Model Based Statistics in Biology.
Part IV. The General Linear Model. Multiple Explanatory Variables.
Chapter 13.6 Nested Factors (Hierarchical ANOVA

ReCap.	Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap	Part III (Ch 9, 10, 11)
ReCap	Multiple Regression (Ch 12)
13.1	Fixed Effects ANOVA (no interactive effects)
13.2	Fixed Effects ANOVA (interactive effects)
13.3	Fixed*Random Factors (Paired t-test)
13.4	Fixed*Random Factors (Randomized Block)
13.5	Fixed*Random Factors (Repeated Measures)
13.6	Nested Random Factors (Hierarchical ANOVA)
13.7	Random within Fixed (Hierarchical ANOVA)
13.8	More Than Two Factors (to be written)

on chalk board

Fly wing length data
Sokal and Rohlf Box 10.1
Ch13.xls

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning is based on models, including statistical analysis based on models.
ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to declare a decision.
Estimation is concerned with the specific value of an unknown population parameter.
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12) GLM with more than one regression variable (multiple regression)
ReCap (Ch 13) GLM with more than one categorical variable (ANOVA).
Two fixed factors (Ch 13.1, Ch13.2)
One fixed and one random factor (Paired t-test, Randomized block),
One random and one or more fixed factors (Repeated measures)

Today: Special case of two factor ANOVA: Hierarchical ANOVA
Both factors random.
The logical relation of one factor to another is hierarchical:
one factor is nested within another.
This is in contrast to crossed designs, such as two-way ANOVA and randomized blocks.

Wrap-up. Comparison of hierarchical with two-way ANOVA.

Two-way ANOVA has an interaction term. Testing starts with this term.

In randomized blocks, the interaction term is present logically, but assumed to be zero if treatments were assigned randomly to blocks.

Hierarchical ANOVA differs from crossed designs. The interaction term is known to be zero, because units of analysis cannot be matched across treatments.

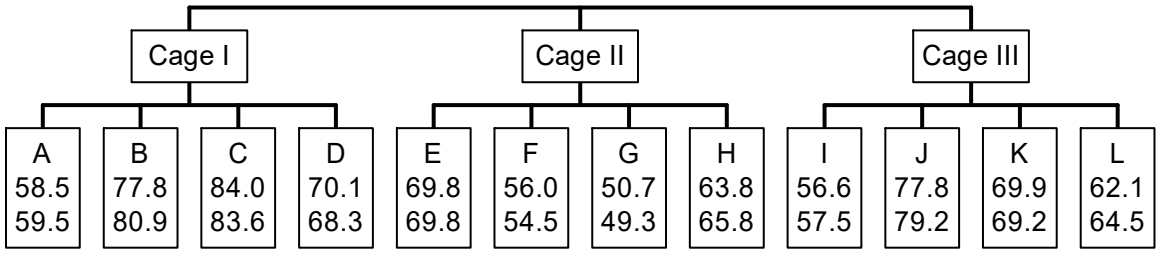
Introduction.
In the examples so far we analyzed variation in the response variable relative to crossed factors. The levels within one explanatory variable (e.g. food type) could be matched with levels in a second explanatory variable (e.g. sex). Consequently, we can display the data in a two-way table. In such a table the numbers in the table are the response variable. The column labels are one factor. The row labels are the other factor.
Here are 4 examples—2 with fixed factors and two with a random and a fixed factor.

Response variable	Explanatory variable	Explanatory variable
Limpet respiration	Species (fixed)	Salinity (fixed by experiment)
Rat weight gain	Protein source (fixed)	Protein Level (fixed)
Hours of extra sleep	Subject (random)	Drug (fixed)
Tribolium dry weight	Block (random)	Genotype (fixed)

Here is an example with 2 explanatory variables that are nested, not crossed.
Sokal and Rohlf (1995 Box 10.1) report wing lengths of 4 flies in 3 cages.

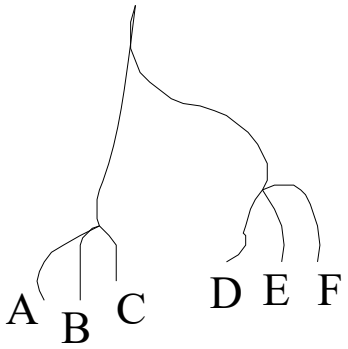
Response variable	Explanatory variable	Explanatory variable
Fly wing length	Cage (random)	Fly (random)

We have no information to match a fly in Cage I (Fly A, B, C, D) to a fly in Cage II (Fly E, F, G, H). Here is a diagram.



