EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

March 5, 2020

Statistical power

- The concept of *statistical power* is crucial for both designing a study and for interpreting a study that has already been conducted.
- *Power* is (informally) defined as the ability to detect non-zero effects (true positives)
- The *power*, or *sensitivity*, of a test is defined as

  $$\Pr(p - value < \alpha \mid H_0 \text{ false}) = 1 - \beta,$$

  where α is the *significance level* set by the researcher/journal and used to declare p-values "significant" or not under the traditional threshold approach.
- Good studies will strive to have $1 - \beta \geqslant 0.80$. Most studies will have much lower power.

Statistical power

	H_0 true	H_0 false
data inconsistent with H_0	Type I error *false positive*	Correct decision *true positive*
data consistent with H_0	Correct decision *true negative*	Type II error *false negative*

	Given H_0 true	Given H_0 false
Pr(data inconsistent with $H_0 \mid \cdots$)	α	$(1-\beta)$
Pr(data consistent with $H_0 \mid \cdots$)	$(1-\alpha)$	β

Statistical power

- Statistical power is a function of many things:
  - Sample size (increasing sample size automatically increases power)
  - Population variability (less variation means more power)
  - Overall distribution of random phenomenon of interest (average effects in clustered or multi-modal distributions can be difficult to detect)
  - Type I error rate, α (increasing α automatically increases power)
  - *True, unobserved effect size* (bigger effect sizes are easier to find)
  - Type of statistical test/procedure used (e.g. nonparametric or robust procedures can be more powerful when data are non-normal)
  - Measurement error (noisier measurements produce more variability, so lead to less power)

Statistical power

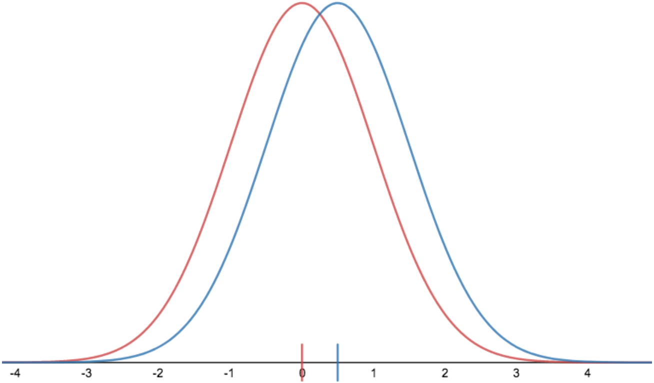
Remember:

- If you design a study that has a poor chance of detecting what you are trying to find, then why bother doing the study at all?
- If your study has low power, but you end up finding a significant non-zero effect anyway, *it is likely that you are making a type I error*.
- If your study has low power but you end up finding a significant non-zero effect anyway, your *effect estimates will be overinflated*, sometimes massively (Type M error).
- If your study has low power but you end up finding a significant non-zero effect anyway, your *effect estimates are likely to be in the wrong direction* (Type S error).

Examples of study situations with different powers

Note: “small” and “large” are *relative* terms

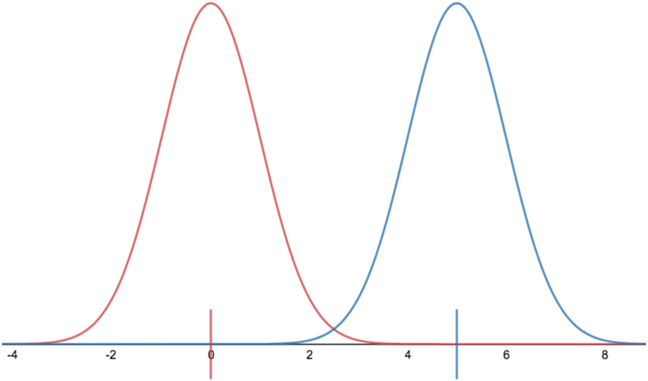
- Low power
- Small true effect size, or small sample size, and/or large pop. variance
- Very hard to distinguish the null distribution (red) from reality (blue)



Examples of study situations with different powers

Note: “small” and “large” are *relative* terms

- High power
- Big true effect size, or large sample size, and/or small pop. variance
- Able to distinguish the null distribution (red) from reality (blue)



Effects of low power on interpretation of analytical output

Low power can come from many different sources. In practice, the three most common are:

- Small sample sizes (overall, or within groups).
- Large variability (overall, or within groups, or due to noisy measurements).
- Small *true* effect sizes.

