EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

Power analysis

Recall last time we discussed some do's and don't's about performing and reporting a power analysis. Here are some varying examples from the literature:

- Sinaki et al. (2002)
- Rogers et al. (2018)
- Rogers et al. (2019)
- Mirenda et al. (r&r)

Optimizing resources/design when sample units are scarce

Small sample sizes are often a hard reality of a lot of experimental/clinical work in the social and health sciences. But there are many ways to maximize the quality of information you get from even small samples:

- Always aim for *balanced* group sizes (recruit more people than you need, usually, due to expected drop-out and/or lack of fidelity).
- Do NOT randomly sample; i.e., have clear inclusion/exclusion criteria to:
  (1) maximize interesting differences between units (e.g., treatment effects across groups of differential interest) and
  (2) minimize nuisance differences between units (confounding effects).
- Restrict randomization (e.g., blocking) in the design and allocation to treatment phase of your study to accomplish these goals.
- Consider nonparametric analyses if group sample sizes are very small and/or response data are very non-normal within subgroups.

Unbalanced ANOVA

A study design or analysis is called *unbalanced* when sample sizes are not equal across all identified groups/subgroups (i.e. across all factor levels of the categorical explanatory variable(s)). Unbalanced designs suffer from several problems:

- Harder to perform/assess model diagnostics.
- Harder to estimate within and between subject variability.
- *Lower power* to detect non-zero effects than balanced designs (usually, power is a function of the smallest group sample size).
- Lack of balance may induce a *confounding* effect (see following examples)

The more unbalanced the groups, the worse these problems become.

Unbalanced ANOVA

Iron concentration in the blood dependent on blood type and medication:

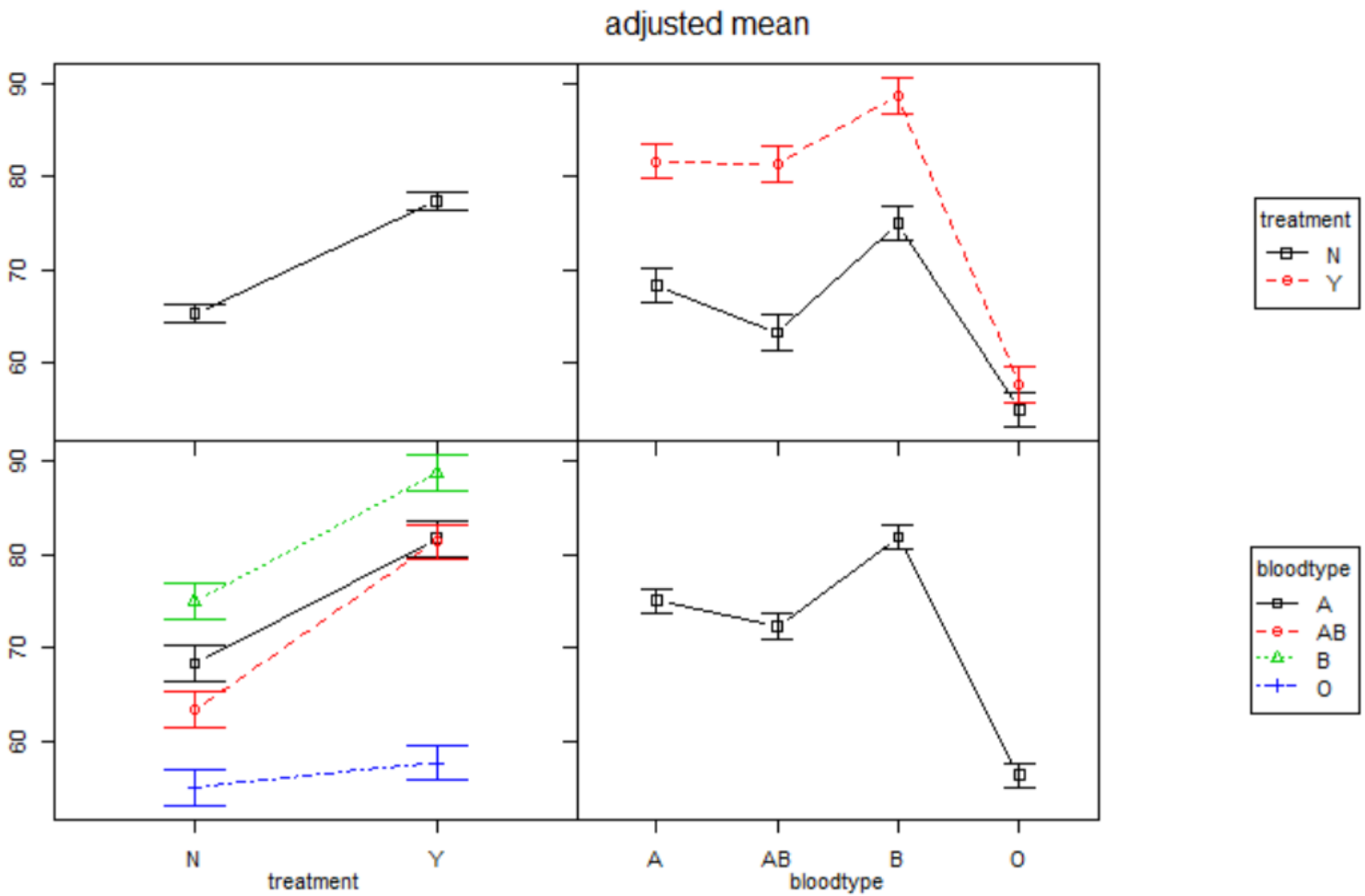
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
treatment	1	852.0	852.0	79.260	1.35e-07	***
bloodtype	3	2098.1	699.4	65.058	3.50e-09	***
treatment:bloodtype	3	191.5	63.8	5.937	0.00639	**
Residuals	16	172.0	10.7			

