Designs to Study Variances II

Sean Hellingman ©

Design for Data Science (ADSC2030)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Another purpose of experimentation is to study the sources of variability in the response.
- Understanding where the variability comes from, allows for more focused designs.
- Some sampling experiments use a natural hierarchical design.
    - Experiments with more than one factor that use nested factors.

Nested Sampling Experiments

- In a **Nested Sampling Experiment** (NSE) there is a hierarchical design to the multiple factors.
- The levels of a **nested factor** are physically different depending on the level of the factor it is nested in.
- So far we have only covered designs to study variances that use *crossed factors*.
  - The factors are uniquely defined separate of the other factors.

Road Map

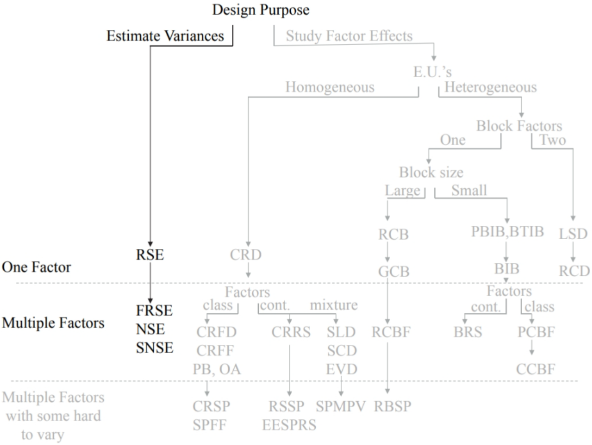

Figure: Source: (1)

Illustrative Example 1 A

- The Experiment covered in Example 5 of Part I has *crossed factors*
  - *Each operator measured each part; therefore, the operator number was uniquely defined and referred to the same operator regardless of which part they measured.*
- By changing to an experiment where n parts were selected and each part was measured by two operators, the operator factor becomes a nested factor.
  - It does not have to be the same two operators measuring each part.
- More convenient to design in this manner if measurements are to be taken over a long period of time.

Illustrative Example 1 B

- As the operators differ depending on the part number being measured, the operator is a **nested factor** within each part.
  - The first operator measuring a specific part is not necessarily the same person as the first operator measuring a different part.

- If each operator must measure a different set of parts, and the part becomes the nested factor (within each operator).
  - *Destructive measurements*
  - The first part measured by the first operator is not physically the same as the first part measured by subsequent operators.

Nested Factors

- We have already seen an example of nested factors when examining the error term (ϵ_{ij}).
  - Represents the effect of the jth replicate EU.
  - Because different EUs are used for each factor level (or combination), the EU is always nested within another factor level (or combination) in the design.

Two-Stage Nested Design Model

- When two factors are *crossed* their interaction can be included in the model.
    - When *nested*, we cannot include the interaction because the nested factor includes the degrees of freedom that could be taken by the interaction.
- Model for two-staged nested design (B nested in A):

$$y_{ijk} = \mu + a_i + b_{(i)j} + \epsilon_{ijk} \tag{1}$$

- *Nested or hierarchical designs can easily be extended to include several stages or factors.*

Two-Stage Nested Design in R

- Assume that we have a nested process with three factors (C nested in B, and B nested in A) and the following model:

$$y_{ijkl} = \mu + a_i + b_{(i)j} + c_{(ij)k} + \epsilon_{ijkl}.$$

- To obtain the REML estimates in R:
  - `model <- lmer(response ~ 1 + (1|A) + (1|A:B) + (1|A:B:C), data = data)`
  - `summary(model)`

Example 1 Preliminaries I

- Four-stage nested sampling study on the variability of properties of crude rubber (1954).
- A sample of four batches of rubber were taken from each of four suppliers.
  - The first batch obtained from the first supplier is not the same as the first batch taken from the second supplier (batch nested within supplier).
- Two sample mixes were made from each batch.
  - Since the two sample mixes for one batch are physically different than the sample mixes for any other batch, the sample mix is nested within the batch.
- Three replicate tests were performed on each mix to determine the elasticity.

