Statistical methods for corpora (using R)

CLiF 2016

Introduction

Amir Zeldes
amir.zeldes@georgetown.edu

Organization

- Contact: amir.zeldes@georgetown.edu

- Please send me a quick e-mail with any questions
- If not using a workstation – please bring laptops to all sessions!
  - Morning sessions – theory and discussion
  - Afternoons – practical sessions with more R
  - Data and slides at: http://corpling.uis.georgetown.edu/amir/CLIF_2016/
- Software
  - R for your platform should be installed
  - I also recommend RStudio and a good text editor:
    - Notepad++ for Windows
    - TextWrangler for Mac

Topics and other stats intros

In this course we will talk about more or less the same statistical procedures found in standard textbooks (t, chi-square...), but:

- Less mathematical theory, more applications (no proofs etc., aim for **understanding reasoning**)
- Focus on statistical evaluation and visualization of **corpus data, exploratory methods**
- Using the freely available and extensible "R" framework: http://www.r-project.org/

Preliminary plan

A very intensive course:

- Monday: Introduction + R
- Tuesday: Descriptive Statistics
  (variables, estimation, variance)
- Wednesday: Analytic Statistics
  (significance, effect size, simple tests)
- Thursday: Multifactorial methods
  (ANOVA, logistic regression)

So what's this all about?

- We are surrounded by statistics, in daily life as in science
- Quantitative statements are often accepted without being seriously questioned or understood
- Few opportunities to really gain insights using statistics for the "uninitiated"
- Very difficult to gain access as a humanities scholar (out of experience…)

So what's this all about?

- However, we cannot avoid quantitative data
- Nor should we want to!
  - Many aspects of language are quantitative
  - If you study corpus data, you want to generalize about what's reliable, abstract away from blips
- A corpus-based rephrasing of competence and performance:
  - Competence – the reliable, recurrent patterns of the language examined
  - Performance – individual variation we do not wish to capture

After this course

You will know what these terms mean and how they can be used for corpus linguistics:

- Binomial distribution, t-test, chi square
- Vectors, matrices, $\Sigma(x_i)$, $\Pi(x_j)$
- Variance, standard deviation, significance
- Correlation, regression, confidence interval
- ...

After this course

But more importantly we will talk about theory:

- How does language performance data relate to "grammar"?
- What are the fallacies of corpus=language? How do we avoid them?
- How to recognize patterns in noisy data
- How to avoid overfitting

And we will learn to recognize these things in work we are reading

Why is this important?

It's not just linguistics. Consider a typical newspaper report:

> **Violence on the Rise**
>
> *"The violent trend is continuing. The number of cases of aggravated assault has risen **substantially**", a police spokesperson told reporters. Although data for three more counties is still missing the **statewide tendency** is clear.*

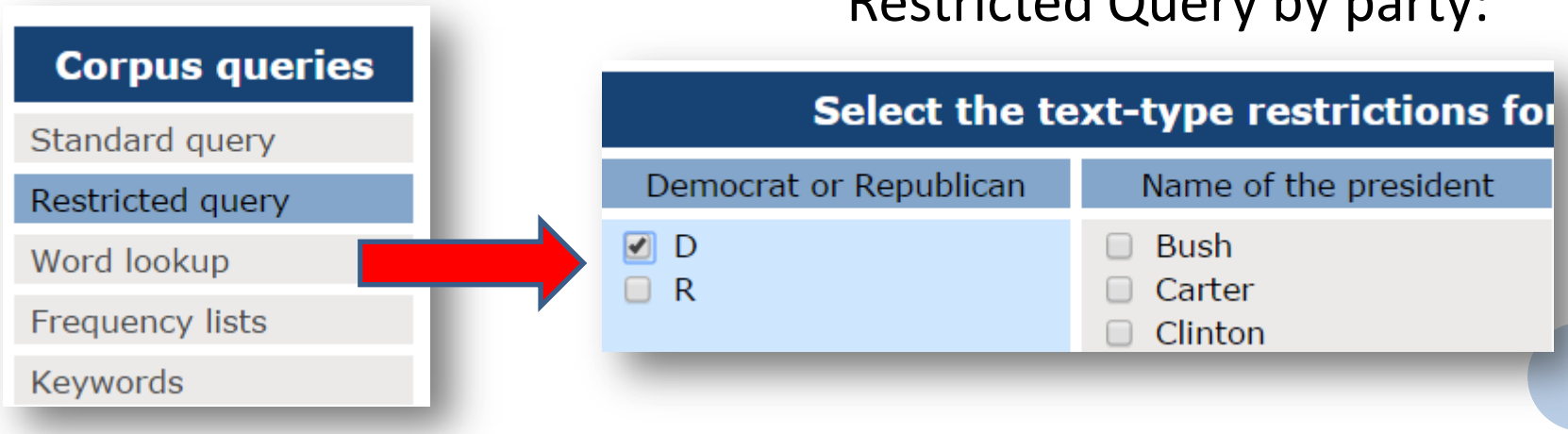
- What was the number of assaults previously?
- What is a "substantial" rise? Is double the cases a rise? Two more cases?
- Does it matter if it's from 1 to 2 or from 100 to 200? Why?
- How can we establish a statewide tendency with missing data? Or at all?

