Introduction to statistical methods for language and literature

Diana Santos

ILOS

d.s.m.santos@ilos.uio.no

Spring 2018, week 1

Three goals

This course has three subjects or learning arenas:

- to present statistical reasoning
- teach you to use the basics of statistics in R
- give an overview of the use of statistics in the humanities

Plan for working in this discipline: In addition to the lectures, I provide

- One article per week for you to read
- Exercises in R
- Further additional literature for deepening or explanation of some subjects in depth

There will be three qualification exercises.

Temporal progression

1. First part: the most important concepts, learn to use R, descriptive statistics, basic visualization, hypothesis testing, ANOVA.
2. Second part: use of exploratory methods
3. Third part: more linguistic-minded application examples: geometric models, dative alternation, translation studies, information retrieval evaluation

What you are going to learn

- statistical thinking
- statistical terminology
- the style and textual genre of statisticians
- develop a sense for numbers, tables, graphs, and for graphical presentation of data
- who is who in statistics

Probability and statistics

Difference between probability theory and statistics according to Julian Simon

- Probability theory is a part of mathematics
- Inferential statistics is an art, or an engineering tool, which uses computation of probabilities

In another way: Inferential statistics is probability models and probability estimation PLUS rules for which models one can use in one practical situation PLUS interpretation principles for what the model gives us.

- Probability theory is mathematics. It deals with situations about which you know the nature, and you estimate probabilites for a system to yield one or more spesific results.
- In inferential statistics, on the contrary, one does not know for sure the nature of the system one is studying, and has to find it on the basis of the data alone. (Therefore inferential statistics is appropriate to research in science.)

Concepts: appetizer

- **probability**: a number between 0 and 1 which measures/shows possibility
- **probability distribution**: together it amounts to one, distributed over all possible outcomes in a discrete situation
- **probability density**: the same for a dense/real variable
- **cumulative distribution**: for the cases $<= x$
- **surprise value**: which cases are different from what we expect. The more surprising, the least probable it is by chance.
- **independence**: the same result together or alone each for itself

Concepts: appetizer (2)

- **null hypothesis,** H_0: what one tries to show that it is false (one tries to reject it)
  One does not reject H_0 when the probability of a result given our model is not small enough. If the probability for the result by chance is very small, then one accepts the alternative hypothesis H_a
- **significance**: how easy it is to get the same result. It is measured through type I error, *alpha*: the probability to reject a true hypothesis. It is represented by p. It is common to see the requirements $p < 0.05$, or $p < 0.01$...
- **power of a test**: $1 - \beta$. β is the probability to commit an error of type II, that is, the probability to accept the null hypothesis when it is not true.
- **confidence intervals**: with a particular probability the answer is inside the interval
- **quantiles**: intervals related to a percentage: deciles, percentiles, quartiles

Platonic approaches?

- In a large part of statistics one deals with a model (theory) and empirical observations, and the whole point isto find the best model to account for the observations
- “Reality/universe” versus “sample”: Usually we only have some samples, but want to estimate what really happens … in the whole population
- To make things more complicated, the names and the symbols one uses are very similar: *sample/empirical* mean vs. *population/theoretical mean*, and sample/empirical variance vs. population/theoretical variance, etc.
- There are the designations **descriptive statistics** and **theoretical statistics** to distinguish between the two kinds of numbers.

Hypotheses

- Hypotheses are usually about numbers which estimate a relationhip, or about differences between one or several numbers
- So the art of statistics is to abstract, reduce or translate interesting hypotheses to numbers (or differences between numbers)
- And to count we need to categorise (and see similarities in different tokens) – **and that is where each science comes in**

Kinds of information

Just like we have adjectives, nouns and verbs in natural language (with different properties and use patterns), there are also different kinds of types in programming languages, and in statistics there are different methods for different kinds of information.

- list of categories (names of trees, names of cards, gender, nationality, ...): categorical variables, also called **nominal** variables
- categories which can be ranked (a lot, little, nothing; child, youth, grown up) – **ordinal** variables
- continuous variables (weight, temperature, age,...): it is possible to do arithmetic with them
  - interval variables, in that the numbers in themselves have no meaning: like temperature (in C or Farenheit) or school grades
  - meaningful numbers, where the zero means something: only in those cases one can use ratios, and talk about three times as much

It is important to understand these differences, although every of these categories can be represented with numbers. But one has to interpret them carefully.

Last words

There are many traps in statistics. Some of the best known are:

- to confuse significance with importance. There is no relationship between the two concepts!
- to believe that using statistics is good science. Only if it is well used, not if it is misused!
- to confuse correlation with causality. Two things that happen at the same time do not mean that one causes the other!
- to believe that there are linguistic methods and statistical methods. No! There are statistical **tools** which can be used in linguistics. But not using statistics does not make a method more (or less) linguistic!

Homework, first week

- Try out the two files to get familiar with R
  - instrSPR4104
  - warming up
- Read the paper (optional)
  - Karlgren129