Repeated Measures Design

Repeated Measures ANOVA

- Instead of having one score per subject, experiments are frequently conducted in which multiple scores are gathered for each case
- Repeated Measures or Within-subjects design

When to Use

- Measuring performance on the same variable over time
    - for example looking at changes in performance during training or before and after a specific treatment
- The same subject is measured multiple times under different conditions
    - for example performance when taking Drug A and performance when taking Drug B
- The same subjects provide measures/ratings on different characteristics
    - for example the desirability of different characteristics of interpersonal relationships
- Note how we could do some RM as regular between subjects designs
    - Ex. Randomly assign to drug A or B

Advantages to RM design

- Design – nonsystematic variance (i.e. error, that not under experimental control) is reduced
  - Take out error variance that's due to individual differences, resulting in more sensitivity/power for the treatment main effect
- Efficiency – fewer subjects are required

Independence in ANOVA

- Analysis of variance as discussed previously assumes cells are independent
- But here that is not going to be the case
  - For example, those subjects who perform best in one condition are likely to perform best in the other conditions
- One thing to note though, observations still need to be independent from one case to the next

Partialling out dependence

- Our differences of interest now reside within subjects and we are going to partial out differences between the subjects
  - This removes the dependence on subjects that causes the problem mentioned
- For example:
  - Subject 1 scores 10 in condition A and 14 in condition B
  - Subject 2 scores 6 in condition A and 10 in condition B
- In essence, what we want to consider is that both subjects score 2 less than their own overall mean score in condition A and 2 more than their own overall mean score in condition B

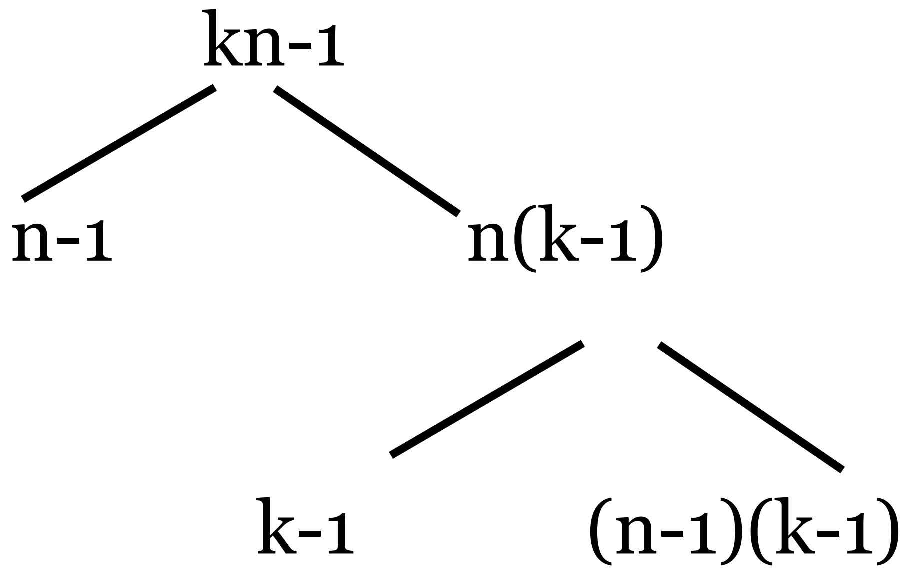
Partition of SS

SS_{total}

$SS_{b/t\ subjects}$ $SS_{w/in\ subjects}$

$SS_{treatment}$ SS_{error}

Partitioning the degrees of freedom



Understanding the Uncertainty

- SS_{total}
  - Deviation of each individual score from the grand mean
- $SS_{b/t\ subjects}$
  - Deviation of subjects' individual means (across treatments) from the grand mean.
  - In the RM setting, this is typically deemed uninteresting, as we can pretty much assume that 'subjects differ'[1]
- $SS_{w/in\ subjects}$: How Ss vary about their own mean, breaks down into:
  - $SS_{treatment}$
    - As in between subjects ANOVA, is the comparison of treatment/measure means to each other (by examining their deviations from the grand mean)
    - However this is now a partition of the within subjects variation
  - SS_{error}
    - Variability of individuals' scores about their treatment/measure mean