The first two sources are easy to see. The last (small true effect sizes) is difficult and subjective, but absolutely crucial.

Effects of low power on interpretation of analytical output

True effect sizes are *unobserved*, but crucial to interpretation:

- We never actually know the *true* effect size (if we did, we wouldn't have to perform a study to estimate it).
- A plausible true effect size depends on the *prior believability of a particular alternative hypothesis*.
- In social science, many of our effects of interest will be small, *especially when compared to the effects of other variables of little or no interest*.
- *Evaluating the power of a study retrospectively requires an informed assessment of how plausible you would find certain effect sizes.*
- **Note:** some applied practitioners and software (e.g. SPSS) will talk about "retrospective power" or "post hoc power analysis"; they do *not* mean what we are talking about (usually, they mean gibberish).

Unbalanced ANOVA

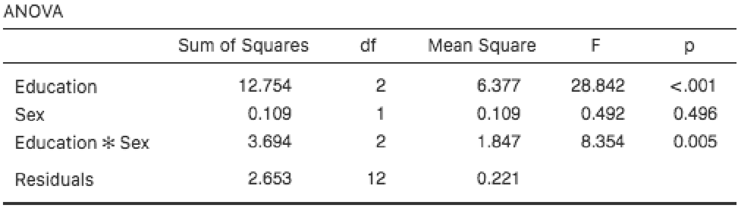
A study design or analysis is called *unbalanced* when sample sizes are not equal across all identified groups/subgroups (i.e. across all factor levels of the categorical explanatory variable(s)). Unbalanced designs suffer from several problems:

- 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

Recall Anxiety vs. Education and Sex two-way ANOVA:



ANOVA

	Sum of Squares	df	Mean Square	F	p
Education	12.754	2	6.377	28.842	<.001
Sex	0.109	1	0.109	0.492	0.496
Education ✻ Sex	3.694	2	1.847	8.354	0.005
Residuals	2.653	12	0.221		

This was a *fully balanced, 3×2 factorial* design:

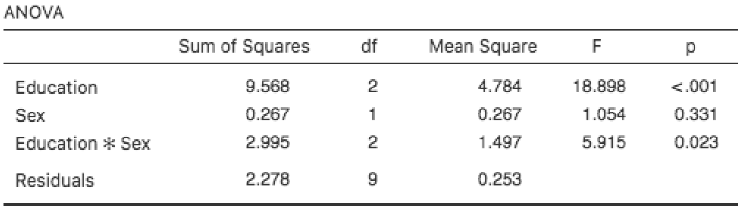
- total sample size $= 18$
- 3 Education levels, sample sizes $= 18/3 = 6$
- 2 Sex levels, sample sizes $= 18/2 = 9$
- $3 \times 2 = 6$ Education×Sex levels, sample sizes $= 18/6 = 3$

Unbalanced ANOVA

But suppose three of our male respondents refused to answer (maybe because they were too stressed out), so that now:

- total sample size $= 15$
- 3 Education levels, sample sizes $= 6{,}5{,}4$
- 2 Sex levels, sample sizes $= 9{,}6$
- $3 \times 2 = 6$ Education×Sex levels, sample sizes $= 3{,}3{,}3{,}2{,}3{,}1$

This is now an *unbalanced* design:



