Completely Randomized Design and Random Effects

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

March 8, 2007

The Big Picture

- The primary goal of experimental designs is to *compare different treatments*.
- *Experimental units* are individuals to which *treatments* are applied.
- Elements of design include:
  - *Replication* — to assess variability.
  - *Randomization* — to control selection bias.
  - *Blocking* — to control known sources of variability.
- In a *completely randomized design*, experimental units from a single homogeneous group are assigned at random to treatments.
- In a *randomized complete block design*, experimental units are grouped with similar units into blocks and then assigned at random to treatments within blocks.

Fixed and Random Effects

- When a blocking variable (or in general, a categorical variable) consists of *all groups of interest*, it is appropriate to model the effects of this variable as *fixed*.
- When a blocking variable (or in general, a categorical variable) is thought of as *a sample of groups from some larger population of possible groups*, it is appropriate to model the effects of this variable as *random*.
- *Random effects models* involve *multiple sources of random variation*.
- We will begin with a review of a fixed effects models and then show how the model changes with random effects.

Data

- A categorical variable has k levels (or treatments).
- There are n_i observations in the ith level for $i = 1, \dots, k$.
- The jth observation in the ith level is y_{ij} for $j = 1, \dots, n_i$.
- The model for the observed data is

$$y_{ij} = \mu_i + e_{ij}, \quad e_{ij} \sim \text{iid} N(0, \sigma_e^2),$$

  where μ_i is the population mean for the ith treatment group, $j = 1, \dots, n_i, i = 1, \dots, k$.
- An alternative expression of the model is

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad e_{ij} \sim \text{iid} N(0, \sigma_e^2),$$

  where μ is the grand population mean $\mu = \frac{1}{k} \sum_{i=1}^{k} \mu_i$ and $\alpha_i = \mu_i - \mu$ is the difference between the ith trt mean and the grand mean.
- Note that $\sum_{i=1}^{k} \alpha_i = 0$.
- This parameterization is not the default in R.

Balanced Designs

- In a *balanced design*, sample sizes are equal in each treatment group ($n_i = n$ for all i).
- Equations associated with ANOVA F tests are simpler in balanced designs.
- The hypothesis of equal treatment means is equivalent to the hypothesis that all treatment effects are zero.

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k \text{ versus } H_a\colon \text{not all } \mu_i\text{'s are equal}$$

$$H_0\colon \alpha_i = 0 \text{ for all } i \text{ versus } H_a\colon \text{not all } \alpha_i = 0$$

- Sums of squares have these formula:
  - $\mathrm{SSTot} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$ on df $= kn - 1$.
  - $\mathrm{SSTrt} = \sum_{i=1}^{k} n(\bar{y}_{i.} - \bar{y}_{..})^2$ on df $= k - 1$.
  - $\mathrm{SSErr} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$ on df $= k(n - 1)$.

  where $\bar{y}_{..}$ is the grand mean and $\bar{y}_{i.}$ is the ith group mean.

ANOVA for Balanced Designs

- The ANOVA table for a balanced design is:

Source	df	SS	MS	F
Trt	$k-1$	SSTrt	MSTrt	MSTrt/MSErr
Error	$k(n-1)$	SSErr	MSErr	–
Total	$kn-1$	SSTot	–	–

- Under model assumptions, the F statistic has an F distribution with $k - 1$ and $k(n - 1)$ degrees of freedom.
- The estimated variance is $\hat{\sigma}_e^2 = \mathrm{MSErr}$.

Expected Mean Square (EMS)

- Both MSTrt and MSErr are random quantities.
- Each has a distribution and an expected value.
- Facts for balanced designs:

$$E(\mathrm{MSErr}) = \sigma_e^2$$

$$E(\mathrm{MSTrt}) = \sigma_e^2 + \frac{n\sum_{i=1}^k \alpha_i^2}{k-1}.$$

- It follows for balanced designs that

$$\frac{E(\mathrm{MSTrt})}{E(\mathrm{MSErr})} = 1 + \left(\frac{1}{\sigma_e^2} \times \frac{n\sum_{i=1}^k \alpha_i^2}{k-1}\right).$$

- For unbalanced designs, there is a messier formula, but the same basic principle holds.
- When treatment effects (α_i) are not all zero, the ratio of expected values is greater than one.
- The F-test is significant when the ratio is large enough.

Randomization

- Randomization provides a way to control potential selection bias.
- If unknown factors affect the response variable, with randomization these factors are likely to be fairly balanced by the treatment allocation.
- If known factors affect the response, these factors should be measured and included in the model or in the design (with blocking).

Randomization in R

- The `sample` function in R can be used to allocate individuals to treatment groups at random.
- Here is an example to allocate 20 individuals into treatment groups A, B, C, and D equally.
- The function `rep` repeats the first argument some number of times.
- The function `sample` with only one argument places the elements of the argument in a random order.

```
> trt = rep(c("A", "B", "C", "D"), each = 5)
> trt

 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C" "D" "D" "D" "D"
[20] "D"

> trt = sample(trt)
> trt

 [1] "C" "C" "A" "B" "B" "A" "B" "A" "D" "C" "A" "C" "D" "D" "B" "C" "D" "D" "B"
[20] "A"
```


Motivating Example

- *Example 1*:
  - Compare the average reading skill of students in 8 *specific* second grade classes in Madison.
  - Take a random sample of 10 students from each class and measure their reading skills.
- *Example 2:*
  - Of interest is the reading skill among *all* second grade classes in Madison.
  - Take a random sample of 8 classes.
  - Within each class, take a random sample of 10 students and measure their reading skills.
- Note the difference in the objectives.
  - In Example 1, we want to compare 8 specific classes and hence the class effect is *fixed*.
  - In Example 2, we want to assess the variabilities among classes and use a random sample of the classes to make inference about all the classes in Madison. Hence the class effect is *random*.

Models

- *Fixed effect model*:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad e_{ij} \sim \mathrm{iid} N(0, \sigma_e^2)$$

  where $\sum_{i=1}^{k} \alpha_i = 0$.
- *Random effect model:*

$$y_{ij} = \mu + a_i + e_{ij}, \quad j = 1, \dots, n_i, i = 1, \dots, k,$$

  where a_i is the class effect with

$$a_i \sim \mathrm{iid} N(0, \sigma_a^2),$$

  and e_{ij} is the error due to variability from student to student with

$$e_{ij} \sim \mathrm{iid} N(0, \sigma_e^2),$$

  and the $\{a_i\}$ are independent of the $\{e_{ij}\}$.
- Often σ_a^2 and σ_e^2 are called *variance components*.