

Unit 8: Non-Randomness of Corpus Data & Generalised Linear Models

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

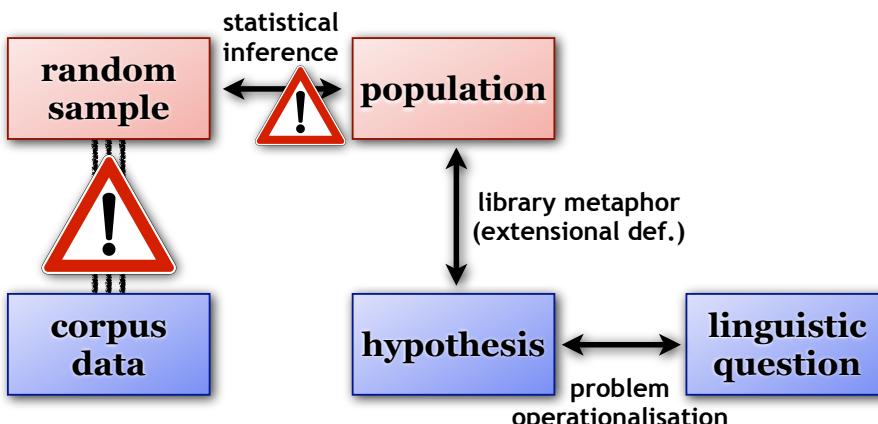
¹Center for Mind/Brain Sciences, University of Trento

²Institute of Cognitive Science, University of Osnabrück

Introduction & Reminder

2

Problems with statistical inference



3

Mathematical problems: Significance

- ◆ Inherent problems of particular hypothesis tests and their application to corpus data
 - X^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

Mathematical problems: Effect size

- ◆ 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 χ^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

Mathematical problems: Multiple hypothesis tests

- ◆ Typical situation e.g. for collocation extraction
 - test whether word pair co-occurs 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. Unit 5 on **word frequency distributions with zipfR**

7

Mathematical problems: 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



Why a corpus isn't a random sample

8

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)

9

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

10

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

11

Unit of sampling

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

12



Unit of sampling

- ◆ Random sampling in the library metaphor
 - walk to a random shelf ...
 - ... select a random book ...
 - ... open it on a random page ...
 - ... and pick a random sentence from the page
 - repeat n times for sample size n
- ◆ Corpus = random sample of books, not sentences!
 - we should only use 1 sentence from each book
 - 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 sentences from each book
- ◆ Do you see why this is wrong?

14

Pooling data

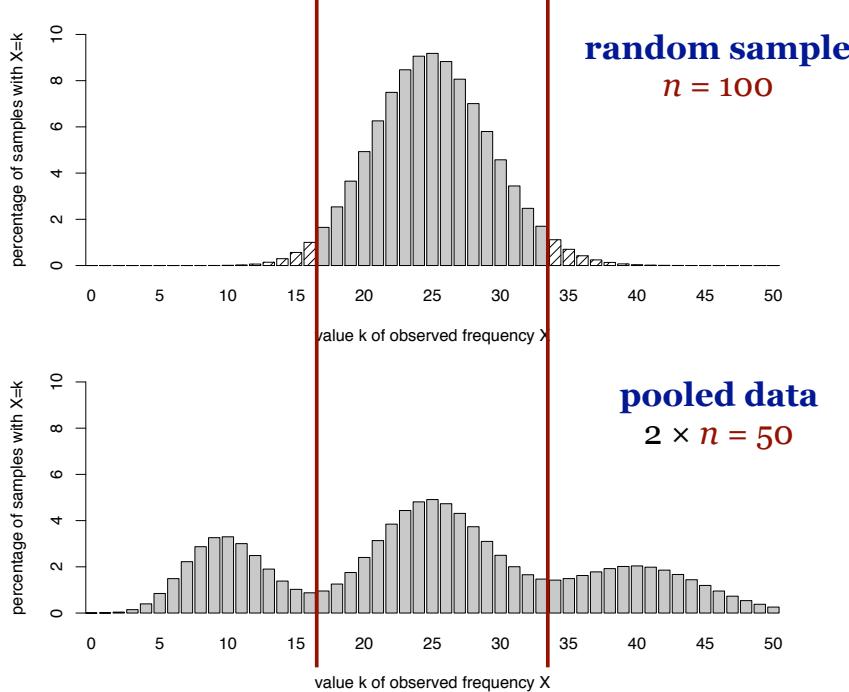
- ◆ Books aren't random samples themselves
 - each book contains relatively homogeneous material
 - but much larger differences *between* books
- ◆ Therefore, the pooled data do not form a random sample from the library
 - for each randomly selected sentence, we co-select a substantial amount of very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