* For those interested in newspapers and statistics: see "Innumeracy" and "A Mathematician Reads the Newspaper" by John Allen Paulos

A more linguistic example

- We can make clickbait out of frequencies easily:
  - "Democratic party biased about guns, corpus study shows"
- We can use the State of the Union corpus in CQP:
  - https://corpling.uis.georgetown.edu/cqp/stateoftheunion/
- How often is ***[lemma="gun"]*** used?

Restricted Query by party:



Clear results?

- **D → guns:**
  - Descriptively adequate (report on presidents to date, no inferential statistics)

  ? How sure are we that there's a difference? **(57 hits!)**

  ? Is this difference really big? **(58/73 don't mention guns)**

  ? Is data distributed evenly? **(25/57 hits in 2000)**

Daniel Mauser was only 15 years old when he was gunned down at Columbine . He was an amazing kid

Professional presentation of quantitative data

In the letters written by wives to their husbands, we find a high frequency of both *I* and *you* …. In the 18th century data women use *I* significantly more than men (**p < 0.003**).

To discover statistically significant differences between categories, we use the **Wilcoxon rank-sum tes**t … the **p value** obtained from a **significance** test refers to the likelihood of observing a single measurement under the **null hypothesis..**

the relationship categories … [are not] **balanced** for gender … the proportion of son-father letters increases from 11% to 24% from 1600–1679 to 1680–1800.

[from Vartiainen et al. 2013]

- What does ***significantly more*** mean?
- What exactly is ***p value, null hypothesis, test …***?
- Is this an interesting result? How interesting?

The bottom line

- We cannot rely on intuition alone to understand and compare quantities
- Many empirical questions cannot be answered without statistics
- We cannot understand the numbers reported in literature without proper training:
  - not to trust our own first impressions blindly
  - to use numbers appropriately in our research
  - to understand and critique use of numbers by others

Quantities and linguistics

- What is the difference between spoken and written language?
- Different genres? Male and female speakers? ...
- How similarly are two words or constructions used? What is usage?
- How productive are different word formation processes? When is something "lexicalized"?
- Which constructions do language learners find particularly difficult?

➢ We need operationalizations, numbers and tools to process them

R

- Available from: http://cran.r-project.org/
- Command line based
- Expressivity of a full fledged programming language
- Look and feel very different from most Windows/Mac programs like Excel or SPSS
- Steep learning curve at first

Why R?

- A practical reason: R is free (SPSS isn't)
- Unparalleled graphical capabilities
- Very flexible, applicable to linguistics needs
- Extensible using thousands of modules
- Runs under Linux/Mac/Windows
- Especially common in the corpus/computational linguistics community (text processing modules)

Recommended reading

- Gries , S. Th. (2009) *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.
- Gries, S. Th. (2013) *Statistics for Linguistics With R: A Practical Introduction*. Berlin & New York: Mouton de Gruyter.
- Oakes, M. P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Dalgaard, P. (2008) *Introductory Statistics with R*. New York: Springer.
- Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.
- Rietveld, T. & van Hout, R. (2005) *Statistics in Language Research: Analysis of Variance*. Berlin/New York: Mouton de Gruyter.

Libraries to install

- We'll go over this in the afternoon too…
- But in case you're not there: Libraries for this course: (in R: packages -> install packages)
  - lsr - Learning Statistics with R (Navarro 2015)
  - rms – Regression Modelling Strategies (Harrell 2001)

Statistics and Probabilities

Square one: Probability

- Statistical analysis is firmly grounded in the concept of probability
- It's important to understand what probability means
- Surprisingly, this basic concept is difficult to define and full of potential dangers...
- This is why it took centuries to define!

In the beginning there was the game

- The mathematical notion of probability did not exist before the 17th century
- Originally the main application or problem related to games of chance:
  - How can we divide the pot in a game if a round is interrupted before completion?
  - Old problem, first documented in the 1500's
  - Formal solution discussed in 1660 by Pascal and Fermat

- One clear notion about probability: it's all about discussing what would happen in **the future**

The classic definition of probability

- First formulated in 1814 by Laplace:

  The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a **fraction** whose **numerator is the number of favorable cases and whose denominator is the number of all the cases possible**

  [A Philosophical Essay on Probabilities, 1814]

- what does this mean?

