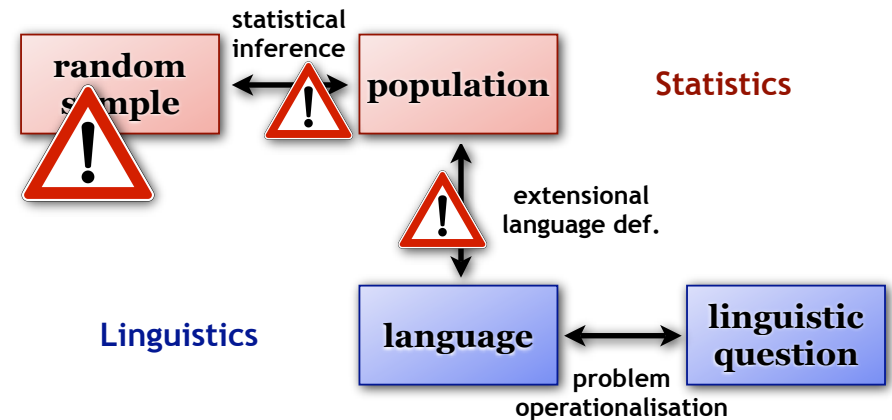


The role of statistics

Statistical Analysis of Corpus Data with R The Limitations of Random Sampling Models for Corpus Data

Marco Baroni¹ & Stefan Evert²
<http://purl.org/stefan.evert/SIGIL>

¹Center for Mind/Brain Sciences, University of Trento
²Institute of Cognitive Science, University of Osnabrück



2

Problem 1: Extensional language definition

- ◆ Are population proportions meaningful?
 - data from the BNC suggests ca. 9% of passive VPs in written English, little more than 2% in spoken English
 - note the difference from the 15% mentioned before!
- ◆ How much written language is there in English?
 - if we give equal weight to written and spoken English, proportion of passives is 5.5%
 - if we assume that English is 90% written language (as the BNC compilers did), the proportion is 8.3%
 - if it's mostly spoken (80%), proportion is only 3.4%

3

Problem 2: Statistical inference

- ◆ Inherent problems of particular hypothesis tests and their application to corpus data
 - χ^2 overestimates significance if any of the expected frequencies are low (Dunning 1993)
 - various rules of thumb: multiple $E < 5$, one $E < 1$
 - especially highly skewed tables in collocation extraction
 - G^2 overestimates significance for small samples (well-known in statistics, e.g. Agresti 2002)
 - e.g. manual samples of 100–500 items (as in our examples)
 - often ignored because of its success in computational linguistics
 - Fisher is conservative & computationally expensive
 - also numerical problems, e.g. in R version 1.x ☹️

4

Problem 2: Statistical inference

- ◆ Effect size for frequency comparison
 - not clear which measure of effect size is appropriate
 - e.g. **difference** of proportions, **relative risk** (ratio of proportions), **odds ratio**, logarithmic odds ratio, normalised X^2 , ...
- ◆ Confidence interval estimation
 - accurate & efficient estimation of confidence intervals for effect size is often very difficult
 - exact confidence intervals only available for odds ratio

5

Problem 3: Multiple hypothesis tests

- ◆ Each individual hypothesis test controls risk of type I error ... but if you carry out thousands of tests, some of them *have* to be false rejections
 - recommended reading: *Why most published research findings are false* (Ioannidis 2005)
 - a monkeys-with-typewriters scenario

6

Problem 3: Multiple hypothesis tests

- ◆ Typical situation e.g. for collocation extraction
 - test whether word pair cooccurs significantly more often than expected by chance
 - hypothesis test controls risk of type I error *if applied to a single candidate selected a priori*
 - but usually candidates selected *a posteriori* from data → many “unreported” tests for candidates with $f = 0$!
 - large number of such word pairs according to **Zipf's law** results in substantial number of type I errors
 - can be quantified with LNRE models (Evert 2004), cf. session on *word frequency distributions with zipfR*

7

Corpora

- ◆ Theoretical sampling procedure is impractical
 - it would be very tedious if you had to take a random sample from a library, especially a hypothetical one, every time you want to test some hypothesis
- ◆ Use pre-compiled sample: a **corpus**
 - but this is not a random sample of tokens!
 - would be prohibitively expensive to collect 10 million VPs for a BNC-sized sample at random
 - other studies will need tokens of different granularity (words, word pairs, sentences, even full texts)

8

The Brown corpus

- ◆ First large-scale electronic corpus
 - compiled in 1964 at Brown University (RI)
- ◆ 500 samples of approx. 2,000 words each
 - sampled from edited AmE published in 1961
 - from 15 domains (imaginative & informative prose)
 - manually entered on punch cards

