Random 1: Random Effects Models

Outline:

0. Some background
1. One-factor random effects designs
2. Two-factor random effects designs
3. Nested random effects designs

Examples:

1. Grass Example: Random effects one-way
2. DNA Example: Random effects two-way
3. Pharmaceutical Example: Nested Factor

0. Some Background

Up to this point in the course, all models we worked with were considered **Fixed Effects Models**. All terms were considered "fixed" except the error term: $\varepsilon \sim N(0, \sigma^2)$.

In this section (Random 1), we will work with **Random Effects Models** where all terms will be considered "random" except the intercept (μ).

In the next section (Random 2), we will work with **Mixed (Effects) Models** where some terms are fixed and others are random. Mixed models commonly arise in research analysis!

Fixed effects correspond to factors that have a predetermined set of levels and the only inferences are for the levels of the factors actually used in the experiment. We are interested in making inference about means (or differences between means).

Random effects correspond to factors that have levels that are randomly selected from a population of possible levels. The inferences are for all levels of the factor in the population (not just the levels used in the experiment). We are interested in identifying sources of variability and making inference about variance components.

Note: Random effects are usually factors (categorical variables). Numerical random effects are much less common and not discussed in STAT512.

To decide whether a term is fixed or random, consider:

1. How were the levels selected?
    - Fixed effects have levels that were chosen for scientific interest. It is of interest to compare these specific levels.
    - Random effects have levels that were chosen at random (or for convenience). It is not of interest to compare these specific levels.
2. How were these levels selected?
    - Fixed effects have all of the levels of interest.
    - Random effect have a subset of levels of interest.
3. Consider if someone repeated the experiment again.
    - For fixed effects, we could run the experiment again with these levels.
    - For random effects, we could not rerun the experiment again with these levels.

(Mixed) Example: A study was done to compare 3 fertilizer treatments on 4 varieties of barley. A blocked design was used where each of the 12 treatment combinations (3 fertilizers x 4 varieties) were observed within each of 10 fields (with random assignments). In other words, a 3x4 factorial with RCB design.

Fertilizer is a fixed effect. It has "specially chosen" levels. It is of research interest to compare mean response for the fertilizers.

Variety is a fixed effect for similar reasons. However, in some cases where you might argue that variety should be treated as a random effect.

Field (block) considered a random effect. Fields are not of research interest. Instead we just want to account for field to field variability.

In this group of notes:

- We will look only at random effects models. In practice, mixed models are much more common.
- We will do lots of testing of variance components. In practice, this is not a common analysis for addressing research questions.
- We will consider two estimation approaches: ML and REML. In practice, REML estimation is the standard. When we move to Random2 and 3, we will always use method=REML (default).

This group of notes will help to illustrate how the analysis and tests change when we have random effects in the model. I think of this group of notes as a transition between fixed effects models and mixed models.

1. One-factor random effects designs

Example 1: Randomly select *t* grass samples from a large area. We have the idea that each sample represents a genotype that exists in that area. Select *n* tillers from each sample, and measure chlorophyll content on each tiller. Consider the one-factor model.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1,\ldots,t \text{ grasses}$$
$$j = 1,\ldots,n \text{ tillers}$$

The model looks the same as the one-way ANOVA model considered before. In that model the α's were fixed, unknown parameters, and we were interested in testing hypotheses about the α's, or forming confidence intervals relating to the α's (or means).

In this example, we are not interested in these particular the α's. We are interested in the population from which they came. We want to know the variance of that population.

This model uses the same formula as the "fixed effects" (with fixed α's) model, but different assumptions are made about the α's:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\text{Assume}: \alpha_i \sim N(0, \sigma_\alpha^2) \text{ and } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

The α's are independent of the ε's.

Primary interest in this problem is on the parameters:

σ_α^2 = variance from grass to grass

σ_ε^2 = variance from tiller to tiller (within the same genotype)

We may also be interested in functions of these parameters.