A simplified explanation...

- To define the probability of an event...
  - We need to decompose a situation into all possible outcomes
  - These should be 'equally possible' i.e. equally probable
  - **Recall we have no idea what probability is**!
- Then probability would be:

$$P(\text{event}) = \frac{\text{Number of outcomes instantiating our event}}{\text{Number of outcomes in total}}$$

A (non-linguistic) example

- We have an urn with
  7 **black** und 3 white stones
- We are interested in the event:
  "drawing a white stone"
- There are 10 equally possible outcomes
- Our event corresponds to 3 outcomes
- So the probability is:

$$P(\text{white}) = 3/10$$

Criticism of the classical probability definition

- The definition is circular:
  - We're supposed to define probability
  - We decompose an event into all possible outcomes
  - These are then assumed to be **equally probable**
- Only applicable if there is a finite number of outcomes

Criticism of the classical probability definition

- What about continuous result sets?
  (e.g. relative frequency)
- The observer's point of view is referred to in the definition (we believe something is equally probable)
- In many cases we can't say what all possible outcomes are exactly (think of vocabulary!)

- *scene*
- *pool*
- *concigliere*
- *Gorbaphile*
- *depleted-uranium*
- *neo-Eisenhoweresque*
- *…*

Frequentist probability definition

- In order to get rid of these problems, later mathematicians (e.g. Venn) defined:

$$P(event) = \lim_{n \to \infty} \frac{event\ occurs}{number\ of\ trials}$$

- This means that relative frequency **converges** towards the actual probability:
  - if we throw dice 1 million times, we get a 6 almost exactly 1/6 * million times
- This is the most widely used definition of probability today

Discussion

- Is the frequentist probability definition applicable to language data?
- Do linguistic categories have probabilities?
  - words
  - phrases
  - constructions
  - styles
  - topics
  - …

- Communicative intent?
- Contextual conditions?
- Residual variation?
- …

- p(article) > p(superlative)?
- p(*declare*) > p(*say*)?
- p(declarative) > p(imperative)?
- p(noun) > p(pronoun)?
- …

Consequences for language dependence?

- Consider the ‘Sapir-Whorf’ hypothesis
- Many formulations, e.g.:
  - *No two languages are ever sufficiently similar to be considered as representing the same social reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached.* (Sapir 1929)
- Data from different languages, different text types, different datasets inevitably differs
- The labels we give and examine determine our results (cf. Anke)

➢ Should we give up?

Making it worse to make it better: Multiple events

Let's forget words again for a second:

- Suppose we know the chance of rolling a 6 is: P(6)=1/6…
- And also the chance of getting 1: P(1)=1/6
- What is the probability of rolling a 6 and then a 1? P(6&1) = ?

One more step: Multiple events

- Answer: only in every sixth case do we get the chance to fulfill the second condition
- Our chances are lowered by 1/6!
- Therefore P(6&1) = (1/6)*(1/6) = P(1)*P(6)
- And in general, for two independent events: P(A&B) = P(A)*P(B)

➢When are events independent?

Multiple events in language data

- Given: (data from COCA spoken, Davies 2008)
  - p(*want*) = 0.0019
  - p(*to*) = 0.0237
- What is p(*want*, *to*)?
  - Expected: 0.0019 * 0.0237 = 0.000045
  - Observed: 0.001326

Independence and conditional probability

- Two events A and B, for which:
  - P(A) = P(A|B) (=probability of A, given that B has happened)

    and
  - P(B) = P(B|A)

  are called **independent**, i.e. the fact that one event occurred does not alter the chances of the other occurring
- For example, if you draw cards from a deck:
  - Event A: you draw a heart
  - Event B: you draw a king

  are independent (even within the same draw!)

- Very many statistical procedures assume that our observations are independent – but this is very often not true!

