DESIGN AND MIXED-MODEL ANALYSIS OF EXPERIMENTS

VII. Determining the analysis of variance table

(Brien (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.)

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? The method embodied in the following 7 steps is one that works for the vast majority of experimental designs used in practice:

A. Description of pertinent features of the study
B. The experimental structure
C. Terms derived from the structure formulae
D. Degrees of freedom
E. The analysis of variance table
F. Maximal expectation and variation models
G. The expected mean squares.

VII.A The procedure

A. Description of pertinent features of the study

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

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

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

For example, a plot in a survey or a rat in an experiment.

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

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

Definition VII.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. ■

Rule VII.1: To determine the unrandomized factors, ask the following question of each factor:

> For each 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.

Components of experiments

Example VII.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.**

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.**

In summary, the components 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 VII.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 components of the study?

1. Observational unit ____

   Variables (including factors) are?

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

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

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

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

		Source
unrandomized	➔	Plots
randomized	➔	Variety
		Residual

Example VII.3 Wheat samplers

Consider the experiment to investigate the error in sampling wheat plants using four different experienced samplers. Four intervals and four areas are employed in this study with samplers being assigned to a particular interval-area combination according to a Latin Square. The arrangement and data are as follows:

Results for the 4×4 Latin square experiment

		Area			
		1	2	3	4
	I	A	B	D	C
		6	11	5	10
	II	D	C	A	B
		8	11	5	12
Interval	III	B	D	C	A
		0	–2	1	1
	IV	C	A	B	D
		2	0	5	5

Sampler (A–D)

What are the components of the study?

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

Components of surveys

Example VII.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 components of the study?

1. Observational unit ______________________________

Variables (including factors) are?

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

Example VII.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 components 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

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

Then 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 VII.2: Determine the experimental structure by

1. describing the crossing and nesting relationships between the unrandomized factors in the experiment,

2. describing the crossing and nesting 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.

Definition VII.6: Two factors are **crossed** if a level of one factor has a characteristic associated with the factor in common across all levels of the other factor. ■

Definition VII.7: Two factors are **nested** if the same level of the nested factor from different levels of the nesting factor have no characteristic in common. ■

An asterisk (‘*’) will be placed between two factors to indicate that they are crossed and a slash (‘/’) to indicate that they are nested. Two other operators that sometimes occur in structure formulae are the dot (‘.’) and the plus (+). A dot placed between two factors signifies all combinations of the two factors and a plus would indicate independence between two factors.

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).

Example VII.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 Students, with 3 levels. Students is nested within Sex because, for example, student 1 from the males is a completely different student to student 1 from the females; there is no characteristic that these two students have in common.

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

Example VII.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. These two factors are crossed because the male and female in the first age group have in common that they are both 18.

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

Example VII.1 Calf diets (continued)

The factors were designated:

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

So are the unrandomized factors crossed or nested? That is, 'Do we have information that connects one of the calves from each 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 VII.3 Wheat samplers (continued)

The factors were designated:

3. Unrandomized factors – Interval, Area
4. Randomized factors – Samplers

Are the unrandomized factors crossed or nested? That is, 'Is there anything that connects one of the four intervals of each area?' **Answer** Yes, one of the four from each area is the same interval. They are crossed and this is written symbolically as Interval*Area.

Thus the experimental structure for this experiment is:

Structure	Formula
unrandomized	*4* Interval**4* Area
randomized	*4* Samplers

Notes:

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

- Genstat calls the unrandomized factors BLOCK factors and the randomized factors TREATMENT 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 areas are used in each interval, the relationship would be nested if the samplers to be used had been randomized to the intervals within each area. 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 had been employed.

		Interval			
		1	2	3	4
Area	I	B	A	D	C
	II	A	D	C	B
	III	A	C	D	B
	IV	B	A	D	C
		Sampler (A–D)			

In this design, Areas form the blocks and each sampler is used in an area once and only once. The same cannot be said of Intervals. The appropriate unrandomized structure for this design is Area/Interval.

- 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. Terms 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 VII.3: The rules for expanding structure formulae involving two factor A and B are:

A*B = A + B + A.B
A/B = A + A.B.