9

The British National Corpus

- ◆ 100 M words of modern British English
 - compiled mainly for lexicographic purposes: Brown-type corpora (such as LOB) are too small
 - both written (90%) and spoken (10%) English
 - XML edition (version 3) published in 2007
- ◆ 4048 samples from 25 to 428,300 words
 - 13 documents < 100 words, 51 > 100,000 words
 - some documents are collections (e.g. e-mail messages)
 - rich metadata available for each document

10

Problem 4: Coverage & representativeness

- ◆ **Coverage:** does corpus include all material that falls under our extensional language definition?
 - some genres problematic for legal or practical reasons (e.g. private letters, conversation, printed books)
 - opportunistic data collection for large corpora: newspapers, parliamentary debates, Web as corpus
- ◆ **Representativeness:** different genres, speakers, etc. included in appropriate proportion?
 - you may not agree with 10% of spoken English in BNC
 - can be corrected for if problem is known and sufficiently detailed meta-information is available

11

Problem 5: Non-randomness

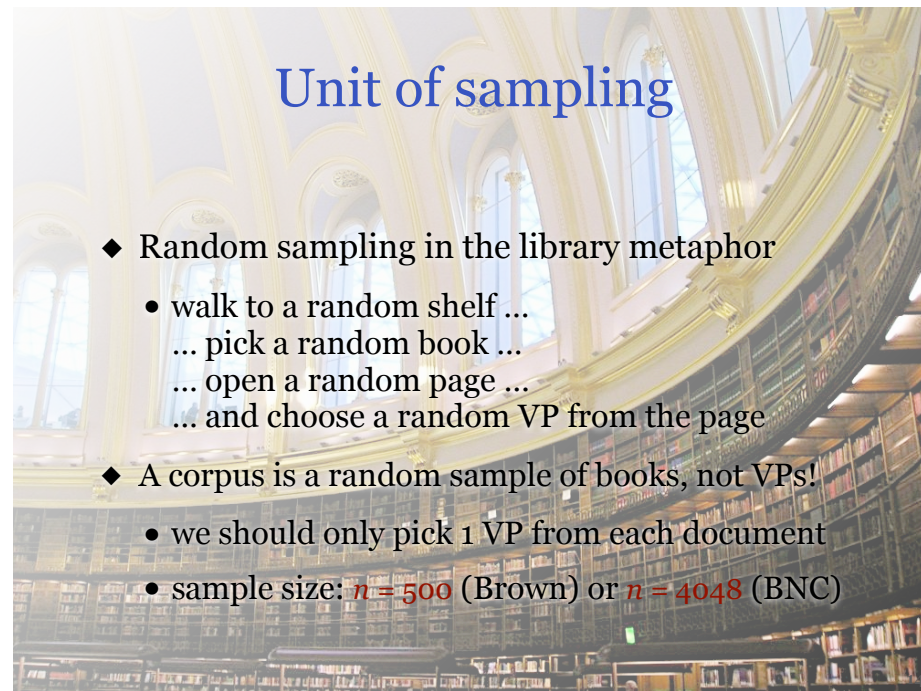


12

Unit of sampling

- ◆ Key problem: **unit of sampling** (text or fragment) \neq **unit of measurement** (e.g. VP)
 - recall sampling procedure in library metaphor ...

13



Unit of sampling

- ◆ Random sampling in the library metaphor
 - walk to a random shelf ...
 - ... pick a random book ...
 - ... open a random page ...
 - ... and choose a random VP from the page
- ◆ A corpus is a random sample of books, not VPs!
 - we should only pick 1 VP from each document
 - sample size: $n = 500$ (Brown) or $n = 4048$ (BNC)

Pooling data

- ◆ In order to obtain larger samples, researchers usually **pool** all data from a corpus
 - i.e. they include all VPs from each book
- ◆ Do you see why this is wrong?

