Statistics II
Repeated Measures ANOVA

Çağrı Çöltekin

ideas/examples/slides from
John Nerbonne & Hartmut Fitz

University of Groningen, Dept of Information Science

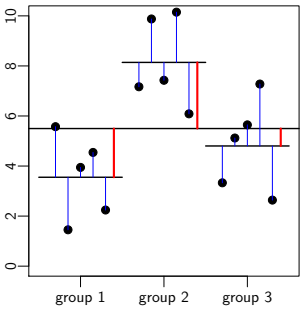
May 22, 2013

Single ANOVA: step by step

ANOVA is applicable when you have numeric observations on more than two independent groups.

- Collect your data: observations should be independent!
- Plot your data: typically, using box and whisker plots (box plots)
- Check for assumptions:
  - observations within each group should be approximately normal
  - the variances of the observations in each group should be approximately equal
- (optionally) set your prior contrasts
- calculate F and associated p-value (run ANOVA in a statistical software)
- (optionally) run pairwise comparisons between each group

Logic of ANOVA



$$F = \frac{MS_{between}}{MS_{within}}$$

$$F = \frac{\frac{SS_{between}}{DF_{between}}}{\frac{SS_{within}}{DF_{within}}}$$

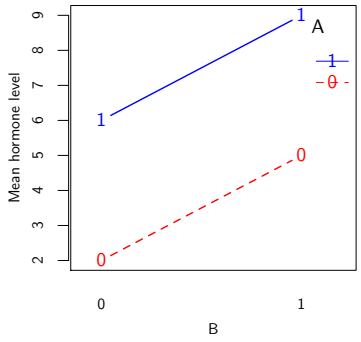
$$DF_{between} = k - 1$$
$$DF_{within} = n - k$$

where k is the number of groups, and n is the number of observations.

Factorial ANOVA

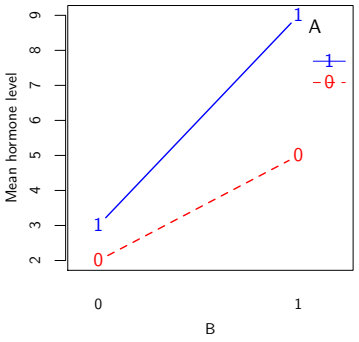
- Factorial ANOVA is a generalization of single ANOVA (or t-test).
- Compare groups along more than one dimension.
- Assumptions: the response variable in all groups
  - is (approximately) normally distributed
  - have (approximately) equal variances
- Efficient in use of subjects.
- Allows to investigate interaction.

Types of interaction (1)



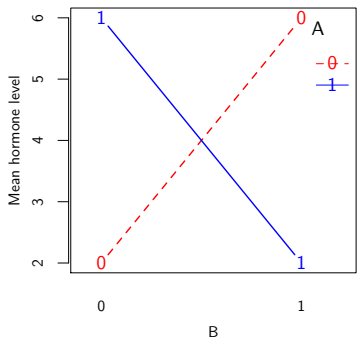
- both drugs have positive effects
- combined effect is additive
- no interaction

Types of interaction (2)



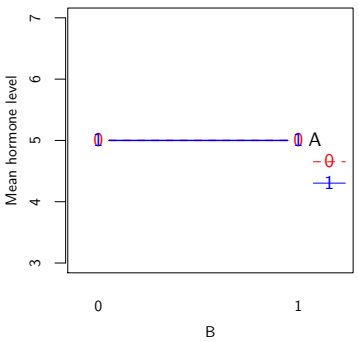
- both drugs have positive effects
- combined effect is stronger than sum of separate effects
- interaction

Types of interaction (3)



- both drugs have positive effects separately
- combination cancel out each other's effect
- interaction

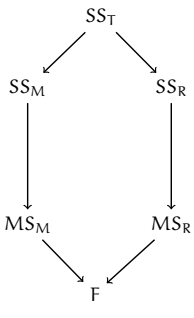
Types of interaction (4)