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 + $\mathcal{L}$.M

where $\mathcal{L}$ is the term formed from the combination of all factors in L. ■

Example VII.1 Calf diets (continued)

Dam/Calf = Dam + Dam.Calf

Example VII.3 Wheat samplers (continued)

Interval*Area = Interval + Area + Interval.Area

D. Degrees of freedom

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

Definition VII.8: For two terms each consisting of a set of factors, the first term is **marginal** to the second if for the **X** matrices corresponding to the terms, $\mathbf{X}_1$and $\mathbf{X}_2$ say, $\mathcal{C}(\mathbf{X}_1) \subseteq \mathcal{C}(\mathbf{X}_2)$, that is if the columns of $\mathbf{X}_1$ can be written as linear combinations of the columns of $\mathbf{X}_2$. ■

This definition is similar to definition III.4 that defines the marginality of two models.

Rule VII.4: One term is marginal to another if the factors specifying the marginal term are a subset of the factors specifying the term to which it is marginal. ■

Of course, a term is marginal to itself.

Example VII.3 Wheat samplers (continued)

Consider the terms Interval and Interval.Area from the Latin square example. These two terms are from the first structure of the experiment and the **X** matrices corresponding to them are:

$$\begin{array}{rc} \text{Interval} & \begin{bmatrix} \mathrm{I} & \mathrm{II} & \mathrm{III} & \mathrm{IV} \end{bmatrix} \\ \mathbf{X}_\mathrm{I} = & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$\begin{array}{rl}
\text{Interval} & \left[\qquad \text{I} \qquad\qquad \text{II} \qquad\qquad \text{III} \qquad\qquad \text{IV} \qquad\right] \\
\text{Area} & \left[1\ \ 2\ \ 3\ \ 4\ \ 1\ \ 2\ \ 3\ \ 4\ \ 1\ \ 2\ \ 3\ \ 4\ \ 1\ \ 2\ \ 3\ \ 4\right] \\
\mathbf{X}_{\text{I.A}} = & \begin{bmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1
\end{bmatrix}
\end{array}$$

The four columns of $\mathbf{X}_{\text{I}}$ are the sums of columns 1–4, 6–8, 8–12 and 15–16, respectively, of $\mathbf{X}_{\text{I.A}}$. Also, the factor in the term Interval is a subset of those in Interval.Area. Hence, Interval is said to be marginal to Interval.Area.

Rule VII.5: The Hasse diagrams for Term Marginalities for a structure formula is formed by placing terms above those to which they are marginal and connecting them by a vertical line. A term for the grand mean is included at the top of the diagram. To the left of a term is written the total number of levels combinations observed for that term. To the right of the term is written the degrees of freedom for that term. It is calculated as the difference between the left-hand entry for that term and the sum of the degrees of freedom of all the terms marginal to the current term.■

Rule VII.6: When all the factors are crossed, the degrees of freedom of any term can be calculated directly. The rule for doing this is:

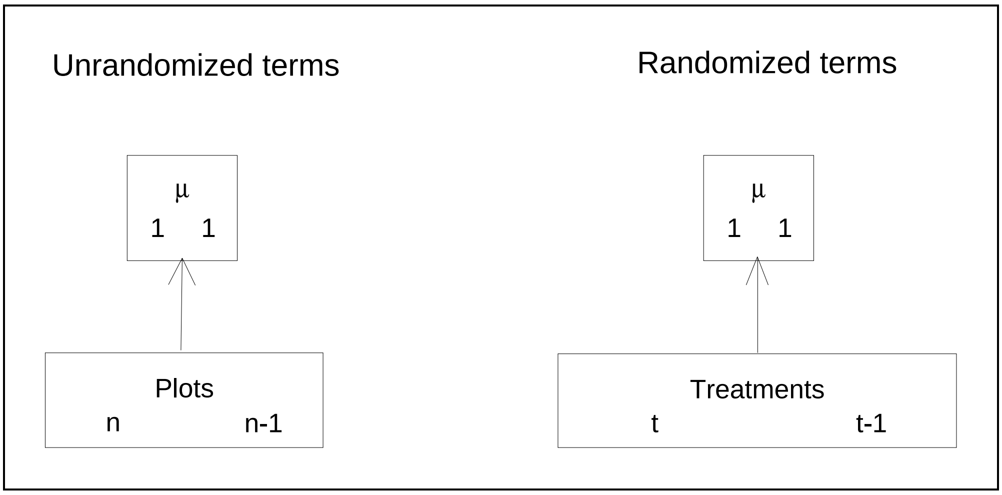
> For each factor in the term, calculate the number of levels minus one and multiply these together. ■