15

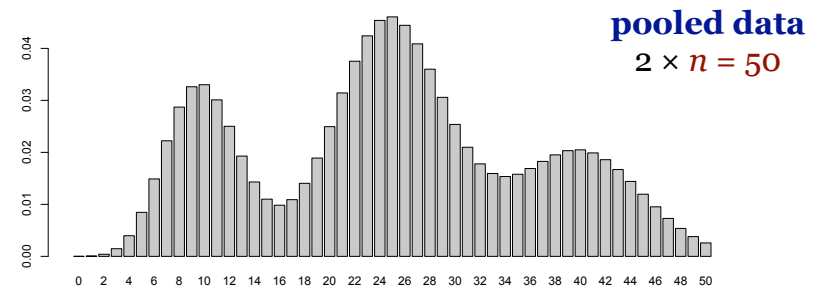
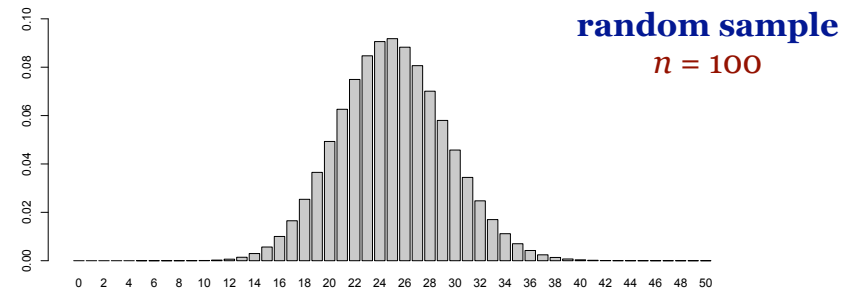
Pooling data

- ◆ Books aren't random samples themselves
 - each book contains relatively homogeneous material
 - much larger differences between books
- ◆ Therefore, pooled data isn't a random sample from the library
 - for each randomly selected VP, we co-select a substantial amount of very similar material
- ◆ Consequence: sampling variation increased

16

Pooling data

- ◆ Let us illustrate this with a simple example ...
 - assume library with two sections of equal size
 - population proportions are 10% vs. 40%
→ overall proportion of 25% in the library
- ◆ Compare sampling variation for
 - random sample of 100 tokens from the library
 - two randomly selected books of 50 tokens each
 - book is assumed to be a random sample from its section



17

18

Problem 5A: Duplicates

- ◆ Duplicates = extreme form of non-randomness
 - Did you know the British National Corpus contains duplicates of entire texts (under different names)?
- ◆ Duplicates can appear at any level
 - *The use of keys to move between fields is fully described in Section 2 and summarised in Appendix A*
 - 117 (!) occurrences in BNC, all in file HWX
 - very difficult to detect automatically
- ◆ Even worse for newspapers & Web corpora
 - see Evert (2004) for examples

19

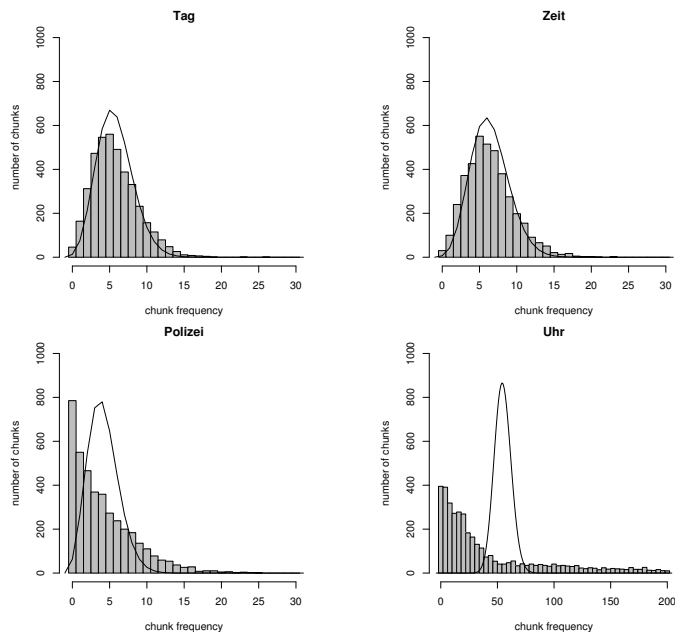
Problem 5B: (Lexical) specialisation

- ◆ Illustrated by data pooling example
 - true population proportions usually different in distinct sections of the library (e.g. spoken vs. written English, different genres, registers, domains, ...)
 - if you pick just a few books, it is likely that some sections will be seriously over-represented
- ◆ Specialisation increases sampling variation
 - even if each book is a random sample from its section!

20

Problem 5B: Lexical specialisation

- ◆ Particularly serious (and well-known) problem for lexical phenomena (words, collocations, ...)
- ◆ Specialisation wrt. domain and topic
 - a book about a football team will use an entirely different vocabulary than a statistics textbook or a romantic novel
 - usually not enough meta-information about topics available to split corpus into homogeneous sections
- ◆ See e.g. Baayen (1996)



Data from *Frankfurter Rundschau* corpus, divided into 10,000 equally-sized chunks

21

23

Problem 5C: Term clustering

- ◆ If a “content” word occurs once in a document, it is very likely to occur again
 - *The chance of two Noriegas is closer to $p/2$ than p^2* (Church 2000; also Church & Gale 1995, Katz 1996)
 - i.e. documents are *not* random samples
- ◆ Two complementary effects:
 - specialisation = non-randomness between documents
 - term clustering = non-randomness within documents

