Statistical Methods in Linguistics

Linguistics Tutorial Spring 2017
Tuesdays 3-5pm, Emerson 318

Zuzanna Fuchs
zuzannafuchs@fas.harvard.edu
Office hours by appointment

Course Description

Experimental and corpus work is rapidly gaining popularity in linguistics, and with such research comes lots and lots of data. This data needs to be prepared, analyzed, and presented appropriately in order to accurately represent the facts. The goal of this class is therefore to increase the students' understanding of statistical methods in linguistics. Upon completing the course, students will be able to understand data and results presented in experimental and corpus work, and to take the necessary steps to choose and apply statistical methods in their own work, should they choose to pursue such research.

The course also aims to expose students to the kind of linguistic work in which data analysis is necessary. Thus, lectures will include examples of experimental and corpus work in linguistics that involves the relevant methodology. In addition, students will have the opportunity to work with real linguistic data in their problem sets, in order to practice newly learned skills using the types of data they may themselves encounter.

Importantly, this course is not about learning as many statistical tests as possible. Rather, it is about understanding what we as linguists need to know about our data and our research questions in order to choose the appropriate test, based on the assumptions that must be met by the data in order to use a given test. Some basic tests will indeed be introduced, but more importantly, this course will allow the students to go beyond what is taught in the class, giving them the knowledge and background they need to read about and understand many more statistical tests encountered in the literature.

Course Requirements

Problem sets (4)	60% (15% each)
Participation	15%
Final project	25%

Problem sets – There will be four problem sets assigned. The purpose of these problem sets is for students to enhance their understanding of the concepts discussed in class through applying them on their own to real linguistic data. Students will have a week for each problem set (homework due at the beginning of class each week), and no late assignments will be accepted.

Participation – Students are expected to actively participate in class discussions. We will be working through problems together, analyzing statistical methods detailed in papers, and learning some basic R programming. For this reason, students are expected to bring their laptop to each class. Attendance is mandatory.

Suggested readings – There are no required readings for this class. The suggested readings posted for each class are supplementary reading to help review and understand the concepts discussed in class and/or papers presented in lecture that involve statistical analysis of linguistic data.

Any readings from the Baayen textbook are available here: http://www.sfs.uni-tuebingen.de/~hbaayen/publications/baayenCUPstats.pdf. Those who would like to learn more about statistics should consider purchasing the book.

Final project – Each student will be asked to submit a final project in which he or she demonstrates understanding of statistical concepts learned in class. In Week 5, all students should meet individually with the instructor to discuss their choice of project. Projects are due the last day of class.

> Option 1: Choose a paper that presents some statistical analysis, from a preselected list. Write a review (5-6 pages) of the paper, paying special attention to the statistics.
>
> Option 2: Choose a data set, from a preselected list of data sets. Write a report (5-6 pages) on how you prepare and analyze the data, discussing all choices you made along the way. Include visual representations of the data and results.

Schedule

Week 1: Introduction; the basics (mean, median, standard deviation, IQR, outliers)

Problem set 1 assigned
Required homework: datacamp.com R bootcamp
Optional reading: Baayen Chapter 1
Strickland et al. 2015 on sign language semantics

Week 2: R work; distributions, visualization of data (graph types, confidence intervals, good practices)

Problem set 1 due
Problem set 2 assigned
Optional reading: Baayen Chapter 2
Conroy et al. 2009 on child acquisition of Principle B

Week 3: Intro to statistical tests – setting hypotheses, what does a p-value actually mean, understanding assumptions and types of error, power and sample size

Problem set 2 due
Problem set 3 assigned
Optional reading: Baayen Section 4.1
Benor and Levy 2006 on order in English binomials

Week 4: Statistical tests: t-test, pairwise t-tests, linear models part I

Problem set 3 due
Problem set 4 assigned
Optional reading: Baayen Section 4.2 & 4.3
Stowe 1991 on filled gap effects

Week 5: linear models part II

Problem set 4 due
Meet with instructor regarding final project
Optional reading: Sell and Kaschak 2012 on “up is more”

Week 6: Wrap up, tools for further work and understanding

Final project due