Completely Randomized Designs

Gary W. Oehlert

School of Statistics
University of Minnesota

January 18, 2016

Definition

A completely randomized design (CRD) has

- N units
- g different treatments
- g known treatment group sizes $n_1, n_2, \ldots, n_g$ with $\sum n_i = N$
- Completely random assignment of treatments to units

Completely random assignment means that every possible grouping of units into g groups with the given sample sizes is equally likely.

This is the basic experimental design; everything else is a modification.[1]

The CRD is

- Easiest to do.
- Easiest to analyze.
- Most resilient when things go wrong.
- Often sufficient.

Consider a CRD first when designing.

[1] "God invented the integers, the rest is the work of man." Leopold Kronecker

Examples

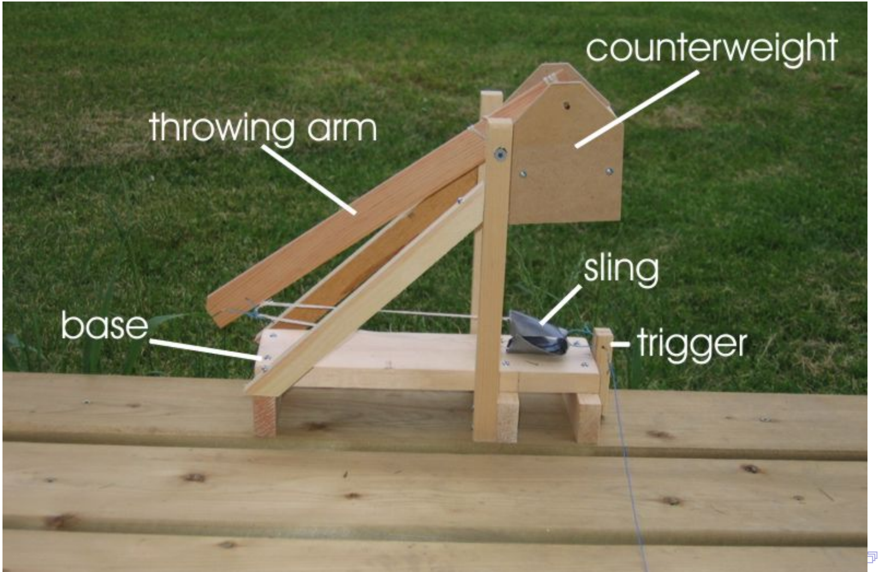
1. Does a wood board .625 inches thick have the same strength as a .75 inch thick wood board with a notch cut to .625 thickness? Twenty-six 2.5” by .75” by 3 foot boards. Half are chosen at random to be notched in the center. Response is load at failure in horizontal bending.

2. Do the efflux transporters P-gp and/or BCRP affect the ability of a certain chemotherapy drug to cross the blood-brain barrier. We will make 30 in-vitro measurements of chemo accumulation in cells. Ten will be done with wild type cells, 10 with cells that over-express P-gp, and 10 with cells that over-express BCRP. The efflux transporters (or not) are randomly assigned to the trials.

3. Do xantham gum and/or cinnamon affect the sensory quality of gluten-free cookies? Eight batches of cookies will be made, with two of the eight batches assigned to each of the four combinations of low/high gum and low/high cinnamon. The response is a sensory score.

4. How do sling length and size of counterweight affect the throw distance of a trebuchet? Randomly assign 27 throws to the nine combinations of three lengths and three weights, with three throws per combination. The response is the distance of the projectile.

Experiment like this:



Build like this?

Inference

Most of our inference is about treatment means:

- Any evidence means are not all the same?
- Which ones differ?
- Any pattern in differences?
- How can differences be described succinctly?
- Estimates/confidence intervals of means and differences.

Variability and other aspects may be of interest in specific cases.

Models

We seek the simplest model consistent with the data.

“All treatments have the same mean” is simpler than “Each treatment has its own mean.” If we cannot say that the complicated model is needed, we take the simple model.

Sometimes we seek a more explanatory model.

“Treatment means vary linearly with temperature” is simpler than “Each treatment has its own mean” or even “Treatment means vary quadratically with temperature.” An explanatory model (especially a simple one) helps us understand the data.

> All models are wrong; some models are useful. — George Box

We might not believe that the simple model can be completely true in some infinitely precise sense, but if the data are consistent with it, we use it.

Comparing models

We gauge model fit by looking at the sum of squared residuals.

We usually choose model parameters so as to minimize the sum of squared residuals.

The total sum of squares in the data SS_T is the sum of the model or explained sum of squares SS_M plus the error or residual sum of squares SS_E. For a fixed set of data, if you change the model making one SS bigger, then the other must get smaller.

$$SS_T = SS_M + SS_E$$

“All treatment means are the same” is a special case of “Each treatment has its own mean.” “Treatment means vary linearly with temperature” is a special case of “Treatment means vary quadratically with temperature” and, indeed, of “Each treatment has its own mean” as well.