Moving forward

- Statistics is all about probabilities, in the main:
  - Predicting outcomes about the future based on past experience
  - Estimating influence of variables on each other
- In linguistics, as in many fields, nothing is absolute:
  - Rank probabilities in different settings – variation studies
  - Usage based grammar:
    - Learn p(utterance|meaning, context) and vice versa (Manning 2003)
    - Language as a set of instructions 'how to mean' (Halliday 1977)
- We want the tools to learn what significantly influences linguistic choices in natural language data

Installing R

- If you haven't already, please download R from:
  - http://www.r-project.org/
- Install the program following its instructions
- Start the program (Rgui.exe/app etc.)
- You will see a very bare bones command line interface which will prompt you like this:

  >

- (you may also want to install the program RStudio, a comfortable environment to use R in: http://rstudio.org/)

First steps in R

- You can input commands in R or 'ask R a question':

```
> 2+2
```

- R will answer your question and wait for further orders:

```
> 2+2
[1] 4
>
```

- You can ignore the [1] for now
  its meaning will become clear soon enough...

R as a calculator

- The operators: `+ – * /`

  are easy to figure out: plus, minus, multiplication, division
- The command: `2^3`

  means $2^3 = 2 \cdot 2 \cdot 2$
- All other common (and uncommon!) mathematical operations are supported. Some more complex examples look like 'functions':

  `sqrt(4)` = √4 = 2

  `log(1)` = 0

  …

A quick exercise

- Calculate the following:

$$\frac{1}{5} - 2$$

$$\frac{1}{5-2}$$

$$(2+3)^2$$

$$2 \cdot 3^2$$

$$2^{5-2}$$

$$\sin(3.1415)$$

A quick exercise

- Calculate the following:

$\frac{1}{5}-2$	`1/5 - 2`
$\frac{1}{5-2}$	`1/(5-2)`
$(2+3)^2$	`(2+3)^2`
$2\cdot 3^2$	`2*3^2`
2^{5-2}	`2^(5-2)`
$\sin(3.1415)$	`sin(3.1415)`

Saving your results: defining variables

- We often have to repeat the same calculation
- and keep calculating using our subtotals
- This is what variables are for:
  - `a <- 3` generates a placeholder or 'variable' called `a`
  - This variable has now been given the value `3`
  - If you type `a` and press enter you will get the current value of `a`
  - The symbols `<-` belong together and mean 'fill this variable with...' (assign `a` the value...)
  - You can combine variables into new variable, e.g. `c <- a+b`

Saving your results: defining variables

- Variable names may not contain spaces or special characters except:
  - _ (underscore)
  - . (period)
  - - (dash)
- Variable names may not begin with a number: 3rd_result
- What does R tell you if you give the variable `mil` the value `1000000` and type in its name again? What does this mean?

More about variables

- It's good practice to give your variables comprehensible names, and not just letters like a, b, c

  ➔ For example:

  - `average_length` or
  - `syllable.count`

  (I will use a, b, c sometimes for the sake of brevity)

More about variables

- You can update the value of variables
- Take a look at what happens here:

```
> a <- 5
> a <- a+1
> a
[1] 6
```

Some practical tips

- As we progress our commands will get longer and more complex but they also recur
- With the **arrow up** key (↑) you can see previous commands:
  - You can edit and re-use these commands
  - You can see all the commands you've used in this session
- With the **tab** key you can auto-complete the names of variables and functions

Logical operations

- You can also ask R if something is **true**:

```
> 3 > 1
[1] TRUE
> 10 > 100
[1] FALSE
```

- And you can save the result in a variable

```
> a <- -1 < 0
> a
[1] TRUE
```

Logical operators and an exercise

Operator	Meaning
>	Greater than
>=	Greater or equals
<	Less than
<=	Less or equals
==	Equals
!=	Does not equal

- Give the variable x the **truth value** of the statement: √4 = 2

Vectors

- We run statistics on bunches of numbers – **"vectors"**.
- For example, the number of imperatives in four genres in the GUM corpus (http://corpling.uis.georgetown.edu/gum)

```
> imperatives <- c(11,0,32,329)
```

