

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

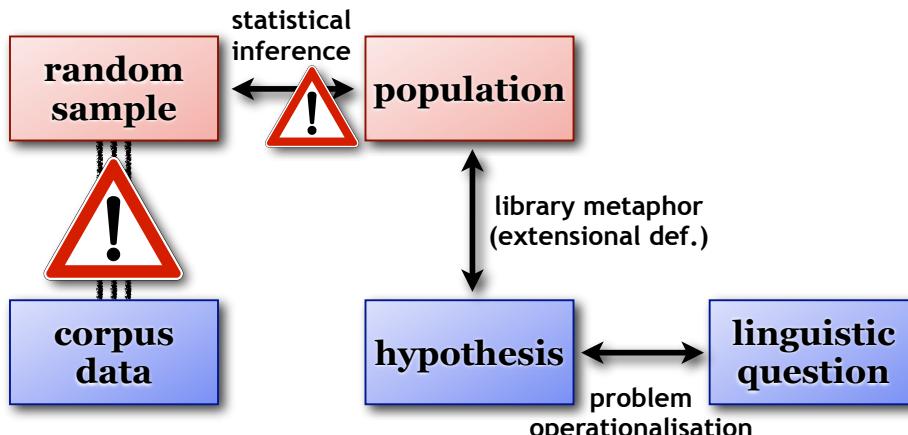
Marco Baroni¹ & Stefan Evert²
<http://purl.org/stefan.evert/SIGIL>

¹Center for Mind/Brain Sciences, University of Trento

²Corpus Linguistics Group, FAU Erlangen-Nürnberg

Introduction & Reminder

Problems with statistical inference



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 😞

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

- ◆ 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

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



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

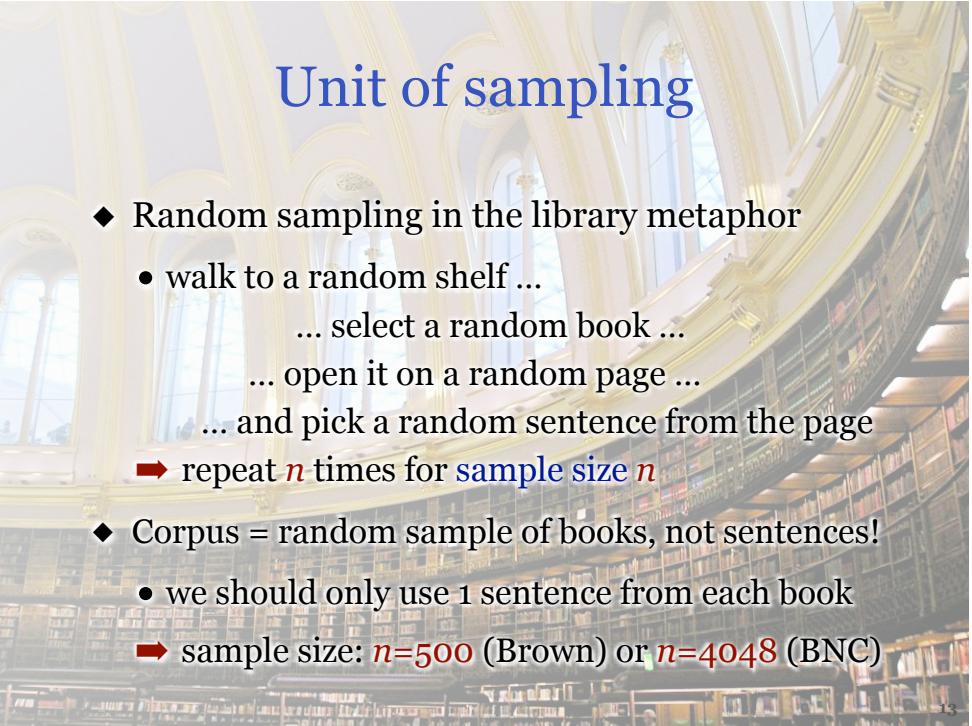
- ◆ 100 M words of modern British English
 - compiled mainly for lexicographic purposes: Brown-type corpora (such as LOB) are too small
 - both written (90%) and spoken (10%) English
 - XML edition (version 3) published in 2007
- ◆ 4048 samples from 25 to 428,300 words
 - 13 documents < 100 words, 51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