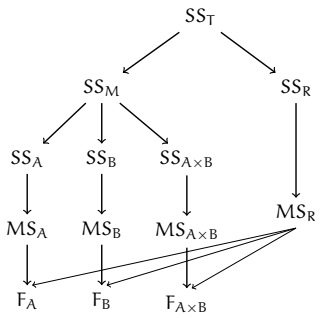
- drugs show no effect
- either separately or in combination
- null hypothesis is true
- no interaction

Partitioning variance in ANOVA

Single ANOVA



Two-way ANOVA



Factorial ANOVA: an example

We will study (yet) another semi-hypothetical example

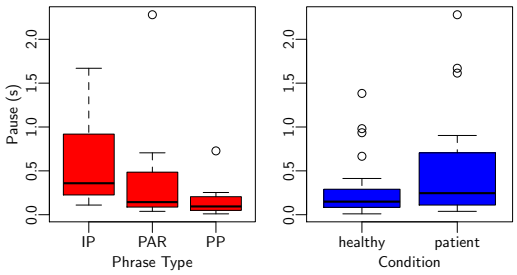
- A linguist wants to know whether three different theoretical constructs, namely, intonational phrases (IP) phonological phrases (PP) and parenthetical expressions (PAR), differ with respect to being isolated from the sentence they are in.
- She also wants to know whether healthy adults and adults diagnosed with aphasia differ in processing of these sentences.
- She prepares a sentences for each phrase type, and records 10 people for each combination of phrase type and health (3x2 design).
- She measures the duration of pauses after the phrase of interest (an obvious sign of isolated phrases) in each recorded utterance.

Factorial ANOVA example: data

We have a numeric response (pause) and two categorical predictors(phrase type and health condition). The data is organized as follows:

participant	pause (seconds)	phrase	condition
1	0.29	IP	healthy
2	0.55	IP	patient
⋮	⋮	⋮	⋮
20	0.09	PP	healthy
21	0.13	PP	patient
⋮	⋮	⋮	⋮
59	0.20	PAR	healthy
60	0.25	PAR	patient

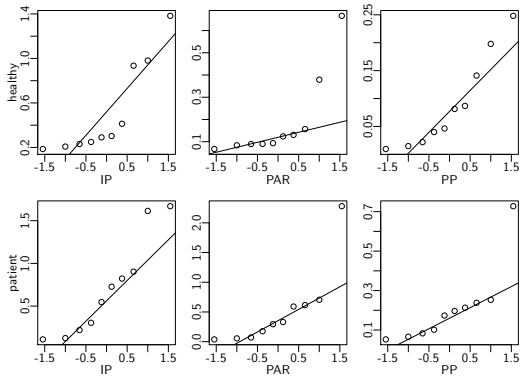
Factorial ANOVA example: data



Anything wrong?

Factorial ANOVA example: normality check

Normal Q-Q plots



Factorial ANOVA example: assumptions

- The data does not seem to be normally distributed.
- The variances seem to differ too (from the box plots). Also:

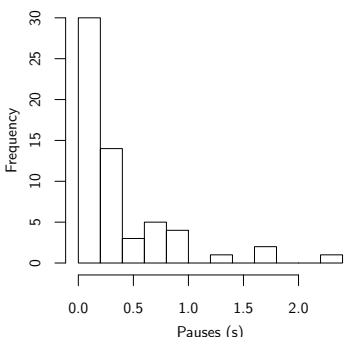
```
Levene's Test for Homogeneity of Variance (center = median)
      Df    F value  Pr(>F)
group  5,54   2.473 0.04341 *
```

Now what?

Factorial ANOVA: when assumptions are violated

- For one-way ANOVA, we'd use a non-parametric alternative (Kruskal–Wallis test).
- Factorial ANOVA does not have a straightforward non-parametric alternative.
- There is one more possibility: transforming your data.
- It is not always easy to find a simple data transformation that corrects the problems. But, when you do, you do not also lose poser as in non-parametric tests.

