An introduction to statistical data analysis (Winter 2018)
Lecture notes

Taught by Shravan Vasishth [vasishth@uni-potsdam.de]

Last edited: October 18, 2018

Contents

Chapter 1

What this course is about

This is a graduate level course in linguistics that introduces statistical data analysis to people who have presumably never done any data analysis before. Only high school pre-calculus mathematics is presupposed, and even there not much is needed beyond basic math skills like addition, subtraction, multiplication, and division.

The goal of this course is to prepare students to understand and use the most commonly deployed statistical models in psycholinguistics. The course is designed to bring people to terms with the linear mixed model framework. We ignore ANOVA in this course because there is not enough time to cover it, and besides, ANOVA is now only a historical artefact of a pre-computing era (in my opinion). We also limit the discussion to two commonly used distributions: the binomial and normal distributions.

The most frequent question people tend to have in this class is: **why do I need to study all this stuff**? The short answer is that linguistics is now a heavily experimental science, and one cannot function in linguistics any more without at least a basic knowledge of statistics. Because time is short in this course, I decided to drastically limit the scope of the course, so that we cover only a small number of topics; these will be the most frequently used tools in linguistics.

By the end of the course you should know the following:

- **Basic** usage of the R language for data analysis.
- Basic understanding of the logic of null hypothesis significance testing.
- The meaning of confidence intervals, p-values, z- and t-values, Type I and II error, Type M, S error, Power.
- Linear models (including simple multiple regression), basic contrast coding for 2×2 repeated measures designs.
- Basics of fitting linear mixed models and presenting results.

Many people come to this course expecting to become experts in data analysis. No one course can achieve that. The present course should be seen as an introduction; for your own research problems, be prepared to study further.

1.1 Quiz: Do you need this course?

You should take this quiz on your own to decide whether you need this course. If you can answer (almost) all the questions correctly, you are in pretty good shape. If you made more than one mistake or don't know the answer to more than one question, you should probably do this course. The solutions are at the end of the book.

Instructions: choose only one answer by circling the relevant letter. If you don't know the answer, just leave the answer blank.

1. Standard error is

   a the standard deviation of the sample scores

   b the standard deviation of the distribution of sample means

   c the square root of the sample variance

   d 2 times the standard deviation of sample scores

2. If we sum up the differences of each sample score from the sample's mean (average) we will always get

   a a large number

   b the number zero

   c a different number each time, sometimes large, sometimes small

   d the number one

3. As sample size increases, the standard error of the sample should

   a increase

   b decrease

   c remain unchanged

4. The 95% confidence interval tells you

   a that the probability is 95% that the population mean is equal to the sample mean

   b that the sample mean lies within this interval with probability 95%

   c that the population mean lies within this interval with probability 95%

   d none of the above

5. The 95% confidence interval is roughly equal to

   a 0.5 times the standard error

   b 1 times the standard error

   c 1.5 times the standard error

   d 2 times the standard error

6. The 95% confidence interval is — the 90% confidence interval

   a wider than

   b narrower than

   c same as

7. A p-value is

   a the probability of the null hypothesis being true

   b the probability of the null hypothesis being false

   c the probability of the alternative hypothesis being true

   d the probability of getting the sample mean that you got (or a value more extreme) assuming the null hypothesis is true

   e the probability of getting the sample mean that you got (or a value less extreme) assuming the null hypothesis is true

8. If Type I error probability, alpha, is 0.05 in a t-test, then

   a we have a 5% probability of rejecting the null hypothesis when it is actually true

   b we have a 95% probability of rejecting the null hypothesis when it is actually true

   c we necessarily have low power

   d we necessarily have high power

9. Type II error probability is

   a the probability of accepting the null when it's true

   b the probability of accepting the null when it's false

   c the probability of rejecting the null when it's true

   d the probability of rejecting the null when it's false

10. When power increases

    a Type II error probability decreases

    b Type II error probability increases

    c Type II error probability remains unchanged

11. If we compare two means from two samples, and the p>0.05 (p is greater than 0.05), we can conclude

    a that the two samples comes from two populations with different means

    b that the two samples comes from two populations with identical means

    c that we don't know whether two samples comes from two populations with identical means or not

1.2 How to survive and perhaps even enjoy this course

I have been teaching this course for several years now, and one reaction that I get quite often is fear, panic, and even anger. A common reaction is: Why do I have to learn all this? How can I do all this programming? And so on.

If you have such questions popping up in your mind, you have to stop and consider a few things before continuing with this course. Linguistics at Potsdam is a very empirically driven program. It is impossible to get through a master's degree in linguistics without coming into contact with data, even in formerly non-experimental disciplines like syntax or semantics. If you are at Potsdam, you are automatically committed to an empirically driven education.

More broadly, there is a widespread misunderstanding that statistics is something that can be outsourced to a statistician. It's true that if you have a non-standard statistical problem you probably need to talk to a professional. But for the kinds of methods used in linguistics, you are personally responsible for the analyses you do, and so you are going to have to learn something about the methods. Many scientists do not understand that in experimental science, the statistics is inseparable from the science, it is not an unnecessary add-on. This is because our statistical inferences inform our scientific conclusions, and if the inferences are wrong, so will be the conclusions. I will provide some examples in this course.

In order to pass this course, you have to understand that **you have to read the lecture notes**, and that it is not enough to just passively read these lecture notes. You have to play with the ideas by asking yourself questions like "what would happen if…", and then check the answer right there using R. That's the whole point of this approach to teaching statistics, that you can verify what happens under repeated sampling. There is no point in memorizing formulas; focus on developing understanding. The concepts presented here require nothing more than middle or high school (pre-calculus) mathematics. The ideas are not easy to understand, but simulation is a great way to develop a deeper understanding of the logic of statistical theory.

Many students students worry that they might make a mistake. It is normal to make mistakes. One should use mistakes as a learning tool. A mistake is a very useful message telling you that you need to think about the material again. A good book that talks about how to learn and solve problems is *The five elements of effective thinking*. I highly recommend this book as preparation for the present course.

1.3 Installing R and learning basic usage

You should google the word CRAN and RStudio and install R and RStudio on your machine. We will explain basic R uage in class; also see the introductory notes on using R released on Moodle. You should spend some time on the CRAN website looking at the information available on R there. Look especially at the section called Contributed (navigation panel on the left on the main page).

A good first introduction to R for non-programmers is: https://rstudio-education.github.io/hopr/

Bibliography

[1] G. Jay Kerns. *Introduction to Probability and Statistics Using R*. 2010.