This is a *fully balanced, 4×2 (blood type × treatment) factorial* design with 3 people in each cross-group:

- total sample size $= 24$
- 4 blood type levels, sample sizes $= 24/4 = 6$
- 2 treatment levels, sample sizes $= 24/2 = 12$
- $4 \times 2 = 8$ blood type $\times$ treatment levels, sample sizes $= 24/8 = 3$

Unbalanced ANOVA

Iron concentration in the blood dependent on bloodtype and medication:



Unbalanced ANOVA

But suppose one of our O bloodtype participants in the placebo group drops-out of the study, so that now:

- total sample size = 23
- 4 blood type levels, sample sizes = 6,6,6,5
- 2 treatment levels, sample sizes = 12,11
- $4 \times 2 = 8$ blood type $\times$ treatment levels, sample sizes = 3,3,3,3,3,2

This is now an *unbalanced* design with considerably less power to detect the interaction:

```
                            Df Sum Sq Mean Sq F value   Pr(>F)
treatment[-24]               1  713.7   713.7  64.491 8.22e-07 ***
bloodtype[-24]               3 2076.3   692.1  62.539 1.04e-08 ***
treatment[-24]:bloodtype[-24]  3  142.0    47.3   4.278   0.0227 *
Residuals                   15  166.0    11.1
---
```

Unbalanced ANOVA

The problem gets worse if two O bloodtype participants in the placebo group drop-out::

