The Theory of Designed Experiments

11. Special Topics

UNESP-Botucatu, August 2010

Special Topics

In this last chapter we will briefly look at several design and analysis situations which are outside the scope of what we have already discussed (or obvious extensions of it).

One thing that many of these have in common is that it is no longer appropriate to make the simple assumption of treatment-unit additivity, i.e. the model $y_{i(r)} = \mu + t_r + e_i$ is no longer valid, or has to be modified.

In other cases, it is impossible to do a valid randomization.

I will try to make some recommendations, often trying to retain as much of the logic of the randomization theory as possible.

Complex responses

Textbooks almost always present designed experiments in the context of a single, univariate, continuous response measured from each experimental unit. This is rare in practice.

Responses might be:

- discrete;
- repeated measurements;
- multivariate;
- functional;
- compositional.

How can we deal with these situations?

Discrete data

Consider comparing several pesticide spray treatments in a completely randomized design with several replicates. The response is number of infected plants out of 25.

Given the possible outcomes, the original model,

$$y_{i(r)} = \mu + t_r + e_i,$$

makes no sense.

Instead a reasonable assumption seems to be

$$Y_{i(r)} \sim Binom(25, \pi_{i(r)}),$$

where $\pi_{i(r)}$ depends on the plot and the treatment.

Discrete data

We have assumed a distribution, so have we completely abandoned the ideas of randomization analysis?

No, we can apply them to the unobservable linear predictor, e.g. assume

$$\log\left(\frac{\pi_{i(r)}}{1-\pi_{i(r)}}\right) = \eta_{i(r)} = \mu + t_r + e_i.$$

Under complete randomization, this becomes

$$\eta_{i(r)} = \mu + t_r + \epsilon_i,$$

where ϵ_i is a random effect, as before.

If we approximate ϵ_i as being Normally distributed, we have a generalized linear *mixed* model (GLMM).

We can include block effects in the usual way.

Thus generalized linear models *have no place* in the analysis of data from designed experiments.

Longitudinal data

Consider an experiment to compare two growth hormones for cattle. We measure the weight of each animal each week for 24 weeks. The experimental unit is the animal, but we have repeated measurements on each unit.

As with discrete data, we can first model the responses from each experimental unit, e.g. if y_{ij} is the jth measurement from animal i, assume

$$Y_{ij} = \beta_{0i} + \beta_{1i} v_{ij} + \epsilon_{ij},$$

where v_{ij} is the time at which the jth measurement is taken on animal i, $E(\epsilon_{ij}) = 0$ and

$$Var(\boldsymbol{\epsilon}_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

Longitudinal data

Simple analysis can be done by fitting a simple linear regression to each animal and using $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ as responses. This works well for balanced data sets.

Otherwise combine this model with the usual randomization model by assuming that the unobservable β_{0i} and the unobservable β_{1i} are additive in terms of unit and treatment effects.

Under randomization we get

$$Y_{ij(r)} = (\mu_0 + t_{0r} + \alpha_i) + (\mu_1 + t_{1r} + \gamma_i)v_{ij} + \epsilon_{ij},$$

where α_i and γ_i are random effects. This is the *random slopes* model, very commonly used to analyse longitudinal data.

This could apply to repeated measurements in space, rather than time.

Multivariate responses

In many (most?) experiments there is more than one response variable. Usually, we analyse each one separately.

Consider the experiment to compare the effects of several factors on baked pastry products. Two responses are the cross-sectional expansion index and the longitudinal expansion index.

The multivariate model with additive treatment and unit effects is

$$\begin{bmatrix} y_{1i(r)} \\ y_{2i(r)} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} t_{1r} \\ t_{2r} \end{bmatrix} + \begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix}.$$

Multivariate responses

Under randomization, this becomes

$$\begin{bmatrix} Y_{1i(r)} \\ Y_{2i(r)} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} t_{1r} \\ t_{2r} \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix},$$

where ϵ_{1i} and ϵ_{2i} are correlated because they are randomized to the same experimental unit. This is a standard multivariate linear model.

This model implies that the treatments could have completely different effects on different responses.

A reasonable alternative model is

$$\begin{bmatrix} Y_{1i(r)} \\ Y_{2i(r)} \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} t_{1r} \\ \phi t_{1r} \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix}.$$

This can be fitted by nonlinear least squares (the simplified linear model is nonlinear).

Carry-over effects

In cross-over designs (row-column designs with subjects and periods) allowance is sometimes made for carry-over, i.e. the response in one period might be affected not only by the treatment applied in that period, but also by the treatment applied in the previous period.

Designs are often used which are balanced for carry-over, i.e. each drug (say) follows each other treatment an equal number of times.

	Period			
Subject	I	II	III	IV
I	1	2	3	4
II	3	1	4	2
III	2	4	1	3
IV	4	3	2	1
			⋮	

Carry-over effects

The only possible randomization is of subjects to subject labels. From the randomization viewpoint, subjects are the experimental units, *sequences of drugs* are the treatments and there are multivariate responses.