Example

- Effectiveness of mind control for different drugs

Prisoner	Placebo	DrugA	DrugB	DrugC	Mean
Henchman	3	4	6	7	**5**
Cobb	0	3	3	6	**3**
Dutton	2	1	4	5	**3**
The Rook	0	1	3	4	**2**
#6	0	1	4	3	**2**
Mean	**1**	**2**	**4**	**5**	***3***

Calculating SS

- Calculate SS_{within}
- Conceptually it reflects the subjects' scores variation about their own individual means
- $(3-5)^2 + (4-5)^2 \ldots (4-2)^2 + (3-2)^2 = 58$
- This is what will be broken down into treatment and error variance

Calculating SS

- Calculate SS_{treat}
- Conceptually it is the sum of the variability due to all treatment *pairs*
  - If we had only two treatments, the F for this would equal t^2 for a paired samples t-test
  - $SS_{b/t} = SS_{treat} = \dfrac{n\sum(\bar{T} - \bar{T}')^2}{k}$
- $SS_{treat} = n\sum\left(\bar{T} - \bar{G}\right)^2$
- $SS_{treat} = 5[(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2] = 50$

5 people in each treatment

Treatment means

Grand mean

Calculating SS

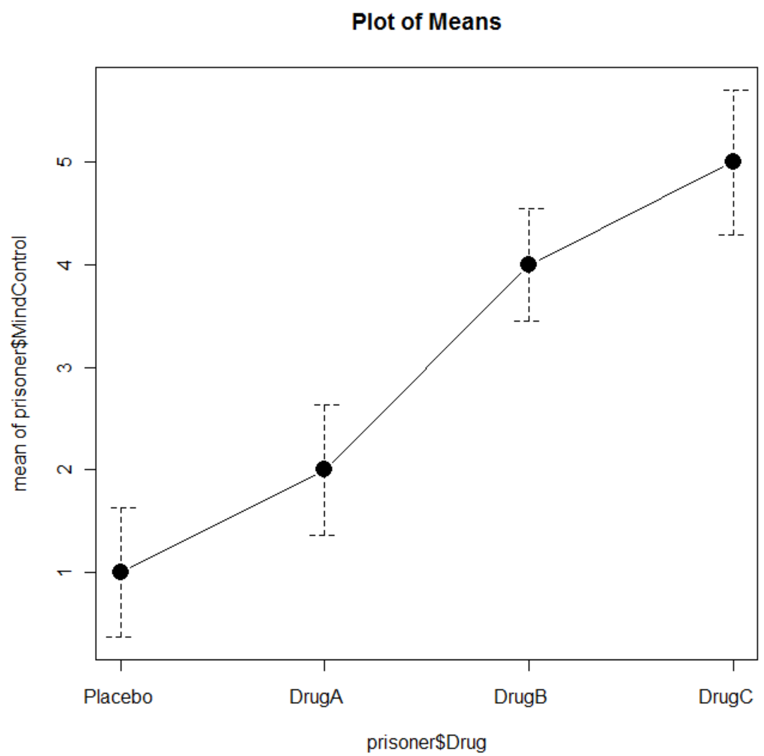
- SS_{error}
- Residual variability
  - Unexplained variance, which includes subject by treatment interaction*
- Recall that $SS_{w/in} = SS_{treat} + SS_{error}$
- $SS_{error} = SS_{w/in} - SS_{treat}$
- 58-50= 8