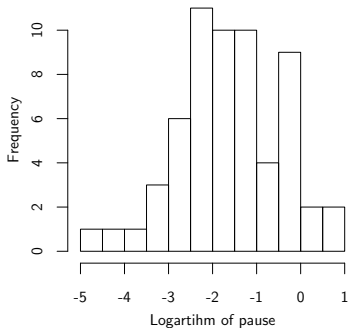
Another look at the data: histograms



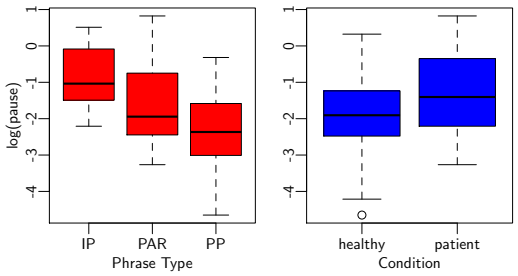
- Only positive values.
- Most data clustered around a narrow range.
- Skewed distribution with a long tail.
- This type of distributions (exponential or power-law) are commonplace in linguistics: (word) frequencies, reaction times . . .

Transforming the data

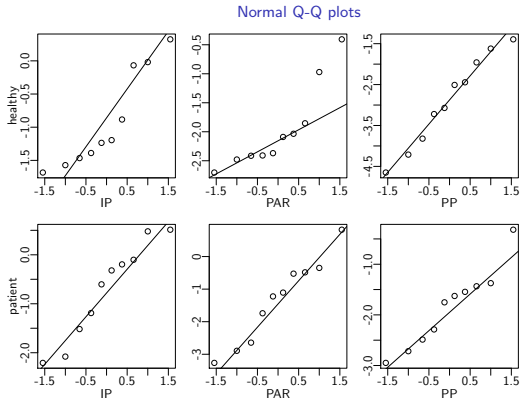
- Log transformation is one of the best common options.
- It generally does a decent job on power-law or exponential distributions.
- For different data you may need other transformations (e.g., square-root, or square) but as common as log-transformation.



Back to the example: box plots



Example: Q-Q plots with log transform



Example: homogeneity with log transform

```
> leveneTest(log(pause)~condition*phrase.type, data=pause.data)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5,54  1.2545  0.297
```

Levene's test confirms the box plots: there is no evidence for non-homogeneity of variances.

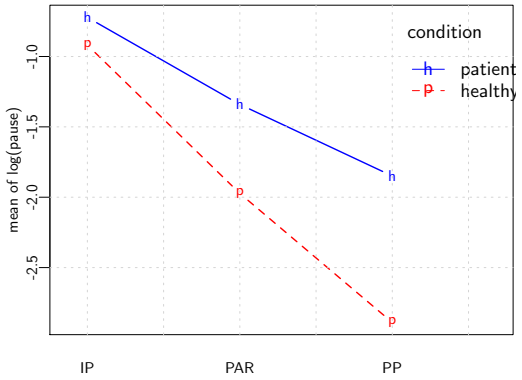
Example: running ANOVA

R output:

```
> summary(aov(log(pause)~condition*phrase.type, data=pause.data))
                  Df Sum Sq Mean Sq F value  Pr(>F)
condition          1   5.83   5.826   6.318   0.015 *
phrase.type        2  24.05  12.024  13.040 2.4e-05 ***
condition:phrase.type  2   1.78   0.891   0.966   0.387
Residuals         54  49.80   0.922
```

- Significant main effects.
- No significant interaction.
- We could run the analysis using contrasts.
- Now, you may want to do post-hoc comparisons.

Visualizing interaction



Repeated-measures ANOVA: motivation

- In (factorial) ANOVA, our observations has to be independent.
- Consider our earlier ANOVA example (now we are only interested in the phrase type).
- If we present the data in this form.

	Phrase		
Subject	IP	pp	PAR
1	0.29		
2	0.55		
⋮	⋮	⋮	⋮
20		0.09	
21		0.13	
⋮	⋮	⋮	⋮
59			0.20
60			0.25

It is clear that not measuring every subject in all three condition is wasteful!