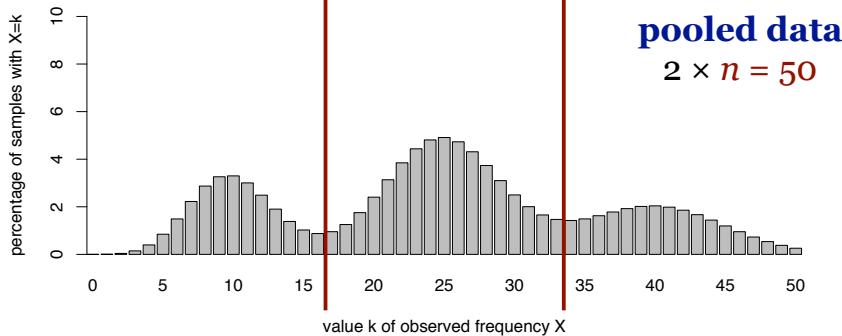
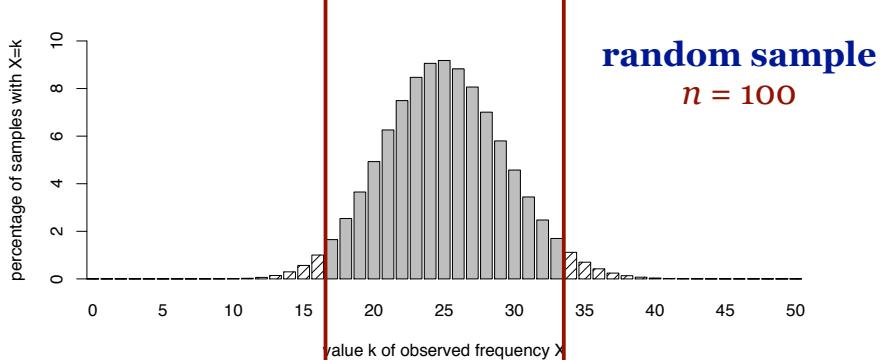
- ◆ Books aren't random samples themselves
 - each book contains relatively homogeneous material
 - but much larger differences *between* books
- ◆ Therefore, the pooled data do not form a random sample from the library
 - for each randomly selected sentence, we co-select a substantial amount of very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

17

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

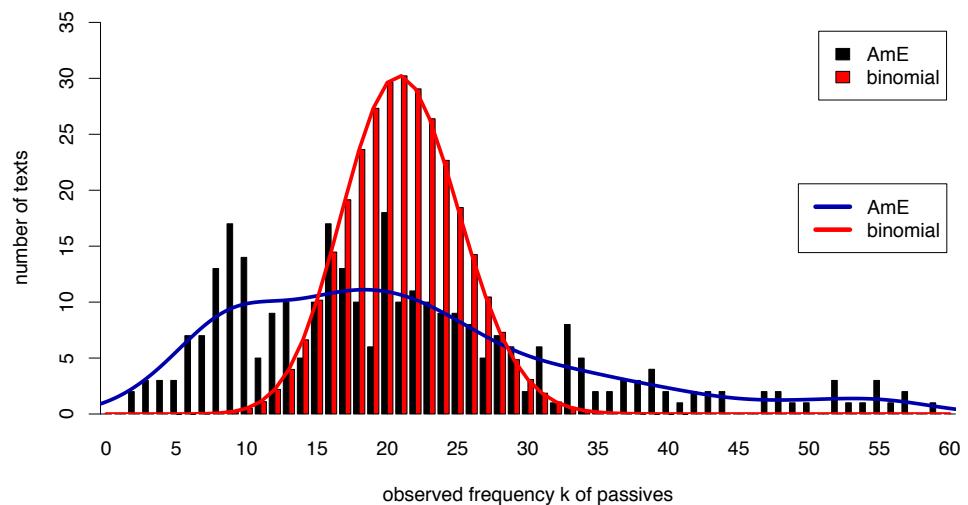
- ◆ The true cause of non-randomness
 - discrepancy between unit of sampling and unit of measurement only leads to non-randomness if the sampling units (i.e. the corpus texts) are not random samples themselves (from same population)
 - with respect to specific phenomenon of interest
- ◆ No we know how 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

