

Note the difference in the denominator of the test for Treatments between the two analyses. For the former analysis it is Treatments#Blocks and the latter analysis it is the Residual. Which analysis is correct depends on the nature of the Block-Treatment interaction. If the Blocks are very different, as in the case where they are quite different sites, and it is anticipated that the treatments will respond quite differently in the different blocks, then it is probable that the Blocks should be regarded as fixed so that the term Treatments#Blocks is also fixed (see Steel and Torrie, sec. 9.8).

■

Example VI.16 Salt tolerance of lizards

To examine the salt tolerance of the lizard *Tiliqua rugosa*, eighteen lizards of this species were obtained. Each lizard was randomly selected to receive one of three salt treatments (injection with sodium, injection with potassium, no injection) so that 6 lizards received each treatment. Blood samples were then taken from each lizard on five occasions after injection and the concentration of Na in the sample determined.

a) Description of pertinent features of the study

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) The experimental structure

Structure	Formula
unrandomized	
randomized	

c) Sources derived from the structure formulae

d) Degrees of freedom and sums of squares

The degrees of freedom for this study can be worked out using the rule for completely crossed structures.

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Seems that Treatments and Occasions should be fixed and Lizards should be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

■

Example VI.17 Eucalyptus growth

An experiment was planted in a forest in Queensland to study the effects of irrigation and fertilizer on 4 seedlots of a species of gum tree. There were two levels of irrigation (no and yes), two levels of fertilizer (no and yes) and four seedlots (Bulahdelah, Coffs Harbour, Pomona and Atherton). Because of the difficulties of irrigating and applying fertilizers to individual trees, these needed to be applied to groups of trees. So the experimental area was divided up into 8 stands of 20 trees, with four stands in one block and the other four in a second block. The four combinations of irrigation and fertilizer were randomized to the four stands in a block. Each stand of 20 trees consisted of 4 rows by 5 columns and the 4 seedlots were randomized to the four rows. The mean height of the five trees in a row was measured.

a) Description of pertinent features of the study

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) The experimental structure

Structure	Formula
unrandomized	
randomized	

c) Sources derived from the structure formulae

d) Degrees of freedom and sums of squares

The Hasse diagrams, with degrees of freedom, for this study are:

In this case we only do the Hasse diagram for the unrandomized factors because the degrees of freedom for the randomized factors can be obtained using the rule for all factors crossed.

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Take all the randomized factors and Blocks to be fixed; the remainder of the factors take as random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagrams, with contributions to expected mean squares, for this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

■

Example VI.18 Eelworm experiment

Cochran and Cox (1957, section 3.2) present the results of an experiment examining the effects of soil fumigants on the number of eelworms. There were four different fumigants each applied in both single and double dose rates as well as a control treatment in which no fumigant was applied. The experiment was laid out in 4 blocks each containing 12 plots; in each block, the 8 treatment combinations were each applied once and the control treatment four times and the 12 treatments randomly allocated to plots. The number of eelworm cysts in 400g samples of soil from each plot was determined.

a) *Description of pertinent features of the study*

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) *The experimental structure*

Structure	Formula
unrandomized	
randomized	

c) *Sources derived from the structure formulae*

d) *Degrees of freedom and sums of squares*

The Hasse diagrams, with degrees of freedom, for this study are:

The Hasse diagrams, with **M** and **Q** matrices, for this study are:

e) The analysis of variance table

Enter the sources for the study, their degrees of freedom and quadratic forms, into the analysis of variance table below.

f) Maximal expectation and variation models

Take the randomized factors to be fixed and unrandomized factors to be random. Hence, the maximal variation and expectation models are:

g) The expected mean squares.

The Hasse diagram, with contributions to expected mean squares, for the unrandomized factors in this study are:

The analysis of variance table is:

Source	df	SSq	E[MSq]

■

Example VI.19 A factorial experiment

An experiment is to be conducted on sugar cane to investigate 6 factor (A, B, C, D, E, F) each at two levels. This experiment is to involve 16 blocks each of eight plots. The 64 treatment combinations are divided into 8 sets of 8 so that the ABCD, ABEF and ACE interactions are associated with set differences. The 8 sets are randomized to the 16 blocks so that each set occurs on two blocks and the 8 combinations in a set are randomized to the plots within a block. The sugar content of the cane is to be measured.

a) Description of pertinent features of the study

1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

b) The experimental structure

Structure	Formula
unrandomized	
randomized	

How many main effects, two factor interactions, three-factor interactions and interactions of more than 3 factors are there? What are the interactions confounded with blocks?

c) Degrees of freedom and sums of squares

What are the degrees of freedom for the unrandomized sources and for the randomized sources?

d) The analysis of variance table

Enter the sources for the study, and their degrees of freedom, into the analysis of variance table below.

e) Maximal expectation and variation models

Seems that randomized factors should be fixed and unrandomized factors should be random. Hence, the maximal variation and expectation models are:

f) The expected mean squares.