Example 1 Preliminaries II

- The model we are interested in for Example 1:

$$y_{ijkl} = \mu + a_i + b_{(i)j} + c_{(ij)k} + \epsilon_{ijkl}.$$

  - y_{ijkl} is the l^{th} elasticity made from the k^{th} sample mix taken by the j^{th} batch from the i^{th} supplier.
  - a_i is the random supplier effect.
  - $b_{(i)j}$ is the random batch effect.
  - $c_{(ij)k}$ is the random sample mix effect.
  - ϵ_{ijkl} is the the random replicate determination effect.

  - $i = 1, ..., 4$; $j = 1, ..., 4$; $k = 1, 2$; $l = 1, ..., 3$

Example 1

- Import the *rubber* data from the *daewr* package.
- Take some time to get to know the data.
- Use the `lmer()` function to estimate the variance components.
- What are your thoughts on the estimates?

Comments on NSE Models

- Recall: *In order to increase the confidence in the variance estimates, the number of random factor levels should be increased.*

- In nested designs, with several stages, increasing the topmost factor greatly increases the overall sample size.
  - From Example 1: Increased the number of suppliers from 4 to 20 to get a more precise $\hat{\sigma}_a^2$, would increase the observations from 96 to 480.

  - Still only $(a - 1) = 19$ degrees of freedom for the supplier effect.

  - $abc(r - 1) = 20(4)(2)(3 - 1) = 320$ degrees of freedom for the random replicate effect.
    - *Really only improving the precision of $\hat{\sigma}^2$ and $\hat{\sigma}_c^2$.*

- **When there are more than three stages, balanced hierarchical designs are usually not recommended.**
  - Staggered nested designs are preferred.

Staggered Nested Designs

- In a **staggered nested design**, only one of the two levels of the succeeding factor leads to the next two-level stage.
    - Unlike the completely nested design where all levels lead into two or more levels.
- Savings in the overall number of observations needed.
- In completely nested designs, the degrees of freedom are concentrated on the lower-tier factors.
    - The information is more balanced in a staggered nested design.

Road Map

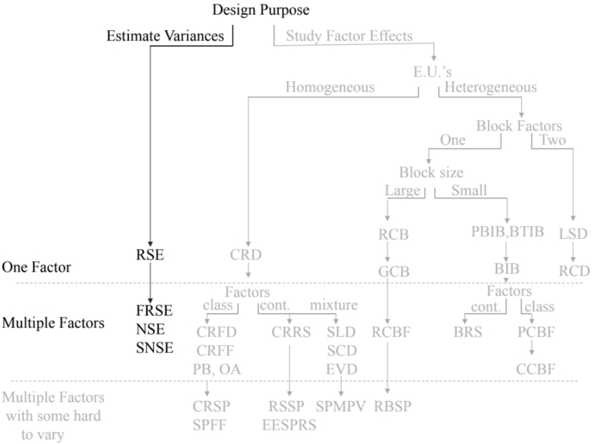

Figure: Source: (1)

Comparing Designs

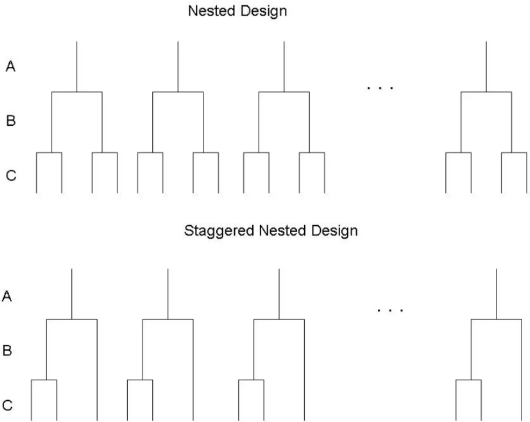

Figure: Source: (1)

Staggered Nested Designs Observations

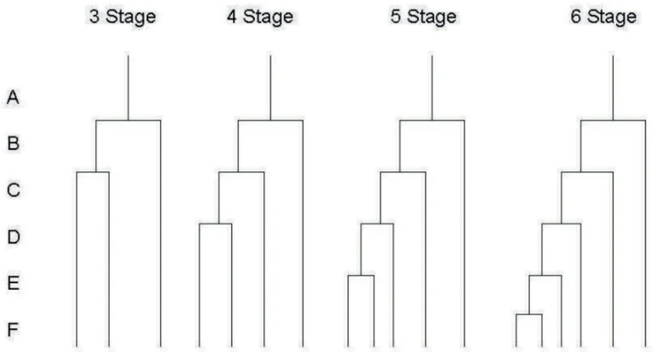

Figure: Source: (1)

