Introduction to statistical methods for language and literature

Diana Santos

ILOS

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

Spring 2016, week 3

On the difference between pDIST and dDIST

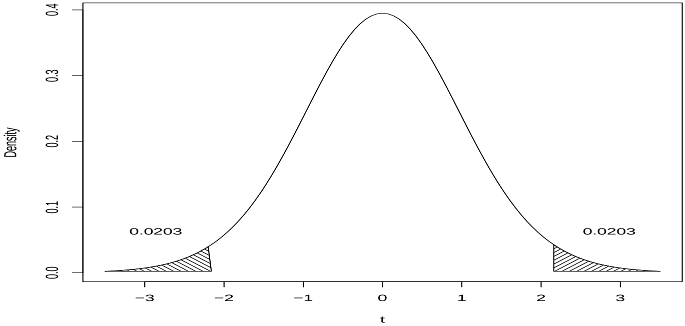
We saw last class that there was a difference between discrete distribtions (example Poisson) and continuous distributions (example: Normal). Therefore, the interpretation of the pDIST and dDIST commands is also different according to this

- In a discrete distribution, `d(x,PARAMETERS)` gives the probability that $X = x$ (how much probability was distributed to the value x)
- In a continuous distribution, `d(x,PARAMETERS)` gives the value of x of the probability density function.
- In a discrete distribution, `p(x,PARAMETERS)` gives the probability that $X <= x$, the value of the cumulative probability distribution, which is in fact the sum of all possible values less or equal than x.
- In a continuous distribution, `p(x,PARAMETERS)` gives the probability that $X <= x$. And this is an area, from minus infinity to x.

How to use a continuous probability distribution: again

The area gives us the probability of the events/facts/objects we got. Less or equal than x, or in the interval $]x_1, x_2[$. Some handy equalities:

- $P(x \in]x_1, x_2[) = P(x <= x_2) - P(x <= x_1)$
- $P(x > x_1) = 1 - P(x <= x_1)$



What is a statistical test?

In frequentist statistics, a statistical test uses numbers obtained from your material to answer questions like:

- are the differences between two samples due to chance?
- does a particular sample belong to the population you expect?
- do some features co-variate in your population (correlation studies)
- which features are more relevant to describe/model a particular situation?

So far, probably so good. The problem is how to operationalize your linguistic questions into these simpler cases.

More specific cases, or standard tests

(what you find in every uninspired introduction to statistics)

- tests for the mean in a normal population with known variance
- tests for the mean in a large sample
- tests for the mean in a normal population with unknown variance
- tests for proportions in a large sample
- tests for proportions in a small sample

Other more specific tests

- test for normality `shapiro.test`
- test for another distribution `ks.test`

The internals of a test

A test consists always of two things:

- one test statistic (to compute from the data)
- one rejection region (the area /values that reject the hypothesis)

The result of a statistic test comes with a p-value: the probability, when the null hypothesis H_0 is true, that the test statistic has this value or is even nearer H0. (it is also called the “observed significance level”)
One can force a p-value before hand, α, called the significance level (“the minimum value that results in rejecting the null hypothesis”), and say that

- You reject H_0 if p-value $<= \alpha$
- You cannot reject H_0 if p-value $> \alpha$

In principle, there are two schools of practical statistics about that.

Two types ot test: one-tailed or two-tailed

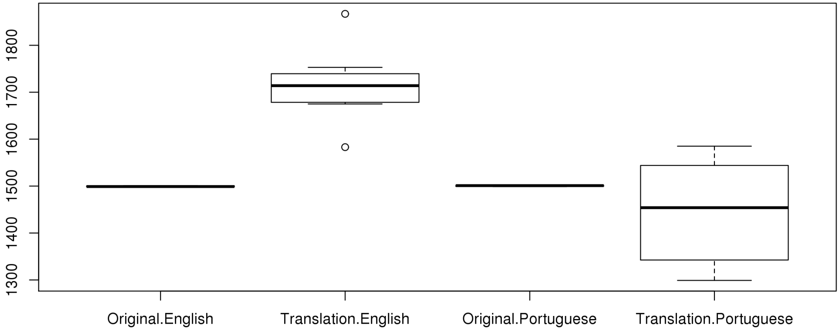
Also called directional or non-directional.

- The non-directional or two-tailed test only tests equality. Side of variation is not important.
- The directional or one-tailed has one hypothesis in one direction, for example that x is greater than 5. Then we only have to test in that direction, and therefore the test is more powerful.

In general, R accepts that information in the specification of the test. For example: $alternative = "one.sided"$ or $alternative = greater"$.

Let's go back to the previous class

- repeat t-tests
- do other tests
- recap the chisq-test



Example with chisq.test

Let us use the `chisq.test`

```
chisq.test(TABELL)
Pearson's Chi-squared test with Yates' continuity correction

data:  tabell
X-squared = 33.112, df = 1, p-value = 8.7e-09
```

This reads as: the null hypothesis (which is, in this case, that the two characteristics are independent) can be rejected at the .01 significance level. (chisq=33.112, p-value=8.7e-09)

Test whether a difference is significant

For a difference to be significant, it has to be established beyond doubt that the differences are not due to chance. The less probable that the differences are due to chance, the more probable that they reflect a difference in reality (and not an artificial difference due to sampling).
Examples: Compare

- numbers from different genres
- numbers from different varieties
- numbers from different authors

How probable is it that the differences are due to chance, but the underlying population is the same? (null hypothesis: no difference due to genre, no different between the authors, etc.)

```
prop.test(PROPORSJON1,PROPORSJON2)
prop.test(c(120,81), c(140,100))
```

Homework for next week

The following dataset has how many words related to colour per author, in a Portuguese corpus. The question is: Are there differences between the authors in this respect?

Dataframe:

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

Hint: visualize first, order, and compare two by two.

Can I be sure?

Can the tests make me sure?
Never! In statistics nothing is sure. What you have is different kinds of errors that you try to minimize:

Decision/Reality	H_0 is true	H_0 is false
Accept H_0	correct	type II error
Reject H_0	type I error	correct

The probability of doing a Type II mistake is called β, of doing a type I mistake is called α.
The power of a test is $1 - \beta$. In general, one can increase the power of a test by increasing the sample size.

Contingency tables (Krysstabeller eller kontingenstabeller)

This is another often used way of presenting material that one wants to compare.
Its is important to understand that a contingency table is not a dataframe.
But there are commands in R that create contingency tables out of dataframes, like `xtabs` (cross-tabulate).
It is over contingency tables that the most common chisq tests work.

Summing up

- You were more formally introduced to the notion of a **statistical test**
- You learned how to do some of the simplest tests, t-test and chisq test
- You heard (again) about p-values, one and two-tailed tests, and paired tests
- You were warned that no certainties exist in statistics
- You were informed about the law of large numbers and some illustrations of it (slides of the previous class)