Example VII.3 Wheat samplers (continued)

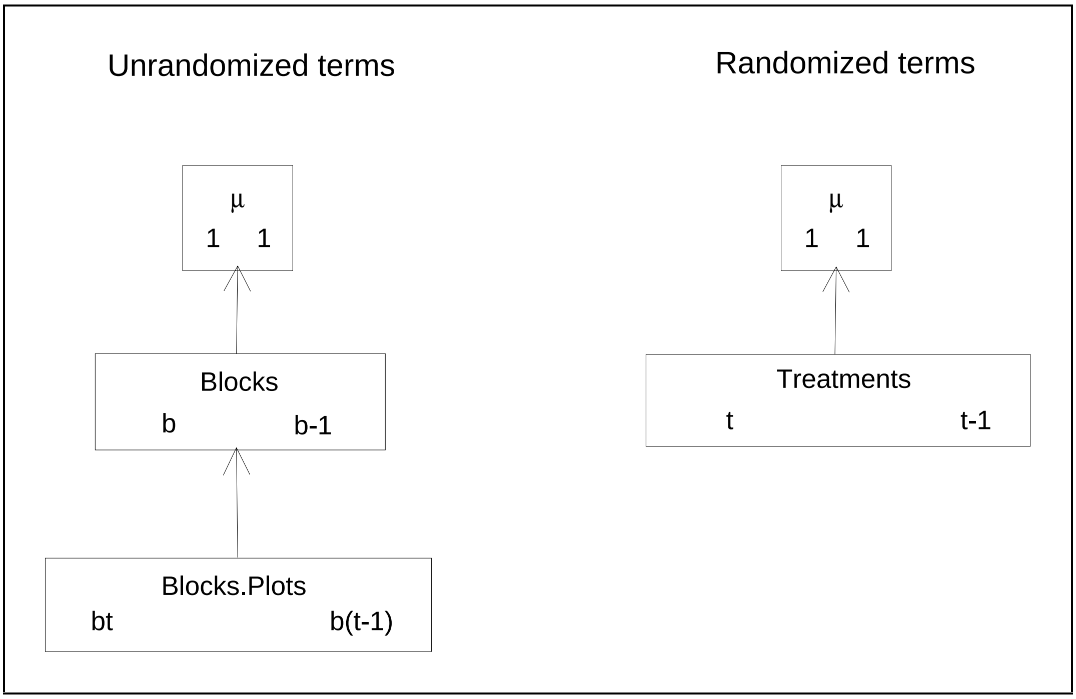
Since both Interval and Area have 4 levels, the degrees of freedom of Interval.Area is $(4–1)(4–1) = 3^2 = 9$.

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

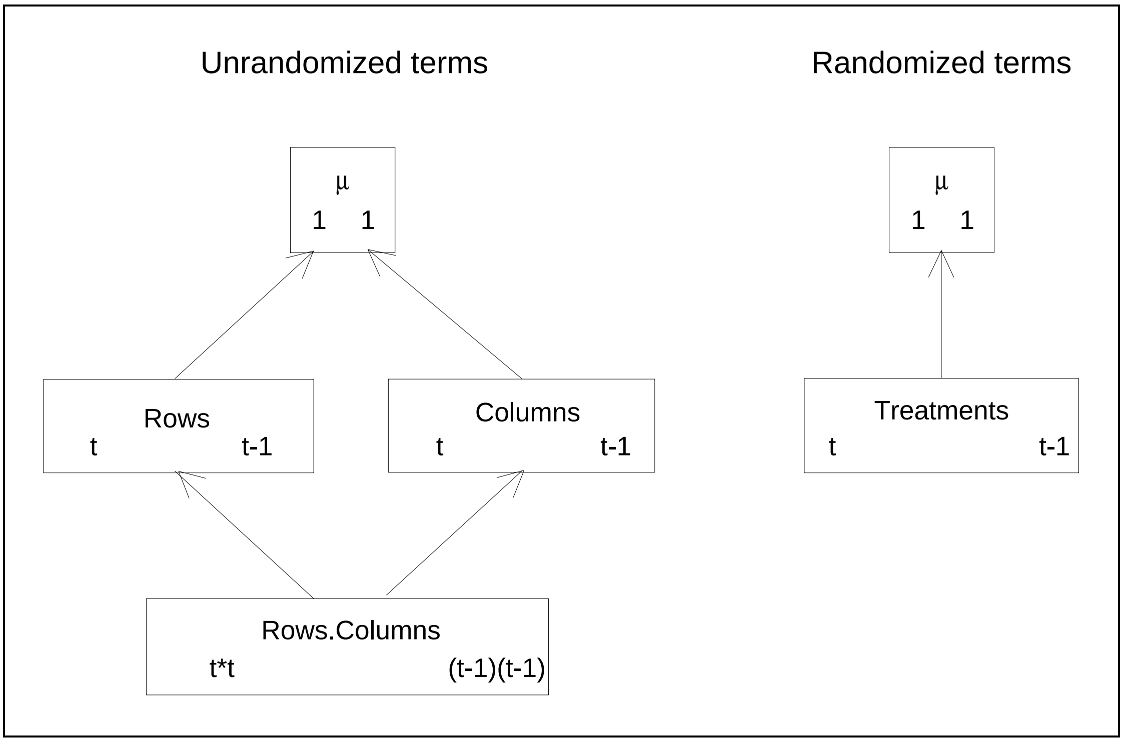
Hasse diagrams for a completely randomized design