22

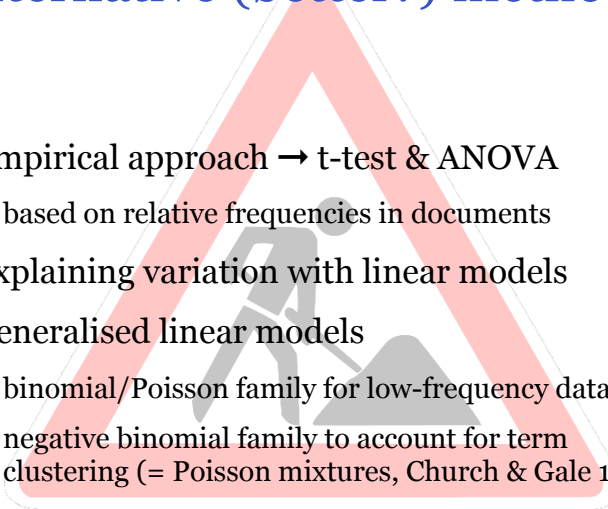
The screenshot shows a web browser window with the address bar containing <http://sigil.collocations.de/wizard.html>. The page title is "SIGIL: Corpus Frequency Wizard". Below the title, there is a section for "SIGIL: Corpus Frequency Test Wizard". The page content is partially obscured by a large red X. Visible text includes "This site provides some online tools for the project Statistical Inference: A Gentle Introduction for Linguistics (SIGIL) by Marco Baroni and Stefan Evert. The main SIGIL page can be found at purl.org/stefan.evert/SIGIL". There are two main sections: "One sample: frequency estimation (confidence interval)" and "Two samples: frequency comparison". The "One sample" section has input fields for "Frequency count" (19) and "Sample size" (100), and a "Calculate" button. The "Two samples" section has input fields for "Sample 1" (Frequency count: 25, Sample size: 100) and "Sample 2" (Frequency count: 25, Sample size: 200), and a "Calculate" button. At the bottom of the page, the URL <http://sigil.collocations.de/wizard.html> is displayed in red text.

Cave canem!

- ◆ Treat statistical methods based on random sampling assumptions with great caution!
 - doesn't mean statistical analysis should be discarded
 - random variation is a lower bound on true variability
- ◆ Can still be useful for the analysis of corpus data, but may also give very misleading answers
- ◆ **Always look at your data!**
 - R helps you to know & understand what you're doing (unlike online wizards and many commercial tools)

25

Alternative (better?) methods

- 
- ◆ Empirical approach → t-test & ANOVA
 - based on relative frequencies in documents
 - ◆ Explaining variation with linear models
 - ◆ Generalised linear models
 - binomial/Poisson family for low-frequency data
 - negative binomial family to account for term clustering (= Poisson mixtures, Church & Gale 1995)

26

*Thank you
for following this course!*

Stefan & Marco

27

References (1)

- ◆ Agresti, Alan (2002). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, 2nd edition.
- ◆ Baayen, R. Harald (1996). *The effect of lexical specialization on the growth curve of the vocabulary*. *Computational Linguistics*, **22**(4), 455–480.
- ◆ Baroni, Marco and Evert, Stefan (2008, in press). *Statistical methods for corpus exploitation*. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 38. Mouton de Gruyter, Berlin.
- ◆ Church, Kenneth W. (2000). *Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2* . In *Proceedings of COLING 2000*, pages 173–179, Saarbrücken, Germany.
- ◆ Church, Kenneth W. and Gale, William A. (1995). *Poisson mixtures*. *Journal of Natural Language Engineering*, **1**, 163–190.
- ◆ Dunning, Ted E. (1993). *Accurate methods for the statistics of surprise and coincidence*. *Computational Linguistics*, **19**(1), 61–74.

28

References (2)

- ◆ Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714.
- ◆ Evert, Stefan (2006). *How random is a corpus?* The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, **54**(2), 177–190.
- ◆ Ioannidis, John P. A. (2005). *Why most published research findings are false*. *PLoS Medicine*, **2**(8), 696–701.
- ◆ Katz, Slava M. (1996). *Distribution of content words and phrases in text and language modelling*. *Natural Language Engineering*, **2**(2), 15–59.
- ◆ Kilgarriff, Adam (2005). *Language is never, ever, ever, random*. *Corpus Linguistics and Linguistic Theory*, **1**(2), 263–276.
- ◆ Rietveld, Toni; van Hout, Roeland; Ernestus, Mirjam (2004). *Pitfalls in corpus research*. *Computers and the Humanities*, **38**, 343–362.