Stat 471/571: Split plot designs and their analysis: part 3

A bit more on Expected Mean Squares

I was asked whether R would compute EMS and I said I knew of nothing that interfaced with lmer(). If you search for EMS computations in R, you will discover the EMSaov library. **DO NOT USE THIS PACKAGE.** Remember that "you get what you pay for". R is open source and most packages are written by volunteers. Many R packages are high quality; some are disasters. EMSaov is a disaster. The rest of this section dives into details to explain my opinion.

There are two systems used to determine Expected Mean Squares. (Yes, statisticians can make anything into something complicated.) The most natural system (in my mind) is that random effects are independent samples from some population. In the PA study, independence means that each school (the main plot error) is an independent draw from some distribution of schools.

The other system goes back to a 1956 paper by Cornfield and Tukey, echoed in many textbooks, including (as far as I can tell) the Montgomery 2008 book cited in the EMSaov documentation. I haven't verified this because I don't have a copy of that text. This system imposes a constraint on the random variables. They sum to 0 within each fixed effect. So in the PA study, the school effects sum to 0 within each treatment group (Intervention or Control). This introduces a negative correlation among the random effects.

Why use this sum-to-zero constraint? Many folks have wondered. Both Cornfield and Tukey are dead, so we can't ask them. The best explanation is that this is a hangover (a very bad one) from the use of a sum-to-zero constraint for fixed effects (i.e., what JMP uses to find estimates in an overparameterized problem). This is reasonable for fixed effects because it doesn't change estimable functions (the quantities that don't depend on the choice of parameterization). In my mind, sum-to-zero isn't reasonable for random effects; it introduces all sorts of unusual behavior. For example, the correlation between schools should be something fixed and characteristic of the population of schools. No correlation, i.e. independent, is the most likely choice. Under Cornfield/Tukey, that correlation depends on the number of schools within a treatment. It's -1/5 when there are 5 schools and -1/10 when there are 10. The stochastic properties of a random variable shouldn't depend on how many are observed!

The choice of system changes the Expected Mean Squares and changes the appropriate denominator for some F tests. EMSaov appears to use the Cornfield/Tukey approach, although I'm not positive because the documentation cites Montgomery 2008. However, what is reported by EMSanova() matches what I expect under the sum-to-zero constraint.

The second reason to not use EMSaov functions is that they require balanced data. That's the easy case. Anyone can compute EMS from balanced data by counting levels of each factor and doing some algebra. The harder, and more relevant, situation is unbalanced data, where the package needs to do some matrix computations to figure out the EMS. The EMSaov functions don't do that. That's a serious red flag in my opinion.

SAS and JMP use the independent random effects system. That's the appropriate choice, in my mind.

R/lmer and R/emmeans use independent random effects. That's what (1|effect) generates. You get correct tests and standard errors/confidence intervals for estimates. You just don't get EMS.

Split-split plot designs

Studies using a split plot design have two sizes of experimental unit. Infrequently, studies have three sizes of eu. That's sometimes called a split-split plot design. Nothing really new. Just figure out the design for each level and knit them together. The split*main interactions go into the split plot level. The split-split*split and split-split*main interactions go into the split-split level.

Here's an example, based on a physical activity study. There are 3 factors being studied: school program (intervention or control), teacher willingness (high, medium, low), and student gender. 10 schools are randomly assigned to the school program, with 5 in each treatment (intervention, control). Data are collected from 3 classrooms in each school, with one from each level of teacher willingness. A random number of students in each classroom are measured, so the numbers of boys and girls per classroom is not fixed.

There 3 sizes of experimental unit: school, assigned to program, classroom "assigned" to teaching willingness, and student "assigned" to boy or girl. The expt. design is a CRD at all levels. Here's what the ANOVA table will look like:

F/R	Source	df
F	Program	1
R	School(program)	8
F	Willingness	2
F	W*P	2
R	Teacher(school, willingness)	16
F	Gender	1
F	G*P	1
F	G*W	2
F	G*W*P	2
R	Residual	many

How to figure out the error df at each level? Count how many eu's there are at each level and subtract df for effects "above" that error, i.e., at and above that level.
School level: 10 schools, 9 df total, 1 df for program, so 8 for the error
Teacher level: 30 teachers, 29 df total, 13 df for everything above, so 16 error df
Student level: unknown number of kids, will be # kids - 1 - everything above

The R lmer model will be compactly written as:
Y ~ Program*Willingness*Gender + (1|School:Program) + (1|Teacher:School:Program:Willingness)

The SAS model will be:
model Y = Program|Willingness|Gender;
random School(Program) Teach(School*Program*Willingness);
The JMP model will include School(program) and Teacher(school*willingness), with both declared as random effects. Although, if school and teacher have unique id's, you don't have to nest/cross them. I prefer to always nest/cross them so I don't get tripped up by shared ids that I don't recognize.

