Statistiek II

John Nerbonne

Dept of Information Science
j.nerbonne@rug.nl
With thanks to Hartmut Fitz for 1st version, still most of this!

February 26, 2014

Last week

Factorial ANOVA:

- used when there are several independent variables (factors)
- allows to study interaction between factors
- assumptions like one-way ANOVA: homogeneity of variance, normality, independence

Today: **repeated measures** ANOVA (aka 'within-subjects'-design)

- one-way repeated measures ANOVA
- factorial repeated measures ANOVA
- mixed factors repeated measures ANOVA

Last week

Last week's 2×2 ANOVA: repetition accuracy of object-relatives

- two factors, two levels each
- factor A: animacy of head noun
- factor B: relative clause subject type
- factors induced four disjoint groups of items (four tokens per type)
- 48 children, dependent measure: averaged repetition accuracy

Conducted factorial ANOVA 'by item', measured whether there was a difference in repetition accuracy between four groups of sentence types (ANP, INP, APro, IPro)

A different way to look at the same data

Could also have looked at repetition accuracy 'by participant'

- same two factors, head noun animacy and relative clause subject type
- average over tokens per type for each participant

	Sentence type			
Child	ANP	INP	APro	IPro
1	0.00	0.00	0.00	0.00
2	0.00	0.00	0.75	0.38
3	0.00	0.50	0.88	0.75
⋮	⋮	⋮	⋮	⋮
48	0.25	0.50	1.00	0.88

Measure participants repeatedly in all conditions, perform 2×2 ANOVA 'by participant' (expect similar main effects)

One-way repeated measures ANOVA

Repeated measures ANOVA:

Like related-samples t-test, but for $\geq$3 conditions A, B, C, etc.

Applications:

- same group of subjects measured under 3 or more conditions A, B, C,...
- matched k-tuples of subjects, one member measured under A, one under B, one under C,...
- in the latter case, matched tuples are treated as one subject

Labels: 'repeated measures' or 'within-subjects design', 'randomized blocks design'

One-way repeated measures ANOVA

Characteristics:

- assumptions like standard ANOVA, but data points **not** independent (repeated measures)
- economical in design because each subject measured under all conditions
- often research question **requires** repeated measures, e.g., longitudinal studies: each sample member measured repeatedly at several ages
- example: children can discriminate many phonetic distinctions across languages without relevant experience; longitudinal study shows there is a decline in this ability (within first year)
- key idea: eliminate variation between sample members (reduces within-groups variance)

Partitioning the variance

One-way **independent samples** ANOVA:

$$\begin{array}{ccccc} \text{SST} & = & \text{SSG} & + & \text{SSE} \\ \textbf{T}\text{otal } \textbf{S}\text{um of } \textbf{S}\text{quares} & = & \textbf{G}\text{roup } \textbf{S}\text{um of } \textbf{S}\text{quares} & + & \textbf{E}\text{rror } \textbf{S}\text{um of } \textbf{S}\text{quares} \end{array}$$

One-way **repeated measures** ANOVA:

- same subjects in each 'group' (i.e., condition)
- determine aggregate **variance among subjects** (SSS):

  SSS $= I \cdot \sum_{j=1}^{N}(\overline{x_j} - \overline{x})^2$ where I number of conditions, $\overline{x_j}$ subject mean (across conditions), and $\overline{x}$ total mean

- remove this effect of **individual differences** from SSE
- determine MSE from SSE*= SSE−SSS

One-way repeated measures example

Experiment: Computational model learns to produce complex sentences from meaning (Fitz, Neural Syntax, 2008).

Task:

- model receives semantic structure of a sentence as input
- tries to produce sentence which expresses this meaning
- production by word-to-word prediction

Example:	Input:	Agent	$\rightarrow$	[DOG]
		Action	$\rightarrow$	[CHASE]
		Patient	$\rightarrow$	[CAT]
Sequential output:		`the dog`		
		`the dog chases`		
		`the dog chases the cat`		

One-way repeated measures example

But how to represent semantic relations for multiple clauses?