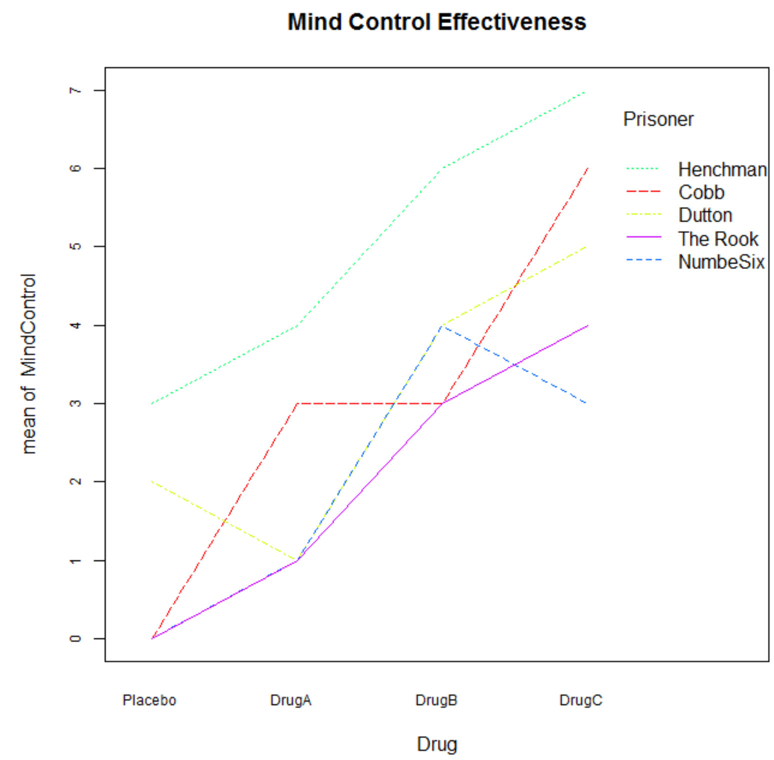
Tests of Within-Subjects Effects

Measure: MEASURE_1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
drug	50.000	3	16.667	25.000	.000	.862
Error(drug)	8.000	12	.667			

Plot of Effects





Effect size

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
drug	50.000	3	16.667	25.000	.000	.862
Error(drug)	8.000	12	.667			

- Note that as we have partialled out error due to subject differences from the overall residual error, our measure of effect here is
  - $SS_{effect}/(SS_{effect} + SS_{error})$
- So this is PES not simply eta² as it was for one-way between subjects ANOVA
- PES = 50/58

Interpretation

- As with a regular one-way Anova, the omnibus RM analysis tells us that there is some difference among the treatments (drugs)
- Often this is not a very interesting outcome, or at least, not where we want to stop in our analysis

- In this example we might want to know which drugs are better than which

Contrasts and Multiple Comparisons

- If you had some particular relationship in mind you want to test due to theoretical reasons (e.g. a linear trend over time) one could test that by doing contrast analyses.
- This table compares standard contrasts available in statistical packages
  - Deviation, simple, difference, Helmert, repeated, and polynomial
- Given the nature of RM design, it is unlikely you wouldn't at least test for a trend of some kind

Deviation	Compares the mean of one level to the mean of all levels (grand mean)
Simple	Compares each mean to some reference mean (either the first or last category e.g. a control group)
Difference (reverse Helmert)	Compares level 1 to 2, level 3 with the mean of the previous two etc.
Helmert	Compares level 1 with all later, level 2 with the mean of all later, level 3 etc.
Repeated	Compares level 1 to level 2, level 2 to level 3, 3 to 4 and so on
Polynomial	Tests for trends (e.g. linear) across levels

Multiple comparisons

- However with our drug example we are not dealing with a time based model and may not have any preconceived notions of what to expect[1]

- So now how are you going to do conduct a post hoc analysis?
- Technically you could flip your data so that treatments are in the rows with their corresponding score, run a regular one-way ANOVA, and do Tukey's etc. as part of your analysis.
- However you would still have problems because the appropriate error term would not be used in the analysis.
  - B/t subjects effects not removed from error term

Multiple comparisons

- The process for doing basic comparisons remains the same
- In the case of repeated measures, as Howell notes (citing Maxwell) one may elect in this case to test them separately as opposed to using a pooled error term[1]
  - The reason for doing so is that such tests would be extremely sensitive to departures from the sphericity assumption
- However using the MS_{error} we can test multiple comparisons and once you have your resulting t-statistic, one can get the probability associated with that

$$\frac{\psi}{\sqrt{\frac{\sum\left(a^2\right)MS_{error}}{n}}}$$

Multiple comparisons

- Example in R comparing placebo and Drug A
  - t = 1.93
  - pt(1.93, 4, lower.tail=F)
  - .063 one-tailed or .126 two-tailed

$$\frac{\psi}{\sqrt{\frac{\sum\left(a^2\right)MS_{error}}{n}}}$$

$$\frac{1}{\sqrt{\frac{2*.667}{5}}} = 1.93$$

MS_{error} from ANOVA table

Multiple comparisons

- While one could correct in Bonferroni fashion, there is a False discovery rate for dependent tests
  - Will control for overall type I error rate among the rejected tests
- The following example uses just the output from standard pairwise ts for simplicity