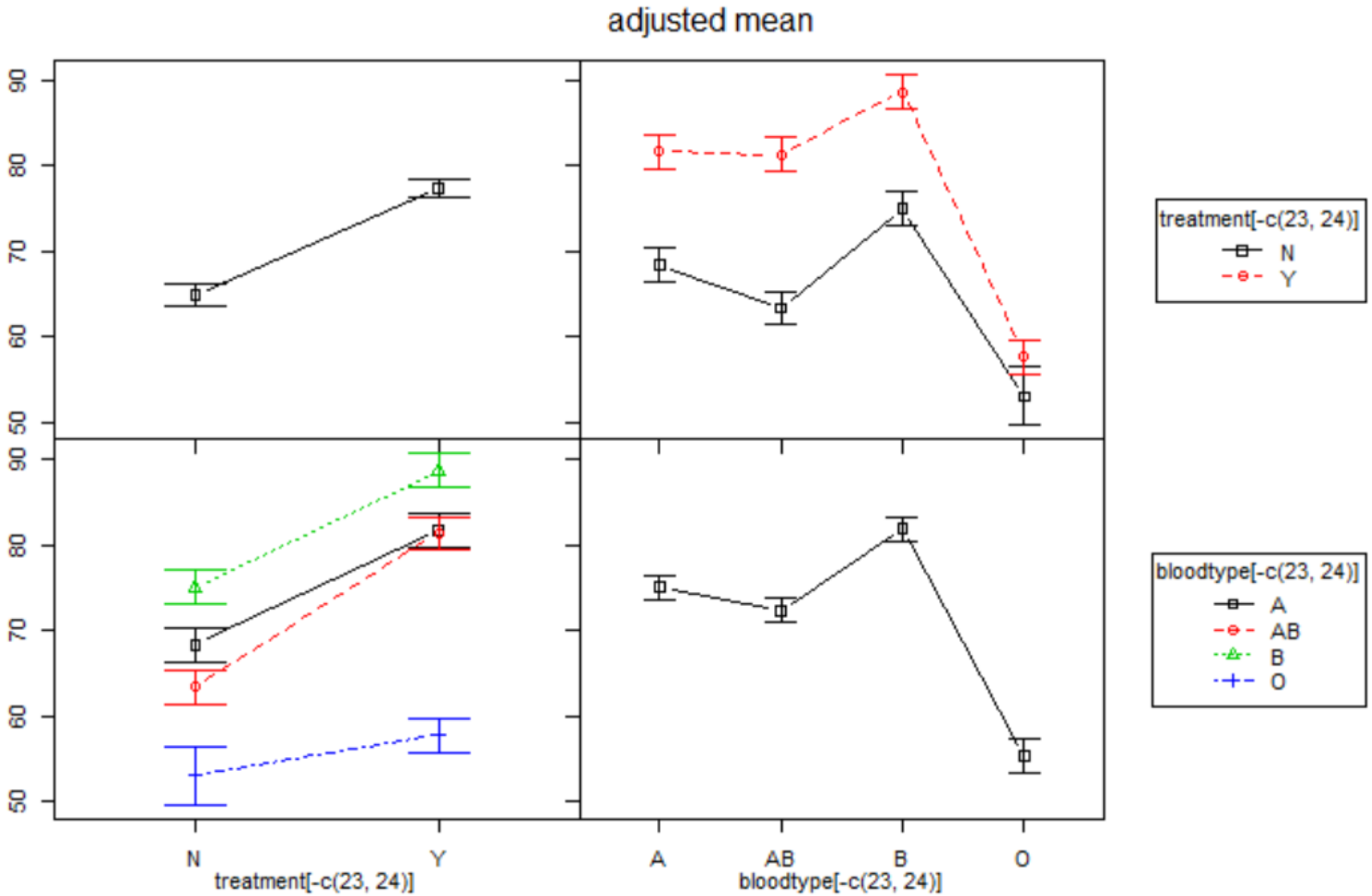
- total sample size = 22
- 4 blood type levels, sample sizes = 6,6,6,4
- 2 treatment levels, sample sizes = 12,10
- $4 \times 2 = 8$ blood type $\times$ treatment levels, sample sizes = 3,3,3,3,3,1

This is now an *unbalanced* design with even less power to detect the interaction:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment[-c(23, 24)]	1	549.1	549.1	46.874	7.98e-06
bloodtype[-c(23, 24)]	3	1993.8	664.6	56.734	4.45e-08
treatment[-c(23, 24)]:bloodtype[-c(23, 24)]	3	89.0	29.7	2.532	0.0992
Residuals	14	164.0	11.7		