ANOVA

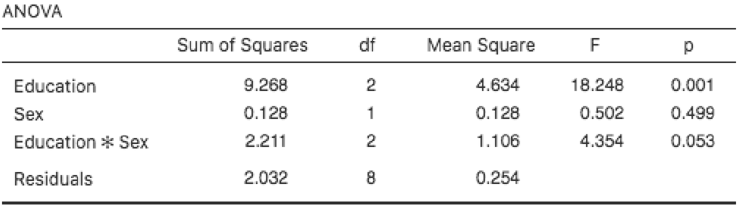
	Sum of Squares	df	Mean Square	F	p
Education	9.568	2	4.784	18.898	<.001
Sex	0.267	1	0.267	1.054	0.331
Education ✻ Sex	2.995	2	1.497	5.915	0.023
Residuals	2.278	9	0.253		

Unbalanced ANOVA

Alternatively, suppose two of our Master's and 2 of our PhD respondents refused to answer (maybe because they were too stressed out), so that now:

- total sample size = 14
- 3 Education levels, sample sizes = 6,4,4
- 2 Sex levels, sample sizes = 7,7
- 3×2 = 6 Education×Sex levels, sample sizes = 3,3,2,2,2,2

This is now an *unbalanced* design:



ANOVA

	Sum of Squares	df	Mean Square	F	p
Education	9.268	2	4.634	18.248	0.001
Sex	0.128	1	0.128	0.502	0.499
Education ✻ Sex	2.211	2	1.106	4.354	0.053
Residuals	2.032	8	0.254		

Complete randomization is not always a good thing

Recall HW1, Q1:

- 27 participants randomly assigned to one of three treatment groups: low, medium, or high social media usage (9 from each baseline)
- Did not assume any restrictions on randomization
- So we *could* have gotten 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 hospitals, but all facilities and therapy regiments are comparable. 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 hospitals × 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:

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

Restricted randomization and blocking

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

	Hospital			
	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* patient group with drug treatment. No way to separate effect of drug from baseline effects of patient group!

Restricted randomization and blocking

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

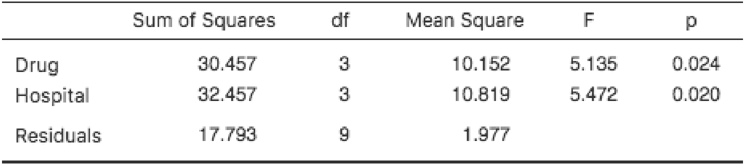
	Hospital			
	I	II	III	IV
Drug treatment	C	A	C	A
	A	A	D	D
	D	B	B	B
	D	C	B	C

- Now we run the experiment:

Restricted randomization and blocking

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

	Hospital			
	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)



	Sum of Squares	df	Mean Square	F	p
Drug	30.457	3	10.152	5.135	0.024
Hospital	32.457	3	10.819	5.472	0.020
Residuals	17.793	9	1.977		

Restricted randomization and blocking

However, the previous design was very inefficient:

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

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

Restricted randomization and blocking

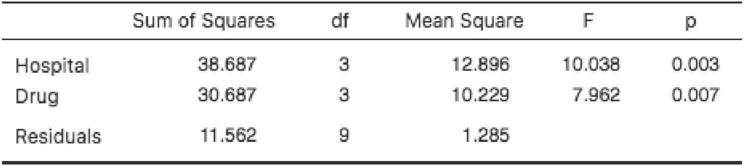
Randomized block design for pain-relieving drug experiment:

	Hospital			
	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 applied the same number of times (once) at each hospital
- All hospitals (blocks) receive all treatments
- No confounding between drug and hospital effects; the SSs capture *only* the marginal variations in the effects

Restricted randomization and blocking



	Sum of Squares	df	Mean Square	F	p
Hospital	38.687	3	12.896	10.038	0.003
Drug	30.687	3	10.229	7.962	0.007
Residuals	11.562	9	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 hospital 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:

	Hospital			
	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).