Pooling data

- ◆ Let us illustrate this with a simple example ...
 - assume library with two sections of equal size
 - e.g. spoken and written language in a corpus
 - population proportions are **10%** vs. **40%**
 - overall proportion of $\pi = 25\%$ in the library
 - this is the null hypothesis H_0 that we will be testing
- ◆ 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

15

16



Duplicates

- ◆ Duplication = 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

18

A sample of random samples is a random sample



Measuring non-randomness

- ◆ Larger unit of sampling is not the original cause of non-randomness
 - if each text in a corpus is a genuinely random sample from the same population, then the pooled data also form a random sample
 - we can illustrate this with a thought experiment

19

20

The random library

- ◆ Suppose there's a **vandal** in the library
 - who cuts up all books into single sentences and leaves them in a big heap on the floor
 - the next morning, the librarian takes a handful of sentences from the heap, fills them into a book-sized box, and puts the box on one of the shelves
 - repeat until the heap of sentences is gone
 - ◆ library of **random samples**
 - ◆ Pooled data from 2 (or more) boxes form a perfectly **random sample** of sentences from the original library!
- 

A sample of random samples is a random sample

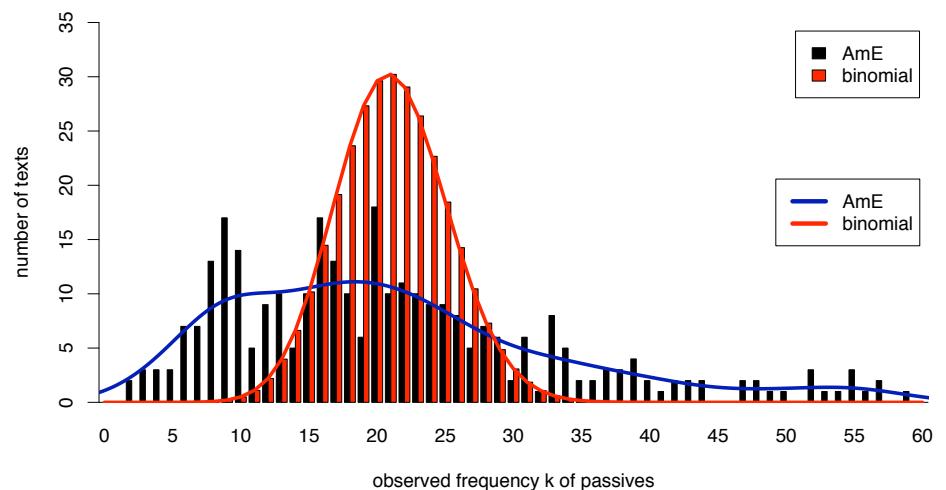
- ◆ The true cause of non-randomness
 - discrepancy between unit of sampling and unit of measurement only leads to non-randomness if the sampling units (i.e. the corpus texts) are not random samples themselves (from same population)
 - with respect to specific phenomenon of interest
- ◆ No we know how to measure non-randomness
 - find out if corpus texts are random samples
 - i.e., if they follow a binomial sampling distribution
 - tabulate observed frequencies across corpus texts