Three semantic conditions:

(a) give more prominence to main clause (order-link)
E.g., **the dog** `that runs` **chases the cat**

(b) mark the topic and focus of both clauses (topic-focus)
E.g., **the dog** `that` [**the dog**] `runs chases the cat`

(c) features which bind topic and focus (binding)
E.g., `the dog that runs chases the cat`, **Agent-Agent**

The model's learning behavior is tested in each of these conditions.

Question: Is model sensitive to different semantic representations?

One-way repeated measures example

Subjects:

- model is randomly initialized
- exposed to 10 different sets of randomly generated training items ($\Rightarrow$ 10 experimental subjects)
- subject = model + fixed parameters + training environment
- each subject tested in conditions (a)–(c) (**repeated measures**)

Dependent variable: mean sentence accuracy after learning phase (on 1000 test items)

Scoring: model produces target sentence *exactly*: 1
any kind of lexical or grammatical error: 0
sentence accuracy: percentage of correct utterances

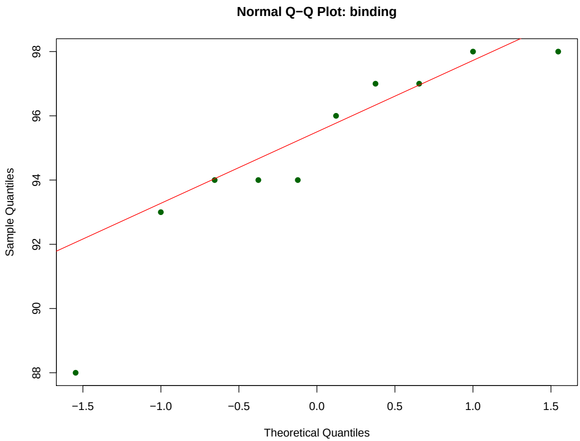
One-way repeated measures example

Data on modelling the acquisition of relative clauses:

Model-subject	Condition			Subject mean
	order-link	topic-focus	binding	
1	80	94	98	90.7
2	73	90	98	87
3	70	98	94	87.3
⋮	⋮	⋮	⋮	⋮
10	71	99	94	88
Mean	76.3	95.8	94.9	89

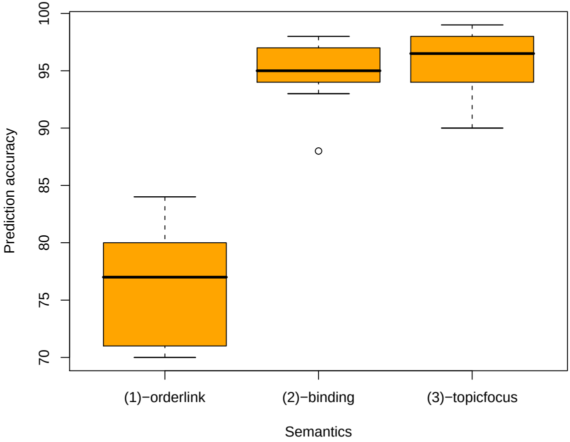
Note: subject means (across conditions) required to compute subject sum of squares (SSS).

Check normality and standard deviations



SDs: order-link: 4.9, topic-focus: 2.66, binding: 3.03 ✓

Visualizing the data



Little skew, different medians, no overlap between (1) and (2) or (3), very likely significant

Computing the error sum of squares

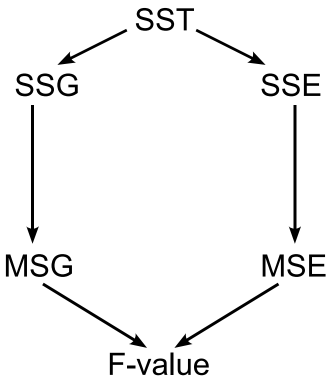
Model-subject	Condition			Subject mean
	order-link	topic-focus	binding	
1	80	94	98	90.7
2	73	90	98	87
3	70	98	94	87.3
⋮	⋮	⋮	⋮	⋮
10	71	99	94	88
Mean	76.3	95.8	94.9	89

$$\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2 = (80 - 76.3)^2 + \ldots + (94 - 94.9)^2 = \underline{362.6}$$

