36-309/749 Experimental Design for Behavioral and Social Sciences

Nov 10, 2015

Lecture 9: Repeat Measures ANOVA

Overview of Repeated Measures

- **Definition**: When more than one measurement is made on the same subject either over time with one treatment or with different treatments we have a ***repeated measures (study) design***. One synonym is ***within-subjects design***. (So every study we looked at until now was a between-subjects design.)*

- **Terminology**: Any factor for which each subject experiences *all* of the levels is a ***within-subjects factor***, i.e., repeated. Any factor for which each subject experiences *only one* of the levels is a ***between-subjects factor***. (So all factors we studied until now were between-subjects factors.)

* Unrelated to within- vs. between-***groups*** SS and MS.

Overview of Repeated Measures, cont.

- New methods are needed that **take into account the correlation** of errors for pairs of measurements made on the same experimental unit (subject, classroom, etc.).
  - Examples of the problem of ignoring correlation:
    - k=3 levels of treatment randomized to classrooms. Ignoring within classroom correlation can greatly increase type-1 error through falsely narrow CIs.
    - k=3 levels of treatment each given to each subject. Ignoring within subject correlation can greatly reduce power through excessively wide CIs.

Overview of Repeated Measures, cont.

- **Advantages of a within subjects design**:
  - more power (through canceling out of subject-to-subject variability and "self-control") and /or reduced number of subjects needed
  - ability to study time trends (including learning).

- **Disadvantages**: possible confounding with previous treatments (possibly fixed with counterbalancing; see below).

- **Example**: Osteoarthritis is a mechanical degeneration of joint surfaces causing pain, swelling and loss of joint function in one or more joints. Physiotherapists treat the affected joints to increase the range of movement (ROM). In this study 10 subjects were each given a trial of therapy with two treatments, TENS (an electric nerve stimulation) and short wave diathermy (a heat treatment), plus control

Overview of Repeated Measures, cont.

- **Appropriate kinds of analysis** for repeated measures fall into four categories

1) ***Response simplification***: e.g. call the difference between two of the measurements for each subject the "response" (DV), and use standard techniques for a between-subjects design. Or use the mean of several responses. This approach does not fully utilize the available information. And it cannot answer some interesting questions.
2) Treat the several responses on one subject as a single "multivariate" response and model the correlation between the components of that response. The main statistics (SS, MS) are now matrices rather than individual numbers. This approach corresponds to results labeled "***multivariate***" under "***repeated measures ANOVA***" for most statistical packages.
3) Treat each response as a separate (univariate) observation, and treat "subject" as a (random) blocking factor. This corresponds to the "***univariate***" output under "***repeated measures ANOVA***". In this form, there are assumptions about the nature of the within-subject correlation that are not met fairly frequently, but standard "adjustments" help.

Overview of Repeated Measures, cont.

- **Appropriate kinds of analysis** for repeated measures fall into four categories

4) Treat each measurement as univariate, but appropriately model the correlations via a "hierarchy" of effects. This is a more modern univariate approach called "***mixed models***" that subsumes a variety of models in a single unified approach. (Covered in Week 11)

Advantages:

- very flexible in modeling correlations
- improved interpretability
- can also be extended to non-normal outcomes
- allows missing data
- allows unequal number and spacing of repeated measurements

A detailed examination of the paired t-test

- A paired t-test is an appropriate analysis for repeated measures with k=2 measurement per subject. E.g., pretend that the osteoarthritis study only compared control to TENS.
- Here are the EDA and (***incorrect***) ANOVA (or independent-samples t-test) results for these data:

Dependent Variable: ROM

			95% Confidence Interval	
Rx	Mean	Std. Error	Lower Bound	Upper Bound
Control	101.900	7.727	85.665	118.135
TENS	84.200	7.727	67.965	100.435

Tests of Between-Subjects Effects
Dependent Variable: ROM

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Rx	1566.450	1	1566.450	2.623	.123
Error	10748.500	18	597.139		
Corrected Total	12314.950	19			

Paired t-test, cont.