22

Measuring non-randomness

- ◆ Tabulate number of texts with ***k*** passives
 - illustrated for subsets of Brown/LOB (310 texts each)
 - meaningful because all texts have the same length
- ◆ Compare with binomial distribution
 - for population proportion $H_0 : \pi = 21.1\%$ (Brown) and $\pi = 22.2\%$ (LOB); approx. $n = 100$ sentences per text
 - estimated from full corpus → best possible fit
- ◆ Non-randomness → larger sampling variation

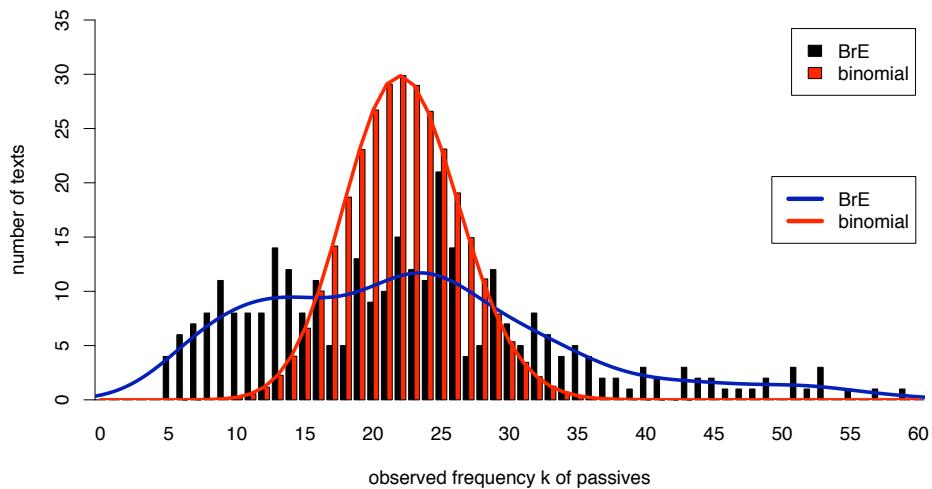
Passives in the Brown corpus



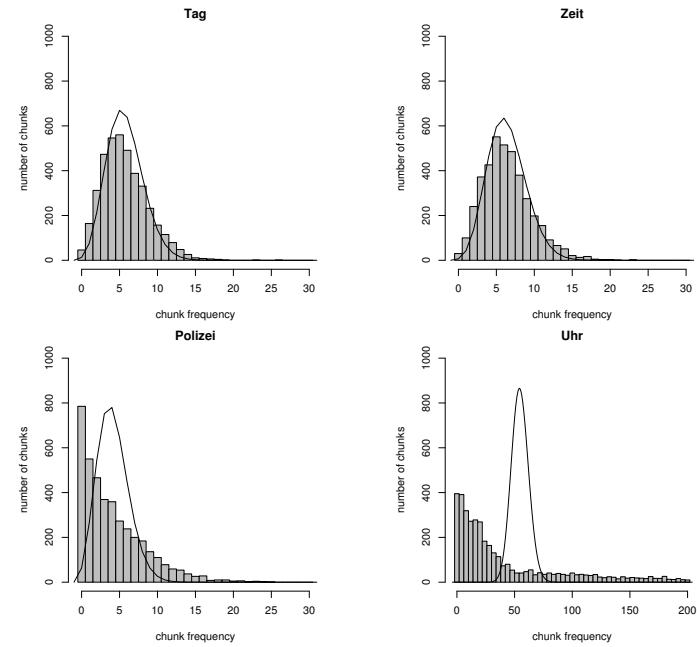
23

24

Passives in the LOB corpus



25



26

Consequences of non-randomness



Consequences

- ◆ Accept that corpus is a **sample of texts**
 - data cannot be pooled into random sample of tokens
 - results in much smaller sample size ...
(BNC: 4,048 texts rather than 6,023,627 sentences)
 - ... but more informative measurements (relative frequencies on interval rather than nominal scale)
- ◆ Use statistical techniques that account for the **overdispersion** of relative frequencies
 - Gaussian distribution allows us to estimate **spread** (variance) independently from **location**
 - Standard technique: **Student's t-test**