Other designs with more than one size of eu's

In a split plot study, it is clear that one size eu is nested in another. Eg., in the PA study, student is nested in school. There is a "larger" and a "smaller" size of eu.

Sometimes you run across designs that clearly have multiple sizes of eu but it isn't clear what is nested in what. Although I had been introduced to one such design in my graduate experimental design class, I had forgotten about it when I first encountered it as a practicing statistical consultant. A reply on a discussion forum set me straight. Here's the most common example of this design.

Fruit trees are commonly planted in neat rows and columns (the ISU Horticultural Research Farm has a lot of these). Imagine a study of two factors: Fertilization (4 levels) and Fungicide amount (2 levels), on apple production. Fertilization is randomly assigned to rows of apple trees; fungicide is randomly assigned to columns of trees. A picture is in the hand-drawn notes. There are 3 rows of trees with each fertilization amount and 4 columns of trees with each fungicide amount. Although this "sort-of" sounds like a Latin Square design, there are two key differences:

- In a Latin Square, rows and columns are blocking variables, used only to remove unwanted variability. Here, rows and columns are randomly assigned to treatment factors. Fertilization to rows; Fungicide to columns.
- Both Fertilization and Fungicide are replicated. The same level is assigned to more than one row / column.

In total, there are data from (4 x 3) x (2 X 4) = 96 trees.

There are three sizes of eu in this study:
row of trees, randomly assigned to fertilization amount
column of trees, randomly assigned to fungicide amount
individual tree (row x column intersection),
randomly assigned to the combination of fertilization and fungicide
The clue that this is not your typical split plot study is that you can't figure out what is the "larger" and what is the "smaller" size eu. In other words, you can't figure out which eu is nested in the other. That's because rows and columns are crossed, not nested.

This design is called a split-block or a strip-plot design. I think strip-plot is more appropriate

because that's what's being done: treatments applied in strips, some horizontal (rows: fertilization) and others vertical (columns:fungicide).

The ANOVA table for the study described above: 4 levels of fertilization replicated in 3 rows each, 2 levels of fungicide replicated in 4 columns each, would be:

Level	F/R	Source	df
Rows	F	Fertilization	3
	R	Row(Fert.)	8
Columns	F	Fungicide	1
	R	Col(Fungicide)	6
Cells	F	Fert x Fung	3
	R	Residual= Tree	74
		c.total	95

You see the downside of this sort of design. Main effects of fertilization and fungicide are tested against the variability between rows or between columns treated alike. Relatively small df. And, if there is no replication (e.g., 4 rows, each with a different fertilization amount), there is no error to assess variability among rows. There one plus side of this design is that there are lots of df to estimate the tree-tree variability (residual error). That's not a bad thing since interaction contrasts have lower power than main effect contrasts.

Gauge repeatability and reproducibility studies

A similar design is used to assess multiple sorts of variability. These are most common in engineering. These days, manufacturing really cares about consistent product. This is assessed by measurements, and the quality of measurements matters. If the measurements are highly variable, it's really hard to assess whether the product is variable or not. In the typical Gauge R&R study, the the concern is with three different sorts of variability:

- Can the same person measuring the same object, get similar results?
- Can different people measuring the same object, get similar results?
- What is the variability between parts, if each could be measured perfectly?

In the engineering literature, repeatability quantifies the first concept (same person, same part). Reproducibility quantifies the second concept (different people, same part). The variability between parts is characteristic of the manufacturing process. The terminology here is confusing, not helped by the overlapping "common sense" interpretations of the words. It doesn't help that repeatability and reproducibility have been defined differently in some literatures.

A typical "gauge R&R" study uses multiple parts and multiple operators. Here's a small example, where three operators measure four parts twice.

	Operator		
Part	A	B	C
1	5.1, 5.0	5.8, 5.7	4.9, 5.1
2	4.9, 4.8	5.7, 5.8	5.0, 5.1
3	5.2, 5.4	5.9, 5.7	5.2, 5.1
4	5.0, 4.9	5.8, 5.7	4.9, 5.0

Because we're interested in variability: between operators, between parts, between measurements, we'll make all factors random. The model is:

$$\begin{aligned} Y_{ijk} &= \mu + P_i + O_j + PO_{ij} + \varepsilon_{ijk} \\ P_i &\sim N(0, \sigma_P^2) \\ O_j &\sim N(0, \sigma_O^2) \\ PO_{ij} &\sim N(0, \sigma_{PO}^2) \\ \varepsilon_{ijk} &\sim N(0, \sigma_e^2) \end{aligned}$$

Repeatability is quantified by σ_e^2; Reproducibility is quantified by $\sigma_O^2 + \sigma_{PO}^2$.

In most situations, fixed effects will be crossed and random effects will be nested, when there are more than one. Gauge R&R studies are one of the few situations where random effects are crossed.