- Here is the paired t-test. [It is equivalent to a one-sample t-test for the within-subject ***difference*** of control vs. TENS with the null hypothesis H_0: $\mu_{\text{difference}}$=0.]

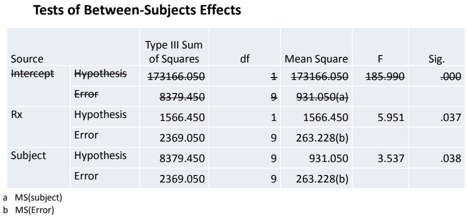
[Note: t = (mean paired difference) / SE(mean paired difference) and df = (#pairs) - 1.]

Paired Samples Test

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Upper	Lower	t	df	Sig. (2-tailed)
control - TENS	17.700	22.945	7.256	1.286	34.114	2.439	9	.037

Paired t-test, cont.

- Here is a an equivalent alternative analysis: two-way ANOVA without interaction and with subject as a ***random effect***:



Tests of Between-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
~~Intercept~~	~~Hypothesis~~	~~173166.050~~	~~1~~	~~173166.050~~	~~185.990~~	~~.000~~
	~~Error~~	~~8379.450~~	~~9~~	~~931.050(a)~~		
Rx	Hypothesis	1566.450	1	1566.450	5.951	.037
	Error	2369.050	9	263.228(b)		
Subject	Hypothesis	8379.450	9	931.050	3.537	.038
	Error	2369.050	9	263.228(b)		

a MS(subject)
b MS(Error)

8379.45+2369.05=10748.50: Four sources of variation decompose to subj.-to-subj. plus others

Paired t-test, conclusions

- Incorrect (assume independent errors) vs. correct analyses (model the errors)
- Relationship between 2-way ANOVA with random subjects (“univariate”) and paired t-test (“response simplification”)
- Paired t-test is a special case of 1-way within-subjects ANOVA (next topic)
- Hidden assumption: no interaction between treatment and subject

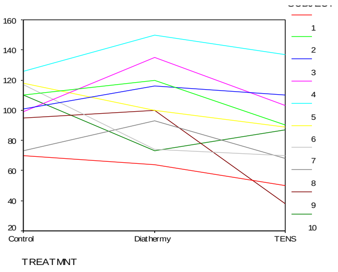
One-way Repeated Measures ANOVA for within-subjects design

- **Problem recognition**: Quantitative outcome and one factor (categorical explanatory variable) with k≥2 levels where “by design” ***every subject receives every level*** of the explanatory variable.
- **Example**: Osteoarthritis problem with ROM outcome, 10 subjects, and two active plus one control treatment (levels of the repeated factor) per subject.
- **Wide format data**: 10 rows with columns for subject id, ROM for the control condition, ROM for Diathermy and ROM for TENS.

1-way RM ANOVA, cont.

- **Model**: For the three treatment levels we can call the population means of the outcome μ_C, μ_T and μ_D.
  - The outcomes for all treatments are normally distributed.
  - The errors are independent ***between (across)*** subjects.
  - The errors are ***correlated*** for the three measurements on any one subject (within subjects).
  - ***Univariate*** approach: The outcomes have equal variance and the errors are positively and equally correlated within subjects.
  - ***Multivariate*** approach: The variance and correlation pattern within subjects is unconstrained (but is the same from subject to subject).
- **Null hypothesis**: $\mu_C = \mu_T = \mu_{D.}$ Alternate hypothesis: at least one population mean differs.

1-way RM ANOVA: EDA



1-way RM ANOVA: Analysis

- In SPSS use General Linear Model / Repeated Measures with data in a “wide format”, i.e. with one column for each level of the within-subjects factor. Both multivariate and univariate analyses are performed, as is a test of the univariate variance/correlation assumption (“sphericity”). For each type of analysis p-values for multiple available choices of a statistic are produced.