We have two wing measurements but no way to match these across flies. Wing length is nested within fly.

Tree. A graphical expression of hierarchical ANOVA is a tree. For example we draw a few branches representing genera, then for each branch draw twigs representing species. The twigs cannot be aligned across branches, so the design is hierarchical.



Mobile. Another visualization is a hanging mobile. Near the top of the mobile are branches that rotate around a balance point. Beneath each branch there are sub-branches, which rotate around a balance point below the branch. The sub-branches beneath one branch rotate independently of those beneath another branch. They cannot be aligned. The design is hierarchical.

Crossed versus nested factors.
Two crossed factors can also be shown as a branching diagram. However a branching diagram does not always show a nested design. To distinguish crossed from nested, we inspect a two way table.

	Fly within Cage											
Cage	I	A	B	C	D							
	II					E	F	G	H			
	III									I	J	K	L

The table has more empty than occupied cells. The interactive effect cannot be estimated. The interactive variability becomes part of the lower level term as follows:

Fly + Cage x Fly - - > Fly(Cage)

More generally: B + A x B - - > B(A)

The symbol B(A) reads naturally from left to right as "B within A." Set notation also reads left to right: $B \subset A$. This notation is readily used in reading, thinking, and writing about nested designs. Most statistical packages use this notation. R does not. It uses Cage/Fly or more generally A/B where B is nested within A.

Here is an accounting of degrees of freedom showing how joining an interaction term with the lower level term produces the df for the nested term.

Crossed	
Source	df
Factor A	3-1 = 2
Factor B	4-1 = 3
A x B	2 x 3 = 6

Nested			
Source	df	Source	df
Factor A	3-1 = 2	Factor A	3-1 = 2
Factor B(A)	3 + 6 = 9	Factor B(A)	(4-1)*3 = 9

B + A × B - - > B(A)
3 + 6 = 9

A nested factor ANOVA can have mixed terms—random within fixed. It can also have all random terms—random within random.

Both Random			
B crossed with A		B nested in A	
A	Random	A	Random
B	Random	B(A)	Random
A × B	Random		

Mixed			
B crossed with A		B nested in A	
A	Fixed	A	Fixed
B	Random	B(A)	Random
A × B	Mixed		

Not all random factors are nested. For example, a latin square design has two random factors that are crossed.

Example Data from Sokal and Rohlf (1995) Box 10.1
Two measurements made on the left wings of 4 mosquitos in each of 3 cages.
This is a nested design. We cannot match flies across cages.
At a lower level, we cannot match wing measurements across flies.

1. Construct model

Verbal model.
Does mosquito wing length vary among cages as well as among mosquitoes?

Graphical model. Plot of the means of the measurements for each fly in each cage.

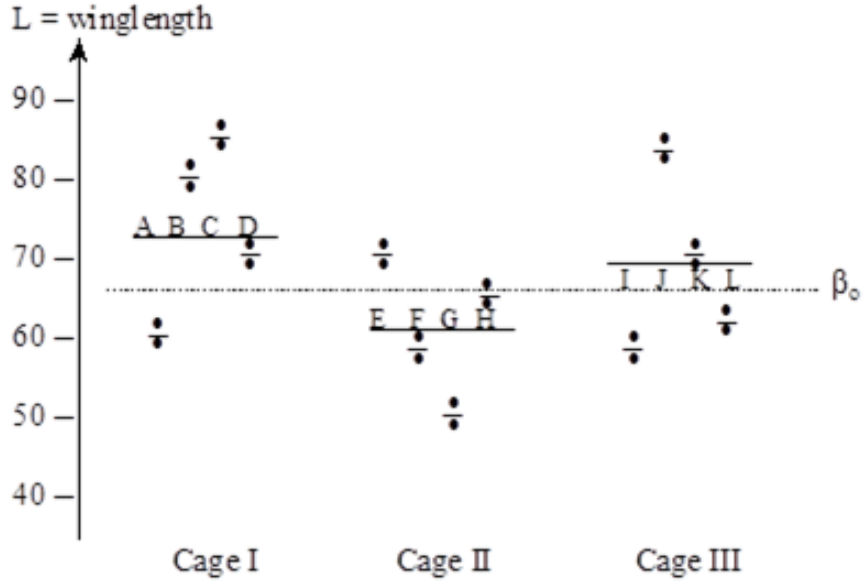


Fig L17b

Response variable is winglength
L = micrometer units
(ratio type of scale)

First explanatory variable is cage X_{cage} (categorical variable)
Second explanatory variable is fly within cage $X_{fly \subset cage}$ (categorical variable)
or equivalently $X_{fly(cage)}$

Each fly gets a different label. We avoid using the label "Fly A" in cage II or III because flies cannot be matched across cages.

