Biostats VHM 802 Course
Winter 2006, Atlantic Veterinary College, PEI
Henrik Stryhn

Notes on Linear Mixed Models

These notes are intended to supplement, *not replace*, material in the textbook [1] of the VHM 802 Advanced Veterinary Biostatistics course. Their purpose is threefold,

1) to introduce linear mixed models and some general concepts/ideas frequently encountered (variance components, nesting, repeatability and reproducibility, to name a few),
2) to review random effects models in 1-way and 2-way factorial designs,
3) to outline a statistical analysis based on the ANOVA table and the so-called expected mean squares for balanced designs.

The present version is a fourth, revised and completed version.

Contents

1 Introduction

Up front we attempt to remedy some common points of confusion. One is the varied and interchangeably used terminology (section 1.1). Another is the different analytical approaches to linear mixed models (section 1.4)

1.1 A word about terminology

Mixed models (for continuous data) are a class of models which contain parameters or effects of two types:

- "fixed", like ordinary regression coefficients,
- "random", referring to the stochastic part of the model (beyond the usual error term)

Although not strictly logical, the term random effects models is usually used to denote such models with both types of effects. Sometimes, though, they are meant to cover models with only categorical predictors, but linear mixed models encompass both continuous and categorical predictors.

Mixed models can be used to take into account that the data have a hierarchical or multilevel or nested structure, and sometimes the models are also referred to by these names. Although other methods exist for hierarchically structured data, the mixed model approach has become a popular choice during the last decade, due to advances in computational power. Mixed models apply also to other data structures, such as longitudinal data with repeated measures on the same observational unit — this data structure only in part falls within the hierarchical data framework.

Variance components are some technical/mathematical constructs used in mixed models (which therefore are also sometimes called variance component models). The main idea is that the variance (variation, variability) in a dataset may be decomposed into (a sum of) several components that can each be given a useful interpretation.

1.2 Data example

The National Environmental Research Institute of Denmark collected (in 1992) data on laboratory measurements of concentrations of different chemical substances. We show here the results for concentrations of 4-methylphenol and 5 laboratories. Six samples were sent to each laboratory, 2 replicates of 3 dilutions. Samples were blinded and the laboratories were not aware of the experimental design.

phenol (μg/l)	Dilution					
Laboratory	1		2		3	
A	5.5	4.7	9.8	10.3	11.6	11.8
B	7.7	7.5	12.4	12.5	16.4	17.0
C	7.4	7.1	12.5	11.8	15.9	16.2
D	6.5	7.1	10.0	9.4	12.6	12.7
E	6.5	7.0	11.0	9.9	13.5	12.7

The purpose of the study was to determine the accuracy of concentrations measured at different laboratories. The participating laboratories used the same analytic procedure and were previously accredited for these analyses.

We first consider only the data at dilution 1, which consists of 2 replicates at each of 5 laboratories and a total of 10 observations. Denote by y_{ij} the concentration of sample j from laboratory i, where i =A,...,E and $j = 1, 2$. The usual 1-way ANOVA model is

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{or} \quad y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{1}$$

where the errors ε_{ij} are independent and $\sim \mathrm{N}(0, \sigma^2)$. The model is appropriate to examine differences among the 5 selected laboratories, and the parameter μ_i corresponds to the level of laboratory i, which would approximately equal the average of many samples from that laboratory. We would, however, like to say something about laboratories in general, and the accuracy on determinations not only at the same laboratory (which we can do from model (1)) but also at different laboratories. Such analysis requires an assumption that the 5 laboratories are representative of a population (of laboratories). With that in place, we change the "fixed effects" α_i in model (1) to "random effects" A_i:

$$y_{ij} = \mu + A_i + \varepsilon_{ij}, \quad \text{where } A_i \sim \mathrm{N}(0, \sigma^2_{\mathrm{A}}). \tag{2}$$

Random means simply that it is modelled as a random variable, in contrast to a (fixed) parameter. The term A_i corresponding to laboratory i being a random variable reflects our perception that the laboratory represents one of the laboratories from the population (ideally, it was randomly selected from the population).

Factors where one is less interested in the specific levels themselves than interpreting them as representing a population are quite common: lots, litters, herds, person, patients.... For any block-type factors (a division of experimental units into homogeneous groups) there may often be more interest in the variation between the groups than the groups themselves. Another angle is that the specific levels may be of interest if they can be used in other studies, whereas levels representing a population are more natural if they are only used in this dataset. Random effects modelling of factors can be justified even if the levels are not drawn randomly from a population. The key assumption is that the levels represent a population, and the focus is on the variability in the population (σ^2_{A}).

Before looking in more detail at this new model, we'll outline how the model could arise as the result of a two-step sampling procedure (sometimes ([1]) called subsampling). For the purpose of determining measurement accuracy among laboratories, the natural procedure would be to submit the same sample to a number of laboratories, each of which would return their (single) measurement. This would give rise to the 1-sample model,

$$y_i = \mu + \epsilon_i \quad \text{where } i \sim \text{laboratory and } \epsilon_i \sim N(0, \sigma^2_\epsilon), \tag{3}$$

where μ and σ_ϵ are the mean and standard deviation in the population of laboratories (more precisely, laboratory measurements of that single sample). Extending the design by sending two identical samples to each laboratory leads to the design discussed above. In the model, we have to additionally include a variation between the two samples at the same laboratory, because if there was no such variation the two analyses at the same laboratory would give the same value. This leads us to the model (2) with two random terms.