Ordinary (between-subjects) ANOVA (requires tall format; *<u>incorrect</u>* because it ignores within-subject correlation):

rom

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2161.800	2	1080.900	1.638	.213
Within Groups	17817.000	27	659.889		
Total	19978.800	29			

1-way RM ANOVA: Analysis, cont.

Repeated Measures Analysis (wide format, correctly takes correlation into account):

Within-Subjects Factors

Measure: rom

Tx	Dependent Variable
1	Control
2	Diathermy
3	TENS

	Mean	Std. Deviation	N
control	101.90	18.586	10
diathermy	102.50	28.025	10
TENS	84.20	29.135	10

1-way RM ANOVA: Analysis, cont.

- **Multivariate Analysis**: [The main calculations are the SSCP (sum of squares and cross products) matrices for treatment and for error. These are constructed for k-1 difference variables. Then MSCP is computed as SSCP/df.] There are four ways of reducing the ratio F matrix ($\mathrm{MSCP}_{\mathrm{treatment}}$ /$\mathrm{MSCP}_{\mathrm{error}}$) to a single F statistic, but ***when we have only a single within-subjects factor, they always agree***.
  - The multivariate approach (shown here) has <u>less strict assumptions</u> than for the univariate approach (see below), but tends to have <u>less power</u>.

Multivariate tests

Effect		Value	F	Hypothesis df	Error df	Sig.
Tx	Pillai's Trace	.549	4.878	2.000	8.000	.041
	Wilks' Lambda	.451	4.878	2.000	8.000	.041
	Hotelling's Trace	1.220	4.878	2.000	8.000	.041
	Roy's Largest Root	1.220	4.878	2.000	8.000	.041

Here p=0.041 is good evidence that the population means for the three treatments are different. (RM contrast testing can be used for more specific null hypotheses.)

1-way RM ANOVA: Analysis, cont.

- ***Mauchly's test of sphericity***: The (uncorrected) ***univariate analysis (next slide) uses an assumption of sphericity,*** which is a slightly more general assumption than the easier-to-understand assumption of "compound symmetry" (the outcomes at each treatment level have the same variance, and all pairs of levels have the same correlation). When there are only 2 levels, there is only one pair of levels so you cannot violate the sphericity assumption.

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
tx	.918	.686	2	.710	.924	1.000	.500

  - A small p-value (≤0.05) suggests violation of the assumption. Unfortunately this assumption test has limited power for small studies, may have too much power for large studies, and can give small p-values when there is good sphericity in the presence of only small to moderate amounts of non-normality. Nevertheless it is commonly used. Here, with p=0.710, we do not reject the sphericity assumption, and we can go with the uncorrected (sphericity assumed) univariate results.

1-way RM ANOVA: Analysis, cont.

- Test of Within-Subjects Effects: This is the ***univariate analysis***. [It is equivalent to two-way ANOVA with subject as a random factor.]

Tests of Within-Subjects Effects [Univariate Analysis]

Source		Type III Sum of Squares	Df	Mean Square	F	Sig.
Tx	Sphericity Assumed	2161.800	2	1080.900	3.967	.037
	Greenhouse-Geisser	2161.800	1.848	1169.749	3.967	.042
	Huynh-Feldt	2161.800	2.000	1080.900	3.967	.037
	Lower-bound	2161.800	1.000	2161.800	3.967	.078
Error(tx)	Sphericity Assumed	4904.200	18	272.456		
	Greenhouse-Geisser	4904.200	16.633	294.851		
	Huynh-Feldt	4904.200	18.000	272.456		
	Lower-bound	4904.200	9.000	544.911		

  - ***If the sphericity assumption is not valid***, a corrected test (or the multivariate test) must be used instead of the "Sphericity Assumed" p-value. I recommend the Huynh-Feldt correction because it is reported to be most robust to non-Normality.
  - Here we reject H_0:μ_C=μ_D=μ_T (F=3.967, with 2 and 18 df, sphericity assumed, p=0.037).

1-way RM ANOVA: Analysis, cont.