27

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

28

A case study: Passives in AmE and BrE

- ◆ Are there more passives in BrE than in AmE?
 - based on data from subsets of Brown and LOB
 - 9 categories: press reports, editorials, skills & hobbies, misc., learned, fiction, science fiction, adventure, romance
 - ca. 310 texts / 31,000 sentences / 720,000 words each
- ◆ Pooled data (random sample of sentences)
 - AmE: 6584 out of 31,173 sentences = 21.1%
 - BrE: 7091 out of 31,887 sentences = 22.2%
- ◆ Chi-squared test (→ pooled data, binomial) vs. t-test (→ sample of texts, Gaussian)



29

A case study: Passives in AmE and BrE

- ◆ Chi-squared test: **highly significant**
 - p-value: .00069 < .001
 - confidence interval for difference: 0.5% – 1.8%
 - large sample → large amount of evidence
- ◆ t-test: **not significant**
 - p-value: .1340 > .05 ($t=1.50$, $df=619.96$)
 - confidence interval for difference: -0.6% – +4.9%
 - H_0 : same average relative frequency in AmE and BrE

30

What are we really testing?

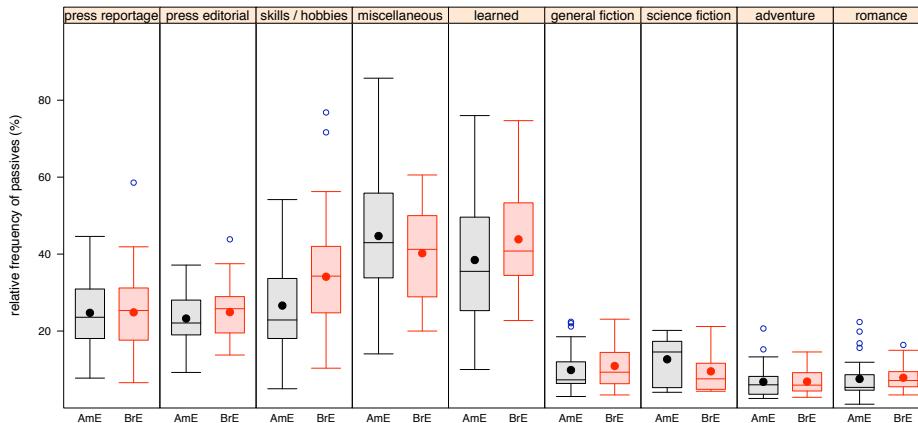
- ◆ Are population proportions meaningful?
 - corpus should be **balanced** and **representative** (broad coverage of genres, ... in appropriate proportions)
 - average frequency depends on composition of corpus
 - e.g. 18% passives in written BrE / 4% in spoken BrE
- ◆ How many passives are there in English?
 - 50% written / 50% spoken: $\pi = 13.0\%$
 - 90% written / 10% spoken: $\pi = 16.6\%$
 - 20% written / 80% spoken: $\pi = 6.8\%$

31

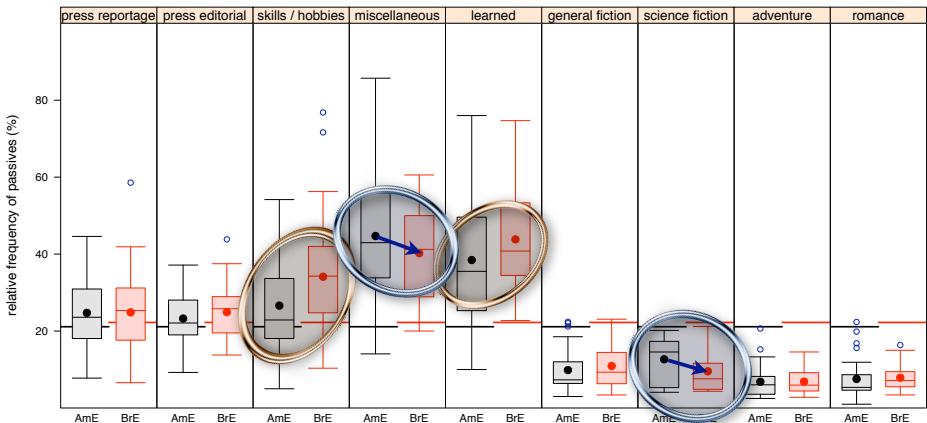
32

Average relative frequency?