Example 2: Randomly select *t* machines of the same type from an assembly area. Select *n* operators per machine to train (for a total of *tn* operators). Measure the productivity of each operator.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \ldots, t \quad \text{machines}$$
$$j = 1, \ldots, n \quad \text{operators}$$

In this example we are not interested in these particular machines; rather, we are interested in the population from which they came. We are interested in σ_α^2, the variance of that population.

Example 3: Randomly select *t* farms from an Extension Service listing of wheat farming cooperators. Sample *n* locations from each farm, and estimate the Russian wheat aphid density at each location. We are interested in σ_α^2, the variance in RWA density from farm to farm in our population.

Example 4 (Sire Model): Randomly select *t* bulls from a large herd. Mate each bull with *n* cows, and measure growth or carcass characteristics of the offspring.

Key point: In each of these examples, we are not interested in the particular individuals that make up the population of α's. We are interested in the populations from which they come. This is called the "random effects" model.

Look at the mean and variance of y_{ij}:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \;\; \alpha_i \sim N(0, \sigma_\alpha^2) \; and \; \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\begin{aligned} E(y_{ij}) &= E(\mu + \alpha_i + \varepsilon_{ij}) \\ &= E(\mu) + E(\alpha_i) + E(\varepsilon_{ij}) \\ &= \mu + 0 + 0 \\ &= \mu \end{aligned}$$

Variance $V(y_{ij}) = V(\mu + \alpha_i + \varepsilon_{ij})$

$= V(\mu) + V(\alpha_i) + V(\varepsilon_{ij})$ [requires independence]

$= 0 + \sigma_\alpha^2 + \sigma_\varepsilon^2$

$= \sigma_\alpha^2 + \sigma_\varepsilon^2$

This value is called the "total variance", and

σ_α^2 and σ_ε^2 are called "components of variance"

Some objectives for Random Effects Models:

1. Estimate the "components of variance" (also called "variance components"), and use those values to estimate the "total variance".

2. Use the estimates of the "components of variance" to estimate other values of interest.

In Example 4:

$$h = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

is called the heritability because it describes the proportion of total variability that is explained by inheritance from the sire. The model in Example 4 is called the "sire model".

3. Test the H0: $\sigma_\alpha^2 = 0$ versus HA: $\sigma_\alpha^2 > 0$.

Estimating the variance components

Method #1 Restricted (Residual) Maximum Likelihood (REML): This is the default with the `lmer(, REML = TRUE)`. This is a type of maximum likelihood estimation that is "restricted" to the residuals. With equal sample size, REML gives unbiased estimates of variances.

Note: In practice, REML estimation is the standard. When we move to Random2 and 3, we will always use method=REML (default).

Estimating the variance components continued

Method #2 Maximum Likelihood Estimation (ML): Can be done using `lmer( , REML = FALSE)`. Not usually used to estimate variance components.

Idea of ML estimation: Estimate the parameters by the values that maximize the probability of observing the data set that was observed. This is a widely used principle, but in this case produces "biased" estimates.

Example: ML estimate of the variance in a single sample problem is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Standard Estimate: $s^2 = \hat{\sigma}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Grass Random One-Way Example: Six grass samples are randomly selected (t=6). Five tillers are analyzed per sample (n=5).

```
> library(lme4)

> library(lmerTest)

> Model2 <- lmer(y ~ (1|grass), data = Grasses)

> summary(Model2)
Random effects:
 Groups   Name        Variance Std.Dev.
 grass    (Intercept) 0.04747  0.2179
 Residual             0.74540  0.8634
Number of obs: 30, groups:  grass, 6

Fixed effects:
            Estimate Std. Error    df t value Pr(>|t|)
(Intercept)    4.776      0.181 5.000   26.39 1.46e-06
```

Notes about lmer():

1. **We will use lmer() from the lme4 package to fit models with random effects.**
2. The syntax (1|grass) is used to indicate a random effect associated with the intercept (1) for each unique level of grass.
3. **In most cases, we will want to use the default REML estimation**! By default lmer() will calculate REML estimates (REML = TRUE).
4. If tests are of interest, the lmerTest package is required. The rand() function will provide likelihood ratio tests for variance components.

emmeans vs blups

1. Estimates of random effects can be obtained using the `ranef()` function. This will cause "Best Linear Unbiased Predictions" (BLUPs) to be printed.
2. The BLUPs are not the same as the emmeans in the fixed effects model.
3. In general, the BLUPs are closer to the middle of the data than the emmeans from the fixed effects model. Using the Grass data as an example, the emmeans range from 4.15 – 5.32. The blups range from 4.63 – 4.91.
4. Notice that for this example, the ranking of the grasses is the same using emmeans or blups.