Hasse diagrams for a randomized complete block design



Hasse diagrams for the Latin Square design



E. The analysis of variance table

Rule VII.7: The analysis of variance table is formed by listing down all the unrandomized terms and then placing the randomized terms indented under the unrandomized terms with which they are confounded. Residual sources are added to account for the left-over portions of unrandomized terms. ■

Example VII.1 Calf diets (continued)

	Source	
unrandomized ➔	Dam	4
	Dam.Calf	5
randomized ➔	Diet	1
	Residual	4

Example VII.3 Wheat samplers (continued)

	Source	
	Interval	3
unrandomized ➔	Area	3
	Interval.Area	9
randomized ➔	Samplers	3
	Residual	6

A question that arises here is what do the various terms represent. The single terms, such as Dam, Interval and Area are straightforward; they represent differences between the means for the different levels of these factors. For example, the difference between the means for each Dam. The terms Dam.Calf and Interval.Area represent the combinations of the levels of the two factors; however, their interpretation depends upon which other single factor terms have been included in the table.

In general, for two factors A and B there are the following possibilities:

i) **A.B only** — A.B measures the differences between the COMBINATIONS OF A AND B

ii) **A and A.B** — measures NESTED EFFECTS OF A AND B, that is, differences between B within A. For example, Dam.Calf measures the difference between calves with the same Dam.

iii) **B and A.B** — as for ii) except A and B interchanged

iv) **A, B and A.B** — measures INTERACTION OF A AND B. For example Interval.Area, which measures the extent to which Area differences change from interval to interval. (More about interaction later.)

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 terms. The rules for obtaining these terms are as follows:

Rule VII.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 the whether a term will contribute to expectation or variation as follows: terms that involve only fixed factors contribute to expectation and terms that contain at least one random factor are variation terms. There must be at least one random factor in an experiment.

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. ■

So the first step in determining the model for an experiment is to classify **all** the factors in the experiment as fixed or random. The definitions for these have been given before.

Definition VII.9: A factor will be designated as **random** if it is considered appropriate to use a probability distribution function to describe the distribution of effects associated with the population set of levels.

Definition VII.10: A factor will be designated as **fixed** if it is considered appropriate to have the effects associated with the population set of levels for the factor differ in an arbitrary manner, rather than being distributed according to a regularly-shaped probability distribution function.

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 occur in expectation model.

Example VII.1 Calf diets (continued)

The usual maximal model for an RCBD is

$$\psi = E\left[\mathbf{Y}\right] = \mathbf{X}_{\mathrm{B}}\beta + \mathbf{X}_{\mathrm{T}}\tau \text{ and } \operatorname{var}\left[\mathbf{Y}\right] = \sigma^2\mathbf{I}.$$

This could be expressed symbolically as

E[Y] = Diet + Dam

and V[Y] = Dam.Calf.

However, this is not the only possible maximal model for the experiment, as outlined in the lecture on the randomized complete block design. In the case of the example, an alternative model would be:

$$E\left[\mathbf{Y}\right] = \mathbf{X}_{\mathrm{T}}\tau \text{ and } \mathbf{V} = \sigma^{\mathbf{2}}\mathbf{I}_n + \sigma_\beta^2\left(\mathbf{I}_b \otimes \mathbf{J}_t\right).$$

This could be written symbolically as

E[Y] = Diet

and V[Y] = Dam + Dam.Calf.

