What to do if your data do not rely on accurate measurements?

Non-parametric tests. Overview and general questions.

Katalin Mády

Pázmány Summer Course in Linguistics for Linguistics Students, Piliscsaba

28 August, 2013

Non-parametric tests

Applications:

- for nominal data (frequencies),
- for ordinal data,
- for metrical data that are not normally distributed,
- for metrical data whose variances are not homogeneous.

There is some discussion about the so-called Likert-scales (e.g. naturalness judgements on a scale from 1 to 5). It is always advisable to do a non-parametric test if you decide to regard the scale as parametric, just in case.

χ^2-test

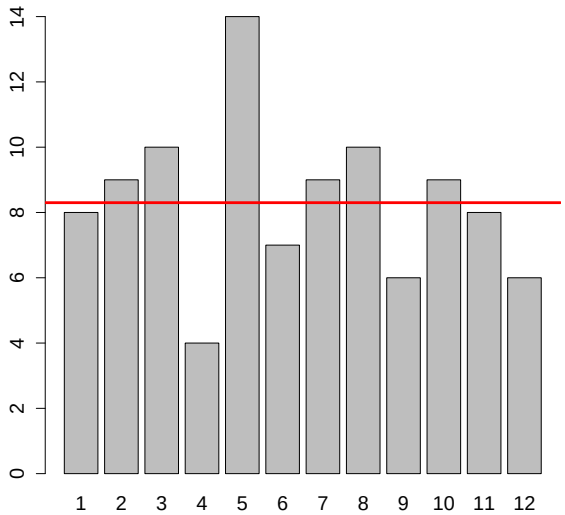
Fits the distribution of frequencies of one or two samples with nominal data to the χ^2 distribution. One sample should contain at least 5 elements.

One sample: χ^2-test of distribution. The frequency of observations is compared to the expected frequency, i. e. n/k.

E.g. is the number of children born in a certain month equally distributed?

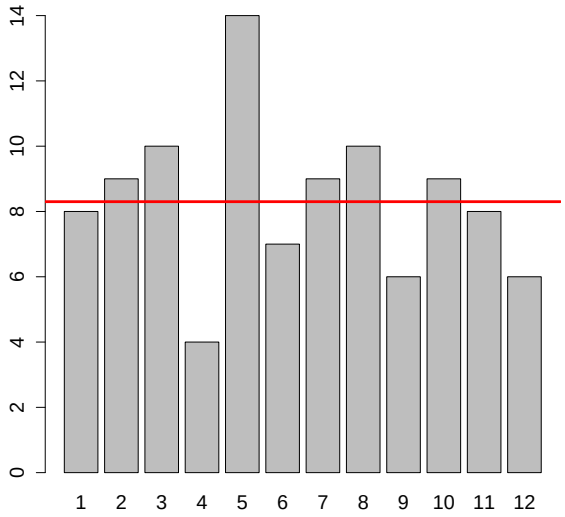
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
obs.	8	9	10	4	14	7	9	10	6	9	8	6
exp.	8,3	8,3	8,3	8,3	8,3	8,3	8,3	8,3	8,3	8,3	8,3	8,3

Example



Can the data be fitted by the expected frequency?

Example



Can the data be fitted by the expected frequency?

Result: $p = 0.6698$.
Since $p > 0.05$, the hypothesis of equal distribution is not rejected.
This is the case even if May is obviously a preferred month by newborns.

If $p < 0.05$, the observed frequencies cannot be fitted to the expected frequencies. Here at least one value has to be extremely high or low (e.g. 20 children in May).

χ^2-test for two samples

χ^2-test for the test of independency: are frequencies independent of the categories of the nominal scale?

Here observed frequencies are not compared to the expected frequencies, but to the other sample.

Example

An English text of ca. 1 million words with part-of-speech tagging. How can we identify typical verb + particle structures?

Procedure: we compute all occurrences of the verb with and without the particle. Then we compute all occurrences with the particle with and without the verb. Frequencies are presented in contingency tables.

observed frequencies for ‘find up’

	up	-up
find	4	1524
-find	2121	1197267

expected frequencies for ‘find up’

	up	-up
find	2.7	1525.3
-find	2122.3	1197265.7

$\chi^2 = 0.6$ which is far above significance level.

observed frequencies for ‘find out’

	out	-out
find	109	1419
-find	2169	1197219

expected frequencies for ‘find out’

	out	-out
find	2.9	1525.1
-find	2275.1	1197112.9

$\chi^2 = 3896.3$ which is far below significance level.

Rank tests

Basic idea: statistics is not calculated from observed measures, but from their rank $\sim$ Spearman's ρ.

Applications:

- if dependent variable is ordinal,
- if metrical data do not have normal distribution.

Requirements: data are comparable, i. e. the density functions have a similar shape, and the equality of the variances is given.