$X_{fly(cage)}$ = Fly A Fly B Fly C Fly D in cage I
$X_{fly(cage)}$ = Fly E Fly F Fly G Fly H in cage II
$X_{fly(cage)}$ = Fly J Fly K Fly L Fly M in cage III

Write GLM:	Y	=	β_o	+	$\beta_{cage}X_{cage}$	+	$\beta_{fly(cage)}X_{fly(cage)}$	+	res
Component	Y_{ijk}	=	μ	+	A_i	+	B_{ij}	+	ε_{ijk}

GLM notation shows parameters. Component notation distinguishes random factors (roman letters) from fixed factors (greek letters).

2. Execute analysis.

Place data in model format:

Column labelled L, with response variable fly wing length
Column labelled X_{cage}, with explanatory variable X_{cage} = I, II, or III
Column labelled $X_{fly(cage)}$ with label (number) for each fly

Code model statement in statistical package according to the GLM

$$Len = \beta_o + \beta_{cage} \cdot X_{cage} + \beta_{fly(cage)} \cdot X_{fly(cage)} + \varepsilon$$

```
MTB > anova 'Len' = 'cage'   'fly'('cage');
SUBC> random 'cage' 'fly'('cage').
```

Minitab

```
FlyMod8<-aov(Len ~ Cage + Error(Cage/Fly), SRBx10_1)
```

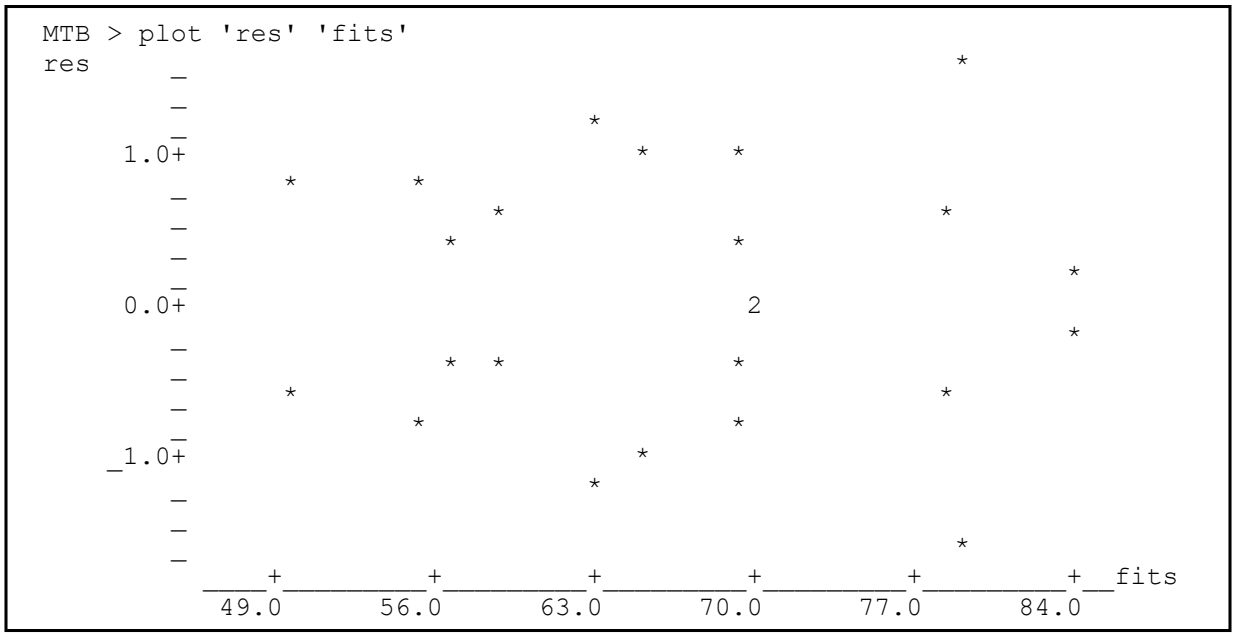
R-code

The code estimates of the overall mean ($\hat{\beta}_o$)
and the mean for each cage ($\hat{\beta}_o + \hat{\beta}_{cage} \cdot X_{cage}$).
The fitted values are the means for each fly ($\hat{\beta}_o + \hat{\beta}_{cage} \cdot X_{cage} + \hat{\beta}_{fly(cage)} X_{fly(cage)}$).