- The ***Tests of Within-Subjects Contrasts*** box shows the "simple" ***planned contrasts*** that I chose in the Contrast dialog. The control vs. TENS have statistically significantly different ROM means (p=0.037) but control is not significantly different from diathermy (p=0.944). Examining the table of means, TENS is ***worse*** than control.

Tests of Within-Subjects Contrasts

Source	Tx	Type III Sum of Squares	df	Mean Square	F	Sig.
Tx	Level 2 vs. Level 1	3.600	1	3.600	.005	.944
	Level 3 vs. Level 1	3132.900	1	3132.900	5.951	.037
Error(tx)	Level 2 vs. Level 1	6196.400	9	688.489		
	Level 3 vs. Level 1	4738.100	9	526.456		

  - The default planned contrast, polynomial, is only appropriate for studying change over time (learning). Alternatives include simple (comparisons to a baseline, shown here) and repeated (comparison of adjacent levels).

1-way RM ANOVA: Analysis, cont.

- Because the only factor is a within-subjects factor, the ***Test of Between-Subjects Effects*** box provides no useful information. As usual, the intercept H_0 is uninteresting.

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Intercept	92544.400	1	92544.400	193.506	.000
Error	4304.267	9	478.252		

1-way RM ANOVA: Analysis, cont.

- The ***post-hoc tests*** in the Pairwise Comparisons box show no additional differences beyond those planned. The Bonferroni (or the less conservative Sidak) correction is used to protect type-one error in the presence of data snooping.

Pairwise Comparisons

(I) tx	(J) tx	Mean Difference (I-J)	Std. Error	Sig.(a)	95% Confidence Interval for Difference(a)	
					Lower Bound	Upper Bound
1	2	-.600	8.298	1.000	-24.939	23.739
	3	17.700	7.256	.112	-3.583	38.983
2	1	.600	8.298	1.000	-23.739	24.939
	3	18.300	6.479	.060	-.705	37.305
3	1	-17.700	7.256	.112	-38.983	3.583
	2	-18.300	6.479	.060	-37.305	.705

Based on estimated marginal means

a Adjustment for multiple comparisons: Bonferroni.

Counterbalancing

- Special ***problem*** of within-subjects design: potential ***confounding with prior treatment effects*** when repeated treatments are administered to the same subject.
- The standard solution is ***counterbalancing*:** each subject is randomly assigned to one of the k! different orderings of treatment.
- Without appropriate counterbalancing, the experiment can be misleading due to either a ***learning effect*** or ***carryover***.
- Counterbalancing for this experiment would involve assigning the six possible orders of the three treatments randomly. Ideally a multiple of 6 subjects would be used with "block randomization" to achieve perfect counterbalancing.
- Additional consideration for this experiment: Because diathermy involves heat while TENS involves nerve stimulation, full subject blinding in not possible, but use of some placebo treatment for the control is possible.

Mixed within- and between-subjects design

- Example: Is it harder to solve a math problem if you are sitting in a room with an uncomfortable temperature? This is a study of the time to solve three problems (per subject) at 3 temperatures (one temperature per subject).
- Repeated measures ANOVA, initial output:

Within-Subjects Factors

problem	Dependent Variable
1	prob1
2	prob2
3	prob3

Between-Subjects Factors

		N
temp	60	15
	70	15
	80	15

Mixed Design, cont.

Multivariate Tests(c)

Effect		Value	F	Hypothesis df	Error df	Sig.
problem	Pillai's Trace	.611	32.217(a)	2.000	41.000	.000
	Wilks' Lambda	.389	32.217(a)	2.000	41.000	.000
	Hotelling's Trace	1.572	32.217(a)	2.000	41.000	.000
	Roy's Largest Root	1.572	32.217(a)	2.000	41.000	.000
problem * temp	Pillai's Trace	.439	5.900	4.000	84.000	.000
	Wilks' Lambda	.569	6.678(a)	4.000	82.000	.000
	Hotelling's Trace	.744	7.441	4.000	80.000	.000
	Roy's Largest Root	.726	15.237(b)	2.000	42.000	.000

a Exact statistic

b The statistic is an upper bound on F that yields a lower bound on the significance level.

c Design: Intercept+temp

Within Subjects Design: problem

- You must pick one test beforehand, e.g., Pillai's trace. ***As usual for ANOVA, ignore main effects in the presence of a significant interaction.*** But, as opposed to between-subjects ANOVA, you cannot drop interaction and re-run the analysis if the interaction is statistically significant.

