

Model Based Statistics in Biology.

Part IV. The General Linear Model. Multiple Explanatory Variables.

Chapter 13.6 Nested Random Effects (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 Effects (Paired t-test)
13.4	Fixed*Random Effects (Randomized Block)
13.5	Fixed*Random Effects (Repeated Measures)
13.6	Nested Random Effects (Hierarchical ANOVA)
13.7	Random within Fixed (Hierarchical ANOVA)
13.8	More Than Two Factors (to be written)

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

on chalk board

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 <u>nested</u> within another.

This is in contrast to <u>crossed</u> designs, such as two-way ANOVA and
--

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

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

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

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

Introduction.

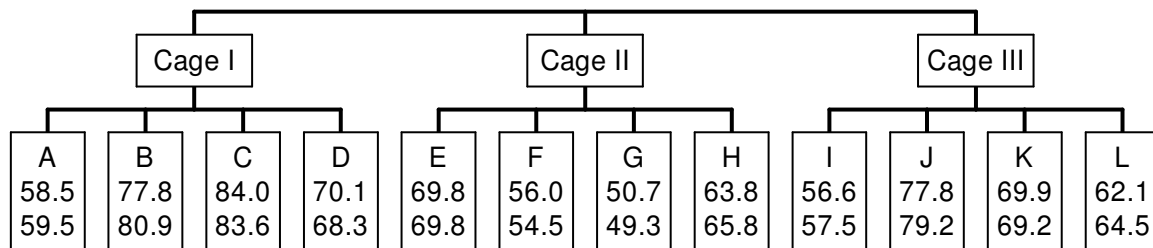
We have used the GLM to analyze a response variable in relation to a single categorical variable (one-way ANOVA, t-test). We then used the GLM to analyze a response variable in relation to two categorical variables (two-way ANOVA). These explanatory variables can be either random or fixed.

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)

In these examples the levels within the first explanatory variable (e.g. food type) could be matched with levels in the second explanatory variable (e.g. sex). Consequently, we can display the data in a two-way table. This is called a crossed design: levels within one factor can be matched with levels in another factor. However, when we have two random factors it is not possible to match levels across factors. In Box 10.1 Sokal and Rohlf (1995) report fly wing lengths of 4 flies in each of 3 cages.

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

In this example we cannot match flies (the second explanatory variable) across cages (the first explanatory variable). Here is a diagrammatic account of the situation.



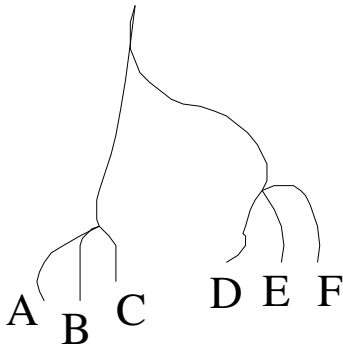
We have no information to match a fly in Cage I (Fly A, B, C, D) to a fly in Cage II (Fly E, F, G, H)

The logical structure is thus hierarchical rather than crossed.

Hierarchical ANOVA has two or more explanatory variables that are categorical and nested one within the next. Both explanatory variables are random factors. The analysis is also called nested ANOVA.

Example. Resting metabolic rate of several species of insect within several different genera. Species are nested within genera. Species cannot be matched up in any way across genera. Relation of the two explanatory variables, species and genus, is hierarchical

Tree. A graphical expression for the idea of hierarchical ANOVA is a tree. One can 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 useful 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, all of 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. This represents a hierarchical design.

As extension of 1-way ANOVA. Hierarchical ANOVA can be viewed as an extension of a one-way ANOVA.

In the case of the insect metabolic rate, we could carry out a one-way ANOVA in many individuals belonging to several different species. We could start with a test for differences in effect among species within one of the genera, using the one-way design. Then we could go on and repeat this for another genus, with another one-way ANOVA. Similarly, an ANOVA could be carried out for a third genus. These three analyses can be combined into a test of differences among genera. The object of this new analysis is to test for differences in metabolic rate at the genus level, above and beyond that observed among species within a genus.

As modification of 2-way ANOVA. Hierarchical ANOVAs can also be viewed as a modification of a two-way (or multiway) ANOVA. In the case of the insect metabolic rate we have two explanatory variables--species and genus. However, there is no way to match species in one genus with a species in another genus, so there can be no interaction term. The logic of the situation compels us to consider species within genus, and not vice-versa. The component of variance that would normally be assigned to the interaction term is assigned to the higher of the two levels, in this case the variance among genera. This is because the variance among the higher level groups includes the variance among the groups within each of the higher level groups. The variance among genera must exceed the variance within genera because it includes the variance within genera.

