What if you want to test different variables within one experiment? How to deal with repetitions?

Analysis of variance (ANOVA), repeated-measures ANOVA.

Katalin Mády

Pázmány Summer Course in Linguistics for Linguistics Students, Piliscsaba

28 August, 2013

Analysis of variance (ANOVA)

Questions: (1) Do groups differ significantly from each other (generalisation of *t*-test)? (2) Does the investigated factor has an effect on the data?

- **One-way ANOVA:** there is one independent variable with more than two levels (e.g. comparison of young, mid and old speakers).
- **Two-way, *n*-way ANOVA:** combination of levels of more than one independent factors.
- **Repeated-measures ANOVA**: if more than one type of data is collected from the same subject/object of investigation.

The above types of ANOVA's are univariate, i.e. there is only one dependent variable.

Applications

- The effect of different types of a given treatment compared to the control group (e.g. higher dose, lower dose, placebo).
- Efficiency of several methods compared to each other and to the control group.
- Effects of independent nominal variables (e.g. the effect of semantic factors on reaction time).

Requirements

- Normal distribution within each group,
- homogeneity of variances,
- independency of observation types from each other (sphericity).

The test of normality is usually not regarded as a necessary condition, because (1) if $n > 30$, data are usually distributed normally anyway, (2) in a sample size of 10 to 20, the deviation from normal distribution is usually small, (3) a sample with $n < 10$ elements does not have a real distribution.

But: homogeneity of variances and the independency of observations (sphericity) are essential requirements, otherwise results are not reliable.

One-way ANOVA

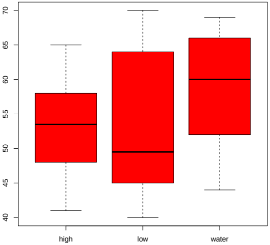
Procedure: the variance of the entire sample is divided to variances within the groups of factor combinations and between these groups → analysis of variances.

1. Variance within the groups is based on the sum of squares of deviations from the mean → same as calculation of variance.
2. Variance is calculated between the group means → estimation of random error (where the regression model is not fitted well), which is the residual variance of the regression model.
3. Variances within and between the groups are compared by an F-test.
4. Decision: if the variance **between the groups** is significantly larger than the variance **within the groups**, then the independent variable has an effect.

Example

The efficiency of a certain fertiliser on the growth of plants is to be tested.

Plants were watered with a high and a low concentration of the fertiliser for a month, the control group with water. Question: does the fertiliser increase the growth of plants?



Example

Test of normality: all distributions are normal.

Test of homogeneity of variances, here: Bartlett-test (more than one group), shows that variances are equal.

F-value for variances between treatments and within treatments is not significant, $p = 0.313 \Rightarrow$ the amount of fertiliser has no effect on growth.

Is this also true for the comparison of the categories 'high concentration' and 'water'?

Post-hoc tests

General problem: with a higher number of comparisons between groups, the probability of α error is also higher.

If we compare two groups (t-test), the probability of the Type I error is 0.05 → the probability of no error is 0.95. For two tests the probability of no error is $0.95 * 0.95 = 0.9025$, for 6 tests it is 0.73, for 10 tests 0.6.

Thus, the probability of getting false-positive significances with 10 tests is $p = 1 - \alpha = 0.4$.

Ten comparisons are easy to reach, i.e. with an independent factor with 5 levels.

Methods to avoid this problem

- Pairwise comparison of the groups by t-tests and application of Bonferroni-correction:

  significance level α has to be divided by $\frac{k(k-1)}{2}$,
  i.e. confidence interval / number of all possible comparisons.

  Disadvantage: if there is a high number of possible comparisons, it is almost impossible to get a significant difference.

- Tukey's Honest Significant Difference which is less conservative.

Post-hoc test

1. Tukey's HSD

low-high	0.9947263
water-high	0.4086354
water-low	0.3572213

The pairwise comparison does not show a significant difference.

2. *t*-test with Bonferroni-correction

We are interested in the comparison of high concentration vs. water. Number of possible combinations = 3, thus confidence interval = 0.05/3 = 0.0167.

p = 0.4462, which is clearly above significance level.

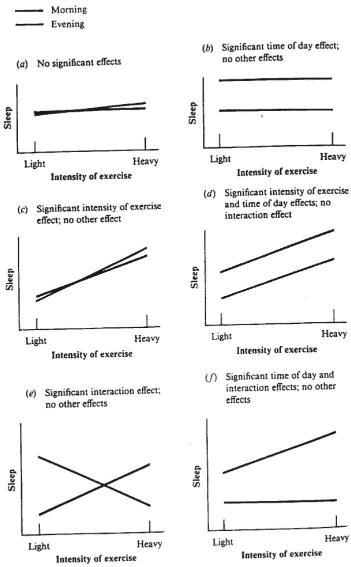
n-way ANOVA

The effect of two or more independent variables on observations.