We say that the special case model is included in the more complicated model, or perhaps that it is a restriction of (a restricted version of) the more complicated model.

We sometimes say that the special case model is nested in the more complicated model, but we will also use the descriptor “nested” in a different way later, so beware.

When we have model A included in model B, then:

1. Model B (fit by LS) always fits at least as well as model A (fit by LS), and usually fits better.
2. The error sum of squares from model B cannot be larger than the error sum of squares from model A, and is usually smaller.
3. Equivalently, the model SS for model B is always at least as large and usually larger than the model SS for model A.
4. The reduction in error SS going from A to B is the same as the increase in model SS going from A to B.

ANOVA

The partitioning of the sums of squares is called Analysis of Variance, or ANOVA.

The special case model never fits as well as the larger model, but how do we decide that it is good enough, that is, is consistent with the data?

The two basic approaches are:

- Significance testing
- Information Criteria

Significance testing

We will make an ANOVA table that has a row for the restricted model, a row for the increment from the restricted model to the larger model, and a row for all of the residual bits.

Each row in the table has a label, a sum of squares, a "degrees of freedom," and a "Mean square."

Degrees of freedom count free parameters. If there are r_1 parameters in the mean structure of the included model, and r_2 parameters in the mean structure of the larger model, then there are $r_2 - r_1$ parameters in the improvement from the small model to the large model, and $N - r_2$ parameters for residuals (error).

An MS is SS divided by DF.

The generic table looks like this (SS_1 is model SS for restricted model, and SS_2 is model SS for the large model):

Source	SS	DF	MS
Model 1	SS_1	r_1	SS_1/r_1
Improvement from Model 1 to Model 2	$SS_2 - SS_1$	$r_2 - r_1$	$(SS_2 - SS_1)/(r_2 - r_1)$
Error	SS_E	$N - r_2$	$SS_E/(N - r_2)$

Notation

There are simple formulae for elements of the ANOVA table for many designed experiments.

Let y_{ij} be the jth response in treatment i. $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, n_i$.

Let

$$\bar{y}_{i\bullet} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

be the mean response in the ith treatment, and let

$$\bar{y}_{\bullet\bullet} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}}{N}$$

be the grand mean response.

Suppose that the restricted model is the model that all treatments have the same mean, and the larger model is the model that each treatment has its own mean. Then:

$r_1 = 1$

$r_2 = g$

$SS_1 = N\overline{y}_{\bullet\bullet}{}^2$

$SS_2 = \sum_{i=1}^{g} n_i \overline{y}_{i\bullet}{}^2$

$SS_2 - SS_1 = \sum_{i=1}^{g} n_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$

$SS_E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2$

and the ANOVA table is . . .

Basic ANOVA

The first four columns of the ANOVA table are:

Source	SS	DF	MS
Overall mean	$N\overline{y}_{\bullet\bullet}{}^2$	1	
Between Treatments	$\sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$	$g-1$	$SS_{Trt}/(g-1)$
Error	$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2$	$N-g$	$SS_E/(N-g)$

and the MS may be denoted MS_E and MS_{Trt}.

In fact, the line for the overall mean is so boring that it is usually left off.

Digression on Pythagorean Theorem

Note that

$$y_{ij} = \overline{y}_{\bullet\bullet} + (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) + (y_{ij} - \overline{y}_{i\bullet})$$

Square both sides and add over all i and j and we get

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 = N\overline{y}_{\bullet\bullet}{}^2 + \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2$$

plus a lot of sums of cross products. All those sums of cross products add to zero (the three components of y_{ij} are perpendicular out in N-dimensional geometry so sums of squares add up).

Probability model

The ANOVA is just algebra, albeit algebra with statistical intent. We need a probability model.

Assume that $y_{ij} \sim N(\mu_i, \sigma^2)$. Then,

$$E(MS_E) = \sigma^2$$

and **if the restricted model is true** we also have

$$E(MS_{Trt}) = \sigma^2$$

If the restricted model is not good enough its expectation is larger than σ^2. This means that

$$F = MS_{Trt}/MS_E$$

is a test statistic for comparing the restricted model to the full model; we reject the null if F is too big.

When the null is true and the normal distribution assumptions are correct, the F-test follows an F-distribution with $g-1$ and $N-g$ df (note df from numerator and denominator MS). Reject the null that the single mean model is true when the p-value for the F-test is too small.

We did the algebra for the single mean model and individual mean model, but the F test is appropriate for general restricted models versus a containing model. It's just that the computations are not always so clean.

Resin example in R.