1.3 Random effects and variance components

The assumptions on the random effects in model (2) are

$$A_i \sim N(0, \sigma^2_{\mathrm{A}}), \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad A_i \text{ and } \varepsilon_{ij} \text{ are independent.}$$

Thus, we assume the impact of laboratory i to be a normally distributed, random fluctuation with mean zero and standard deviation σ_{A}. Therefore the parameter σ_{A}^2 may be interpreted as the overall variation (variance) in measurements between laboratories. The parameter μ is the population mean, and σ^2 is the variation (variance) within a laboratory (between measurements at the same laboratory). Furthermore, we may calculate

$$\mathrm{Var}(y_{ij}) = \mathrm{Var}(A_i) + \mathrm{Var}(\varepsilon_{ij}) = \sigma_{\mathrm{A}}^2 + \sigma^2. \tag{4}$$

In effect, we have decomposed the total variance as a sum of the variance between laboratories and the error variance (or the variance within laboratories). Therefore, the σ^2's are called variance components. As we will see below, the estimates in our example are $\hat{\sigma}_{\mathrm{A}}^2 = 0.852$ and $\hat{\sigma}^2 = 0.138$. From these we can compute interesting quantities, such as

- the proportion of the variance residing at the different levels: $\sigma_{\mathrm{A}}^2/(\sigma_{\mathrm{A}}^2 + \sigma^2)$ and $\sigma^2/(\sigma_{\mathrm{A}}^2 + \sigma^2)$ (in the example, these values are 86% and 14%, respectively),
- the (intraclass) correlation between two observations from the same laboratory: $\rho = \sigma_{\mathrm{A}}^2/(\sigma_{\mathrm{A}}^2 + \sigma^2)$ (in the example, the value is 0.86).

Note finally that in the 1-sample model (3) the error term ϵ_i includes both the variation between laboratories and the variation within each laboratory that have been separated in the 1-way model.

1.4 Scope of statistical approach

The traditional (or classical) method of analysis for random effects model is based on the ANOVA table. For a balanced design, the table differs only from that of the corresponding fixed effects model by the way the F-statistics are calculated. More precisely, tests in random effects models may have another denominator than MSE, reflecting that another variation is appropriate for measuring the effect. Furthermore, the mean values of mean squares can be used to construct estimates of the variance components in the model. Therefore, ANOVA tables for random effects models often have an added EMS (expected mean square) column. Both of these approaches work fine in a balanced dataset and reasonably well in a slightly unbalanced dataset. However, one serious disadvantage of the ANOVA-based analysis is that it is usually up to the analyst to figure out the standard errors of parameter estimates, using formulas pertinent to the actual design. This is because most statistical software using ANOVA-based methods is little helpful on this point (Stata, `anova` command; Minitab; SAS, `glm` procedure). Formulas are provided here for the basic designs, as well as some general principles.

The modern method of analysis is based on the likelihood function, and involves an iterative procedure to obtain the so-called (restricted) maximum likelihood estimates. This method is available in some general statistical packages (SPSS, version 11; SAS, `mixed` procedure; S-Plus and R, `lme` library), as well as in special-purpose packages for multilevel data (MLwiN, HLM). For balanced designs, it gives similar and in many cases identical results to the ANOVA-based method, but with correct standard errors. For (strongly) unbalanced designs, this is the recommended method.

These notes deal only with ANOVA-based methods for balanced designs.

2 1-way ANOVA with random effects

The parameters of model (2) are μ, σ_{A}^2 and σ^2, and the full list of model assumptions is

i) independence of the random terms ε_{ij} and A_i,
ii) fixed part of the model: $\mathrm{E}y_{ij} = \mu$, where $\mathrm{E}y_{ij}$ is the expectation of y_{ij},
iii) homoscedasticity — same variance of all observations y_{ij}, by equation (4),
iv) normal distribution of the random terms ε_{ij} and A_i.

2.1 Statistical analysis

The ANOVA-table for a balanced 1-way model with random effects, a groups and n observations per group (note: we use n instead of N used in [1]) is

Source	DF	SS	MS	EMS	F
A (groups)	$a-1$	$\sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2$	SSA/DFA	$\sigma^2 + n\sigma^2_{\mathrm{A}}$	MSA/MSE
Error	$a(n-1)$	$\sum_{ij}(y_{ij} - \bar{y}_{i.})^2$	SSE/MSE	σ^2	
Total	$an-1$	$\sum_{ij}(y_{ij} - \bar{y}_{..})^2$			

Table 1: ANOVA table for 1-way random effects model.

The F-test is for the hypothesis of no difference between groups, H_0: $\sigma^2_{\mathrm{A}} = 0$, and follows the usual distribution F(DFA,DFE) under H_0. The EMS (expected mean square) column shows how E(MSA) depends on the variation between groups (σ^2_{A}). Just as in ordinary ANOVA tables, the use of a one-sided F-statistic may be justified by referring to that fact that it is a ratio between statistics with the same expected value when H_0 is true, and that the numerator has larger expected value when H_0 is false. Furthermore, we use this column to construct our (unbiased) estimates of the variance components:

$$\begin{aligned} \mathrm{E(MSE)} = \sigma^2 \quad &\Rightarrow \quad \hat{\sigma}^2 = \mathrm{MSE}, \\ \mathrm{E(MSA)} = \sigma^2 + n\sigma^2_{\mathrm{A}} \quad &\Rightarrow \quad \hat{\sigma}^2_{\mathrm{A}} = (\mathrm{MSA} - \mathrm{MSE})/n. \end{aligned}$$

If the value obtained for σ^2_{A} is negative, it is common practice to set it to zero. For estimation of μ we have the formulas:

$$\hat{\mu} = \bar{y}_{..} \quad \text{and} \quad \mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{MSA}/(an)}.$$

This is our first example of using a different variation than MSE. The standard error of $\hat{\mu}$ is computed using MSA, and the degrees of freedom for a confidence interval are accordingly DFA. The estimate itself ($\hat{\mu}$) is the same as in the fixed effects model, but the variation associated with it is different, because in a random effects model *it must also take the variation between groups into account*. For checking the assumptions of the statistical model involving the random effects A_i, we compute the estimates (residuals):

$$\hat{A}_i = \bar{y}_{i.} - \bar{y}..,$$

and compare their distribution to a normal distribution.