Null hypotheses: (1) First factor (independent variable) has no effect on dependent variable. (2) Second (or n^{th}) variable has no effect on dependent variable. (3) The two factors have no effect on each other, i.e. there is no interaction between them.

Procedure: first the interaction between two factors is tested, then their effect one after the other.

Interaction



(*a*) No significant effects

(*b*) Significant time of day effect; no other effects

(*c*) Significant intensity of exercise effect; no other effect

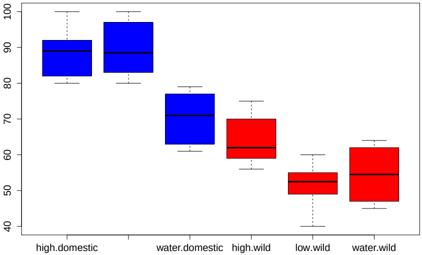
(*d*) Significant intensity of exercise and time of day effects; no interaction effect

(*e*) Significant interaction effect; no other effects

(*f*) Significant time of day and interaction effects; no other effects

Example

Now we test this fertiliser on the wild and the domestic version of another plant.



Two-way ANOVA:

fertiliser	1.20e-07
type	2e-16
fertiliser:type	1.32e-05

Both factors have a significant effect, and they are in interaction.

Interpretation

Decision for H_1: the fertiliser leads to significantly faster growth for both sorts of plants.
Question: do we achieve more growth when using the high concentration?

Interpretation

Decision for H_1: the fertiliser leads to significantly faster growth for both sorts of plants.
Question: do we achieve more growth when using the high concentration?

Procedure: comparison of effect of water and low concentrated fertiliser on 1st and 2nd plant by Tukey's HSD. There is no way to compare only these, due to the α multiplication problem.

TukeyHSD

2 × 3 factor design has 15 possible combinations, but we are only interested in the following:

water:domestic-low:domestic	0.0000005
high:domestic-low:domestic	0.9979237
water:wild-low:wild	0.9352339
high:wild-low:wild	0.0014372

The low concentration of the fertiliser enhances the growth of the domestic plant significantly, but not that of the wild plant. For the wild plant a high concentration is needed to achieve significantly more growth.

Repeated-measures ANOVA

The above methods assume that data are collected from independent groups. This is seldom the case in linguistics where we normally collect several kinds of data from the same participant. Here normal ANOVA is not appropriate. The setting is equivalent to independent and paired *t*-tests.

The equivalent method to the paired *t*-test is the repeated measures ANOVA.

Important: repeated measures do not refer to the fact that we require the same kind of data several times from the same subject (e.g. a repeated reading of sentences), but that repeated measures are performed with **one and the same person**.

E.g. treatments in medical sciences: the effect of a given drug before treatment, two weeks after first treatment, one month after first treatment etc.

Procedure

Testing one dependent and one or more independent variables where the differences between the **within**-factors (people, plants, objects that were used for repeated measures) are regarded as a **random** factor → *within*-subjects factor.

The method also allows for a comparison between two groups, e. g. speakers of different languages, different sorts of the same plant etc. This is the *between*-subjects factor.

The random error that is caused by subject variation is added to the formula of ANOVA.

Requirements

- At least five subjects.
- One single datum per factor combination: that means that if there were repeted observations for the same cell with the same subject (i.e. five repetitions of the same words), the mean has to be calculated for each subject and cell.
- Balanced design: if one factor has two levels, then the same levels have to be investigated for the other factor.

Disadvantages

- Cell means have to be calculated by hand – or by a self-written function in R.
- Since only the means are taken into account, variances within cells cannot be taken into account.
- It is not possible to combine several within-subject factors, e.g. subjects and words.
- This method can only be applied if sphericity is given. Otherwise a repeated-measures multivariate ANOVA has to be applied.
- There is no post-hoc test, only *t*-tests with Bonferroni-correction can be used.

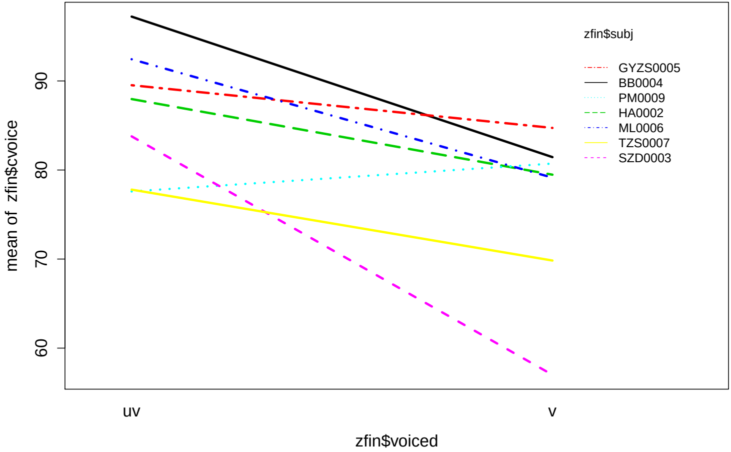
Example

The voiced proportion of the sounds /s/ and /z/ were measured in sentence-final position in Hungarian words. Question: are sentence-final /z/ sounds more voiced than sentence-final /s/ sounds? (The answer is not obvious, due to a large amount of phrase-final devoicing.)