Staggered Nested Design in R

- Assume that we have a nested process with three factors (C nested in B, and B nested in A) and the following model:

$$y_{ijkl} = \mu + a_i + b_{(i)j} + c_{(ij)k} + \epsilon_{ijkl}.$$

- To obtain the REML estimates in R:
  - `model <- lmer(response ~ 1 + (1|A) + (1|A:B) + (1|A:B:C), data = data)`
  - `summary(model)`

- **The only difference in the applications will be the degrees of freedom.**

Example 2 Preliminaries I

- Staggered nested design was used to estimate sources of variability in a continuous polymerization process (1989).
  - Polyethylene pellets produced in lots of 100 000 lbs.
- A four-stage design was used to partition the source of **variability in tensile strength**:
  - Between lots
  - Within lots
  - Due to the measurement process
- The model:

$$y_{ijkl} = \mu + a_i + b_{(i)j} + c_{(ij)k} + \epsilon_{ijkl}.$$

Example 2 Preliminaries II

- Thirty lots were sampled at random
  - Lot represents the topmost factor (source of variability A)

- From each lot, two boxes of pellets were randomly selected
  - Box of pellets represents the second stage (source of variability B)

- From the *first* box selected from each lot, **two** preparations were made for strength testing.
- From the *second* box selected from each lot, **one** preparation was made
  - The preparations represent the third stage (source of variability C)

- Finally, two repeat strength tests were made from the first preparation from box one, while only one strength test was made from the other three preparations.

Example 2 Preliminaries III

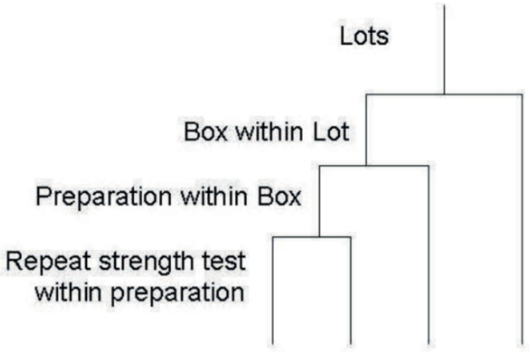

Figure: Source: (1)

Example 2

- Import the *polymer* data from the *daewr* package.
- Take some time to get to know the data.
- Use the `lmer()` function to estimate the variance components.
- What are your thoughts on the estimates?
    - How much of the total variation is due to variability among lots?
- *Comment on the degrees of freedom.*

Confidence Intervals in R

- To estimate the asymptotic confidence intervals using R:
  - `pr1 <- profile(lmer(response ~ 1 + (1|A) + (1|A:B)+ (1|A:B:C), data = data))`
  - `confint(pr1)`
- *Should have at least 30 observations.*

Example 3

- Estimate the asymptotic confidence intervals using R for the variances you obtained in Example 2.

Staggered Nested Design Comments

- The degrees of freedom are approximately the same for all stages.
- **To determine desired sample size: use the methodology from *Sample Size for One-Factor Sampling Studies* (Designs to Study Variances I).**
- Fewer observations needed overall compared to *Nested Designs.*

Fixed and Random Factors

- Sometimes when fixed treatment factors are being studied, random factors are also introduced in the model.
- The random factors are introduced through the way the experiment is collected.
- This leads to a *mixed model* with fixed and random factors.

Example 4 Preliminaries I

- Designed an experiment to compare different formulations and methods of applying pesticide to the leaves of cotton plants.
    - **Goal to increase the amount of active pesticide remaining on the plant one week after application.**
- Two different formulations and two different application methods: 2^2 factorial experiment.
- The experimental unit was a row of cotton plants called a plot.
    - Eight plots were selected and two were randomly assigned to each of the four treatment combinations ($r = 2$).
- There was too much plant material to send to the lab for analysis.
- One week after application, two samples of leaves were sent to the lab from each plot.
    - An amount that was suitable to be studied in the laboratory.