Note 2.1. As a more technical note, we demonstrate how to determine $\mathrm{SE}(\hat{\mu})$. From the model formula (2), we calculate

$$\bar{y}_{..} = \mu + \bar{A}_{.} + \bar{\varepsilon}_{..} \quad \text{and} \quad \mathrm{Var}(\bar{y}_{..}) = \mathrm{Var}(\bar{a}_{.}) + \mathrm{Var}(\bar{\varepsilon}_{..}) = \sigma^2_{\mathrm{A}}/a + \sigma^2/(an).$$

From the last formula it follows that $\mathrm{Var}(\bar{y}_{..}) = \mathrm{E(MSA)}/(an)$, so the natural estimate for the standard error is $\sqrt{\mathrm{MSA}/(an)}$.

Example 2.1. One dilution measured at multiple laboratories
For the data example with 5 laboratories and 2 replicates within each laboratory, we get the table:

Source	DF	SS	MS	EMS	F	P
Laboratories	4	7.37	1.84	$\sigma^2 + 2\sigma^2_{\mathrm{A}}$	13.35	0.007
Error	5	0.69	0.138	σ^2		
Total	9	8.06				

and the parameter estimates:

$$\begin{aligned} \hat{\sigma}^2 &= \mathrm{MSE} = 0.138, \\ \hat{\sigma}^2_{\mathrm{A}} &= (\mathrm{MSA} - \mathrm{MSE})/2 = (1.84 - 0.138)/2 = 0.852, \\ \hat{\mu} &= 6.70 \quad \text{and} \quad \mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{MSA}/10} = \sqrt{0.184} = 0.43. \end{aligned}$$

The data show clear evidence of a variation between laboratories, and that variation seems in fact to much larger than the variation within laboratories. □

2.2 Repeatability and reproducibility in laboratories

Loosely speaking, repeatability refers to the agreement between two measurements made using the same method and under the same circumstances, and reproducibility to the agreement between measurements from similar (but not the same) circumstances. The meaning of "similar" depends on the context; some examples are *(i)* different laboratory, *(ii)* same laboratory, but different day and/or technician and/or equipment. Depending on the type of measurement and design, the repeatability and reproducibility can be quantified in different ways. We describe here a method from the international standard ISO 5725, based on random effects models, see also [2].

The key idea is to quantify agreement not on a relative scale (like a correlation) but as the value two observations would differ at most by when taken under the same or similar conditions. Since our models are based on the normal distribution (which extends infinitely), such a value can only be given with a certain confidence, and the usual confidence chosen is 95%. That is,

- the *repeatability*, $\hat{r}$, is the value not exceeded with probability 95% by the difference between two measurements taken under the same conditions,
- the *reproducibility*, $\hat{R}$, is the value not exceeded with probability 95% by the difference by two measurements taken under similar but not the same conditions.

In our example, repeatability refers to the variation within laboratories and reproducibility refers to the variation of two measurements taken in different laboratories. For the 1-way model, the formulas are:

$$\hat{r} = 2\sqrt{2} \times \sqrt{\hat{\sigma}^2} \quad \text{and} \quad \hat{R} = 2\sqrt{2} \times \sqrt{\hat{\sigma}^2_{\mathrm{A}} + \hat{\sigma}^2},$$

which for the data example gives the values

$$\hat{r} = 2.83\sqrt{0.138} = 1.05 \quad \text{and} \quad \hat{R} = 2.83\sqrt{0.852 + 0.138} = 2.82.$$

Our interpretation is that we can be 95% confident that the difference of two values from the same laboratory does not exceed 1.05, whereas two values from different laboratories with 95% do not differ more than 2.82. The large variation between laboratories makes in this case the reproducibility considerably larger than the repeatability.

Note 2.2. As a technical note, we show how the factor $2\sqrt{2}$ arises. Recall that in a normal distribution with mean μ and standard deviation σ, a central 95% (prediction) interval of the distribution is $(\mu - 1.96\sigma, \mu + 1.96\sigma)$. The variable in question here is a difference between two observations. For r these are from the same laboratory, say y_{11} and y_{12}, and for R from different laboratories, say y_{11} and y_{21}. In both cases, the differences follow a normal distribution with mean zero (because the observations have the same fixed model term, μ). Furthermore, using the model equation (2),

$$\begin{aligned}\mathrm{Var}(y_{11} - y_{12}) &= \mathrm{Var}(\mu + A_1 + \varepsilon_{11} - (\mu + A_1 + \varepsilon_{12})) = \mathrm{Var}(\varepsilon_{11} - \varepsilon_{12}) = 2\sigma^2,\\ \mathrm{Var}(y_{11} - y_{21}) &= \mathrm{Var}(\mu + A_1 + \varepsilon_{11} - (\mu + A_2 + \varepsilon_{21})) = \mathrm{Var}(A_1 - A_2 + \varepsilon_{11} - \varepsilon_{12}) = 2(\sigma_{\mathrm{A}}^2 + \sigma^2).\end{aligned}$$

Taking the square-root of these equations gives us the standard deviations of the zero-mean normal distributions, and the prediction intervals for the differences will be symmetric around 0 and extending 1.96 times the standard deviation to both sides. That is, the differences are (with probability 95%) within 1.96 times the standard deviation, or approximately 2 times the standard deviation — which are our formulas.

3 2-way ANOVA with random effects

We consider here a balanced 2-way factorial with factors A (a levels) and B (b levels) and replications per (A,B)-combination (n replicates). The full dataset of our example is one example of such a layout, with A $\sim$ laboratories ($a = 5$) and B $\sim$ dilutions ($b = 3$) and $n = 2$ replicates. The 2-way models contain 3 terms: the main effects of A and B as well as their interaction. As each of these terms can in principle be either fixed or random, the number of possible models increases considerably. We review the most commonly encountered models and their interpretation, and give the ANOVA tables and some hints for the statistical analysis.