2. Two-way random effect designs

DNA Example: Sample 3 subjects (factor A). Sample 3 analysts (factor B). Y = DNA content of n=2 samples from all 9 combinations of subjects and analysts.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$\alpha_i \sim \mathrm{N}(0, \sigma^2_\alpha) \qquad \beta_j \sim \mathrm{N}(0, \sigma^2_\beta)$$

$$(\alpha\beta)_{ij} \sim \mathrm{N}(0, \sigma^2_{\alpha\beta}) \qquad \varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2_\varepsilon) \quad \text{All independent!}$$

Objectives: (1) Estimate the 4 variance components (2) Test hypotheses about the variance components.

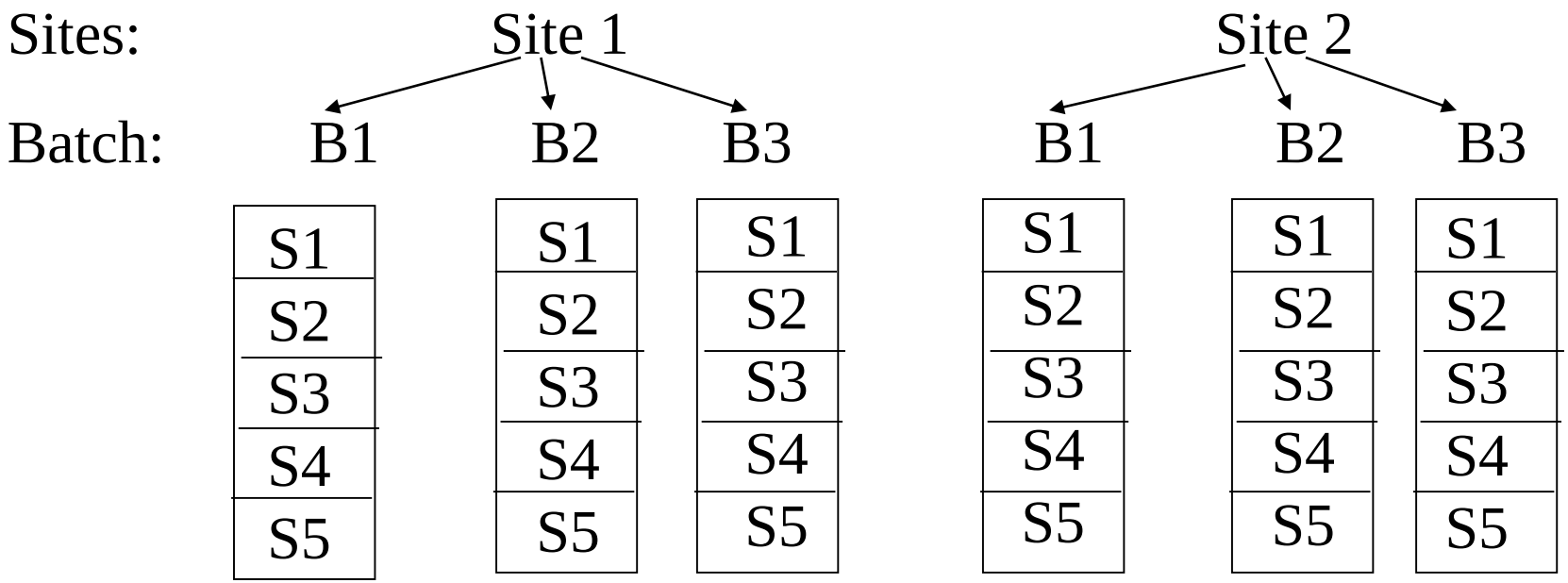
In this example, we find that two of the variance components (for analyst and subject*analyst interaction) are estimated to be zero. Those terms can be dropped from the model, but it does not change the results.

```
> Model1 <- lmer(DNAcont ~ (1|subject) +
(1|analyst) + (1|subject:analyst), data = DNA)
>  summary(Model1)

Random effects:
 Groups           Name        Variance  Std.Dev.
 subject:analyst (Intercept) 0.0000    0.0000
 analyst         (Intercept) 0.0000    0.0000
 subject         (Intercept) 4.6217    2.1498
 Residual                    0.1491    0.3862
Number of obs: 18, groups:  subject:analyst, 9;
analyst, 3; subject, 3
```