Average relative frequency?

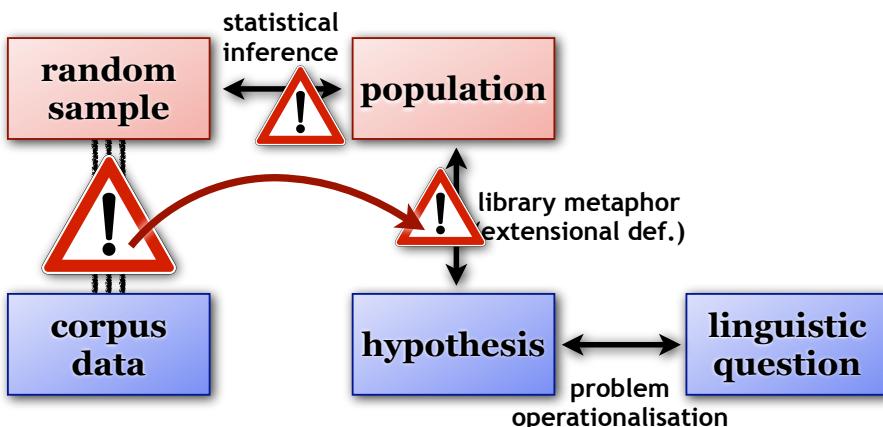


33



34

Problems with statistical inference



35

5

Rethinking corpus frequencies

36

Studying variation in language

- ◆ It seems absurd now to measure & compare relative frequencies in “language” (= library)
 - proportion π depends more on composition of library than on properties of the language itself
- ◆ Quantitative corpus analysis has to account for the variation of relative frequencies between individual texts (cf. Gries 2006)
 - research question → one factor behind this variation

37

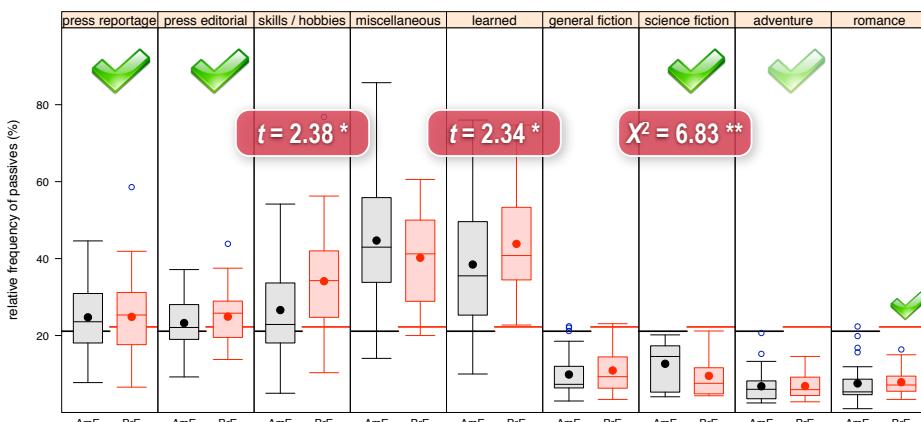
Studying variation in language

- ◆ Approach 1: restrict study to sublanguage in order to eliminate non-randomness
 - data from this sublanguage (= single section in library) can be pooled into large random sample
- ◆ Approach 2: goal of quantitative corpus analysis is to explain variation between texts in terms of
 - random sampling (of tokens within text)
 - stylistic variation: genre, author, domain, register, ...
 - subject matter of text → term clustering effects
 - differences between language varieties