3.1 Model types

When building 2-way factorial models, the two factors A and B may be taken as fixed or random independently of each other. The rule for interactions is that they must be taken as a random effect when at least one of the factors is random. However, it is possible for an interaction to be a random effect even if all factors are fixed effects. Therefore, the most interesting models for a 2-way factorial are as listed below. We denote our data by y_{ijk}, where $i = 1, \ldots, a \sim$ factor A, $j = 1, \ldots, b \sim$ factor B, and $k = 1, \ldots, n \sim$ replicates.

$$\begin{aligned}\text{I:}\quad & y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} && \text{– all fixed effects,} && (5)\\ \text{II:}\quad & y_{ijk} = \mu + \alpha_i + \beta_j + AB_{ij} + \varepsilon_{ijk} && \text{– random interaction,} && (6)\\ \text{III:}\quad & y_{ijk} = \mu + A_i + \beta_j + AB_{ij} + \varepsilon_{ijk} && \text{– random effect A and A}*\text{B,} && (7)\\ \text{IV:}\quad & y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk} && \text{– all random effects.} && (8)\end{aligned}$$

The models assume all random variables to be independent and distributed as follows: $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$, $AB_{ij} \sim \mathrm{N}(0, \sigma_{\mathrm{AB}}^2)$, $A_i \sim \mathrm{N}(0, \sigma_{\mathrm{A}}^2)$ and $B_j \sim \mathrm{N}(0, \sigma_{\mathrm{B}}^2)$. Note that we use the combined notation AB_{ij} or $\alpha\beta_{ij}$ for the interaction terms instead of introducing a new variable name like γ_{ij}, in order to make it more transparent from which variables the interaction is formed.

Example 3.1. Several dilutions measured at multiple laboratories
In our data example we have so far only considered the first dilution. Including several dilutions (within the normal range) increases the scope of the test, and enables a more detailed comparison of the variations. As already noted, the full dataset has a two-say layout with factor A $\sim$ laboratories, factor B $\sim$ dilutions and two replicates. By analysing all dilutions together we assume the variation to be constant across the range of dilutions. This can be checked by comparing estimated variance components from separate analyses of the 3 dilutions. These data possibly show a slight indication of higher variation between laboratories at dilution 3, but nothing to stop us from analysing the data together.

To choose the statistical model we consider which of the two factors should be fixed and random. As before, we model the laboratories by random effects because our interest is in the variation between them. The dilutions were probably set by the person responsible for the test, to cover roughly the range of values of interest. They can not be considered as a sample from a population, and there is no interest in the variation between dilutions. Therefore, dilutions should be taken as a fixed effect. By the above interaction rule, the interaction should be a random effect, so we choose model III in (7). The three variance parameters have the following interpretations:

- σ^2: the variation within laboratories at the same dilution,
- σ^2_{A}: the overall variation between laboratories,
- σ^2_{AB}: the variation between laboratories specific to different dilutions.

The total variance is separated into these variance components,

$$\mathrm{Var}(y_{ijk}) = \sigma^2_A + \sigma^2_{AB} + \sigma^2.$$

□

3.2 Statistical analysis

The ANOVA tables of the four 2-way models (5)–(8) differ only in their EMS and F columns, shown in table 2. In the EMS columns, we denote by σ^2_α the variation between the α_i's in a fixed effects model for factor A. It is positive, unless there are no differences between the factor levels (H_0: $\alpha_1 = \ldots = \alpha_a$ is true). We similarly use σ^2_β and $\sigma^2_{\alpha\beta}$ for fixed effects of factors B and A∗B.

Source	DF	EMS(I)	EMS(II)	EMS(III)	EMS(IV)	F(I)	F(II–IV)
A	$a-1$	$\sigma^2+\sigma^2_\alpha$	$\sigma^2+n\sigma^2_{\mathrm{AB}}+\sigma^2_\alpha$	$\sigma^2+n\sigma^2_{\mathrm{AB}}+bn\sigma^2_{\mathrm{A}}$	$\sigma^2+n\sigma^2_{\mathrm{AB}}+bn\sigma^2_{\mathrm{A}}$	$\frac{\mathrm{MSA}}{\mathrm{MSE}}$	$\frac{\mathrm{MSA}}{\mathrm{MSAB}}$
B	$b-1$	$\sigma^2+\sigma^2_\beta$	$\sigma^2+n\sigma^2_{\mathrm{AB}}+\sigma^2_\beta$	$\sigma^2+n\sigma^2_{\mathrm{AB}}+\sigma^2_\beta$	$\sigma^2+n\sigma^2_{\mathrm{AB}}+an\sigma^2_{\mathrm{B}}$	$\frac{\mathrm{MSB}}{\mathrm{MSE}}$	$\frac{\mathrm{MSB}}{\mathrm{MSAB}}$
A∗B	$(a-1)(b-1)$	$\sigma^2+\sigma^2_{\alpha\beta}$	$\sigma^2+n\sigma^2_{\mathrm{AB}}$	$\sigma^2+n\sigma^2_{\mathrm{AB}}$	$\sigma^2+n\sigma^2_{\mathrm{AB}}$	$\frac{\mathrm{MSAB}}{\mathrm{MSE}}$	$\frac{\mathrm{MSAB}}{\mathrm{MSE}}$
Error	$ab(n-1)$	σ^2	σ^2	σ^2	σ^2	–	–
Total	$abn-1$						

Table 2: Condensed ANOVA table for four balanced 2-way factorial models I–IV.

The ANOVA table shows how F-statistics for the main effects change between the fixed effects model (I) and the random effects models: by substituting for MSE the MSAB. The denominator

degrees of freedom of the F-distribution changes accordingly to $(a-1)(b-1)$. The table also shows how the variance components enter the EMS column, from which their estimates are computed:

$$\begin{aligned}
\mathrm{E(MSE)} = \sigma^2 &\quad\Rightarrow\quad \hat\sigma^2 = \mathrm{MSE},\\
\mathrm{E(MSAB)} = \sigma^2 + n\sigma^2_{\mathrm{AB}} &\quad\Rightarrow\quad \hat\sigma^2_{\mathrm{AB}} = (\mathrm{MSAB}-\mathrm{MSE})/n,\\
\mathrm{E(MSA)} = \sigma^2 + n\sigma^2_{\mathrm{AB}} + bn\sigma^2_{\mathrm{A}} &\quad\Rightarrow\quad \hat\sigma^2_{\mathrm{A}} = (\mathrm{MSA}-\mathrm{MSAB})/bn,\\
\mathrm{E(MSB)} = \sigma^2 + n\sigma^2_{\mathrm{AB}} + an\sigma^2_{\mathrm{B}} &\quad\Rightarrow\quad \hat\sigma^2_{\mathrm{B}} = (\mathrm{MSB}-\mathrm{MSAB})/an.
\end{aligned}$$

Again, a negative value for a variance component would usually be set to zero.