Multiple comparisons

- Output from t-tests

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Placebo - DrugA	-1.000	1.414	.632	-2.756	.756	-1.581	4	.189
Pair 2	Placebo - DrugB	-3.000	.707	.316	-3.878	-2.122	-9.487	4	.001
Pair 3	Placebo - DrugC	-4.000	1.225	.548	-5.521	-2.479	-7.303	4	.002
Pair 4	DrugA - DrugB	-2.000	1.225	.548	-3.521	-.479	-3.651	4	.022
Pair 5	DrugA - DrugC	-3.000	.707	.316	-3.878	-2.122	-9.487	4	.001
Pair 6	DrugB - DrugC	-1.000	1.414	.632	-2.756	.756	-1.581	4	.189

Multiple comparisons

- Output from R. P-values from previous slide have been ordered

	rawp	Bonferroni	Holm	Hochberg	SidakSS	SidakSD	BH	BY
[1,]	0.001	0.006	0.006	0.005	0.00598502	0.005985020	0.003	0.00735
[2,]	0.001	0.006	0.006	0.005	0.00598502	0.005985020	0.003	0.00735
[3,]	0.002	0.012	0.008	0.008	0.01194016	0.007976032	0.004	0.00980
[4,]	0.022	0.132	0.066	0.066	0.12494948	0.064558648	0.033	0.08085
[5,]	0.189	1.000	0.378	0.189	0.71547193	0.342279000	0.189	0.46305
[6,]	0.189	1.000	0.378	0.189	0.71547193	0.342279000	0.189	0.46305

- The last column is the Benjamini and Yekutieli correction of the p-value that takes into account the dependent nature of our contrasts
- A general explanation might lump Drug A as ineffective (not statistically different from the placebo), and B & C similarly effective

Assumptions for RM

- Standard ANOVA assumptions
  - Homogeneity of variances
  - Normality
  - Independent observations

- For RM design we are looking for homogeneity of covariances among the treatments e.g. t1,t2,t3
  - Special case of HoV
  - Spherecity
    - When the variance of the difference scores for any pair of groups is the same as for any other pair

Sphericity

- Observations may covary across time, dose etc., and we would expect them to. But the degree of covariance must be similar.
  - If covariances are heterogeneous, the error term will generally be an underestimate and F tests will be positively biased
- Such circumstances may arise due to carry-over effects, practice effects, fatigue and sensitization

Sphericity

- Suppose the repeated measure factor of TIME had 3 levels – before, after and follow-up scores for each individual
- RM ANOVA assumes that the 3 correlations
  - r (Before-After)
  - r (Before-Follow up)
  - r (After-Follow up)
- Are all about the same in size
  - i.e. any difference due to sampling error

Sphericity

- Tests can be run to show if the are not
  - Mauchly’s test of Sphericity generates a significant chi square
  - Again, when testing assumptions we typically hope they do not return a significant result
- A correction factor called EPSILON is applied to the degrees of freedom of the error term when calculating the significance of F
  - This is default output for some statistical programs

Sphericity

- If the Mauchly Sphericity test is significant, then use the Corrected significance value for F
  - Different approaches will apply different corrective factors[1]
- If there are only 2 levels of the factor, why is sphericity is not a problem?