research question

38

Eliminating non-randomness



39

Explaining variation

- ◆ Statisticians explain variation with the help of linear models (and other statistical models)
 - linear models predict response (“dependent variable”) from one or more factors (“independent variables”)
 - simplest model: linear combination of factors
- ◆ Linear model for passives in AmE and BrE:

$$p_i = \beta_0 + \beta_1(\text{genre}) + \beta_2(\text{AmE/BrE}) + \epsilon_i$$

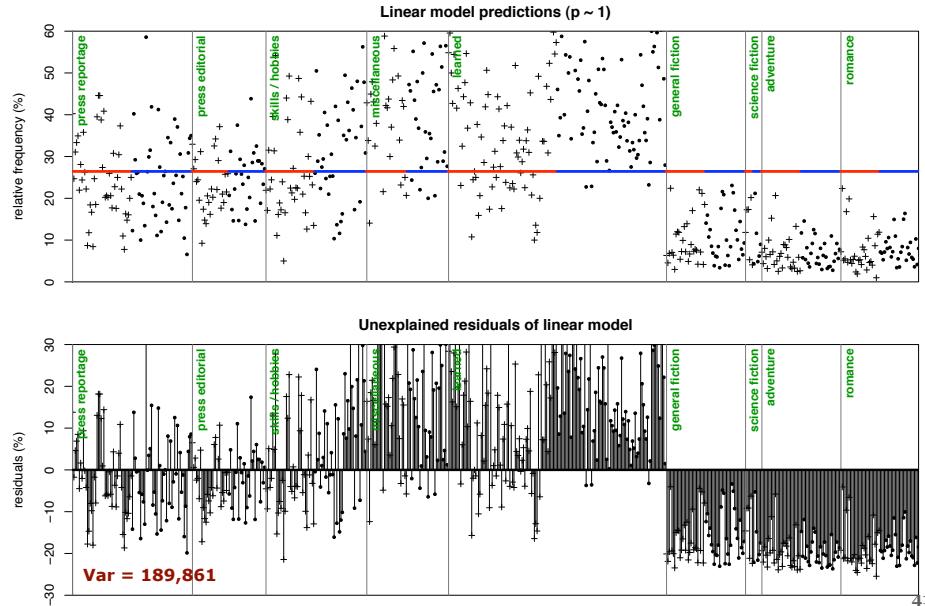
relative frequency in text i

overall average
 “intercept”
 unexplained “residuals”
 + sampling variation

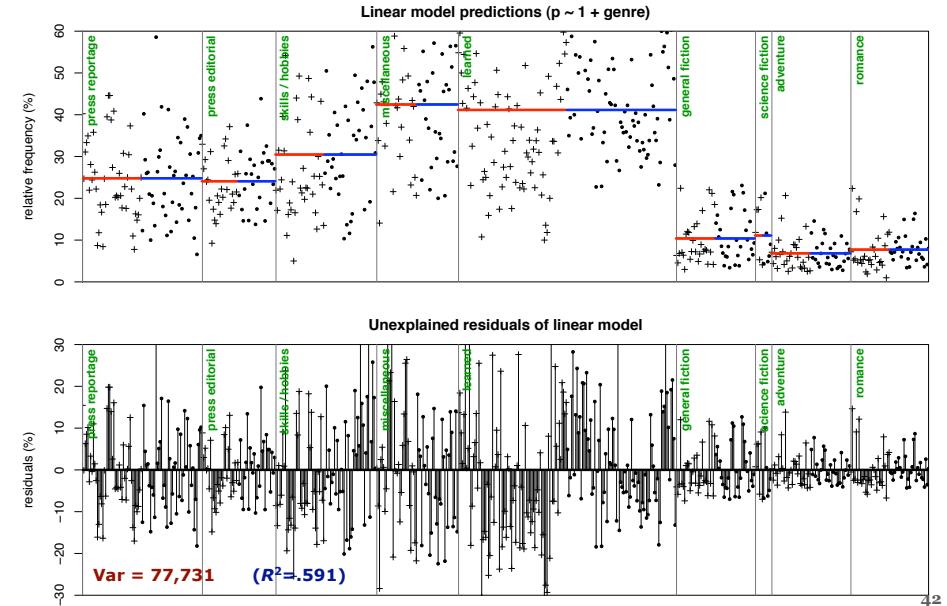
I'm just an ANOVA ...