Note 3.1. Standard errors of fixed effect estimates and contrasts are generally more difficult to compute than in random effects models. The additional question (relative to fixed effects models) is which random variations to involve. For the (balanced) 2-way ANOVA models II and III with one fixed effect (B), the general rule is that contrasts (including pairwise comparisons) for B use the MSAB. The procedure for group level means for B is different in the two models because the random effects of factor A in model III must be taken into account. The following calculations, which are typical for random effects models, show how the appropriate variations are determined:

$$\begin{aligned}
\text{II \& III:}\quad \mathrm{Var}(\bar y_{\cdot 1\cdot} - \bar y_{\cdot 2\cdot}) &= \mathrm{Var}(\overline{AB}_{\cdot 1} - \overline{AB}_{\cdot 2} + \bar\varepsilon_{\cdot 1\cdot} - \bar\varepsilon_{\cdot 2\cdot})\\
&= 2(\sigma^2_{\mathrm{AB}}/a + \sigma^2/(an)) = 2\,\mathrm{E(MSAB)}/(an),\\
\text{II:}\quad \mathrm{Var}(\bar y_{\cdot 1\cdot}) &= \mathrm{Var}(\overline{AB}_{\cdot 1} + \bar\varepsilon_{\cdot 1\cdot}) = \sigma^2_{\mathrm{AB}}/a + \sigma^2/(an) = \mathrm{E(MSAB)}/(an),\\
\text{III:}\quad \mathrm{Var}(\bar y_{\cdot 1\cdot}) &= \mathrm{Var}(\bar A_{\cdot} + \overline{AB}_{\cdot 1} + \bar\varepsilon_{\cdot 1\cdot}) = \sigma^2_{\mathrm{A}}/a + \sigma^2_{\mathrm{AB}}/a + \sigma^2/(an).
\end{aligned}$$

It follows that in both models the MSAB is used for pairwise comparisons (and contrasts). It also follows that in model II the MSAB is used as well for the standard error of group means of factor B. In model III, no MS-value has an expected value proportional to the group mean variance. Still, the standard error we may compute simply by inserting the estimated variance components, but no degrees of freedom is available. Several approximations exist: a conservative one is to use the smallest DF among the random effects involved (in this case, DFA), and very liberal ones are to use the largest DF or a standard normal.

Example 3.2. Several dilutions measured at multiple laboratories (cont.)
The ANOVA table for model III applied to the full dataset is as follows:

Source	DF	SS	MS	EMS	F	P
Laboratories	4	47.70	11.93	$\sigma^2 + 2\sigma^2_{\mathrm{AB}} + 6\sigma^2_{\mathrm{A}}$	8.59	0.005
Dilutions	2	271.70	135.8	$\sigma^2 + 2\sigma^2_{\mathrm{AB}} + \sigma^2_{\beta}$	98.0	< 0.001
Lab. $*$ Dil.	8	11.11	1.389	$\sigma^2 + 2\sigma^2_{\mathrm{AB}}$	8.61	< 0.001
Error	15	2.42	0.161	σ^2		
Total	29	332.93				

The table shows all effects to be clearly significant. Estimates of the variance components are

$$\hat\sigma^2 = \mathrm{MSE} = 0.161, \quad \hat\sigma^2_{\mathrm{AB}} = \frac{\mathrm{MSAB}-\mathrm{MSE}}{2} = 0.614, \quad \hat\sigma^2_{\mathrm{A}} = \frac{\mathrm{MSA}-\mathrm{MSAB}}{6} = 1.756.$$

From these values, we may recompute estimates of repeatability and reproducibility based on the full dataset. As previously, the repeatability refers to the variation within a laboratory and at the same dilution, and is estimated by

$$\hat r = 2\sqrt{2} \times \sqrt{\hat\sigma^2} = 2.83\sqrt{0.161} = 1.14.$$

Thus, we are 95% confident that two measurements at the same laboratory differ no more than 1.1 in their value. This value is pretty close to the one from the analysis of one dilution only. The reproducibility refers to variation between laboratories for measurements on the same dilution, and must therefore include all variance components. Note that we include σ^2_{AB} as well, because we do not restrict our statement to be valid for one dilution only (for which R could be obtained as in the previous analysis) but for any of the 3 dilutions. Thus, two values from different laboratories and at any (same) dilution differ with 95% probability no more than

$$\hat{R} = 2\sqrt{2} \times \sqrt{\hat{\sigma}^2_{\text{A}} + \hat{\sigma}^2_{\text{AB}} + \hat{\sigma}^2} = 2.83\sqrt{1.756 + 0.614 + 0.161} = 4.50.$$

This value is quite a bit larger than the one obtained for dilution 1, indicating as already noted that the variation between laboratories seems to increase with the dilutions. □

In fixed effects ANOVA models, main effects for factors involved in significant interactions are often of little interest, and should be interpreted with care. For example, when there is a strong interaction the main effects of A may be small (non-significant) because of opposite effects at the different levels of B. In such a case it is clearly wrong to conclude that A has no effect. In random effects models, there is greater freedom to examine main effects even in the presence of an interaction. In our example, the main laboratory effect expresses the overall variation between laboratories (the variation that is common to all dilutions), and this quantity is meaningful no matter the magnitude of the interaction. It therefore makes sense to both estimate σ^2_{A} (as we already did) and to test whether it is zero. The denominator of the F-test is MSAB; thus, we measure the variation between laboratories not relative to the variation within laboratories but to the variation between laboratories at the different dilutions. As MSAB is usually larger than MSE and has less degrees of freedom, we have less power to detect such differences between laboratories. Similar considerations apply to dilution effects. Accepting the presence of random fluctuations in dilution effects across laboratories, we may test for overall dilution effects — relative to the variations across laboratories. The conclusion of the analysis in the example was that the overall dilution effects were huge, even relative to the variations across laboratories.