- 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)

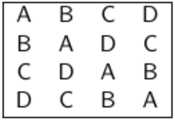
To eliminate possible confounding due to severity of cancer, we can block again; i.e. *block over Hospitals and block over Severities*

	Hospital			
	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)

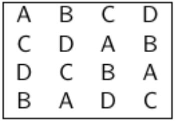
- Now, each treatment appears once and only once *in each row and in each column*
- This experimental design is called a *Latin square* or *orthogonal array*
- Interestingly, *there is still randomization here*; i.e. there are many different ways to construct Latin squares of various dimensions (just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

Restricted randomization and blocking (Latin squares)

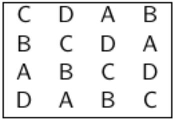
There are 576 different Latin squares of order 4 (i.e. 4 treatments × 4 hospitals × 4 severities). For example:



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

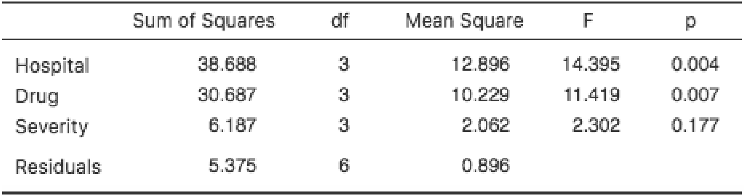


A	B	C	D
C	D	A	B
D	C	B	A
B	A	D	C



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

Restricted randomization and blocking (Latin squares)



	Sum of Squares	df	Mean Square	F	p
Hospital	38.688	3	12.896	14.395	0.004
Drug	30.687	3	10.229	11.419	0.007
Severity	6.187	3	2.062	2.302	0.177
Residuals	5.375	6	0.896		

- The SSs are still accurate for Hospital 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 Doctors. This could create a 4 Drug × 4 Hospital × 4 Severity × 4 Doctor 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 in 4 Hospitals 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.

Repeated measures

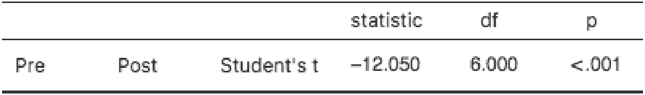
- When you have more than one observation on the *same* sample unit, the experiment is said to contain *repeated measures*.
- Ubiquitous in the health and social sciences.
- Classic example is measuring the effect of an intervention *pre* and *post* application. In this case, average treatment effect can be quantified with a (paired) t-statistic.
- But you may want to measure the effect of an intervention at *many* points in time over the *same* sample units. This suggests an ANOVA framework.
- A repeated measures design is a special case of a *nested* design.
- It is also a special case of a *blocked* design.

Repeated measures ANOVA

Consider the following example quantifying the physical strength of seven subjects before and after a specified 2 month fitness regimen.

Subjects	Pretest	Posttest
1	100	115
2	110	125
3	90	105
4	110	130
5	125	140
6	130	140
7	105	125

Could quantify and test the average treatment (fitness regimen) effect using a paired *t*-test:



			statistic	df	p
Pre	Post	Student's t	−12.050	6.000	<.001

Repeated measures ANOVA

Alternatively, we can think of this experiment as a *two-way randomized complete block design* where measurements (pre or post) are *blocked* within subjects; i.e.

- Each block (subject) gets assigned both “treatments” (pre or post) exactly once
- There are then 2×7 different factor levels, and each factor level has only *one* observation.

In this framework, it may be easier to think of the data as follows:

Subjects	Measurement	Response
1	Pre	100
1	Post	115
2	Pre	110
2	Post	125
⋮	⋮	⋮

Repeated measures ANOVA

Just as in the two-way blocked design from before, the ANOVA model we can estimate is:

$$Y = \mu + \tau_{subject} + \tau_{measurement} + \varepsilon$$

Notice, again, there is no way to estimate an interaction term. Why?