The Hasse diagram, with contributions to expected mean squares, for the unrandomized factors in this study are:

The analysis of variance table is (just give the interactions confounded with blocks and the numbers of main effects, two-factor, three-factor and other interactions):

Source	df	E[MSq]
Blocks	15	
ACE	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{ACE}}(\psi)$
ADF	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{ADF}}(\psi)$
BCF	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{BCF}}(\psi)$
BDE	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{BDE}}(\psi)$
ABCD	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{ABCD}}(\psi)$
ABEF	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{ABEF}}(\psi)$
CDEF	1	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}} + q_{\mathbf{CDEF}}(\psi)$
Residual	8	$\sigma^2_{\text{BP}} + 8\sigma^2_{\text{B}}$
Plots[Blocks]	112	
main effects	6	$\sigma^2_{\text{BP}} + q_{\mathbf{i}}(\psi)$
2-factor interactions	15	$\sigma^2_{\text{BP}} + q_{\mathbf{i.j}}(\psi)$
3-factor interactions	17	$\sigma^2_{\text{BP}} + q_{\mathbf{i.j.k}}(\psi)$
other interactions	19	$\sigma^2_{\text{BP}} + q_{\mathbf{i.j.k.l+}}(\psi)$
Residual	55	σ^2_{BP}

■

VI.F Exercises

VI.1 The following data are from a Latin square experiment designed to investigate the moisture content of turnip greens. The experiment involved the measurement of the percent moisture content of five leaves of different sizes from each of five plants. The treatments were time of measurement in days since the beginning of the experiment.

				Plant		
		1	2	3	4	5
	A	5 6.67	2 5.40	3 7.32	1 4.92	4 4.88
	B	4 7.15	5 4.77	2 8.53	3 5.00	1 6.16
Leaf Size	C	1 8.29	4 5.40	5 8.50	2 7.29	3 7.83
(A = smallest,	D	3 8.95	1 7.54	4 9.99	5 7.85	2 5.83
E = largest)	E	2 9.62	3 6.93	1 9.68	4 7.08	5 8.51

What are the features of this experiment?

1. Observational unit ____________________
2. Response variable ____________________
3. Unrandomized factors ____________________
4. Randomized factors ____________________
5. Type of study ____________________

What is the experimental structure for this experiment?

Structure	Formula
unrandomized	
randomized	

What are the Hasse diagrams of generalized-factor marginalities, with degrees of freedom and with **M** and **Q** matrices, for this study?

Derive the maximal expectation and variation models for this study?

Determine the expected mean squares for this study, using where appropriate the Hasse diagrams of generalized-factor marginalities.

Give the analysis of variance table, including the degrees of freedom, sums of squares and expected mean squares.

Source	df	SSq	E[MSq]
Total	24		

VI.2 A chemist has four different containers of soil. He wants to determine whether the moisture contents of these four soils differs. He randomly selects 10 samples from each container and determines the moisture content of each sample.

What are the features of this experiment?

1. Observational unit ______________________________
2. Response variable ______________________________
3. Unrandomized factors ______________________________
4. Randomized factors ______________________________
5. Type of study ______________________________

What is the experimental structure for this experiment?

Structure	Formula
unrandomized	
randomized	

What are the Hasse diagrams of generalized-factor marginalities, with degrees of freedom and with **M** and **Q** matrices, for this study?

Derive the maximal expectation and variation models for this study?

Determine the expected mean squares for this study, using where appropriate the Hasse diagrams of generalized-factor marginalities.

Give the analysis of variance table, including the degrees of freedom, sums of squares and expected mean squares.

Source	df	SSq	E[MSq]
Total			

VI.3 In an experiment to investigate the yield (ℓ/hr) of machine producing a chemical, four randomly selected machines were operated at five different temperatures for an hour and the yield measured. The order in which each machine was operated at the different temperatures was randomized for each machine.

What are the features of this experiment?

1. Observational unit ____
2. Response variable ____
3. Unrandomized factors ____
4. Randomized factors ____
5. Type of study ____

What is the experimental structure for this experiment?

Structure	Formula
unrandomized	
randomized	

What are the Hasse diagrams of generalized-factor marginalities, with degrees of freedom and with **M** and **Q** matrices, for this study?

Derive the maximal expectation and variation models for this study?

Determine the expected mean squares for this study, using where appropriate the Hasse diagrams of generalized-factor marginalities.

Give the analysis of variance table, including the degrees of freedom, sums of squares and expected mean squares.

Source	df	SSq	E[MSq]
Total			

VI.4 A study is to be conducted to compare two methods of measuring the concentration of a certain component of a liquid product. Three factories are selected from those that routinely determine the concentration of the component. A sample of the product is obtained and divided into 3 lots of 4 portions and each lot is randomly assigned to be sent to one of the factories. At each factory the concentration of their 4 portions is determined using both methods. The order in which a portion is tested using a particular method is completely randomized.

What are the features of this experiment?

1. Observational unit ______________________
2. Response variable ______________________
3. Unrandomized factors ______________________
4. Randomized factors ______________________
5. Type of study ______________________

What is the experimental structure for this experiment?

Structure	Formula
unrandomized	
randomized	

What are the Hasse diagrams of generalized-factor marginalities, with degrees of freedom and with **M** and **Q** matrices, for this study?

Derive the maximal expectation and variation models for this study?

Determine the expected mean squares for this study, using where appropriate the Hasse diagrams of generalized-factor marginalities.

Give the analysis of variance table, including the degrees of freedom, sums of squares and expected mean squares.

Source	df	SSq	E[MSq]
Total			

So which ratio of mean squares would be used to test for an average difference between the two methods?

VI.5 Adapt the R expressions for randomizing a randomized complete block design to obtain a randomized layout and dummy analysis for the experiment described in exercise VI.4.