40

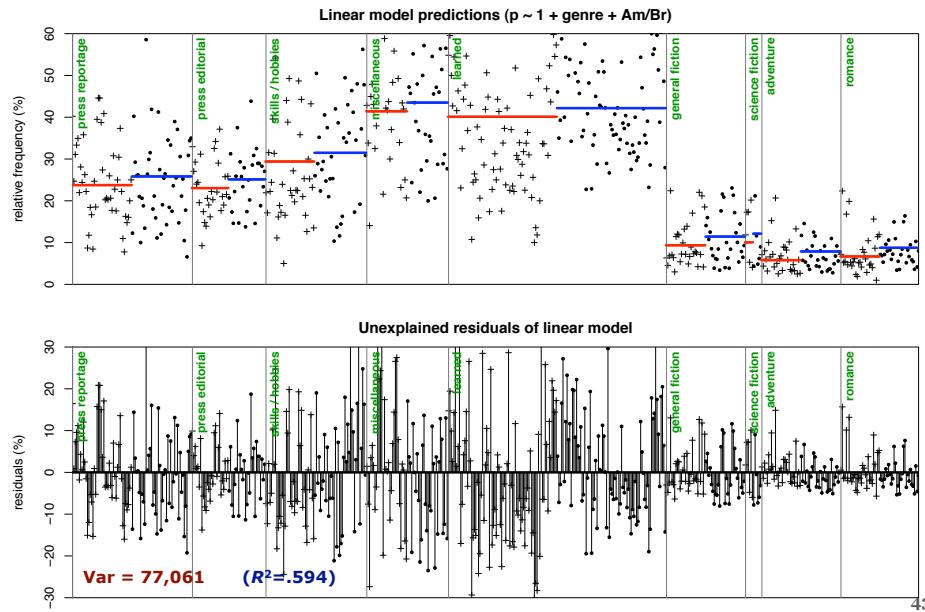
Linear model for passives



Linear model for passives



Linear model for passives



Linear model for passives

◆ Goodness-of-fit (analysis of variance)

- total variance (sum of squares): **189,861**
- explained by genre***: **112,113 (= 59.0%)**
- explained by AmE/BrE*: **687 (= 0.4%)**
- unexplained (residuals): **77,061 (= 40.6%)**

◆ Is variance explained well enough?

- binomial sampling variation: **ca. 10,200 (= 5.4%)**

Linear model for passives

- ◆ F-tests show significant effects of genre ($p < 10^{-15}$) and AmE / BrE ($p = .0198$)
- ◆ 95% confidence intervals for effect sizes:
 - AmE / BrE: 0.3% ... 3.8%
 - genre = learned 13.4% ... 19.3%
 - compared to “press reportage” genre as baseline
 - genre = romance -20.8% ... -13.4%
 - genre = ...

45



Linear models are not appropriate!

46

Why linear models are not appropriate for frequency data

- ◆ Binomial sampling variation not accounted for
- ◆ Normality assumption (error terms)
 - Gaussian approximation inaccurate for low-frequency data (with non-zero probability for negative counts!)
- ◆ Homoscedasticity (equal variances of errors)
 - variance of binomial sampling variation depends on population proportion and sample size
 - different sample sizes (texts in Brown/LOB: 40 – 250 sentences; huge differences in BNC)
- ◆ Predictions not restricted to range 0% – 100%

47

Generalised linear models

◆ Generalised linear models (GLM)