Taking this line of reasoning a bit further leads to a justification of the model (II) in (6) with fixed main effects and a random interaction. When a significant interaction is encoutered in an ordinary fixed effects model, and the interaction cannot be given a clear interpretation but seems mostly to be random fluctuations around the parallel curves (in the interaction plot), one may decide to "make" the interaction random and thereby measure the two main effects against the interaction variation. They will be significant in the analysis if their effects are stronger than the interaction. Before doing this, one should however always make sure that the interaction does not contain interesting information in itself.

4 Nested 2-way ANOVA

It is common usage to call one factor B nested within another factor A, if there is no link between the observations at the same B-level across different A-levels. This contrasts the usual situation, which is sometimes called two crossed factors, where all observations at the same level of one factor do share the corresponding common feature. We illustrate these ideas by an example.

Example 4.1. Pig breeding data
A breeding experiment involving 5 sires, 2 dams per sire and 2 pigs per litter recorded the weight gain (over a certain period) of the piglets (data from [3]).

Sire	1		2		3		4		5	
Dam	1	2	1	2	1	2	1	2	1	2
weight	2.77	2.58	2.28	3.01	2.36	2.72	2.87	2.31	2.74	2.50
gain	2.38	2.94	2.22	2.61	2.71	2.74	2.46	2.24	2.56	2.48

Here the dams are nested within sires, because the dams used for different sires are completely unrelated. That is, a total of 10 dams were used, and they would in principle be randomly selected from a population of dams. If the two factors were to be crossed, there would only be 2 different dams in the experiment, and both would have been used with each of the 5 sires. □

An alternative notion for designs with a nesting is that they have a hierarchical structure (next section). For the analysis of a design with nested factor(s), the simple rule is that a nested factor should never be allowed a main effect (because the factor levels are meaningless when viewed alone) and that the factor is therefore represented solely by its interaction with the factor into which it is nested. In the example, the combined factor from the factors Sire and Dam has 10 levels, corresponding exactly to the 10 dams in the dataset. The interaction is often taken as a random effect; the general considerations for random effects apply. In the ANOVA table, the row of the main effect of the nested factor cancels, and the SS and DF are pooled into the interaction. It is worthwhile knowing that many software packages have a special notation for nesting: the most common one is B(A), for the factor B being nested within A.

Example 4.2. Pig breeding data (cont.)
For the pig data, we take the effect of sires to be fixed (assuming that there is specific interest in comparing the performance of the 5 sires) and the effect of dams to random (assuming that the dams represent a population of dams). Note that if dams are taken as fixed effects, we cannot examine main effects of sires in presence of a dam effect. Thus, the statistical model is (with y_{ijk} denoting the weight gain of pig k bred by dam j and sire i),

$$y_{ijk} = \mu + \alpha_i + AB_{ij} + \varepsilon_{ijk}, \tag{9}$$

where $AB_{ij} \sim \mathrm{N}(0, \sigma^2_{\mathrm{AB}})$ and $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$. The ANOVA table is shown below.

Source	DF	SS	MS	EMS	F	P
Sires	4	0.100	0.025	$\sigma^2 + 2\sigma^2_{\mathrm{AB}} + \sigma^2_\alpha$	0.22	0.92
Dams (Sires)	5	0.564	0.113	$\sigma^2 + 2\sigma^2_{\mathrm{AB}}$	2.91	0.071
Error	10	0.387	0.039	σ^2		
Total	19	1.050				

The table shows that there is only a weak (just above statistical significance at the 5% level) effect of dams, with an estimated variance component of

$$\hat{\sigma}^2_{\mathrm{AB}} = (\mathrm{MSAB} - \mathrm{MSE})/2 = (0.113 - 0.039)/2 = 0.037.$$

In addition, there are absolutely no statistically interesting differences to be seen between sires. □

5 Split-plot designs and hierarchical structures

This section supplements the comprehensive discussion of split-plot designs in the textbook ([1]) with a discussion of data structures akin to a split-plot design and of the rationale behind a split-plot design. Finally, we give models and ANOVA tables for both of the two basic versions of split-plot designs: with and without blocks. We also summarise the computation of standard errors for fixed effects estimates and contrasts, although these are discussed in detail in the textbook.

5.1 Clustering derived from data structure

In our usage, clustering means that some observations share some common features (that is not explicitly taken into account by explanatory variables in a model). We discuss here clustering as a result of sharing a common environment, physical clustering in space and repeated measurements within the same individual. Cows within a herd, puppies within a litter, quarters within a cow are all examples of clustering in environment. We usually assume that the degree of similarity among all pairs of observations within such a cluster are equal. Such clustering is not necessarily restricted to a single level. For example, pigs may be clustered within a litter which may be clustered within a pen of pigs, which may be clustered in a farm which may be clustered in a region, as shown in the Fig. 5.1. Such data are called hierarchical or multilevel data. The structure shown in Fig. 1 is a 5-level structure. In practice, we deal more often with data that have a 2-level or 3-level structure.

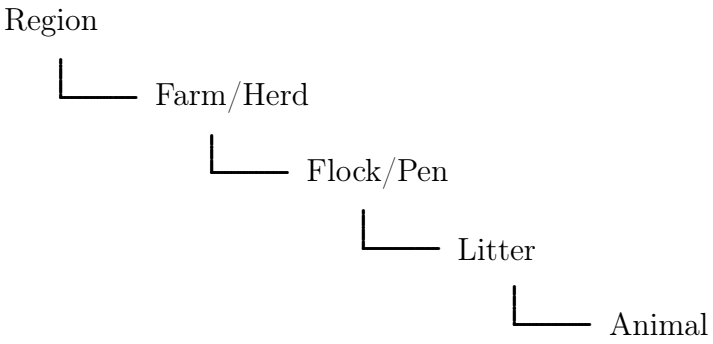

Figure 1: A typical hierarchical data structure in veterinary epidemiology.