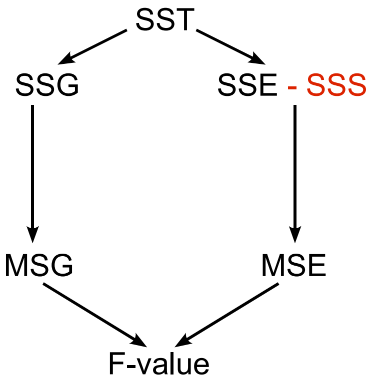
Key idea of repeated measures

Because subjects are measured in all conditions: remove variability due to individual differences from SSE!

Independent samples:



Repeated measures:



Computing the subject sum of squares

Subject **S**um of **S**quares: aggregate measure of between-subjects variability

$$
\begin{aligned}
\text{SSS} &= I \cdot \sum_{j=1}^{N} (\overline{x_j} - \overline{x})^2 \\
&= 3 \cdot (90.7 - 89)^2 + 3 \cdot (87 - 89)^2 + \ldots + 3 \cdot (88 - 89)^2 \\
&= \underline{86}
\end{aligned}
$$

Adjust error sum of squares:

$\text{SSE*} = \text{SSE} - \text{SSS} = 362.6 - 86 = \underline{276.6}$

Computing the mean squared error

SSE*: usual SSE **minus** between-subjects sum of squares (SSS)

Recall different degrees of freedom:

$$\begin{aligned} \text{DFT} &= N - 1 = 30 - 1 = 29 && \text{(total)} \\ \text{DFG} &= I - 1 = 3 - 1 = 2 && \text{(group)} \\ \text{DFE} &= N - I = 30 - 3 = 27 && \text{(error)} \end{aligned}$$

Subject degrees of freedom (corresponding to SSS):

DFS $=$ Number of subjects in each group $-1 = 10 - 1 = 9$

Remove this component from DFE, and what remains is:

DFE* $=$ DFE$-$DFS $= 27 - 9 = 18$

R output

Manually: $\text{MSE*} = \frac{\text{SSE*}}{\text{DFE*}} = \frac{276.6}{18} = 15.37$

$$F\text{-value: } F = \frac{\text{MSG}}{\text{MSE*}} = \frac{1211.7}{15.37} = 78.83$$

R output:

Error: subject

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	9	86.00	9.55		

Error: subject:semantics

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
semantics	2	2423.40	1211.70	78.85	1.2428e-09 ***
Residuals	18	276.60	15.37		

—

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .

Reject null hypothesis H_0, i.e., conclude that difference in semantic representations **does** affect the model's learning behavior

Post-hoc tests*

Tukey's **H**onestly **S**ignificant **D**ifferences test

- suitable for multiple comparisons when ANOVA is significant
- requires equal group sizes!
- based on Studentized range statistic Q

SPSS doesn't do HSD for repeated measures (use Bonferroni)

Compute HSD manually: $q^* = \frac{\mu_i - \mu_j}{\sqrt{\frac{\text{MSE}^*}{N}}}$

Null-hypothesis H_0: $\mu_i = \mu_j$
Alternative hypothesis H_a: $\mu_i \neq \mu_j$

Reject H_0 if $q^* \geq q$ (check table)

Applying Tukey HSD*

Test difference between ‘topic-focus’ and ‘binding’ condition in the example:

$$q^* = \frac{95.8 - 94.9}{\sqrt{\frac{15.37}{10}}} = \frac{0.9}{\sqrt{1.537}} = 0.73$$

q has two degrees of freedom: group size (here 9), and
DFE* (here 18)

$q(9, 18) = 6.08$ (from table for Studentized range statistic)

Hence, $q^* \leq q$, do not reject H_0 (at $\alpha = 0.01$).