- Only one observation per 2 × 7 factor levels; so no variation at the interaction level to explain.
- Equivalently, all degrees of freedom would be used up, so none leftover for residuals (so no F-tests!): 1 df + 6 df + 6 df = 13 df

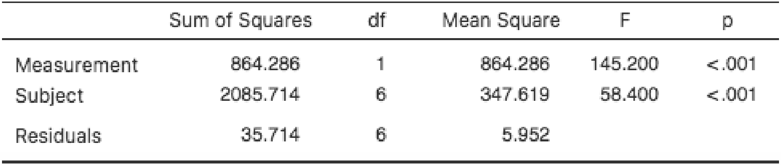
This model is sometimes written as follows:

$$Y = \mu + \tau_s + \tau_{m(s)} + \varepsilon$$

This form has the advantage of explicitly signaling that the measurements (m) are *nested* within subjects (s).

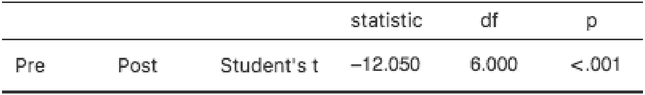
Repeated measures ANOVA

Running the ANOVA proceeds *exactly* as usual. The repeated measures design is now *built into the ANOVA model* we specified.



	Sum of Squares	df	Mean Square	F	p
Measurement	864.286	1	864.286	145.200	< .001
Subject	2085.714	6	347.619	58.400	< .001
Residuals	35.714	6	5.952		

Compare with paired t-test from before:

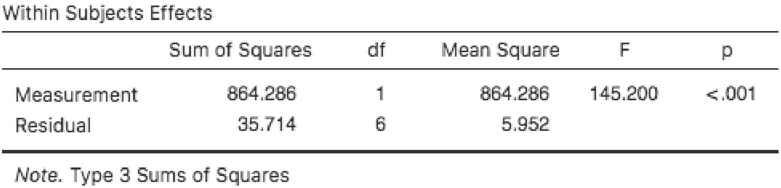


			statistic	df	p
Pre	Post	Student's t	−12.050	6.000	< .001

Note that $(-12.05)^2 = 145.20$; so our t-test is the same as the F-test in this two group (pre vs. post test) [This is true in general: t-tests are equivalent to F-tests on two groups].

Repeated measures ANOVA in Jamovi

- In Jamovi, there is a special “Repeated Measures ANOVA” option that is convenient to use.

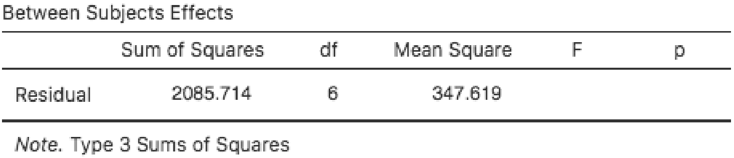


Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Measurement	864.286	1	864.286	145.200	< .001
Residual	35.714	6	5.952		

Note. Type 3 Sums of Squares

[3]



Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residual	2085.714	6	347.619		

Note. Type 3 Sums of Squares

Repeated measures ANOVA in Jamovi

Note the special terminology, very common to repeated measures analyses:

- The effect attributable to each sample subject is typically referred to as the "between subject residuals".
  - Think: we expect there to be differences between sample subjects, but we don't really care about these differences; they are essentially *baseline* differences.
  - Unfortunate terminology to call them "residual effects," but very common (sadly).
- The overall ANOVA model residuals (i.e. the leftover variation after accounting for the explanatory variables in the ANOVA model) are typically referred to as the "within subject residuals".
  - Think: we already know that ANOVA model residuals are unique to each observation, and here the observations are "within subject."

Assumptions of repeated measures ANOVA

The assumptions for a repeated measures ANOVA are a bit different:

- Independence of observations *between* subjects/factors only (obviously, observations *within* subjects are related).
- Equality of variances (homoskedasticity) over all levels of *between* subject factors.
- Normality assumption over all levels of *between* subject factors.
- Equality of variances and normality assumption *within* factors when *more* than two repeated measurements (time points): variances of the *differences* between all adjacent pairs of repeated measurements must be the same over all adjacent time points, and variances of the *differences* between all other possible pairs of repeated measurements must be the same over all possible pairs of time points, in addition to multivariate normality. This assumption is called *sphericity*.