RM ANOVA: motivation

- In repeated-measures ANOVA, we measure each subject (participant) in each condition.
- Independence of observations is not required (or desired).
- A lot more economical in experiment design.
- More powerful, since individual variation is not a problem for RM ANOVA.
- A generalization of paired t-test to multiple groups.
- Probably, most common analysis method in psycholinguistics and in general experimental sciences.
- Power comes with a price: more strict assumptions.

Example applications

Examples will be similar to single or factorial ANOVA. Repeated measures can be

over time: testing effects of treatment, teaching method or just time. Typically you get more than two pretests – post-tests to

- make sure the pre- or post- tests are stable.
- check short- and long-term effects.
- check whether/how effect diminishes after the intervention.

no time related. Examples:

- reaction time for different sort of stimuli
- left or right side of the brain in an ERP study
- measurements taken in the same city/region/country

beware of carry-over effects!.

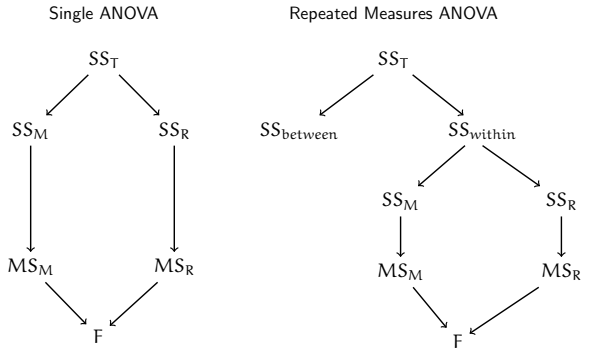
Why not always use RM?

- Carry-over effects: learning or fatigue effects during the experiment.
- Some conditions are not easy or even possible to repeat: think about healthy vs. aphasia patient in our example, or, gender differences.
- RM ANOVA assumptions/requirements are stricter. You rarely see RM ANOVA used in non-experimental studies.

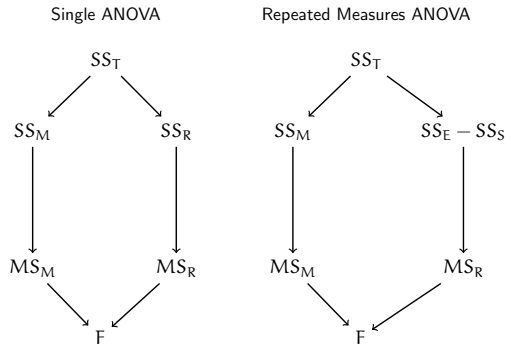
Between subjects and within subjects variance

- A between subjects variance is the variation you observe due to differences between individuals.
- In independent (single or factorial) ANOVA, all variation observed is between subjects.
- A within subjects variation is due to variation observed in repeated measurement over the same subject.
- In a purely repeated design ANOVA, all experimental effect is confined in within-subjects variance.

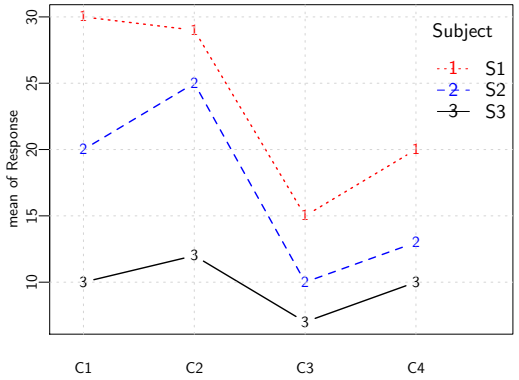
Partitioning variance in RM ANOVA



Partitioning variance in RM ANOVA (2)



Another look at between and within subject variation



Degrees of freedom for RM ANOVA

For n observations, m subjects, and k groups (conditions):

$$\begin{aligned} DF_T &= DF_M + DF_R \\ n-1 &= k-1 + n-k-(m-1) \\ n-1 &= k-1 + (k-1)\times(m-1) \end{aligned}$$

Note: residuals are the interaction between subjects and the experimental conditions.