Conclude: the model learns complex sentences equally well in the ‘topic-focus’ and ‘binding’ condition

Applying Tukey HSD*

Test difference between 'binding' and 'order-link' condition in the example:

$$q^* = \frac{94.9 - 76.3}{\sqrt{\frac{15.37}{10}}} = \frac{0.9}{\sqrt{1.537}} = 15.0$$

q has two degrees of freedom: group size (here 9), and
DFE* (here 18)

$q(9, 18) = 6.08$ (from table for Studentized range statistic)

Hence, $q^* \geq q$, reject H_0 (at $\alpha = 0.01$).

Conclude: the model learns complex sentences more reliably in the 'binding' than in the 'order-link' condition.

Repeated measures in factorial design

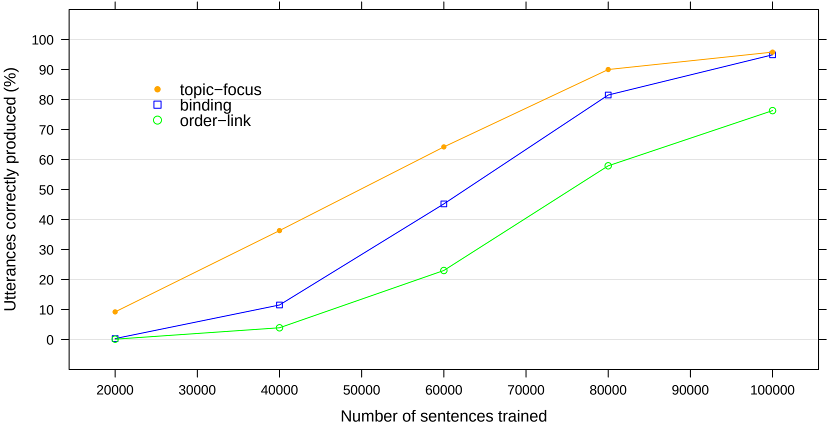
Note: repeated measures—i.e., within-subjects factors—can also be used in factorial ANOVA

Example:

- in previous experiment include **time** as another within-subjects factor
- test whether model learns better (averaged over time) with any one semantics
- test whether model learns **faster** with any one semantics

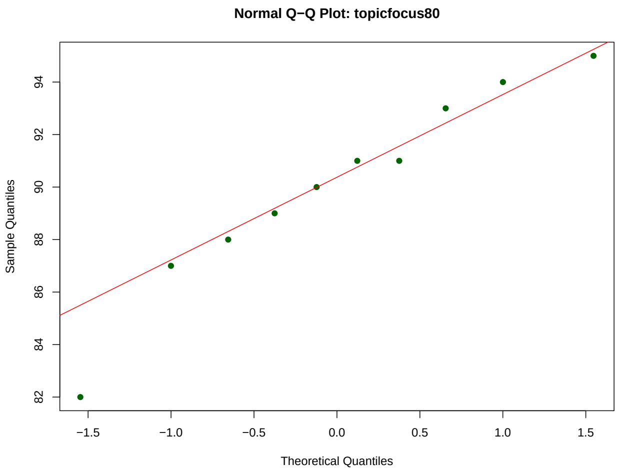
A positive answer is strongly suggested when looking at the model's performance over time, the learning trajectories

Repeated measures in factorial design



Model performance over time (for the three semantics)

Check normality



Check normality and standard deviations for 2×5 subgroups!