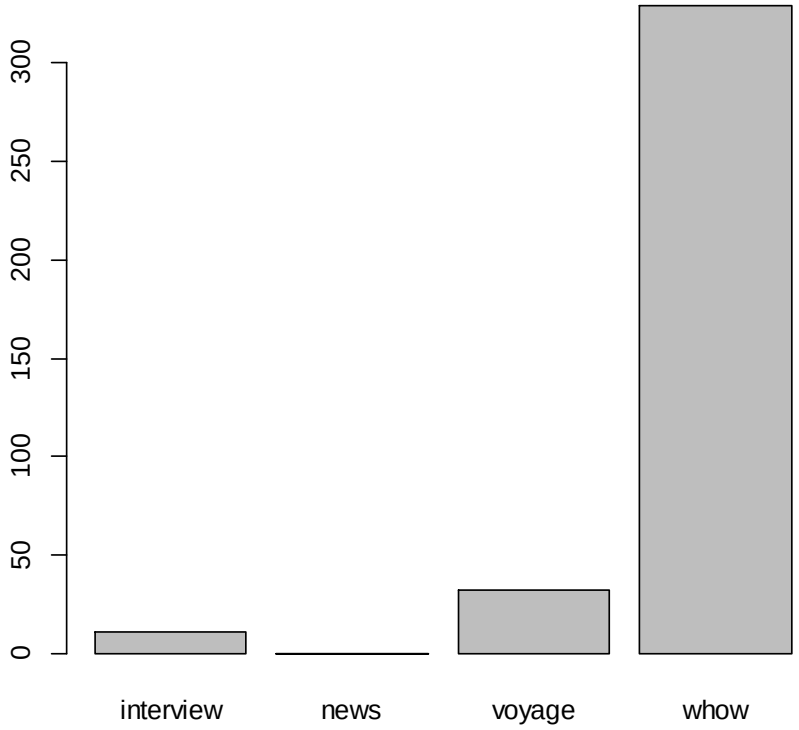
- We can also add names corresponding to each subcorpus:

```
> names(imperatives) <-
c("interview","news","voyage","whow")
```

Vectors

- And plot them like this:

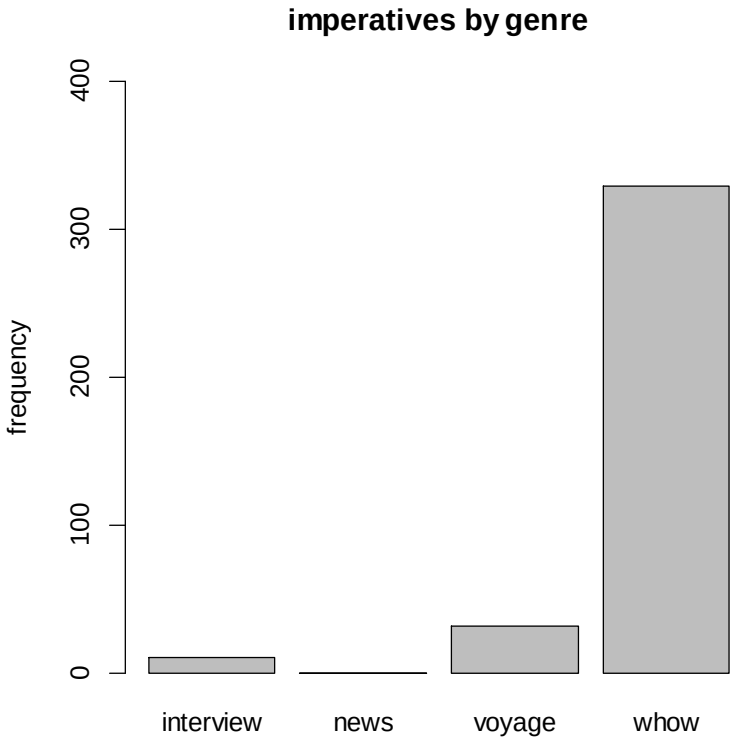
```
> barplot(imperatives)
```



Vectors

- You can change a lot of things about plots:

```
> barplot(imperatives, main="imperatives by
genre", ylim=c(0,400), ylab="frequency")
```



Vectors

- We can perform most mathematical operations on vectors just like numbers
  - Let's add a vector for indicative declaratives*:

```
> declaratives <- c(522,325,329,300)
> imperatives+declaratives
interview      news    voyage      whow
      533       325       361       629
> imperatives/declaratives
 interview       news     voyage       whow
0.02107280 0.00000000 0.09726444 1.09666667
```

*The full data to play with is in GUM_stypes.tab

Vectors

- You can also get the sum and length of a vector like this:

```
> length(imperatives)
[1] 4
> sum(declaratives)
[1] 1476
> sum(imperatives+declaratives)
[1] 1848
```

Using scripts

- It’s a good idea to not just type commands into the terminal
  - Write them down in a **script file**
  - Easier to reproduce and alter your analysis
  - Add comments like this: # Here I’m trying to…
- If you’re using RStudio, you can look at your current environment in detail
- Corresponding terminal functions also exist (e.g. ls() lists currently defined variables)

Functions so far

- c(), sum(), length(), names(), ls()
- barplot(data,main="x",ylim=c(x,y),ylab="freq")
- sqrt, log, sin, …

- You can learn more about a functions arguments and usage using e.g.: help(barplot)

Self-paced exercises

- We now turn to our first R worksheet
- The worksheet is available online
- The solution is posted as well
- Work at your own pace!
  - Feel free to ask questions
  - Work in groups if you like
  - Typing everything yourself once into your own computer is still recommended ☺