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 hierarchical design because the 4 mosquitoes in cage A are not matched up in any way with the 4 mosquitoes in cage B, or in cage C.

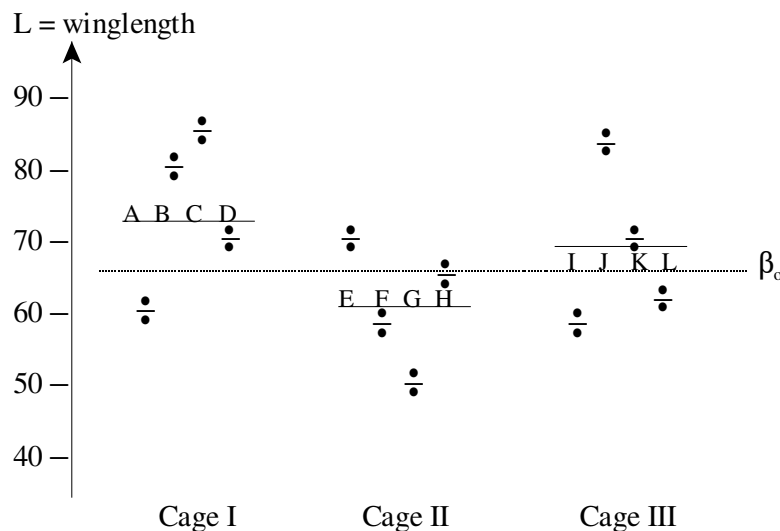
1. Construct model

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)

Explanatory variable is cage $X_{\text{cage}} = \text{I II III}$ (categorical variable)

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

$X_{\text{fly} \subset \text{cage}}$ the subscript is read “fly within cage”

Each fly gets a different label. We have avoided using the label “Fly A” in cage II.

There is no way to match a fly in cage II with Fly A in cage I.

$X_{\text{fly} \subset \text{cage}} = \text{Fly A Fly B Fly C Fly D in cage I}$

$X_{\text{fly} \subset \text{cage}} = \text{Fly E Fly F Fly G Fly H in cage II}$

$X_{\text{fly} \subset \text{cage}} = \text{Fly J Fly K Fly L Fly M in cage III}$

Write GLM: $Y = \beta_0 + \beta_{\text{cage}} X_{\text{cage}} + \beta_{\text{fly} \subset \text{cage}} X_{\text{fly} \subset \text{cage}} + \text{res}$
 S & R $Y_{ijk} = \mu + A_i + B_{ij} + \epsilon_{ijk}$

The symbol $X_{\text{fly} \subset \text{cage}}$ is read ‘fly within cage.’

2. Execute analysis.

Place data in model format:

Column labelled L, with response variable fly wing length

Column labelled X_{cage} , with explanatory variable $X_{\text{cage}} = \text{I, II, or III}$

Column labelled $X_{\text{fly} \subset \text{cage}}$ with label (number) for each fly

Code model statement in statistical package according to the GLM

$$L = \beta_o + \beta_{\text{cage}} \cdot X_{\text{cage}} + \beta_{\text{fly} \subset \text{cage}} X_{\text{fly} \subset \text{cage}} + \epsilon$$

```
MTB > anova 'wlength' = 'cage' 'fly'('cage');  
SUBC> random 'cage' 'fly'('cage');  
SUBC> fits c4;  
SUBC> residuals c5.
```

This produces estimates of the overall mean ($\hat{\beta}_o$)

and the mean for each cage ($\hat{\beta}_o + \hat{\beta}_{\text{cage}} \cdot X_{\text{cage}}$).

The fitted values are the means for each fly ($\hat{\beta}_o + \hat{\beta}_{\text{cage}} \cdot X_{\text{cage}} + \hat{\beta}_{\text{fly} \subset \text{cage}} X_{\text{fly} \subset \text{cage}}$).

$$\hat{\beta}_o = \text{mean}(L) = 24^{-1} \cdot 1599.2 = 66.633 \text{ } \mu\text{m}$$

$$\hat{\beta}_o + \hat{\beta} \cdot X_{\text{cage}} = \begin{cases} \text{mean}(L_{\text{cage=I}}) = 8^{-1} \cdot 582.7 = 72.84 \text{ } \mu\text{m} \\ \text{mean}(L_{\text{cage=II}}) = 8^{-1} \cdot 479.7 = 59.96 \text{ } \mu\text{m} \\ \text{mean}(L_{\text{cage=III}}) = 8^{-1} \cdot 536.8 = 67.10 \text{ } \mu\text{m} \end{cases}$$