Information criteria

Akaike introduced the first information criterion, AIC.

Later Bayesians added a second one, BIC.

Now there are several more.

Information criteria include a measure of how well the data fit the model (smaller being better) plus a penalty for using additional parameters.

Models with smaller values of AIC or BIC are better models.

Let L be the maximized likelihood for the data. This is the “probability” of the data under the model, with the parameters chosen to make the probability as high as possible. This likelihood model has k parameters that we can choose. Typically these parameters are things like treatment means, or regression coefficients, or residual variances.

We’ll say a lot more later, but for now suffice it to say that big L is good.

$$AIC = -2\ln(L) + 2k$$

$$BIC = -2\ln(L) + \ln(N)k$$

Choose a model with smaller AIC (or BIC).

In general, AIC tends to choose models with more parameters than we get from significance testing, i.e., some things in the selected model might be “insignificant.” The reverse tends to be true for BIC, especially for big data sets.

Except for very small data sets, BIC penalizes additional parameters more than AIC. BIC thus tends to choose smaller models than AIC.

AIC tends to work better when all candidate models are approximate; BIC tends to work better in large samples when one of the candidate models is really the right model.

Resin example in R, continued.

Parameters

You have an apartment in SE Minneapolis. You can locate it by

- Latitude and longitude;
- Street address (note, streets in SE are not oriented NS/EW, so this is different than lat/long);
- Walking directions from here;
- Distance and direction from here.

Four completely separate ways to identify the same place. In fact, walking directions are not even unique!

Mean parameters suffer the same issue: there are many ways to describe/parameterize the same set of means. Sometimes one is better than another in a particular context. Sometimes one is more understandable than another.

It is an embarrassment of riches, but as long as the parameters describe the same means, we are OK.

They can all be different yet still correct, but you need to know which ones you're working with.

Consider the resin example.

Trt ($^o C$)	175	194	213	231	250	All data
Average	1.933	1.629	1.378	1.194	1.057	1.465
Count	8	8	8	7	6	37

If we have a single mean model, the only parameter is the overall mean μ. Our estimate would be $\widehat{\mu} = \overline{y}_{\bullet\bullet} = 1.465$.

In the separate means model, parameters are the group means, and the estimates would be $\widehat{\mu}_1 = \overline{y}_{1\bullet} = 1.933$ and so on.

Sometimes we want to write

$$\mu_i = \mu + \alpha_i$$

Where μ is some kind of “central value” and α_i is a treatment effect.

We always have $\alpha_i = \mu_i - \mu$ and $\widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu}$, but how do we define μ?

Like the walking instructions, there are many, many ways, but there are three semi-standard ways.

Define μ	Equivalent constraint
$\mu = \mu_1$	$\alpha_1 = 0$
$\mu = \frac{\sum_i \mu_i}{g}$	$\sum_i \alpha_i = 0$
$\mu = \frac{\sum_i n_i \mu_i}{N}$	$\sum_i n_i \alpha_i = 0$

The first is the default in R, I find the second more interpretable, and the third is useful in hand calculations.

The important things ($\mu_i - \mu_j = \alpha_i - \alpha_j$) are the same in all versions.

Care about μ in the single mean model; care about μ_i and $\alpha_i - \alpha_j$ in the separate means model.

What about polynomial models? Let z_i be the temperature treatment for group i. Here are some models

$$\begin{aligned}
\mu_i &= \beta_0 \\
\mu_i &= \beta_0 + \beta_1 z_i \\
\mu_i &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 \\
\mu_i &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 \\
\mu_i &= \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 + \beta_4 z_i^4
\end{aligned}$$

The first is the same as the single mean model, the last fits the same means as the separate means model, and the others are intermediate.

Note that equivalently written parameters have different meanings (and different values) in different models.

Note that we maintain hierarchy.

But we don't even leave polynomials in peace. Consider

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1[z_i - 210.0811] \\ &\quad + \beta_2[z_i^2 - 422.9z_i + 44043.5] \\ &\quad + \beta_3[z_i^3 - 636.4z_i^2 + 133812.3z_i - 9294576.3]\end{aligned}$$

This is equivalent to the cubic model on the last slide, but here the β_i retain values and meanings as we change linear to quadratic to cubic (and you can go higher). These are *orthogonal polynomials*.

The moral of the story is that

- Parameters are tricksy and can often be defined in many ways within a single mean structure.
- We usually only use parameters as a means to an end.
- Most parameters are arbitrary, so inference on parameters (as opposed to model comparison or comparison of means) is also somewhat arbitrary.

R will compute the estimates as well as standard errors for various parameterizations, polynomials, orthogonal polynomials, trigonometric series, and so on. They are done correctly, but they retain the arbitrariness of their definition.

Back to resin.