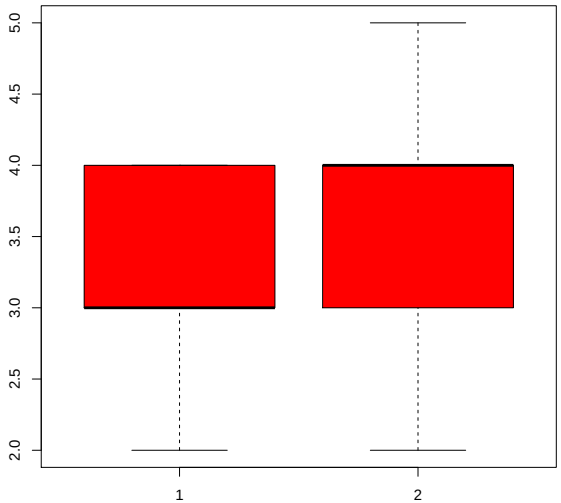
Types of tests

- Mann-Whitney test/U-test: equivalent of two-sample t-test: two ordinal or not normally distributed metrical samples.
- Wilcoxon test: equivalent of paired t-test: two ordinal or not normally distributed metrical observations with the same subjects.
- Kruskal-Wallis test, H-test: equivalent of one-way ANOVA: one independent variable with more than two levels.

Example: Mann-Whitney test

A Hungarian teacher is not satisfied with her class in this year. She thinks that the parallel class has much better grades in geography. She compares their grades. Let's assume that 1 is the worst and 5 is the best grade, for simplicity's case. Both classes have 30 students.

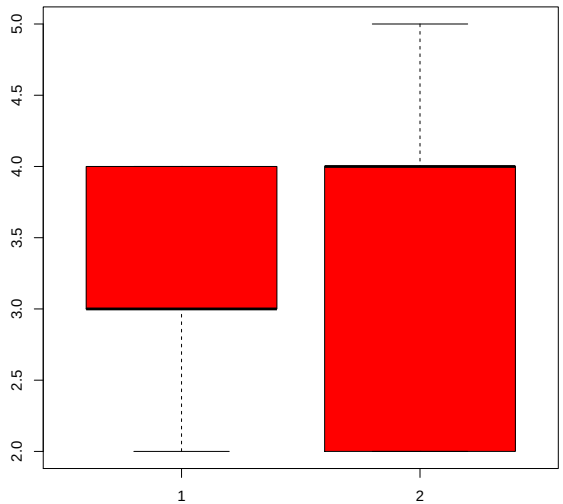
Since school grades are ordinal data, a t-test is not appropriate here.



Mann-Withney test: $p = 0.02993$, thus H_0 can be rejected: the other class's grades are indeed better than those of this class.

Example: Wilcoxon test

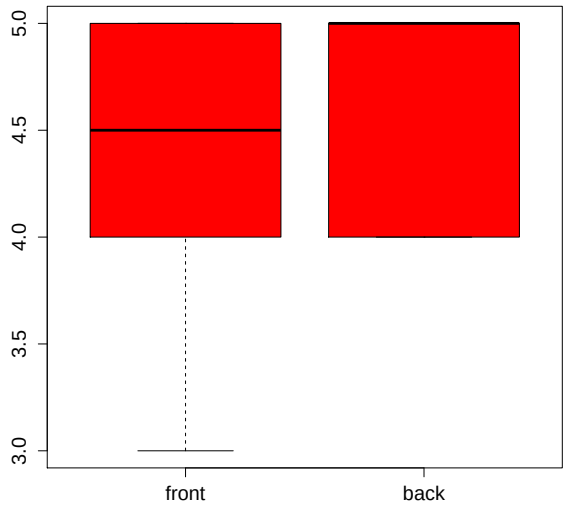
The teacher realises that her class has just received a new teacher in geography this year. She wants to know whether this has an impact on the grades of students compared to last year.



Wilcoxon test: $p = 0.2515$, thus H_0 cannot be rejected: the students did not like geography more last year than now.

Example: Wilcoxon test

How acceptable are the forms *hotelbe* and *hotelba* in Hungarian? These are alternative forms that are theoretically both correct, but some speakers prefer one over the other. Acceptability judgements on a 5-point scale from 10 subjects:



Since subjects gave scores on both forms, this is a paired sample, thus the Wilcoxon-test is appropriate here.