Assumptions of RM ANOVA

- Normality of response variable in each group.
- Sphericity: variances of pairwise differences between each experimental condition must be approximately equal.
- Each subjects has to be tested in all conditions.
- RM ANOVA is sensitive to missing values, unequal group (cell) sizes.

Sphericity

Sphericity states that for all levels of within-subjects predictor (the experimental condition in RM ANOVA), variances of the pairwise differences of the response variable should have (approximately) equal variances.
If we had three conditions A, B and C:

$$\text{Homogeneity:} \quad \sigma^2_A \approx \sigma^2_B \approx \sigma^2_C$$
$$\text{Sphericity:} \quad \sigma^2_{A-B} \approx \sigma^2_{A-C} \approx \sigma^2_{B-C}$$

The test for sphericity is called 'Mauchly's sphericity test'. As tests for homogeneity and normality, significant p-value means violation of the assumption.

An example

We will use the earlier example on pauses and phrases, and turn the experiment into a RM ANOVA design:

- A linguist wants to know whether three different theoretical constructs, namely, intonational phrases (IP) phonological phrases (PP) and parenthetical expressions (PAR), differ with respect to being isolated from the sentence they are in.
- She prepares a sentences for each phrase type, recruits 30 participants.
- Each participant is recorded with all three sentences.
- She measures the duration of pauses after the target phrase in each recorded utterance.

Example: the data

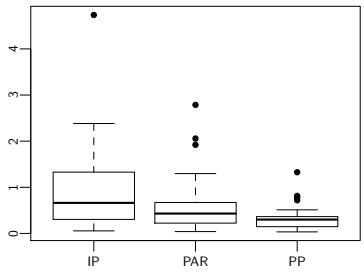
	Phrase		
Subject	IP	PAR	PP
1	1.70	0.18	0.30
2	2.14	1.30	0.35
⋮	⋮	⋮	⋮
29	1.38	0.44	0.36
30	0.86	0.55	0.81

Note: we have three times the observations we'd get otherwise.

Example: the data

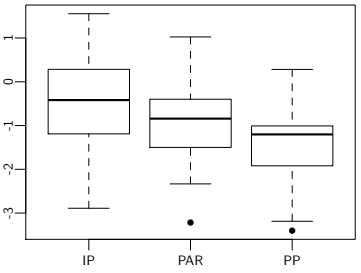
subject	pause (seconds)	phrase
1	1.70	IP
1	0.18	PAR
1	0.30	PP
⋮	⋮	⋮
30	0.86	IP
30	0.55	PAR
30	0.81	PP

Example: box plots



Familiar?

Example: box plots (after log transform)



Example: assumptions

Normality check using Q-Q or P-P plots.

Sphericity Check using Mauchly's Test (after log transform)

```
'Mauchly's Test for Sphericity'
  Effect         W         p p<.05
2 phrase 0.9522092 0.5037942
```

If sphericity assumption fails, we need to use corrected F scores (e.g., Greenhouse–Geisser correction).

Example: ANOVA result

```
> summary(aov(log(pause) ~ phrase + Error(subj), data=d))
Error: subj
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 29  32.69   1.127

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
phrase     2  12.82   6.412   7.391 0.00138 **
Residuals 58  50.32   0.868
```

Next: pairwise comparisons (unless we answered our questions with prior contrasts).

RM ANOVA: summary

- RM ANOVA is applicable when using multiple correlated observations (with experimental manipulation).
- RM ANOVA is more efficient:
  - Reduced residual variance by accounting for subject variation.
  - More efficient use of subjects (multiple observations per subject).
- RM ANOVA also has more strict requirements, and more sensitive to unbalanced data.
- RM ANOVA can be factorial
  - Multiple within subject predictors.
  - within subject and between subject predictors (mixed-ANOVA design).

Next week: Logistic regression.