More Examples

Another RM example

Before	During	After
14	15	14
34	33	22
26	28	25
30	33	18
35	33	30
16	31	23
27	28	28
28	33	19
28	34	27
25	27	25

- Students were asked to rate their stress on a 50 point scale in the week before, the week of, or the week after their midterm exam

Analysis

- For comparison, first analyze these data as if they were from a between subjects design
- Then conduct the analysis again as if they come from a repeated measures design

Comparing outputs

- Between-subjects

Source	df	SS	MS	F	p
time	2	204.8	102.4	2.981	.0676
error	27	927.5	34.352		

+

- Within subjects

Source	df	SS	MS	F	p
Subject	9	654.3	72.7		
time	2	204.8	102.4	6.747	.0065
error	18	273.2	15.178		

Comparing outputs

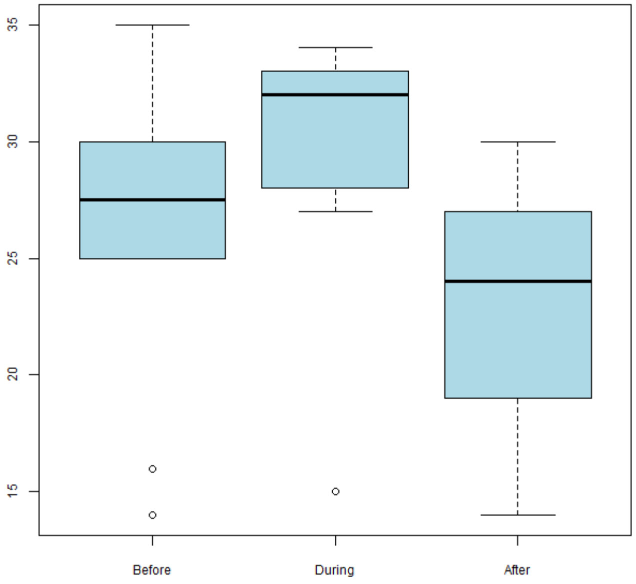
- SS due to subjects has been discarded from the error term in the analysis of the treatment in the RM design
- Same b/t treatment effect
- Less in error term
- More power in analyzing as repeated measures
- Given outliers, robust check is below using trimmed means

```
[1] "The number of groups to be compared is"
[1] 3
$test
[1] 7.11542
$df
[1] 1.229382 6.146911
$siglevel
[1] 0.03272527
```

Trend Contrast

- It would be unlikely to no expect differences in a certain pattern, in particular, we’d expect a hump like the one portrayed in the data

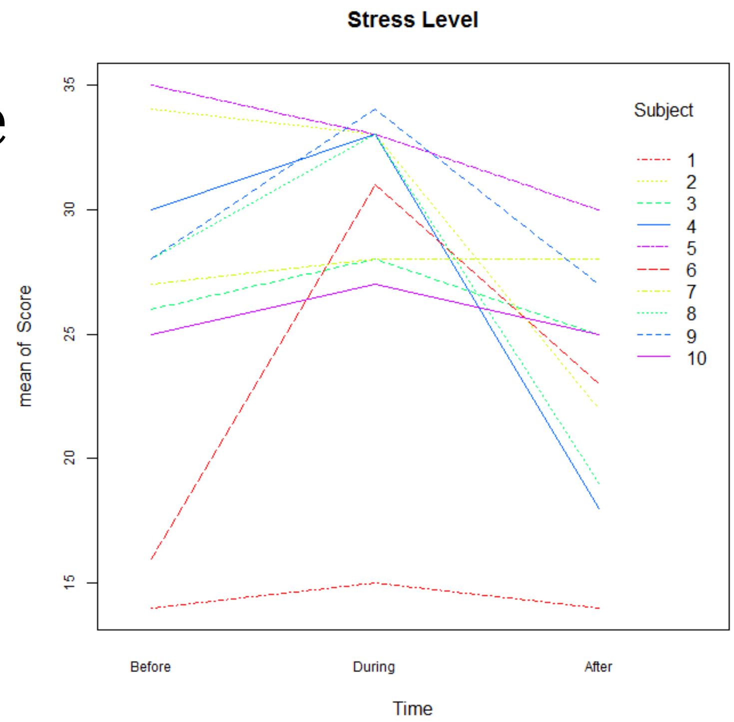
- Testing for a quadratic trend would tell us if in fact this were the case



Trend Contrast

- Plot of subject differences
- A quadratic trend does appe best

```
 Df  Sum Sq Mean Sq F value   Pr(>F)
Time               2 204.800 102.400  6.7467 0.006508 **
  Time: Linear     1  51.200  51.200  3.3734 0.082830 .
  Time: Quadratic  1 153.600 153.600 10.1201 0.005172 **
Residuals         18 273.200  15.178
```