3. Nested random effects designs

Example: (O&L, ex 17.11) A large company wants to study the uniformity of content of a drug that is produced at several sites. Two sites are selected (we will treat these as randomly selected from a larger population of sites). Three batches are randomly selected within each site. Five samples are analyzed within each batch.



What makes this example different from previous examples is that batches are randomly selected within sites. It would not make sense to fit a batch effect that averages over sites, because they are different batches at the different sites. We say that "batches are nested within sites".

To see that batches are nested within sites:

1. Notice how the sampling was done. Batches were randomly selected within a site.
2. Notice that we could rename the batches without corrupting the data interpretation.
3. Notice that each batch uniquely represents a single site.

Definitions of Crossed and Nested (from Ott and Longnecker):

In a factorial experiment, the factors A and B are said to be **crossed** if the physical properties of the b levels of factor B are identical for all levels of factor A.

Factor B is said to be **nested** within the levels of factor A if the physical properties of the b levels of factor B vary depending on which level of factor A it is associated with.

Definitions of Crossed and Nested (from West):

When a given level of a factor (fixed or random) can be measured across multiple levels of another factor, one factor is said to be **crossed** with the other.

When a particular level of a factor (B, fixed or random) can only be measured within a single level of another factor (A) and not across multiple levels, the levels of the B are said to be **nested** within the levels of the A.

Returning to the Pharmaceutical Example:

Let y_{ijk} = content for the i^{th} site, i=1,..,a

j^{th} batch, j=1,..,b

k^{th} sample, k=1,..,n

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

$$\alpha_i \sim N(0, \sigma_\alpha^2) \qquad \beta_{j(i)} \sim N(0, \sigma_{\beta(\alpha)}^2)$$

(The $\beta_{j(i)}$ is read as: "beta j within i".)

σ_α^2 represents the site to site variability.

$\sigma_{\beta(\alpha)}^2 = \sigma_\beta^2$ represents the batch to batch variability (within a site).

Returning to the Pharmaceutical Example:

```
> Model1 <- lmer(content ~ (1|site/batch),
data = Tablet)
> summary(Model1)

Random effects:
 Groups     Name        Variance Std.Dev.
 batch:site (Intercept) 0.01647  0.1283
 site       (Intercept) 0.00000  0.0000
 Residual               0.01209  0.1100
Number of obs: 30, groups:  batch:site, 6;
site, 2
```

Notes about nesting:

1. In R, to specify that factor B is nested within factor A we use the notation A/B. This is equivalent to A + A:B.

2. It does not make sense to include an "additive" or "main effect" for batch because batches are different at each site. Hence it does not make sense to average batches over sites.

3. The error term in the pill example is a nested replication (samples are nested within batch), but that is usually ignored in the notation.

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{k(ij)}$$

4. When there is nesting, I prefer to work with unique names to prevent confusion/mistakes. For the pharmaceutical example, instead of labeling batches 1 - 3 within each site, we can label batches 1 - 6!

5. If you are having trouble deciding whether one factor is nested within another, try this diagram:

Batch	Site 1	Site 2
1		
2		
3		

If you can re-arrange rows within a column without corrupting the data interpretation, then then the row factor is "nested within" the column factor.

6. Compare the RCB model to the CRD model:

A) In the RCB design reps are crossed with trts:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

B) In the CRD design reps are nested within trts:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$