Mixed Design, cont.

Mauchly's Test of Sphericity

Measure: time

					Epsilon		
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Huynh-Feldt	Lower-bound	Greenhouse-Geisser
problem	.972	1.149	2	.563	.973	1.00	.500

Tests of Within-Subjects Effects [Univariate Analysis]

Measure: time

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
problem	Sphericity Assumed	1545.486	2	772.743	28.891	.000
	Greenhouse-Geisser	1545.486	1.946	794.093	28.891	.000
	Huynh-Feldt	1545.486	2.000	772.743	28.891	.000
	Lower-bound	1545.486	1.000	1545.486	28.891	.000
problem * temp	Sphericity Assumed	885.974	4	221.494	8.281	.000
	Greenhouse-Geisser	885.974	3.892	227.613	8.281	.000
	Huynh-Feldt	885.974	4.000	221.494	8.281	.000
	Lower-bound	885.974	2.000	442.987	8.281	.001
Error(problem)	Sphericity Assumed	2246.740	84	26.747		
	Greenhouse-Geisser	2246.740	81.742	27.486		
	Huynh-Feldt	2246.740	84.000	26.747		
	Lower-bound	2246.740	42.000	53.494		

Mixed Design, cont.

Tests of Within-Subjects Contrasts [not useful with I/A]

Measure: time

Source	problem	Type III Sum of Squares	df	Mean Square	F	Sig.
Problem	Linear	1363.445	1	1363.445	51.220	.000
	Quadratic	182.040	1	182.040	6.774	.013
problem * temp	Linear	74.478	2	37.239	1.399	.258
	Quadratic	811.496	2	405.748	15.098	.000
Error(problem)	Linear	1118.021	42	26.620		
	Quadratic	1128.719	42	26.874		

Tests of Between-Subjects Effects [not useful with I/A]

Measure: time

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	60750.744	1	60750.744	2399.114	.000
temp	1356.385	2	678.193	26.783	.000
Error	1063.531	42	25.322		

Math example: Conclusions

- Both problem and temperature affect solution time in a complex (non-additive) way (from the small p-value for interaction from either the multivariate or univariate approach).
- Simple effects contrasts and/or difference of differences would be useful, but are not available. [Use the Contrasts button for additive models.]

Summary

- ***Ignoring the uncorrelated errors assumption gives incorrect conclusions.***
- The ***paired t-test*** is a special case of repeated measures ANOVA for 2 levels of treatment.
- ***One-way repeated measures ANOVA*** is used to study the effects of different levels of one factor on a quantitative outcome when each subject is exposed to all levels of the factor.
- ***Additional power*** derives from the fact that the between-groups SS and MS do not contain the subject-to-subject variability component, so a smaller error MS (with s-to-s variability subtracted out) is appropriately used for the F statistic.
- For one-factor repeated measures analysis with three or more levels of the repeated factor, there is a choice between using ***"multivariate" vs."univariate" analyses***. Luckily they often agree closely, especially when we use the Huynh-Feldt corrected p-value for the univariate analysis in the presence of any clear violation of the sphericity assumption. But if the two analyses disagree, appropriate analysis is not clear, and at a minimum, the EDA should be examined for any unusual aspects of the data.
- A ***"mixed" between/within design*** is probably most common. Each subject "sees" one level of the between-subjects factor and all levels of the within-subjects factor. Interpret similar to 2-way ANOVA: either both factors affect the outcome in a complicated way or interpret each overall null hypothesis separately and check planned contrasts.