Unbalanced ANOVA

Iron concentration in the blood dependent on bloodtype and medication with attrition in one cross-group:



Complete randomization is not always a good thing

- 27 participants recruited with different baseline levels of social media usage: low, medium, or high (9 from each group)
- These 27 participants then randomly assigned to one of three treatment groups that set their social media usage to low, medium, or high for 1 week; then measure anxiety outcomes.
- Do not assume any restrictions on randomization
- So we *could* get unlucky with our randomization and gotten a study design like this:

	Social media assignment		
	L	M	H
Social media baseline use	L	M	H
	L	M	H
	L	M	H
	⋮	⋮	⋮

Complete randomization is not always a good thing

But this would be a terrible design!

- Effect of treatment (our main interest) cannot be separated from baseline effect (confounding)

	Social media assignment		
	L	M	H
Social media baseline use	L	M	H
	L	M	H
	L	M	H
	⋮	⋮	⋮

Instead, we should be able to design a better study by *restricting* the random assignment mechanism carefully.

Restricted randomization and blocking

If you are designing an *experiment*, you should be smart about how you assign your experimental treatments. You want to:

- Maximize information about the treatment effect
- Minimize confounding with other variables
- Ensure no sample unit is going to waste (i.e. maximize power)

Remember:

- Experimental manipulation is the *only* sure way to tease out causal relationships between variables
- Experiments are costly (money and time)

If you are fortunate enough to be running an experiment, you should pick a design that is efficient and effective.

Restricted randomization and blocking

Consider the following example: we have money to run a study to test the effects of four pain-relieving drugs on first-time liver cancer patients who have undergone 2 months of radiation therapy. Patients come from one of four doctors, all with comparable experience. Response of interest is a pain-index compiled from a suite of quantitative and qualitative patient outcomes.

- We only have money for 16 sample units
- 4 doctors $\times$ 4 drug treatments
- So we do *not* have enough data to estimate an interaction effect (3 df + 3 df + 9 df would mean 0 df leftover for residuals!)
- Thus, the only two-way model we can estimate is:

$$Y = \mu + \tau_{doctor} + \tau_{drug} + \varepsilon$$

Restricted randomization and blocking

We (naively) randomize drug assignment (4×4) and get the following design:

	Doctor			
	I	II	III	IV
Drug treatment	A	B	C	D
	A	B	C	D
	A	B	C	D
	A	B	C	D

- This study design would *completely confound* attending doctor with drug treatment. No way to separate effect of drug from baseline effects of attending doctor!

Restricted randomization and blocking

- But that was a very special (and very unlucky) case. We could randomize treatment assignment again and find:

	Doctor			
	I	II	III	IV
Drug treatment	C	A	C	A
	A	A	D	D
	D	B	B	B
	D	C	B	C

- Now we run the experiment:

Restricted randomization and blocking

Response: pain-index outcomes on a 1-20 point composite scale

	Doctor			
	I	II	III	IV
Drug treatment	C(12)	A(14)	C(10)	A(13)
	A(17)	A(13)	D(11)	D(9)
	D(13)	B(14)	B(14)	B(8)
	D(11)	C(12)	B(13)	C(9)

Output of ANOVA:

```
          Df Sum Sq Mean Sq F value Pr(>F)
Drug       3  30.69  10.229   5.174 0.0238 *
Doctor     3  32.46  10.819   5.472 0.0204 *
Residuals  9  17.79   1.977
```

Restricted randomization and blocking

However, the previous design was very inefficient:

- Drug A was never used by Doctor III
- Drug D was never used by Doctor II
- Drug B was never used by Doctor I
- Variation in Drug A may be disproportionally affected by a Doctor II effect (confounding)
- Similar for Drug D and Doctor I, and Drug B and Doctor III (confounding)

A much better experimental design would *remove* this possible confounding by restricting the random drug assignment within each doctor. This process is called *blocking* and the doctors are then called *experimental blocks*.