That is, 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.

As discussed in the lecture the first model corresponds to Dam being fixed, whereas the second model corresponds to Dam random. Which model will we use here?

The two levels of Diet may well be the population set of levels and we anticipate arbitrary differences between them so it is designated to be a fixed factor. However, there is a large number of dams that the five observed dams represent and 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 could be described using a probability distribution with some variance. Thus, it is also appropriate to designate Calf to be a random factor. Diagnostic checking will be used to try and support this last assumption. The terms in the analysis are Dam, Dam.Calf and Diet. As the first two terms contain at least one random factor, they are variation terms. The term Diet consists of only a fixed factor and so it is an expectation term. Hence, the maximal model to be used for this experiment is the alternative model given above:

$$E[Y] = \text{Diet}$$

and $V[Y] = \text{Dam} + \text{Dam.Calf}$.

As outlined in the lecture on RCBD experiments, 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 VII.9: The general rules for constructing the expected mean squares are:

1. Write down a component of variation for each variation term;

2. Determine the multiplier for each variation component. For a particular component it is the **replication** of the combinations corresponding to that component (multiplied by an efficiency factor if the term is nonorthogonal, but balanced).

3. For each variation component, write the component, with its multiplier, in any line in the table which:

   is marginal to the term corresponding to the component

   or

   is indented under a line which is marginal to the term corresponding to the component;

4. For each line in the table whose source is an expectation term, include an expectation component; this component is a function of the expectation vector that depends on the term for the line. ■

VII.B The Latin square example

Example VII.3 Wheat samplers (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 – an interval in an area
2. Response variable – Error
3. Unrandomized factors – Interval, Area
4. Randomized factors – Samplers
5. Type of study – Latin Square

B. The experimental structure

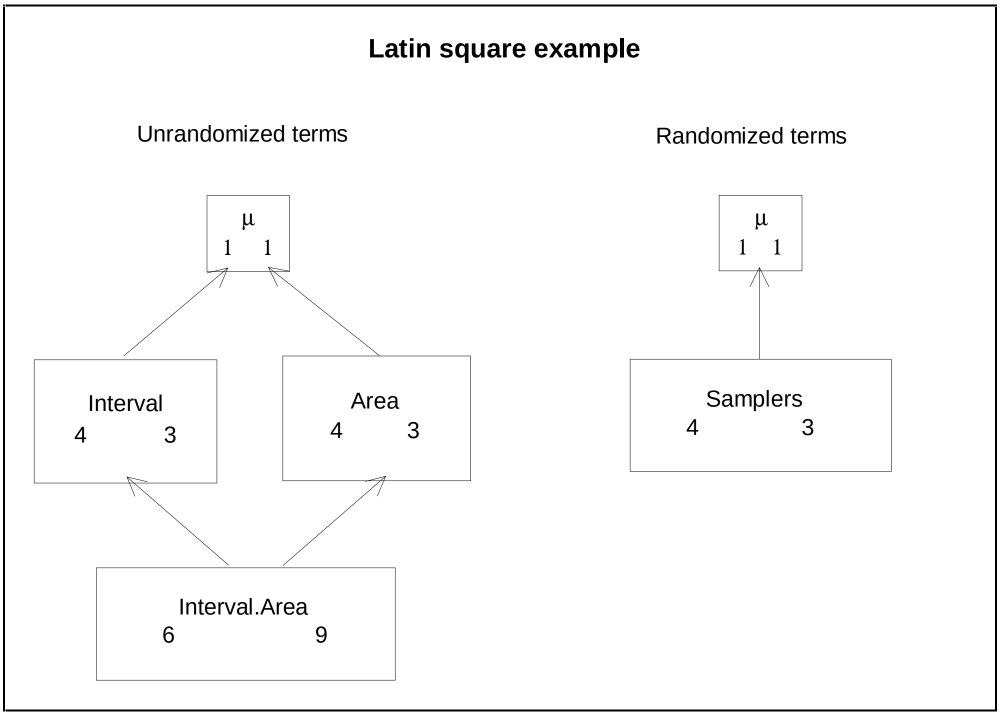
Structure	Formula
unrandomized	*4* Interval**4* Area
randomized	*4* Samplers