As above the model under randomization is

$$\begin{bmatrix} Y_{1i(r)} \\ \vdots \\ Y_{4i(r)} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_4 \end{bmatrix} + \begin{bmatrix} t_{1r} \\ \vdots \\ t_{4r} \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \vdots \\ \epsilon_{4i} \end{bmatrix}.$$

Any further analysis requires additional assumptions, e.g. $t_{jr} = d_s + c_t$, where s is the drug applied in period j and t is the drug applied in period $j-1$.

Similar problems arise in two dimensions in agricultural field trials.

Nonorthogonal block structures

Randomization theory only applies to simple orthogonal block structures and not, for example, to blocks of unequal sizes. These can arise, for example, when blocks are litters of animals.

The reason the theory breaks down is that units in blocks of different sizes are no longer exchangeable.

From the randomization viewpoint, the sets of blocks of unequal sizes are actually different experiments. This tells us how to analyse the data.

Let y_{ijk} be the response from unit k in block j from experiment i. Then, under randomization

$$Y_{ijk(r)} = \mu_i + t_r + \beta_{ij} + \epsilon_{ijk},$$

where $V(\beta_{ij}) = \sigma^2_{bi}$ and $V(\epsilon_{ijk}) = \sigma^2_i$.

Sequential design

Many statisticians recommend sequential design, although it seems to be little used in practice.

Some possible applications:

- Response surface studies, in which a first order design is used, then augmented to get a second order design.
- Nonlinear model fitting, where the optimal design depends on the prior estimate of the parameters, so can be improved as we start to learn about the parameters.
- Clinical trials, where we can stop early for ethical reasons, or adapt the design by dropping inferior treatments.

From the randomization viewpoint, each stage in experimentation is a separate experiment.

Sequential design

Under randomization, the model should be, for example for unit j in stage i with treatment r applied,

$$Y_{ij(r)} = \mu_i + t_r + \epsilon_{ij},$$

where $V(\epsilon_{ij}) = \sigma_i^2$.

Note:

- Pure sequential designs, in which the next treatment is chosen after each unit, have no possible randomization-based analysis.
- Stages with two units require estimation of more parameters than there are units and so should also not be used.
- Very small stages (3 or 4 units?) allow very little randomization and so might lead us to seriously doubt any conclusions drawn.
- The design at each stage should take account of the fact that a combined analysis is going to be performed.

Randomized-not-reset factors

In factorial experiments, the run order is often randomized, but the factors not reset between runs with the same level of a factor.

Say X_1 is not reset in the following design.

Run	X_1	X_2	X_3
1	0	0	1
2	1	1	0
3	1	0	0
4	1	1	1
5	0	0	0
6	0	1	0
7	0	1	1
8	1	0	1

Randomized-not-reset factors

This has been described in the recent literature as *inadvertent split-plotting*, with 4 main plots being defined, consisting of runs (1), (2, 3, 4), (5, 6, 7) and (8). It has been recommended that the corresponding mixed model be used to analyse the data.

However, this is wrong. A split-plot design is defined by a restriction in the randomization. Here the design was completely randomized. Any run order was equally likely to have occurred.

The problem with this design is not that the randomization has been restricted, but that the assumption of additive treatment and unit effects is not believable.

Randomized-not-reset factors

In reality, factor X_1 has six levels, namely low and high levels each set immediately, one run previously or two runs previously.

The design is actually

Run	X_1	X_2	X_3
1	0	0	1
2	1	1	0
3	2	0	0
4	3	1	1
5	0	0	0
6	4	1	0
7	5	1	1
8	1	0	1

Randomized-not-reset factors

Note:

- This is a very poor design for this treatment structure.
- This treatment structure does not correspond to the objectives of the experiment.
- The choice of treatments is defined by the outcome of the randomization. How does this affect the analysis?

Data from randomized-not-reset experiments should be thrown on the scrap-heap and proper multi-stratum designs used instead.

Computer experiments

In many scientific (and other) disciplines, large-scale deterministic models can be evaluated numerically using very expensive computer programmes, e.g. computational fluid dynamics uses numerical solutions to differential equations.

The aim of an "experiment" is to find a relatively simple model which will give a good approximation to the true model over a range of interest of several input variables.

The choice of which inputs to use for each computer run is very much like an experimental design problem.

However, different "experimental units" with the same treatment will always give identical responses.

Hence, the model is $y_{i(r)} = \mu + t_r$ and it is clear that randomization has no place in this setup.

Computer experiments

Various methods are used to model the data, e.g. kriging, which interpolates the responses with a function which is (according to some criterion) as smooth as possible.

Various ideas for designing the experiments have been used, such as:

- Space-filling designs, which attempt to spread the points as evenly as possible in the input space.
- Latin hypercube designs, which break the input space into equally sized hypercubes and pick a point at random from within each.
- Bayesian designs which minimize something like mean squared prediction error.

Final comments

I have emphasised the randomization approach, not as the only one, but as the basis for everything else.

My hopes:

- In consulting, you will use this way of thinking when advising clients how to analyse their data.
- Researchers in statistical modelling will pay attention to the design structure when developing and using their methodology.
- Researchers in the design of experiments will develop methods for planning efficient experiments for these more complex kinds of modelling.
- More people will start doing research in design of experiments!