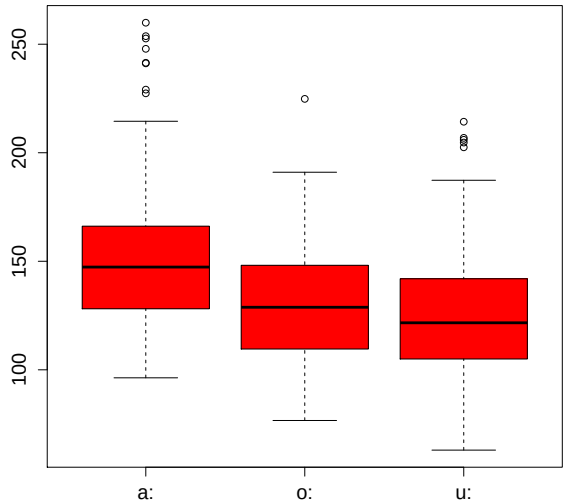
$p = 0.017$, thus judgements can be regarded as significantly different.

Example: Kruskal-Wallis test

Data set longvow.RData on clara.nytud.hu/∼mady, seminar 7. Durations of long Hungarian /a o u/ are listed. Do durations differ significantly?

First, normality assumption has to be tested.

Test of normality gives highly significant p for /u:/ and /a:/, thus the assumption for an ANOVA is not met $\rightarrow$ it's safer to use the Kruskal-Wallis test.



$p < 0.0001$, thus vowel durations differ significantly from each other.

How to find the appropriate test?

You find a cheet sheet here:
http://clara.nytud.hu/~mady/courses/statistics/scills2013/statchs_balint_en.pdf

Should I use statistics?

- Many linguists have spent happy lives without it – you can do the same.

Should I use statistics?

- Many linguists have spent happy lives without it – you can do the same.
- But if you are young, you are likely to need it, since it's a current trend that will hopefully hold on ☺.

Should I use statistics?

- Many linguists have spent happy lives without it – you can do the same.
- But if you are young, you are likely to need it, since it's a current trend that will hopefully hold on ☺.
- If you are a corpus linguist, there is no way around it.

Should I use statistics?

- Many linguists have spent happy lives without it – you can do the same.
- But if you are young, you are likely to need it, since it's a current trend that will hopefully hold on ☺.
- If you are a corpus linguist, there is no way around it.
- If you are a speech scientist, you will never get a job without experience in statistics.

Should I use statistics?

- Many linguists have spent happy lives without it – you can do the same.
- But if you are young, you are likely to need it, since it's a current trend that will hopefully hold on ☺.
- If you are a corpus linguist, there is no way around it.
- If you are a speech scientist, you will never get a job without experience in statistics.
- If you are a phonologist, it enriches your life and makes you cool.

Should I use statistics?

- Many linguists have spent happy lives without it – you can do the same.
- But if you are young, you are likely to need it, since it's a current trend that will hopefully hold on ☺.
- If you are a corpus linguist, there is no way around it.
- If you are a speech scientist, you will never get a job without experience in statistics.
- If you are a phonologist, it enriches your life and makes you cool.
- If you are a sociolinguist, your research will be far more relevant.

If you like statistics. . .

. . . but statistics does not like you yet:

- Get a good book that explains things in a simple way (books for pschycologists contain most relevant methods for linguistics).
- Think of a simple experiment and try to do it yourself.
- Ask someone from your university (in statistics, psychology, phonetics etc.). Don't be shy to contact people outside of your own department, they will be happy to help you.
- Write me an email.

What kinds of experiments can you do?

There are possibilities in all subfields of language and speech research.

- There are areas where you won't or shouldn't get away *without* statistics, such as corpus linguistics, phonetics, modern psycho- and sociolinguistics. These are the ideal fields for using parametric tests since you deal with measurable or countable data.
- Historical linguistics: there are interesting questions on the size of shared vocabulary or language similarity. There is a current trend to use models for language change such as agent-based modelling.
- Syntax and semantics: acceptability scales/naturalness judgements. Quantitative analysis of corpus data. Reaction time experiments. Reaction time and acceptability judgements can be combined.

What designs are appropriate?

Rule nr. 1: Don't go out and start collecting data in a random manner! It will be a desaster to analyse them.

1. Formulate your hypothesis H_1.
2. Check whether you have considered all potentially relevant factors that can have an impact.
3. Chose the statistical method you are going to use before starting to collect data.
4. Avoid answering all possibly relevant questions within the same experiment. Why? A higher number of independent variables and levels reduce your chance to receive significant results where you want to see them.
5. It is legitimate to analyse only one part of your data to reduce too many factor combinations. But it is not legitimate to analyse them within the same model and just ignore the part you are not interested in.