- account for binomial sampling variation of observed frequencies and different sample sizes
- allow non-linear relationship between explanatory factors and predicted relative frequency (π_i)

$$f_i \sim B(n_i, \pi_i) \quad \text{binomial sampling ("family")}$$

$$\pi_i = \frac{1}{1 + e^{-\theta_i}} \quad \text{"link" function}$$

$$\theta_i = \beta_0 + \beta_1(\text{genre}) + \beta_2(\text{AmE/BrE}) \quad \text{linear predictor} \rightarrow \theta_i = \beta_0 + \beta_1(\text{genre}) + \beta_2(\text{AmE/BrE})$$

48

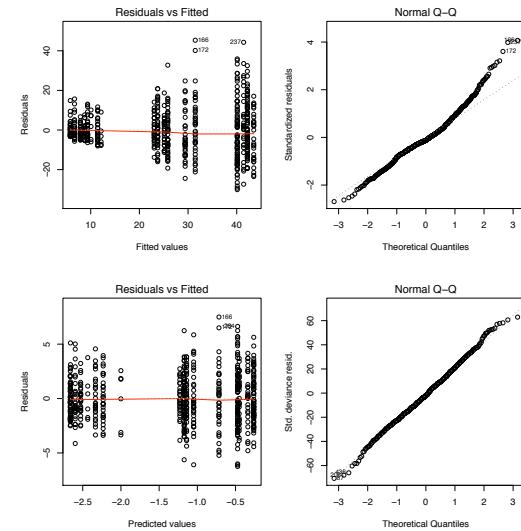
GLM for passives

- ◆ Goodness-of-fit (analysis of deviance)
 - total deviance (“unlikelihood”): **13,265**
 - explained by genre***: **8,275** (= **62.4%**)
 - explained by AmE/BrE***: **36** (= **0.3%**)
 - unexplained (residual deviance): **4,953** (= **37.3%**)
 - binomial sampling variation: $\approx 1,000$ (= **7.5%**)
- ◆ Interpretation of confidence intervals difficult

49

Model diagnostics

Linear Model



Generalised Linear Model

Still no satisfactory explanation for observed variation in frequency of passives between texts!

50

Take-home messages

- ◆ Don't trust statistic(ian)s blindly
 - You know how complex language really is!
 - linguists and statisticians should work together
- ◆ No excuse to avoid significance testing
 - good reasons to believe that binomial sampling distribution is a lower bound on variation in language
- ◆ Needed: large corpora with rich metadata
 - study & “explain” variation with statistical models
 - full data need to be available (not Web interfaces!)

51



52

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). *Statistical methods for corpus exploitation*. In A. Lüdeling and M. Kyö (eds.), *Corpus Linguistics. An International Handbook*, chapter 38. Mouton de Gruyter, Berlin.
- ◆ Church, Kenneth W. (2000). *Empirical estimates of adaptation: The chance of two Noriega's is closer to p/2 than p²*. 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.

53

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.
- ◆ Gries, Stefan Th. (2006). *Exploring variability within and between corpora: some methodological considerations*. *Corpora*, **1**(2), 109–151.
- ◆ Gries, Stefan Th. (2008). *Dispersions and adjusted frequencies in corpora*. *International Journal of Corpus Linguistics*, **13**(4), 403–437.
- ◆ Ioannidis, John P. A. (2005). *Why most published research findings are false*. *PLoS Medicine*, **2**(8), 696–701.

54

References (3)

- ◆ 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.
- ◆ Rayson, Paul; Berridge, Damon; Francis, Brian (2004). *Extending the Cochran rule for the comparison of word frequencies between corpora*. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles* (JADT 2004), pages 926–936, Louvain-la-Neuve, Belgium.
- ◆ McEnery, Tony and Wilson, Andrew (2001). *Corpus Linguistics*. Edinburgh University Press, 2nd edition.
- ◆ Rietveld, Toni; van Hout, Roeland; Ernestus, Mirjam (2004). *Pitfalls in corpus research*. *Computers and the Humanities*, **38**, 343–362.

55