Another example (from Howell)

- An experiment designed to look at the effects of a relaxation technique on migraine headaches
- 9 subjects are asked to record the number of hours of such headaches each week to provide a baseline, then are given several weeks of training in relaxation technique – during which the number of hours of migraines per week were also recorded
- A single within-subject effect with 5 levels (week 1, week 2, week 3, week 4, week 5)

Data

baseline		training		
week1	week2	week3	week4	week5
21	22	8	6	6
20	19	10	4	4
17	15	5	4	5
25	30	13	12	17
30	27	13	8	6
19	27	8	7	4
26	16	5	2	5
17	18	8	1	5
26	24	14	8	9

Results

Source	df	Sum of Squares	Mean Square	F	p
Subject	8	486.711	60.839		
time	4	2449.200	612.300	85.042	.0001
error	32	230.400	7.200		

Note however that we really had 2 within subjects variables condition (baseline vs. training) and time (week 1-5)

How would we analyze that?
Tune in next week! Same bat time! Same bat channel!

Appendix

More stuff that likely won't make the cut at next time

Review - Covariance

- Measure the joint variance of two (or more) variables
- Use cross product of deviation of scores from their group mean
- Formula:

$$\text{covar}_{\text{xy}} = \frac{\sum\left(\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)\right)}{n-1}$$

$$r = \frac{\text{covar}_{\text{xy}}}{s_x s_y}$$

Correcting for deviations

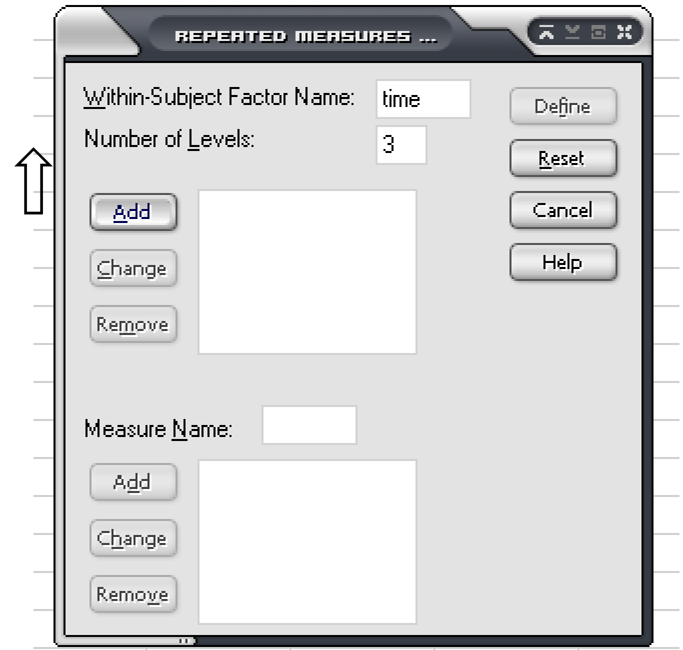
- Epsilon (ε) measures the degree to which covariance matrix deviates from compound symmetry
  - All the variances of the treatments are equal and covariances of treatments are equal
- When $\varepsilon = 1$ then matrix is symmetrical, and normal df apply
- When $\varepsilon = \frac{1}{k-1}$ then matrix has maximum heterogeneity of variance and the F ratio should have 1, n-1 df

Estimating F

- Two different approaches, adjusting the df
  - Note that the F statistic will be the same, but the df will vary
- Conservative: Box's/Greenhouse-Geisser
- Liberal: Huynh-Feldt
- Huynh-Feldt tends to overestimate ε (can be > 1, at which point it is set to 1)
- Some debate, but Huynh-Feldt is probably regarded as the method of choice, see Glass & Hopkins

Conducting RM in SPSS

- “Analyze” ⇒ “General Linear Model” ⇒ “Repeated Measures”
- Type “time” into the within-subject factor name text box (in place of “*factor1*”)
- Type “3” in the Number of Levels box
- Click “Add”
- Click “Define”



Conducting RM in SPSS

- Move the variables related to the treatment to the within subjects box
- Select other options as desired