C. Terms derived from the structure formulae

Interval*Area = Interval + Area + Interval.Area

Samplers = Samplers

D. Degrees of freedom



E. The analysis of variance table

Source	df
Interval	3
Area	3
Interval.Area	9
Samplers	3
Residual	6

F. Maximal expectation and variation models

Take the random factors to be Areas and Samplers and the fixed factor to be Intervals. Then the maximal expectation and variation models are

E[Y] = Intervals and
var[Y] = Samplers + Area + Area.Interval

G. The expected mean squares.

The expectation term is Interval; the variation terms are: Samplers, Area, and Interval.Area. Thus, the variation components will be σ_{S}^2, σ_{A}^2 and σ_{IA}^2, respectively; the multipliers of these components are 4, 4 and 1, respectively

Source	df	E[MSq]
Interval	3	$\sigma_{\mathrm{IA}}^2 + f_{\mathrm{I}}(\psi)$
Area	3	$\sigma_{\mathrm{IA}}^2 + 4\sigma_{\mathrm{A}}^2$
Interval.Area	9	
Samplers	3	$\sigma_{\mathrm{IA}}^2 + 4\sigma_{\mathrm{S}}^2$
Residual	6	σ_{IA}^2
Total	15	

VII.C Rules for determining the analysis of variance table

A. Description of pertinent features of the study

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

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

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

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

Definition VII.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 VII.1: To determine the unrandomized factors, ask the following question of each factor:

> For each 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 VII.2: Determine the experimental structure by

1. describing the crossing and nesting relationships between the unrandomized factors in the experiment,

2. describing the crossing and nesting 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. ■

Definition VII.6: Two factors are **crossed** if a level of one factor has a characteristic associated with the factor in common across all levels of the other factor. ■

Definition VII.7: Two factors are **nested** if the same level of the nested factor from different levels of the nesting factor have no characteristic in common. ■

C. Terms 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 VII.3: The rules for expanding structure formulae involving two factor A and B are:

A*B = A + B + A.B
A/B = A + A.B.

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 + $\mathcal{L}$.M

where $\mathcal{L}$ is the term formed from the combination of all factors in L. ■

D. Degrees of freedom

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

Rule VII.4: One term is marginal to another if the factors specifying the marginal term are a subset of the factors specifying the term to which it is marginal. ■

Definition VII.8: For two terms each consisting of a set of factors, the first term is **marginal** to the second if for the **X** matrices corresponding to the terms, $\mathbf{X}_1$ and $\mathbf{X}_2$ say, $\mathcal{C}(\mathbf{X}_1) \subseteq \mathcal{C}(\mathbf{X}_2)$, that is if the columns of $\mathbf{X}_1$ can be written as linear combinations of the columns of $\mathbf{X}_2$. ■

Rule VII.5: The Hasse diagrams for Term Marginalities for a structure formula is formed by placing terms above those to which they are marginal and connecting them by a vertical line. A term for the grand mean is included at the top of the diagram. To the left of a term is written the total number of levels combinations observed for that term. To the right of the term is written the degrees of freedom for that term. It is calculated as the difference between the left-hand entry for that term and the sum of the degrees of freedom of all the terms marginal to the current term.■

Rule VII.6: When all the factors are crossed, the degrees of freedom of any term can be calculated directly. The rule for doing this is:

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

E. The analysis of variance table

Rule VII.7: The analysis of variance table is formed by listing down all the unrandomized terms and then placing the randomized terms indented under the unrandomized terms with which they are confounded. Residual sources are added to account for the left-over portions of unrandomized terms. ■

F. Maximal expectation and variation models

Rule VII.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 the whether a term will contribute to expectation or variation as follows: terms that involve only fixed factors contribute to expectation and terms that contain at least one random factor are variation terms. There must be at least one random factor in an experiment.

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 VII.9: A factor will be designated as **random** if it is considered appropriate to use a probability distribution function to describe the distribution of effects associated with the population set of levels.

Definition VII.10: A factor will be designated as **fixed** if it is considered appropriate to have the effects associated with the population set of levels for the factor differ in an arbitrary manner, rather than being distributed according to a regularly-shaped probability distribution function.

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 occur in expectation models.

G. The expected mean squares

Rule VII.9: The general rules for constructing the expected mean squares are:

1. Write down a component of variation for each variation term;

2. Determine the multiplier for each variation component. For a particular component it is the **replication** of the combinations corresponding to that component (multiplied by an efficiency factor if the term is nonorthogonal, but balanced).