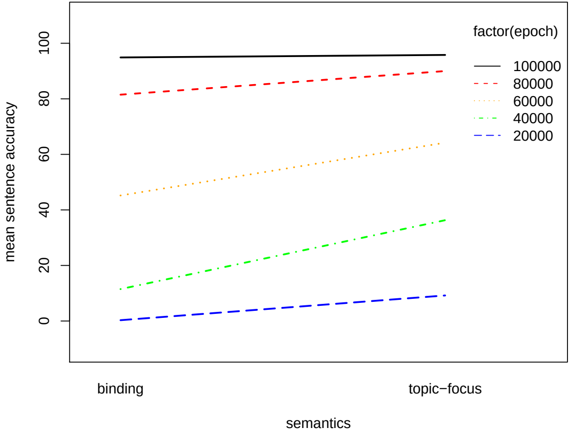
Repeated measures in factorial design

We compare the 'binding' with 'topic-focus' semantics

Conduct a 2×5 repeated measures ANOVA with **time** and **semantics** as within-subjects factors

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
epoch	4	120875.740	30218.935	646.14094	2.22e-16 ***
Residuals	36	1683.660	46.768		
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
semantics	1	3856.4100	3856.4100	13.41262	0.0052167 **
Residuals	9	2587.6900	287.5211		
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
epoch:semantics	4	1785.14000	446.28500	9.49397	2.3996e-05 ***
Residuals	36	1692.26000	47.00722		
—					
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	

Visualizing interaction



Interaction: Although with both semantics model reaches similar proficiency, it learns significantly faster in the topic-focus condition

Mixed factor ANOVA design

Often, subjects divided into separate groups, e.g.,

- gender: male/female
- age: 3/4-year old children
- type of language impairment: Wernicke/Broca aphasia
- mother tongue: Dutch, English, German

but subjects in each group are tested in several conditions

Mixed-factors: n-way ANOVA with between-subjects **and** within-subjects factors

In fact, perhaps the most common ANOVA design (see next example)

Mixed factor ANOVA: example

Withaar & Stowe investigated effects of **syntax** and **phonology** on processing time of relative clauses

Task: read sentences word-by-word on computer screen, press button to see following word. Times between button presses are measured (reading times)

Syntax: difference between relative clause types where

- relative pronouns are understood **subjects**:

  de bakker die de tuinmannen verjaagt

- relative pronouns are understood **objects**:

  de bakker die de tuinmannen verjagen

Phonology: rhyming vs. non-rhyming words in relative clause (Longoni, Richardson & Aiello showed that word lists with rhyming elements take longer to process)

Syntax, rhyme, reaction times

Design: Four kinds of sentences shown, one group of participants per rhymed/non-rhymed, both syntactic structures shown to each group.

		Syntax: within-subjects	
	Phonology	Object Relative	Subject Relative
between-subjects	non-rhym.	non-rhym. obj.-rel.	non-rhym. subj.-rel.
	rhym.	rhym. object-rel.	rhym. subject-rel.

Extras: W&S also controlled for subject's attention span, and for which sentences were shown (no similar sentences shown to same subject)

Measurement: time needed for the last word in relative clause

Data: means and SDs of four groups

rhyming(y/n)	process time obj-rel.	process time subj-rel.
non-rhyming		
Mean	1581.86	1265.90
StdDev	341.82	316.89
rhyming		
Mean	1494.51	1250.55
StdDev	382.45	198.30
Grand Total		
Mean	1538.19	1258.23
StdDev	360.75	261.03

Note: no SD is twice as large as another (but it's close...)
Factorial ANOVA question: are means significantly different?

Sphericity

In repeated measures analyses with **three** or more factors (explanatory variables), the standard deviations/variances (in the repeated measures) have to be comparable per factor.

MAUCHLY'S TEST can be applied to determine if sphericity holds. It's a hypothesis test, so *p*-values below 0.05 indicate that sphericity is violated.

And for only two factors?

Sphericity

In repeated measures analyses with **three** or more factors (explanatory variables), the standard deviations/variances (in the repeated measures) have to be comparable per factor.

MAUCHLY'S TEST can be applied to determine if sphericity holds. It's a hypothesis test, so p-values below 0.05 indicate that sphericity is violated.