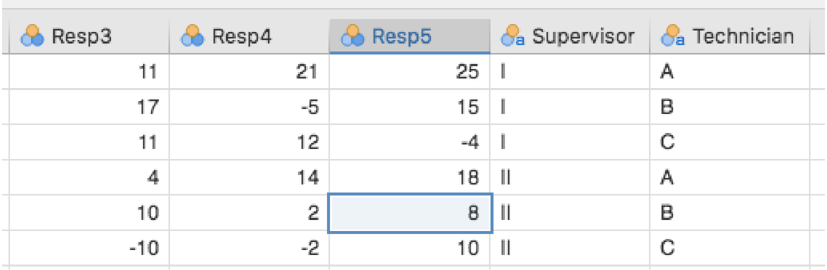
Checking assumptions of repeated measures ANOVA

In Jamovi:

- Equality of variances checked by Levene's test.
- Normality is not separately assessed (annoyingly). One easy way to check normality of between subject factor levels is to fit a bunch of ordinary ANOVAs on the response at *each* time point separately. Ignore the ANOVA output, but examine the QQ-plot.
- Sphericity assumption checked by Mauchly's test and other statistics (only relevant for more than two time points).
- Sphericity is a major practical problem of implementation (when more than two time points in data).

Repeated measures ANOVA, example

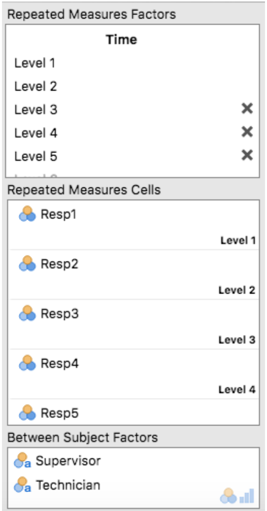
Suppose we have 3 technicians learning how to operate a new piece of machinery. 3 supervisors evaluate their performance at 5 different time points over a one hour period (these evaluations are treated as replications). We thus have a 3 technician × 3 supervisor experiment on 5 repeated points in time. Some sample data are as follows:



Resp3	Resp4	Resp5	Supervisor	Technician
11	21	25	I	A
17	-5	15	I	B
11	12	-4	I	C
4	14	18	II	A
10	2	8	II	B
-10	-2	10	II	C

Repeated measures ANOVA, example

Select “Repeated Measures ANOVA” in Jamovi and specify the repeated measures columns and the between subjects variables:



Repeated Measures Factors

Time

Level 1

Level 2

Level 3 ✕

Level 4 ✕

Level 5 ✕

Repeated Measures Cells

Resp1	Level 1
Resp2	Level 2
Resp3	Level 3
Resp4	Level 4
Resp5	

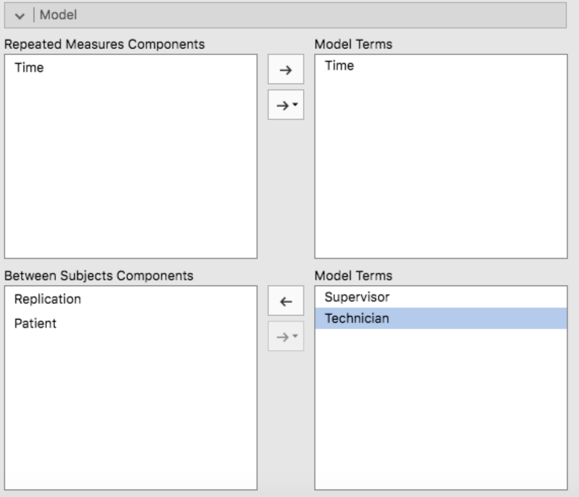
Between Subject Factors

Supervisor

Technician

Repeated measures ANOVA, example

Specify the model that you want fitted:



Model

Repeated Measures Components

- Time

Model Terms

- Time

Between Subjects Components

- Replication
- Patient

Model Terms

- Supervisor
- Technician

Repeated measures ANOVA, example

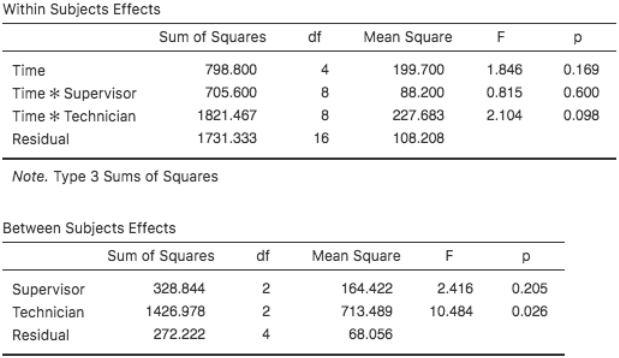
The model Jamovi then fits is:

$$Y = \mu + \tau_{time} + \tau_{sup} + \tau_{tech} + \tau_{time \times sup} + \tau_{time \times tech} + \tau_{sup \times tech} + \varepsilon$$

- In repeated measures ANOVA, we are usually most interested in the ‘time’ effect.
- Here, we probably are not interested in the ‘technician’ effect, since we expect there to be a natural baseline difference between technicians.
- We may be interested in the ‘supervisor’ effect, as it could suggest whether or not supervisors are evaluating technicians consistently.
- The time interactions may be of interest.
- If we specify an interaction between ’supervisor’ and ’technician’, then Jamovi will use up all our degrees of freedom creating a three-way interaction (not mathematically necessary, but default for Jamovi).

Repeated measures ANOVA, example

Examine the output:



Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Time	798.800	4	199.700	1.846	0.169
Time ∗ Supervisor	705.600	8	88.200	0.815	0.600
Time ∗ Technician	1821.467	8	227.683	2.104	0.098
Residual	1731.333	16	108.208		

Note. Type 3 Sums of Squares

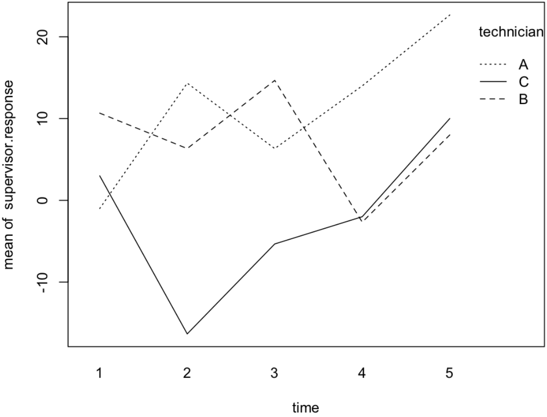
Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Supervisor	328.844	2	164.422	2.416	0.205
Technician	1426.978	2	713.489	10.484	0.026
Residual	272.222	4	68.056		

- Unsurprisingly, the small sample size leads to low power
- No noticeable time effect

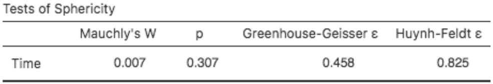
Repeated measures ANOVA, example

Examine an interaction plot:



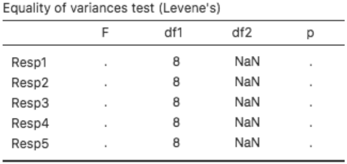
Repeated measures ANOVA, example

Try to assess the assumptions:



Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Time	0.007	0.307	0.458	0.825



Equality of variances test (Levene's)

	F	df1	df2	p
Resp1	.	8	NaN	.
Resp2	.	8	NaN	.
Resp3	.	8	NaN	.
Resp4	.	8	NaN	.
Resp5	.	8	NaN	.

- Too little data for meaningful tests!
- Can still try to assess normality of between subjects factor levels, but so little data will make the assessment difficult.
- In practice, when there are too few data to assess assumptions, *nonparametric* options are preferable.