$$\hat{\beta}_o = \text{Mean(L)} = 24\text{-}1\ 1599.2 = 66.633$$

$$\hat{\beta}_o + \hat{\beta}_{cage} \cdot X_{cage} = \begin{cases} \text{Mean}(L_{cageI}) = 8^{-1}\ 582.7 = 72.84 \\ \text{Mean}(L_{cageII}) = 8^{-1}\ 479.7 = 59.96 \\ \text{Mean}(L_{cageIII}) = 8^{-1}\ 536.8 = 67.10 \end{cases}$$

3. Evaluate the model Plot residuals versus fitted values.



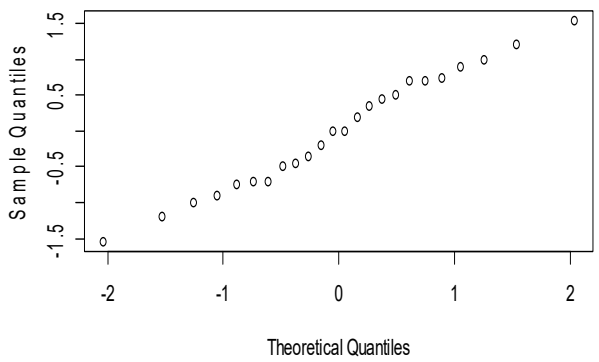
A. No line fitted in model, so skip evaluation of straight line assumption.

B1. Residuals homogeneous? Yes.

No systematic change in residuals with increase in fitted values (*i.e.* no fans).

3. Evaluate the model

B2. Normal?




```
Histogram of res   N = 24
Midpoint   Count
    _1.5       1  *
    _1.0       3  ***
    _0.5       6  ******
     0.0       4  ****
     0.5       5  *****
     1.0       4  ****
     1.5       1  *
```

B2. Normal ? Yes.

B3. Independent?

```
MTB > let c6 = lag('res')
MTB > plot c6 'res'
C6
        _                                  *
        _
        _
     1.0+              *       *
        _                 2
        _    *                          *
        _                 *     *
        _                             *
        _
     0.0+                             *            *
        _                                          *
        _          *                        2
        _             *                          *
        _                                *         *
        _
    _1.0+                                                *
        _                                                   *
        _
        _                                                      *
        _
          ________+_________+_________+_________+_________+________res
              _1.20      _0.60       0.00       0.60       1.20

MTB > corr c5 c6
Correlation of res and C6 = _0.567
```

Lag plot shows a downward trend because the data are organized according to cage and fly. The result is negative correlation of adjacent residuals by definition r = - 0.5 (Anscombe, F.J. Tukey J.W. 1963. The examination and analysis of residuals. *Technometrics* 5: 141-160)

Induced correlation is evident in the downward trend in plot of adjacent values. Lag plots are misleading when replication is less then 4 at the lowest level in the analysis (at the level of fly in this analysis).

The errors are acceptably homogeneous and normal.

4. Partition df and SS according to model. Calculate *LR*

Compute total df: $n-1 = 24 - 1 = 23$

Partition df according to model

Cage term: 3 cages hence $3-1 = 2$ df.

Fly term: 4 flies per cage hence $4-1 = 3$ df per cage.

$3*(4-1) = 9$ df for the fly within cage term.

Calculate $SS_{total} = 23* \text{Var}(Len) = 23*104.43 = 2401.98$

By hand: $SS_{total} = \Sigma Y^2 - n^{-1}(\Sigma Y)^2 = 108962 - 2^{-1} \cdot 1599.2^2 = 2401.98$

In Minitab:

```
MTB > let k1 = ssq('Len')
MTB> print k1
   k1  2401.98
```

In a spreadsheet

```
=DevSQ(A1:A24)
```

In R

```
> sum( (x - mean(x) )^2 )
[1] 2401.98
```

Use statistical software to partition the SS_{total} and produce the ANOVA table.

$$Y - \beta_o = \beta_{cage}X_{cage} + \beta_{fly(cage)}X_{fly(cage)} + res$$

```
MTB> GLM    'wlength' = 'Cage' 'Fly'('cage')
SUBC> random 'cage' 'fly'('cage').
```

The subcommand `SUBC> random` forces Minitab to compute the correct F-ratio.

```
FlyMod8<-aov(WLEN ~ Cage + Error(Cage/Fly),
SRBx10 1)
```

Here is R-code to partition df, SS. The F-ratios are not correct.