3. For each variation component, write the component, with its multiplier, in any line in the table which:

   is marginal to the term corresponding to the component

   or

   is indented under a line which is marginal to the term corresponding to the component;

4. For each line in the table whose source is an expectation term, include an expectation component; this component is a function of the expectation vector that depends on the term for the line. ■

VII.D Determining the analysis of variance table – further examples

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

A. Description of pertinent features of the study
B. The experimental structure
C. Terms derived from the structure formulae
D. Degrees of freedom
E. The analysis of variance table
F. Maximal expectation and variation models
G. The expected mean squares.

Example VII.8 Fertilizer effects on kale

An experiment is to be conducted to investigate the effect of two levels of potassium (K), three levels of nitrogen (N) and two levels of phosphorus (P) on the yield of three varieties of kale (couve). The two levels of potassium were none and 50 kg added, the three levels of nitrogen were none, 25 kg and 50 kg added and the two levels of phosphorus were none and 30 kg added. The 36 treatment combinations are to be applied using a randomized complete block design with 3 blocks of 36 plots.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

Blocks.Plots = Blocks + Blocks.Plots

N*P*K*Varieties

= N + P + K + Variety
+ N.P + N.K + N.Variety + P.K + P.Variety + K.Variety
+ N.P.K + N.P.Variety + N.K.Variety + P.K.Variety
+ N.P.K.Variety

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

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

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

Var[Y] =

E[Y] =

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

	df	E[MSq]
Blocks	2	$\sigma^2_{\mathrm{BP}} + 36\sigma^2_{\mathrm{B}}$
Blocks.Plots	105	
N	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{N}}(\psi)$
P	1	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{P}}(\psi)$
K	1	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{K}}(\psi)$
Variety	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{V}}(\psi)$
N.P	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NP}}(\psi)$
N.K	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NK}}(\psi)$
N.Variety	4	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NV}}(\psi)$
P.K	1	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{PK}}(\psi)$
P.Variety	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{PV}}(\psi)$
K.Variety	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{KV}}(\psi)$
N.P.K	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NPK}}(\psi)$
N.P.Variety	4	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NPV}}(\psi)$
N.K.Variety	4	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NKV}}(\psi)$
P.K.Variety	2	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{PKV}}(\psi)$
N.P.K.Variety	4	$\sigma^2_{\mathrm{BP}} + f_{\mathrm{NPKV}}(\psi)$
Residual	70	σ^2_{BP}

Example VII.9 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]

Example VII.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.

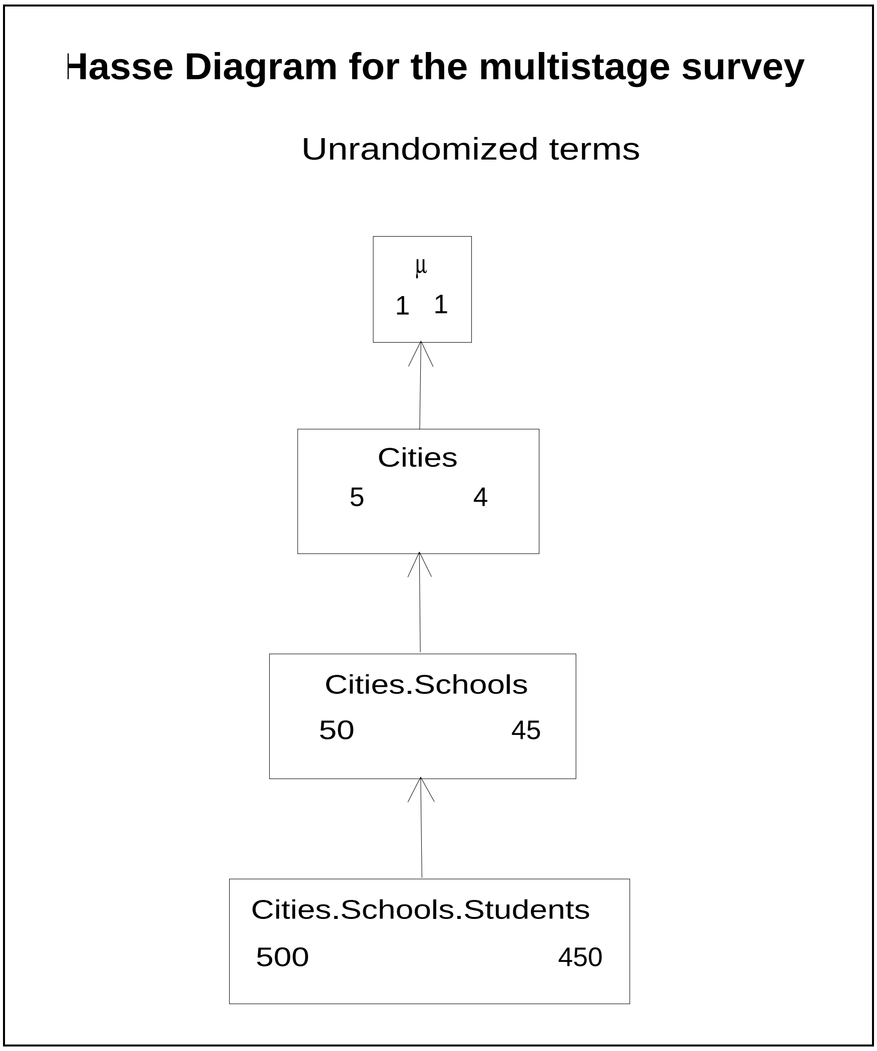
1. the observational unit
2. response variable
3. unrandomized factors
4. randomized factors
5. type of study