- Dependent variable: duration of voicing in consonant.
- Independent variable: voicing.
- Within-subject factor: speaker.
- Between-subject factor: none.

Example

Plotting tendencies within the same subject:



Means for voiced proportion for each subject, uv: unvoiced, v:

Example

Does the voicing of the preceding consonant has an effect on voicing proportion?

Table of means has to be calculated again.
`Error: subj:voiced` voicing durations for each subject, dependent on voicing (p has changed because of recalculated cell means).

`Error: subj:c1` voicing durations for each subject depending in previous consonant.

`Error: subj:voiced:c1` voicing durations for each subject, interaction of voicing and preceding consonant, i.e. does the preceding consonant have an impact on the effect of voicing?

Sphericity

Assumption of sphericity:

Variances of the differences between treatment levels are equal.

This is often not the case in repeated-measures experiments. E.g. if you compare vowel durations for 10 speakers with normal and fast speech rate, then you are likely to get different variations for the two sets.

Test of sphericity by Mauchly's test.

Greenhouse-Geisser-test can be used (which is implemented in SPSS).

Repeated-measures MANOVA

Repeated-measures multivariate ANOVA

It is used when sphericity assumption is violated, or if we don't want to care about it.

MANOVA includes two dependent variables: one is the dependent variable we want to test, the other is the covariance between the different levels of the independent variable.

RM MANOVA is implemented in R and is described in detail in Field et al. (2012).

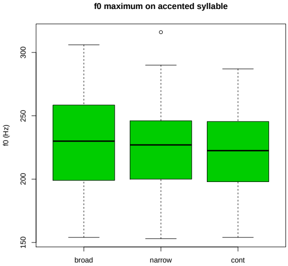
Unsolved problems

Repeated-measures MANOVA is not always applicable:

- If there is more than one *within*-factor (e.g. subject and stimulus word).
- If dependent data are ordinal.
- If there are empty cells, i.e. there are empty combinations of factors or there are missing values.

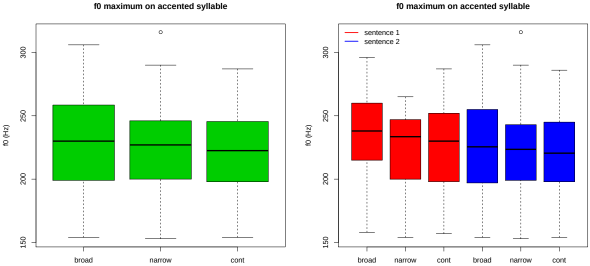
Example

The maximum of fundamental frequency on the accented syllable of a focus word. There are two sentences and 7 speakers.



Example

The maximum of fundamental frequency on the accented syllable of a focus word. There are two sentences and 7 speakers.



There is a tendency to have lower f0 peaks in the order broad > narrow > contrastive focus (this is true data!), but only in the first sentence.

It would be good to treat both subject and sentence as a random factor, but this is not possible in RM ANOVA.

Mixed models

Advantages:

- no restrictions with respect to distribution or the scale of dependent variable,
- several random factors (here: random effects) can be included,
- no necessity to calculate cell means.

Disadvantages:

- The degrees of freedoms cannot be calculated which would be necessary to compute a p value.
- It is only reliable for large samples where > 200.

Procedure

Fixed effects: independent variables.
Random effects: *within*-subject factors.

Method: comparison of fixed effects, assumption: variance of random factors is unknown.

Model is based on residual maximum likelihood. These are estimates of the components of the variance whose goodness is tested.

Results for above example

Fixed effects:

```
                   t value
  focuscontrastive  -2.422
       focusnarrow  -1.825
```

Factor levels are ordered alphanumerically, i. e. the order is broad $<$ contrastive $<$ narrow. All levels are compared to the first one which is taken as the basis of the comparison (here: broad).

The obtained *t* values are relevant for the comparisons broad vs. contrastive, broad vs. narrow. For the contrastive vs. narrow comparison, order has to be changed, and contrastive has to be put on the first place.

Decision

If the significance level is set to 0.05, and the test is two-sided, we need the t value for $p = 0.025$.

But: in order to calculate t, the degree of freedom is needed. But this model does not include df.

Workaround: t changes little from $df \geq 60$. Thus, the t value for this sample size is calculated and is regarded as the critical value, which is $t = 2$.

Note that the t-distribution is symmetrical, thus the absolute values of the t values have to be compared.