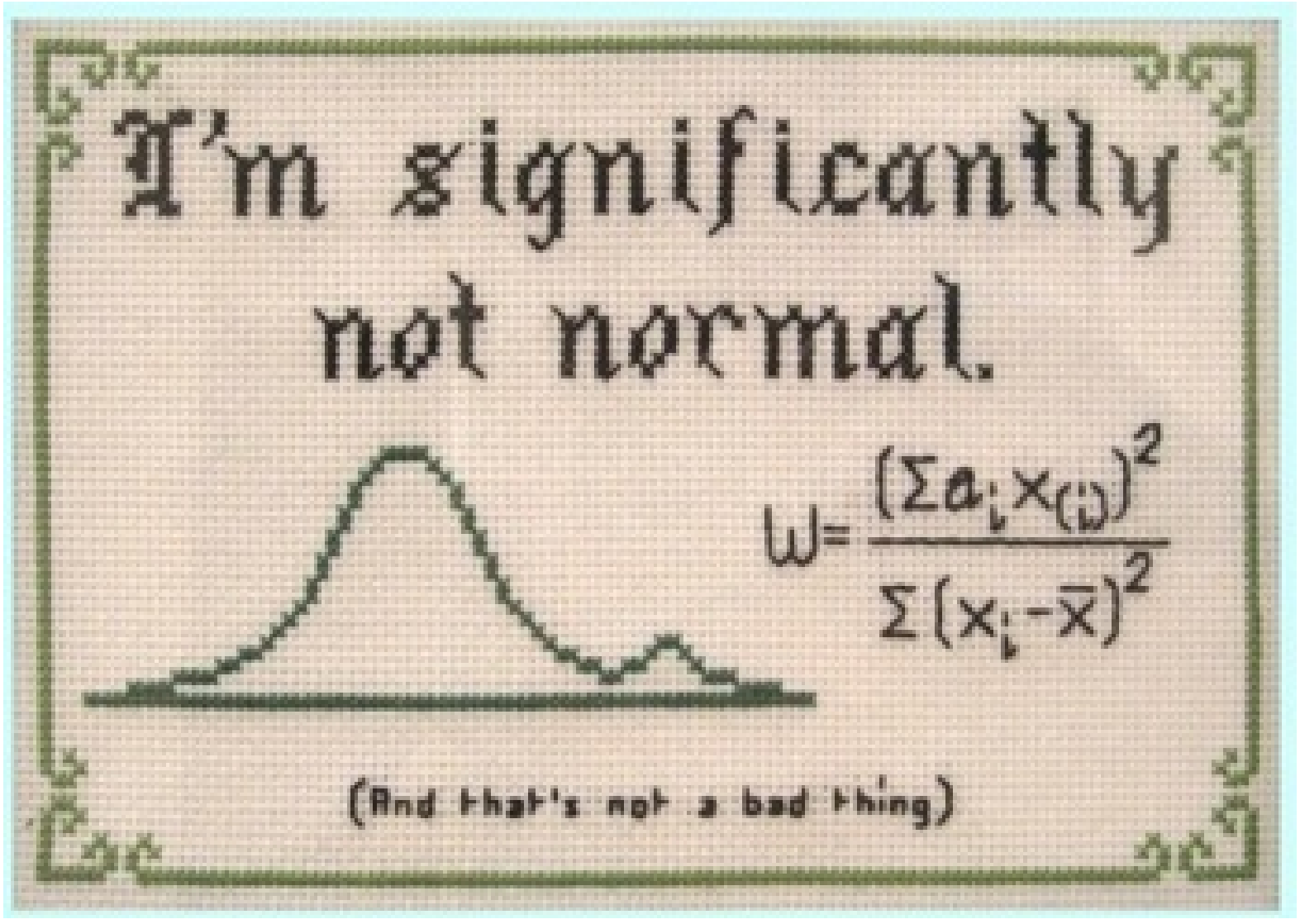
Possible pitfalls

- You use a Likert-scale with 4 points. Remember that these scales can be regarded as parametric of they contain at least 5 points. Even then, it is advisable to use non-parametric tests along with parametric ones.
- Never use a normal ANOVA for data with repeated measures. In fact, it is best to use repeated-measures MANOVA instead.
- Many people think that you can get around all problems with mixed-effect models. But remember: they are only reliable when you have at least 200 observations. The higher the number of factor combinations, the less reliable the outcomes.
- What if you don't get the results you expected? Of course, your hypothesis might just have been wrong. This does happen. But think of the β error. You might have been unlucky. Try to get a larger sample or a new sample with the same design.

How to write up things?

Since your results must be replicable by others, it is essential to describe your methods and materials very carefully. Field et al. (2012) give useful advice for each method.

- Read a few experimental papers in your field to get an impression of style and the way you do it.
- Look at conference papers. Many conferences publish 4-page papers with lots of experiments. Just an example with some syntax-relevant papers: the Speech Prosody conference, the last one in 2012: http://www.speechprosody2012.org/page.asp?id=150#PG_PS1
- Be as exact as possible. Give all details that might be relevant to others.
- Ask someone from your field to read your methods chapter. If something is unclear to her/him, it will be unclear to your readers.


I'm significantly not normal.
$W = \frac{(\Sigma a_i x_{(i)})^2}{\Sigma (x_i - \bar{x})^2}$
(And that's not a bad thing)