The hierarchy in Fig. 1 suggests that farms in the same region are similar. It seems natural to replace or extend this relationship by one where the dependence between farms is directly related to (inversely proportional to) their distance. Spatial models incorporate the actual locations, in this example of the farms but it could also be the actual locations of cows in a tie-stall barn. If such detailed information is not available or detailed spatial modelling is not desirable (eg. due to sparse data), spatial clustering may be accounted for by hierarchical level(s).

Repeated measures arise when a several measurements of a variable are taken on the same animal (or other unit of observation) over a period of time. Daily milk weights in a cow are highly correlated since the level of milk production on one day, is likely to be quite close to the production on the day before and the day after. Multiple measurements of lactation total milk production across lactations within a cow are also repeated measurements, but would not be so highly correlated. We may think of repeated measures as a special type of clustering and for the above examples even add an extra, bottom level in the diagram (Fig. 1) for days or time. However, just as with spatial clustering, several special considerations apply. Observations close together in time are likely to be more highly correlated than measurements with a longer time span between them. Also, the clustering may occur

at any level in the hierarchy, not just a the lowest level. For example, if a study on pig production involved several batches within a farm, the flock/pen level should be replaced by batches, which would then correspond to repeated measures over time on the farm.

Diagrams like Fig. 1 are generally highly recommended to determine and present data structures, only should their defaults with regard to spatial and repeated structures be kept in mind. Note that the data structure pertains not only to the outcome but also to the predictor variables and it is very useful to know whether they vary or were applied at particular levels. We elaborate on this idea in the context of the simplest two-level experimental design: the split-plot design.

5.2 Split-plot design

The split-plot concept and terminology dates back to the early 20th century where statistical methods were developed in the context of agricultural field trials. Consider the planning of an experiment involving two factors A and B with a and b levels, respectively. The special feature of the design is that factor B is practically applicable to smaller units of land (plots) than factor A. In the field trial context, we may think of A as a large-scale management factor such as pesticide spraying by plane and B as a small-scale factor like variety. The experimental units for factor A are called whole plots. The design needs some replication, and we assume to have a total of ac whole plots at our disposal, laid out in c blocks of size a. The blocks would typically be separate pieces of land or experimental sites. A minor modification of the design occurs if the ac whole plots are not laid out in blocks but are just replicates; the same principle applies, but for simplicity we describe the design with blocks only. Within each block, the design would now by laid out in a two-step procedure, as illustrated in Fig. 5.2:

1.) randomly distribute the levels of factor A onto the a whole plots,

2.) divide each whole plot into b subplots, and randomly distribute the levels of factor B onto the subplots.

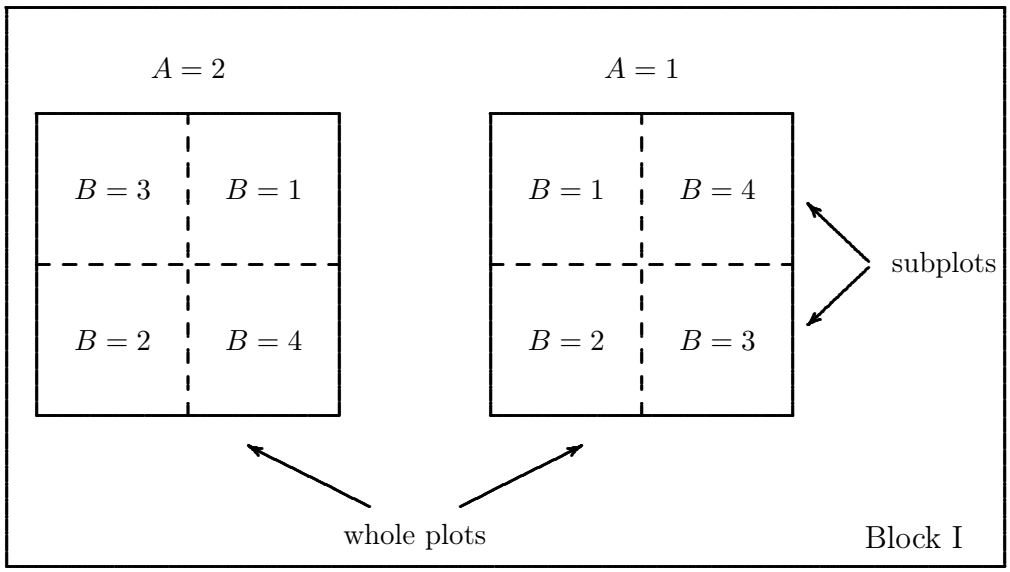

Figure 2: Split-plot layout within one block, with $a = 2$ whole plots and $b = 4$ subplots.

As an animal production example, we might have a herd management factor A and a factor B applicable to individual animals (so that animals in the same herd may have different levels of B).

Thus, the whole plots would be the herds, and the subplots the animals. The blocks would be groups of similar herds, eg. in the same region. Generally, a split-plot design corresponds to a 2-level hierarchy with whole plots as the upper level and subplots as the bottom level.

In the analysis of a split-plot experiment, the two factors cannot be expected to be treated equally because they are applied to different experimental units. In particular, effects of the whole plot factor A should be compared to the variation between whole plots (corresponding to the first step of the design construction), and effects of the subplot factor B to the variation between subplots. It follows that it is necessary (and possible) to split the total variation into variations between and within whole plots. These variations are estimated independently from each other and with different accuracy (degrees of freedom). Usually the whole plot variation will be considerably larger than the subplot variation, and factor A is estimated with less precision than factor B. As will be seen in the ANOVA tables, the interaction between A and B "belongs to" the subplot variation, intuitively because differences between B-levels within any A-level can be determined within the whole plots. This makes the split-plot design particularly attractive in situations where the principal interest is in the interaction and less in the main effect of factor A.

