Introduction to Categorical Statistics

Dylan Glynn
Email: dsg.up8@gmail.com
URL: www.dsglynn.univ-paris8.fr/

Description

In all empirical sciences, the inductive testing of the explanatory power of theories just as the inductive testing of the accuracy of descriptions is fundamental. If there is an exception to this, it is in linguistics, which is probably the last science to adopt these norms. This situation, however, is changing rapidly. Making generalisations based on sample (inductive research), //necessitates// the use of statistics and Linguistics has recently grown to appreciate this fact..

1. Patterns – Identifying complex (multidimensional) structure in data.
Although human cognition is exceptionally good at pattern identification, there are limits to what is feasible using subjective observation and qualitative analysis. Complex pattern identification is one of the most important uses of statistics

2. Representativity – Determining the significance (repeatability) of an observation
Even very large samples are, in fact, small samples with respect to size of our "population" – language. Counting occurrences and determining their proportion is // not// enough to make scientific claims. The probability that the observation in the sample is true for the population is sine qua non in inductive research.

3. Importance – Determining the effect size or the relative importance of different patterns (or parts of the patterns) in explaining the behaviour a linguistic phenomenon is quasi impossible to achieve using qualitative methods. Multivariate statics calculates the relative importance, or contribution, that a given factor or dimension of the analysis has in explaining the object of study.

4. Predictive Accuracy – Calculating the descriptive precision of a model (theory, hypothesis...) is arguably the most important role of statistics in empirical science. It permits one to determine how successful a theory is at explanting observable behaviour, and therefore to compare the success of competing theories. It also allows one to determine the descriptive accuracy of an analysis.

The workshop will introduce statistics using R. No prior knowledge of statistics or R is needed. The workshop will focus on statistics for categorical data (counted observations, corpus linguistics, sociolinguistics, *etc. etc*). The workshop will //not// look at statistics for continuous (measured observations, psycholinguistics, phonetics etc.

The seminar will follow, more or less, the contents of the book Glynn & Robinson (2014), ask you library to buy it. More advanced content, if we have time, will be based on Baayen (2008).

Aims

1. Understand the need for statistics in linguistics
2. Obtain an overview of the kids of things you can do with statistics
3. Gain enough confidence in applying and //interpreting// stats to know that you can do this!

Reading
There is no reading, this is a practical course.

Slides for the first class as well as commands are on line
www.dsglynn.univ-paris8.fr/
student downloads, password: student

Requirements
Sound knowledge of linguistics

Outline
The pace of the course will depend on the participants. I will go as slowly as needed for everyone to keep up. For this reason I will not give a hour-to-hour breakdown.

However, note that this is a practical course – you miss a session, it will be very difficult to catch up. So, make sure to catch up before you come to the next class – you are warned :)

Part 1
First Steps with Quantitative Data

a. Fundamental principles of statistics
b. Getting data into R (from Daniel's corpus course)
c. Chi^2 Test for independence
d. Adding some manually annotated factors
(if time)
e. Spearman's rank correlation – Rho Test

Part 2
Looking for Structure and Mapping It

a. Association plots and Mosaic Plots
b. Multiple Correspondence Analysis
c. Agglomerative Cluster Analysis
d. *K*-Means Cluster Analysis
(if time)
e. Factor Analysis

Part 3
Modelling Structure and Proving It

a. Binary Logistic Regression
b. Log-Linear Analysis
(if time)
c. Mixed-effects Regression
d. Multinomial Fixed-Effects Logistic Regression