Restricted randomization and blocking

Randomized block design for pain-relieving drug experiment:

	Doctor			
	I	II	III	IV
Drug treatment	B(14)	D(11)	A(13)	C(9)
	C(12)	C(12)	B(13)	D(9)
	A(17)	B(14)	D(11)	B(8)
	D(13)	A(14)	C(10)	A(13)

Notice how this design maximizes experimental efficiency:

- Each drug is given the same number of times (once) by each doctor
- All doctors (blocks) receive all treatments
- No confounding between drug and doctor effects; the SSs capture *only* the marginal variations in the effects

Restricted randomization and blocking

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Drug	3	30.69	10.229	7.962	0.00668	**
Doctor	3	38.69	12.896	10.038	0.00313	**
Residuals	9	11.56	1.285			

Looking at the ANOVA output:

- The SSs are accurate (unconfounded) estimates of marginal effects
- Residual variation has been reduced since all data now efficiently measure drug and doctor effects (no confounding)
- Power to detect non-zero effects has increased due to more efficient design

Restricted randomization and blocking (Latin squares)

There are still some potential inefficencies in our randomized block design if we have extra information on patients we would like to account for:

	Doctor			
	I	II	III	IV
Drug treatment	B	D	A	C
	C	C	B	D
	A	B	D	B
	D	A	C	A

Suppose that patients in row 1 have the least aggressive cancers, while patients in row 4 have the most aggressive cancers (rows 2 and 3 contain patients with moderately aggressive cancers); assume it makes clinical sense to categorize "severity" this way.

- Now "severity of cancer" is a potential confounding variable
- But no patients from the high severity group ever receive Drug B.

Restricted randomization and blocking (Latin squares)

To eliminate possible confounding due to severity of cancer, we can block again; i.e. *block over Doctors and block over Severities*

	Doctor			
	I	II	III	IV
Severity 1	C(12)	D(11)	A(13)	B(8)
Severity 2	B(14)	C(12)	D(11)	A(13)
Severity 3	A(17)	B(14)	C(10)	D(9)
Severity 4	D(13)	A(14)	B(13)	C(9)

- Now, each treatment appears once and only once *in each row and in each column*
- This experimental design is called a *Latin square* or *orthogonal array*
- Interestingly, *there is still randomization here*; i.e. there are many different ways to construct Latin squares of various dimensions (just how many is a famous open problem in theoretical mathematics)

Restricted randomization and blocking (Latin squares)

There are 576 different Latin squares of order 4 (i.e. 4 treatments $\times$ 4 doctors $\times$ 4 severities). For example:

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

A	B	C	D
C	D	A	B
D	C	B	A
B	A	D	C

C	D	A	B
B	C	D	A
A	B	C	D
D	A	B	C

Restricted randomization and blocking (Latin squares)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Drug	3	30.69	10.229	11.419	0.00683	**
Doctor	3	38.69	12.896	14.395	0.00378	**
Severity	3	6.19	2.062	2.302	0.17695	
Residuals	6	5.37	0.896			

- The SSs are still accurate for Doctor and Drug because our design still separates (unconfounds) those effects from drug assignment
- Moreover, we have eliminated any potential confounding due to Severity with our design; so all SSs are *unconfounded*
- Residual variation has been further reduced
- Power hasn't changed much (but that's okay)

Restricted randomization and blocking (Latin squares)

But there's no need to stop at 3 effects!

- Maybe the patients are coming from one of four different Locations. This could create a 4 Drug $\times$ 4 Doctor $\times$ 4 Severity $\times$ 4 Location blocking experiment.
- Such a design is called a *Graeco-Latin square*.
- There are also similar designs for *unbalanced* or *incomplete* designs (say, if we were only testing 3 Drugs by 4 Doctors over 4 Severities); this is called a *Youden square*.
- And lots, lots more!

Moral: even if you can only afford a very small sample, you can still design very efficient experiments. Seek out professional advice if unsure of the options.