5.3 Split-plot models and ANOVA tables

The statistical model for a split-plot design with blocks is as follows, where observations are y_{ijk} with $i = 1, \ldots, a \sim$ the whole plot factor A, $j = 1, \ldots, b \sim$ the subplot factor B, and $k = 1, \ldots, c \sim$ blocks, (again in a slightly different notation than in [1])

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + C_k + (AC)_{ik} + \varepsilon_{ijk}, \tag{10}$$

where the block effects $C_k \sim \mathrm{N}(0, \sigma^2_{\mathrm{C}})$, the whole plot errors $AC_{ik} \sim \mathrm{N}(0, \sigma^2_{\mathrm{AC}})$ and the subplot errors $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$, and all errors are independent. The blocks are taken here with random effects, but the analysis with fixed block effects is quite similar. Table 3 outlines the corresponding ANOVA table. As previously, we use the notation σ^2_α, σ^2_β and $\sigma^2_{\alpha\beta}$ to denote fixed effects variations.

Source	DF	SS	MS	EMS	F
A (whole plot factor)	$a-1$	SSA	SSA/DFA	$\sigma^2 + b\sigma^2_{\mathrm{AC}} + \sigma^2_\alpha$	MSA/MSAC
C (blocks)	$c-1$	SSC	SSC/DFC	$\sigma^2 + b\sigma^2_{\mathrm{AC}} + ab\sigma^2_{\mathrm{C}}$	MSC/MSAC
A∗C (whole plot var.)	$(a-1)(c-1)$	SSAC	SSAC/DFAC	$\sigma^2 + b\sigma^2_{\mathrm{AC}}$	MSAC/MSE
B (subplot factor)	$b-1$	SSB	SSB/DFB	$\sigma^2 + \sigma^2_\beta$	MSB/MSE
A∗B	$(a-1)(b-1)$	SSAB	SSAB/DFAB	$\sigma^2 + \sigma^2_{\alpha\beta}$	MSAB/MSE
Error (subplot var.)	$a(b-1)(c-1)$	SSE	SSE/MSE	σ^2	
Total	$abc-1$				

Table 3: ANOVA table for a split-plot model with blocks.

The split-plot design was introduced in the context of a block design for the whole plots where the layout is easiest to describe. However, split-plot structures arises equally in contexts where there are replications instead of blocks for the whole plot factor. In particular, this is a more appropriate

analogue for the hierarchical models. The models and ANOVA tables are very similar to those shown above, with the modifications that follow from the absence of blocks. The model can be written as follows, where index k now corresponds to replications of each whole plot level,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (AC)_{ik} + \varepsilon_{ijk}, \tag{11}$$

and with the same assumptions as above on the random effects. The ANOVA table is given as well. In both versions of the balanced split model, the split-plot variance σ^2_{AC} can be estimated from the respective ANOVA tables as,

$$\hat{\sigma}^2_{\text{AC}} = (\text{MSAC} - \text{MSE})/b.$$

Source	DF	SS	MS	EMS	F
A (whole plot factor)	$a-1$	SSA	SSA/DFA	$\sigma^2 + b\sigma^2_{\text{AC}} + \sigma^2_\alpha$	MSA/MSAC
C(A) (whole plot var.)	$a(c-1)$	SSAC	SSAC/DFAC	$\sigma^2 + b\sigma^2_{\text{AC}}$	MSAC/MSE
B (subplot factor)	$b-1$	SSB	SSB/DFB	$\sigma^2 + \sigma^2_\beta$	MSB/MSE
A∗B	$(a-1)(b-1)$	SSAB	SSAB/DFAB	$\sigma^2 + \sigma^2_{\alpha\beta}$	MSAB/MSE
Error (subplot var.)	$a(b-1)(c-1)$	SSE	SSE/MSE	σ^2	
Total	$abc-1$				

Table 4: ANOVA table for a balanced split-plot model with replications.

Note 5.1. Standard errors of fixed effect estimates and contrasts use different estimates of variation, similar to the discussion for a two-way ANOVA in note 3.1, to which the reader is referred for details. We give only a summary of the results for (both) split-plot models.

- whole plot factor: standard errors for means and contrasts use the whole plot variation MSAC instead of the residual (subplot) MSE,
- subplot factor: standard errors for contrasts use MSE, but standard errors for group means use both variance components: $\text{Var}(\bar{y}_{\cdot 1 \cdot}) = (\sigma^2_{\text{AC}} + \sigma^2)/ac$, see below for approximating the degrees of freedom for the estimated total variance, $\sigma^2_{\text{tot}} = \sigma^2_{\text{AC}} + \sigma^2$,
- interaction: standard errors for means of the combined factor (A×B) use the estimate for σ^2_{tot}, standard errors for contrasts within the same level of the whole plot factor (e.g., $\psi = (\alpha\beta)_{11} - (\alpha\beta)_{12}$) use MSE, whereas standard errors for contrasts involving different levels of the whole plot factor (e.g., $\psi = (\alpha\beta)_{11} - (\alpha\beta)_{21}$) also use the estimated σ^2_{tot}.

The so-called Satterthwaite approximation of the degrees of freedom for the estimated σ^2_{tot} is determined from

$$1/\text{df} = k^2/\text{DFAC} + (1-k)^2/\text{DFE}, \quad \text{where} \quad k = \text{MSAC}/(\text{MSAC} + (b-1)\text{MSE}).$$

References

[1] Christensen, R. (1996), *Analysis of Variance, Design and Regression*, Chapman & Hall / CRC Press, Boca Raton.

[2] Kotz, S. & Johnson, N. (1985). Measurement Error. In *Encyclopedia of Statistical Sciences*, Vol. 5, Wiley.

[3] Snedecor, G. W. & Cochran, W. G. (1967). *Statistical Methods*, 6th ed. Iowa State University Press.