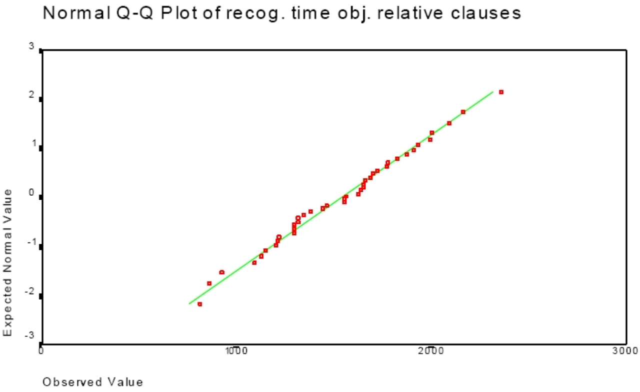
And for only two factors?

Unnecessary!

Normality assumption

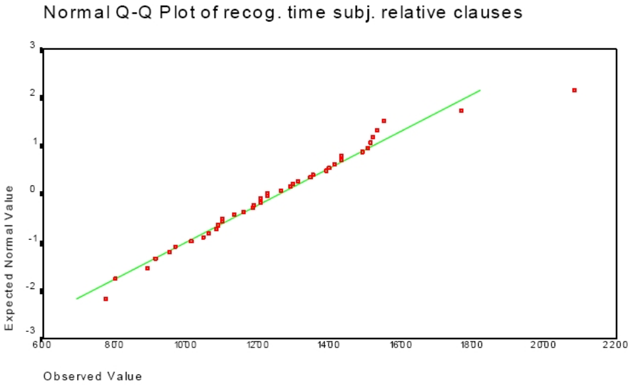
Look at data: are distributions normal?

Rhymed and unrhymed object-relatives



Normality assumption

Rhymed and unrhymed subject-relatives



Remark: longest reaction time good candidate for elimination (worth checking on)

Multiple questions

Again, we ask **two/three** questions simultaneously:

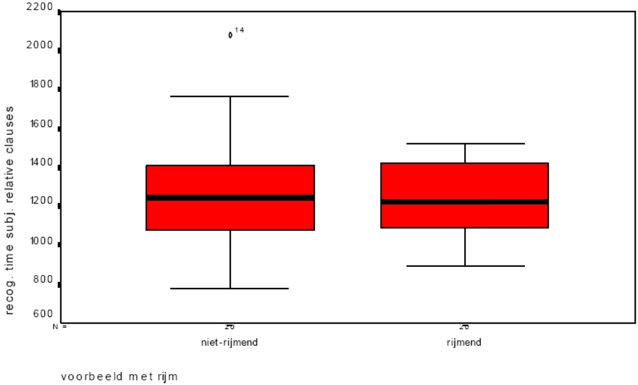
1. Is rhyme affecting word processing time?
2. Do relative clause types affect processing time?
3. Do the effects interact, or are they independent?

Questions 1 & 2 might have been asked in separate one-way ANOVA designs (but these would have been more costly in number of subjects)

Question 3 can only be answered with factorial ANOVA

Visualizing ANOVA questions

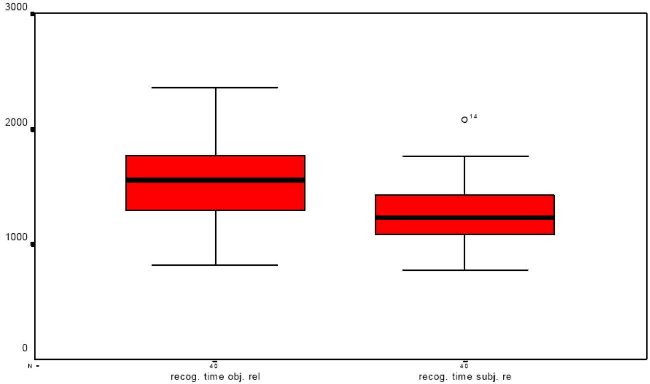
Question 1: Is rhyme affecting processing time?



