Some methods that can be used with corpora

Diana Santos

ILOS

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

September 2013

Some methods that can be used with corpora

Using statistical artillery

1. Comparison of two proportions
2. Correlation of two properties
3. Classifying instances using a set of properties
4. Grouping several elements in sets or clusters based on a set of measures

All this use measures over features, that is, some quantification (counting) and in some cases the notion of geometrical space.
No magic, and one needs a solid linguistic analysis of the features used, in addition to a solid mathematical analysis of the mathemeatical presuppositions.

Some quotes about use of measures, or figures

Stefan Evert, 2006:

> *Statistical methods give only numbers - it is linguistic interpretation that gives them meaning.*

Eugenie C. Scott, 2013 (free rendering):

> *From hypothesis to textbook... the iterative scientific method, get data, test, find alternative explanations (critical thinking), test again, publish, have others testing, have others come with alternative explanations, get some scientific consensus, translate into textbook science to teach to students as scientific discoveries.*

Parametric statistics

Parametric methods use known probability distributions, that only require that we know the values of the parameters. Families of distributions (one element for each parameter value).

Distribution	Parameters
Binomial	p
Poisson	λ
Gaussian	μ, σ
t	ν - degrees of freedom
F	ν_1, ν_2
χ^2	ν - degrees of freedom

If we don't know the distribution function, there are non-parametric methods.

Comparison between two proportions

What could one do about studying the passive in Portuguese (or another language) using corpora?

- Absolute frequencies?
- Relative? (what is the unit? Is this a meaningful proportion?
- Distribution by which feature?
  - Let us study the proportion of passives in different genres
  - Let us study the distribution of passives for different verbs
  - Let us study the distribution of passive for different tenses

Operationalization: how do you really do this?

- First you create a "table", or better a dataframe in R, which is a computational object that includes, for each observation/unit, a set of values, organized in columns, and obtained from your corpus (or from your field observations).
- R provides a lot of machinery to deal with such "tables" (of numbers or values). Both for counting (arithmetics), for visualization, and for statistical processing.
- The tables people usually see in papers are already the result of processing these dataframes, for example contingency tables.

Are there differences among text genres as far as fear is concerned?

The dataframe "medo" contains the number of fear-related words for a set of genres in the NILC/São Carlos corpus of Portuguese. Do the genres differ according to this feature?
Dataframe:

`http://folk.uio.no/dssantos/cursoR/medo.txt`

Hint: First visualise, then sort, and then compare pairwise.

Test whether a difference is significant

```
prop.test(PROPORTION1,PROPORTION2)
prop.test(c(120,81), c(140,100))
```

What is a statistical test?

(in frequentist statistics) A test has always two elements:

1. a test statistic – a function of the sample that we will compute based on our sample);
2. and a rejection region – we will reject the hypothesis if the test statistic lies in that region

The p-value is the probability that the test statistic has this value if the nulll hypothesis, H_0, is true. The lower the p-value (also called the observed significance level), the more comfortable we are in rejecting the null hypothesis.

Correlation

This dataframe include the number of adjetives per author, and the number of colour words per author.

- Is there a correlation between adjective richness and colour richness?
- Choose two authors and check if it can be said that they are significantly different as far as colours are concerned.

```
cor(VEC1,VEC2)
cor.test(VEC1,VEC2)
```

The simplest example

Two books tentatively assigned to Aristotle have the following distribution of the last word in the sentence, in the first 100 sentences. Assuming that this statistic (POS of the last word) is sound, what can we conclude from the table? (A: others, S: nouns, V: verbs)

PoS	S	V	A
Retorics	28	32	40
R. to Aleksander	27	52	21

Examples of use of a t-test

Differences between means: is the mean in our sample just like the known mean (12)?

```
t.test(HEDGES, mu=12)
```

Are F1 frequencies of men and women different?

```
t.test(F1S~gender, paired=FALSE)
t.test(F1S[gender=="M"], F1S[gender=="F"], paired=FALSE)
```

Are the length differences in translation consistently higher?

```
t.test(Length~OrigOrTrans, paired=TRUE)
```

Exploratory methods

There are situations when you don't really know what is happening, what are the possible factors, and therefore your "weakest" hypothesis is: Let us see what happens if I measure everything I can or everything I may have the slightest suspicion of, and see if the method can give me some clues. Three paradigms/examples (in my view):

clustering You have many examples, and want to know if they can be represented by fewer cases. It is classification, or categorization, you are trying to perform. Find categories in your data.

tor analysis You have many classified examples, and you want to see if you find the features that allow you to classify them. Identify what makes an X an X, in terms of smaller, individual properties.

ine learning You want to develop a system that learns from a set of classified examples so that it classifies or clusters best new cases.

Of course, you can use machine learning techniques to do clustering or factor analysis, but its inception was to create intelligent systems.

Several clustering techniques

- Principal components analysis, `prcomp` : find the n "components" that explain the variance better, new dimensions that reduce the need for so many axes. Then one studies in general the first components, or better, tries to interpret what they mean, often by visual inspection and argumentation.
- Factor analysis, `factanal`: In addition to components (now called factors) one allows for error, so one has to choose a priori the number of factors.
- multidimensional scaling, `cmdscale`: new (fewer) dimensions that keep as best as possible the original distances between points. Correspondence analysis, `corres.fnc` is one kind of MDS for two-way contingency tables (counts).
- hierarchical cluster analysis, `diana, hclust`: presents results in tree format.

Classification

Here we know which classifications we want, just not how to produce them to new instances.

- Classification trees, `rpart`: produce a method to classify, based on the features of the instances. In order for it not to be too connected to the data points, one has to prune the tree based on cross-validation.
- Discriminant analysis: find linear discriminants, that is "linear equations" that predict a class. Again, one needs to use cross-validation to evaluate the discriminants.
- Support vector machines, `svm`: find the best way to model the boundary areas between classes, which are called the "support vectors". No way to visualize, one still needs to cross-validate, but considered the best in terms of performance in classification tasks.

Behind the scenes

Devore and Berk (2007) state that

> *the chi-squared, t, and F distributions are "distributions based on a normal random sample"*

- for the distribution of the sample variance, one needs the distribution of sums of squares of normal variables -> the χ^2 distribution
- to use the sample standard deviation in a measure of precision for the mean $\overline{X}$, we need a distribution that combines the square root of a chi-squared variable with a normal variable -> the t distribution
- to compare two independent sample variances, we need the distribution of the ratio of two independent chi-squared variables -> the F distribution

This means that in some cases you may want to give up the normal approximation and simply use non-parametric cases.