Example 4 Preliminaries II

- Formulation, application technique, and their interaction are fixed factors.
- The plot is a random factor nested in the combination of formulation and application technique.
    - Multiple plots per treatment combination were included so that the variance of different plots could be estimated.
- The replicate samples taken from each plot would be classified as sub-samples.

Example 4 Preliminaries III

- The model for data:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + p_{(ij)k} + \epsilon_{ijkl} \tag{2}$$

  - y_{ijkl} pesticide residual on the l^{th} sample, from the k^{th} plot, with formulation i, and application j.
  - α_i the formulation effect.
  - β_j the application effect.
  - $\alpha\beta_{ij}$ the interaction effect.
  - $p_{(ij)k}$ the random plot effect.
  - ϵ_{ijkl} the random sample effect.

- The assumptions for each component (fixed and random) must hold.

Fixed and Random Factors in R

- If it can be assumed that there is correlation between the fixed effects due to the experiment we need to use orthogonal contrasts in R:
  - `library(lme4)`
  - `c1 <- c(-0.5, 0.5)`

  - `model <- lmer(response ~ 1 + factor.A + factor.B + factor.A:factor.B + (1|random.factor:factor.A:factor.B), contrasts = list(factor.A = c1, factor.B = c1), data = data)`
  - summary(model)

Example 4 from slides

- Import the *pesticide* data from the daewr package to complete the following tasks:
  1. Take some time to get to know the data.
  2. Use the `lmer()` function to estimate the model containing fixed and random effects.
     - *In this example, common application of the pesticide to each plot might induce a correlation between sub-samples from the same plot.*
  3. Comment on the fixed and random effects estimates.

Comparing Means in R

- **When using objects created by the `lmer()` function:**
  - For designs with more than two levels in the fixed factors the `estimable()` function from the *gmodels* package as well as the `lsmeans()` function may be used.
- The *lsmeans* package can also compute Tukey's adjusted pairwise comparisons of the means:
  - `lsmeans(model, pairwise ~ factor, adjust = c("tukey"))`
- The `anova(model)` function will produce correct F-tests for the fixed factors.

Checking Model Assumptions

- We will use two visualizations to check the normality assumption of the random effects.

- Note: *The usual assumptions on the fixed effects still apply*
  - Constant variance
  - Normal distribution
    - *Both can be checked through the residuals*

Normality Assumption I

- We can use a histogram of the random intercepts to see if the distribution is approximately normal.
- In R:
    - `intercept.fix = fixef(model)["(Intercept)"]`
    - `est.eff = coef(model)["Factor.A:Factor.B"]`
    - `` hist(est.eff$`Factor.A:Factor.B'[,"(Intercept)"] - intercept.fix) ``
- *Make sure to appropriately label your plots*

Normality Assumption II

- We can check the Q-Q plot of the random effects.
- In R:
  - `qqnorm(ranef(model)$"Factor.A:Factor.B"[[1]], main= "Label for Q-Q Plot", ylab= "EBLUP", xlab = "Normal Score" )`
- EBLUP: Empirical Best Linear Unbiased Predictors (obtained using the `lmer()` function in R).

Example 5

- Check the normality assumption of the random effects of the `lot:box:prep` term from Example 2.

Visualize the Random Effects

- The *sjPlot* package can be used to model the random intercepts estimated by the `lmer()` function:
  - `library(sjPlot)`
  - `plot_model(model, type = "re")`

Example 6

- Use the `plot_model(model, type = "re")` function to visualize the random effects from the model estimated in Example 2.

Exercise 1

- Are there any models from the random blocking portion of the course that could be treated as designs with fixed and random factors?
- If so, estimate an appropriate model for that particular example and check the normality assumptions of the random term(s).

References & Resources

1. Lawson, J. (2014). *Design and Analysis of Experiments with R (Vol. 115)*. CRC press.

- lme4 package
- `lmer()`
- `plot_model()`
- daewr
- `qqnorm()`
- `lsmeans()`