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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



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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is

Source	df	E[MSq]

Example VII.11 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is

Source	df	E[MSq]

Example VII.12 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. 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 4 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	D	B	D	B	C	A	B	A	D	A	C
	II	A	A	D	B	C	C	B	C	D	A	B	D
	III	D	C	C	B	B	C	A	D	A	B	A	D
	IV	B	B	A	D	C	D	B	D	C	C	A	A

In working out the analysis for this experiment include a term for Block × Treatment interaction and assume 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]
Total		

Example VII.13 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]

Example VII.14 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]

Example VII.15 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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:

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]

Example VII.16 Wheat samplers

Suppose a study is to be conducted to investigate whether four samplers differ in the amount by which they differ in the amount of error in their selection of wheat samples. Four areas and four intervals are to be employed in the study and the following Latin Square arrangement is to be used to assign the samplers to the interval-area combinations:

		Area			
		1	2	3	4
	I	A	B	D	C
	II	D	C	A	B
Intervals	III	B	D	C	A
	IV	C	A	B	D

(Samplers A, B, C, D)

1. the observational unit – a interval in a area
2. response variable – Kilometres per litre
3. unrandomized factors – Intervals, Areas
4. randomized factors – Samplers
5. type of study – a Latin Square

The experimental structure is:

Structure	Formula
unrandomized	*4* Intervals**4* Areas
randomized	*4* Samplers

The terms derived from these structure formulae are:

Intervals*Areas = Intervals + Areas + Intervals.Areas

Samplers = Samplers

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

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

Will take the random factors to be Areas and Samplers and the fixed factor to be Intervals. Then the maximal expectation and variation models are

E[Y] = Intervals and
var[Y] = Samplers + Area + Area.Interval

What are the variation components, and their multipliers, for this study?

$$4\sigma_{\mathrm{S}}^2,\ 4\sigma_{\mathrm{A}}^2 \text{ and } 1\sigma_{\mathrm{IA}}^2$$

The analysis of variance table is:

Source	df	E[MSq]
Intervals	3	$\sigma_{\mathrm{IA}}^2 + f_{\mathrm{I}}(\psi)$
Areas	3	$\sigma_{\mathrm{IA}}^2 + 4\sigma_{\mathrm{A}}^2$
Intervals.Areas	9	
Samplers	3	$\sigma_{\mathrm{IA}}^2 + 4\sigma_{\mathrm{S}}^2$
Residual	6	σ_{IA}^2

Suppose that the experiment is to be repeated by replicating the Latin Square twice using the same areas but new intervals on a second occasion. What are the components of the study?

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]

Example VII.17 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.

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

The experimental structure is:

Structure	Formula
unrandomized	
randomized	

The terms derived from these structure formulae are:

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

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

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

What are the variation components, and their multipliers, for this study?

The analysis of variance table is:

Source	df	E[MSq]

Example VII.18 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 8 sets 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.

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

The experimental structure is:

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?

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

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

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

$$\text{Var}[Y] =$$

$$E[Y] =$$

What are the variation components, and their multipliers, for this study?

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

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