Here is the partitioning of the SS_{total}.

GLM:	$Y - \beta_o$	=	$\beta_{cage}X_{cage}$	+	$\beta_{fly(cage)}X_{fly(cage)}$	+ res
Source:	total	=	cage		fly(cage)	res
df	24−1	=	3− 1	+	3*(4− 1)	+ 12
SS	2402	=	665.68	+	1720.68	+ 15.62

$1 - R^2 = (15.62/2401.97) = 0.65\ \%$

$LR = (1-R^2)^{(-24/2)} = 1.7 \times 10^{26}$

The evidence for the overall model is strong so we proceed to a calculation of the likelihood for each component of variation. In a model with random factors the likelihood ratio and the F-ratio are formed relative to the correctly nested ratio of expected mean squares, as described in texts in experimental design (*e.g.* Cochran and Cox 1957, Quinn and Keough, 2002).

4. Calculate *LR*

The likelihood ratios (and F-ratios) for the cage term are *not* formed relative to the residual SS. They are formed relative to the random term below it in the ANOVA table. Here is a step by step procedure to write out the correct LRs and F-ratios.

List the terms in the model, as in the ANOVA table

List the same term horizontally. In each row, show the row term.

		Cage	Fly%in% Cage	Error	
EMS	Cage	Cage			Each EMS includes itself
EMS	Fly%in% Cage		Fly%in% Cage		
EMS	Error			Error	

		Cage	Fly%in% Cage	Error	
EMS	Cage	Cage		Error	Each EMS includes the fixed error term
EMS	Fly%in% Cage		Fly%in% Cage	Error	
EMS	Error			Error	

		Cage	Fly%in% Cage	Error	
EMS	Cage	Cage	Fly%in% Cage	Error	Each EMS includes crossed (or nested) random terms
EMS	Fly%in% Cage		Fly%in% Cage	Error	
EMS	Error			Error	

		Cage	Fly%in% Cage	Error	Correct Denominator MS	
EMS	Cage	Cage	Fly%in% Cage	Error	Fly%in% Cage	Identify the denominator MS for the F-ratio
EMS	Fly%in% Cage		Fly%in% Cage	Error	Error	The denominator MS cancels all but the term of interest
EMS	Error			Error		

		Cage	Fly%in% Cage	Error	Incorrect Denominator MS	
EMS	Cage	Cage	Fly%in% Cage	Error		Cage MS / Error results in *two* uncancelled terms.
	Fly%in% Cage					
EMS	Error			Error	Error	The F-test is ambiguous

df	SS	1+ SSratio			
2	665.68	1.39	= (665.68 + 1720.68)/1720.68	LR = $(1.39)^{-24/2}$	LR = 51
9	1720.68	111.16	= (1720.68 + 15.62) /15.62	LR = $(111.16)^{-24/2}$	LR = 10^{24}
12	15.62				
23	2401.98				

The likelihood ratio for Cage is formed relative to the random term below it.
The likelihood ratio for Fly(Cage) is formed relative to the random term below it.
The evidence for the fly within cage term is far stronger than the evidence for the cage term.

5. State population and whether sample is representative.

This is a laboratory study with a well established measurement protocol that generates data that meets Hacking's (1965) requirement for likelihood inference (Hacking. I. 1965 *The Logic of Statistical Inference*). Inference requires a procedural statement that generates data under reproducible conditions. A better and better estimate of the value of a parameter emerges from the law of large numbers as sample size increases.

Inference is to a population of wing measurements according to a measurement protocol. Inference is to a population of cages similar to those in this study.

5. Choose mode of inference. Is hypothesis testing appropriate?

The goal of the analysis is an estimate of variance at different levels. Estimates such as these are central to statistical design. A judgement relative to Type I error is not required. A decision at a fixed Type I error is not needed. We will calculate and report likelihood ratios as a measure of evidence.

10. Report and interpret variances.

The parameters in this analysis were not of interest. The variance due to each factor is of interest. The interest is in the amount of variation, not the contrasts in means. The amount of variability due to each factor is used to design efficient experiments.

Source	df	SS	%
Cage	2	665.68	27.7
Fly□Cage)	9	1720.68	71.6
Error	12	15.62	0.7
Total	23	2401.97	

Most of the variability (72%) is among flies. As a result, in designing an experiment we would want to use many flies to reduce the mean square at this level. The percent variability is less at the level of cages, so we would not need to use many cages to reduce the mean square at the cage level. The variability due to measurement error is negligible, so we do not need to devote effort to repeating measurements on a single fly.