Note: similar box plots for rhyme in subject-relatives

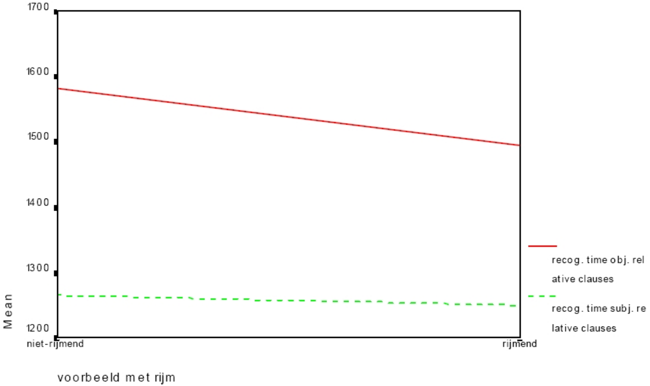
Visualizing ANOVA questions

Question 2: Does relative clause type affect processing time?



Little skew, different medians, large overlap: difficult to tell

Visualizing interaction



If **no** interaction, lines should be parallel. In fact, rhyming speeds processing of object relatives. Multiple ANOVA will measure this exactly.

Mixed-factor ANOVA in SPSS

Syntax: within-subjects factor (repeated measures)
Phonology: between-subjects factor

between-subjects

Phonology	Syntax: within-subjects Object Relative	 Subject Relative
non-rhym.	non-rhym. obj.-rel.	non-rhym. subj.-rel.
rhym.	rhym. object-rel.	rhym. subject-rel.

Invoke: `repeated measures` → define distinct factors → take care not to mix them up!

Mixed-factor ANOVA results

Between-subjects (row) effects (rhyme/no rhyme):

```
* * * * * * Analysis of Variance -- design 1 * * * * * *

Tests of Between-Subjects Effects.

 Tests of Significance for T1 using UNIQUE sums of squares
 Source of Variation        SS      DF      MS      F  Sig of F

 WITHIN+RESIDUAL       6332920     38  166656
 RIJM                    52734      1   52734    .32   .577
```

Hence, rhyme does not significantly affect processing speed

Mixed-factor ANOVA results

Within-subjects (column) effects (object- vs subject-relatives):

```
Tests involving 'SYNTAX' Within-Subject Effect.

 Tests of Significance for T2 using UNIQUE sums of squares
 Source of Variation        SS      DF      MS       F    Sig of F

 WITHIN+RESIDUAL       1321219     38    34769
 SYNTAX                1567532      1  1567532   45.08     .000
 RIJM BY SYNTAX          25917      1    25917     .75     .393
```

Hence, syntax has a profound effect on processing speed; no interaction (in spite of graph!)

Fallbacks

- Suppose sphericity isn't given (Mauchly's test).

Fallbacks

- Suppose sphericity isn't given (Mauchly's test).
  —Greenhouse-Geiser adjusted *p*-values (by lowering the degrees of freedom based on the Mauchly estimation of sphericity)

Fallbacks

- Suppose sphericity isn't given (Mauchly's test).
  —Greenhouse-Geiser adjusted *p*-values (by lowering the degrees of freedom based on the Mauchly estimation of sphericity)
- One-way repeated measures (special case)

Fallbacks

- Suppose sphericity isn't given (Mauchly's test).
  —Greenhouse-Geiser adjusted *p*-values (by lowering the degrees of freedom based on the Mauchly estimation of sphericity)
- One-way repeated measures (special case)
  —Friedman's "ANOVA" uses ranks, like Kruskal-Wallis

Repeated measures ANOVA: summary

Repeated measures ANOVA:

- generalized related-samples t-test
- assumptions like standard ANOVA except for independence
- required whenever a group of subjects measured under different conditions
- eliminates between-subjects variance from MSE
- typical applications:
  - linguistic ability of children measured over time
  - cognitive function in same group of subjects tested under different conditions
  - computational learning models compared for different input environments
- advantage over independent samples: efficient in experimental design

Next week

Next week: correlation and regression