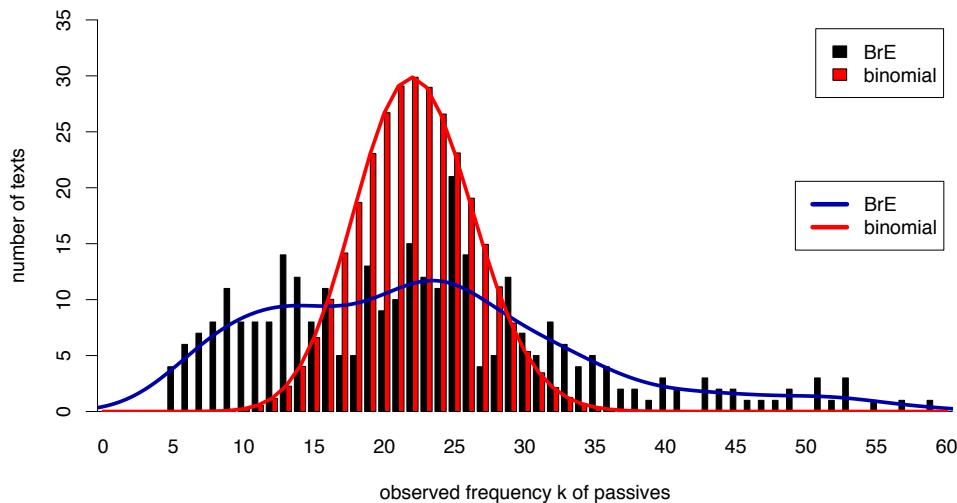
Passives in the Brown corpus



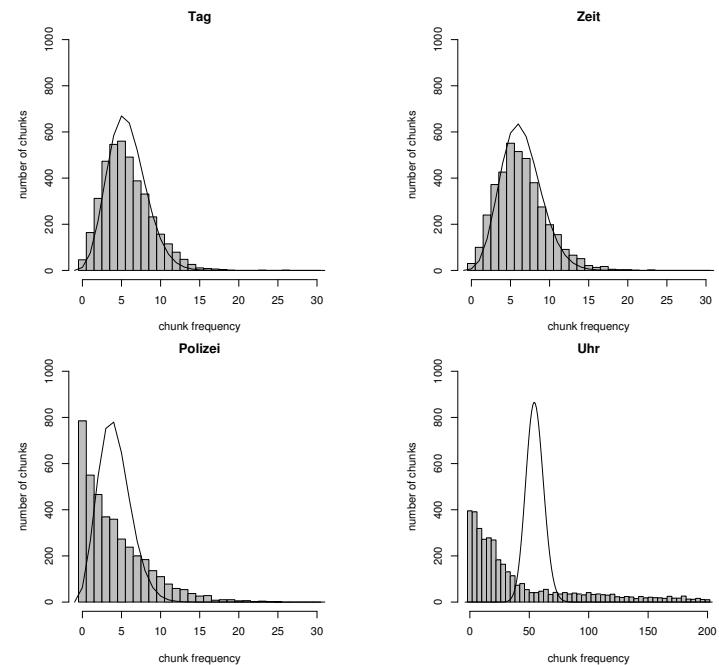
23

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
- ◆ R code: pooled counts + proportions test

```
> passives.B <- sum(Brown$passive)
> n_s.B <- sum(Brown$n_s)
> passives.L <- sum(LOB$passive)
> n_s.L <- sum(LOB$n_s)
> prop.test(c(passives.L, passives.B),
            c(n_s.L, n_s.B))
```

Let's do that in R ...

```
# passive counts for each text in Brown and LOB corpus
> library(SIGIL)
> Passives <- BrownLOBPassives

# display 10 random rows to get an idea of the table layout
> Passives[sample(nrow(Passives), 10), ]

# annotate relative frequency of passives in each file (%)
> Passives <- transform(Passives,
                           relfreq = 100 * passive / n_s)

# split into separate data frames for Brown and LOB texts
> Brown <- subset(Passives, lang == "AmE")
> LOB <- subset(Passives, lang == "BrE")
```

30

A case study: Passives in AmE and BrE

- ◆ 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
- ◆ R code: t.test() function (from Unit #3)

```
> t.test(LOB$relfreq, Brown$relfreq)
# alternative syntax: "formula" interface
> t.test(relfreq ~ lang, data=Passives)
# visualize the distributions being compared
> boxplot(relfreq ~ lang, data=Passives)
```

Why does this put AmE first?

31

32



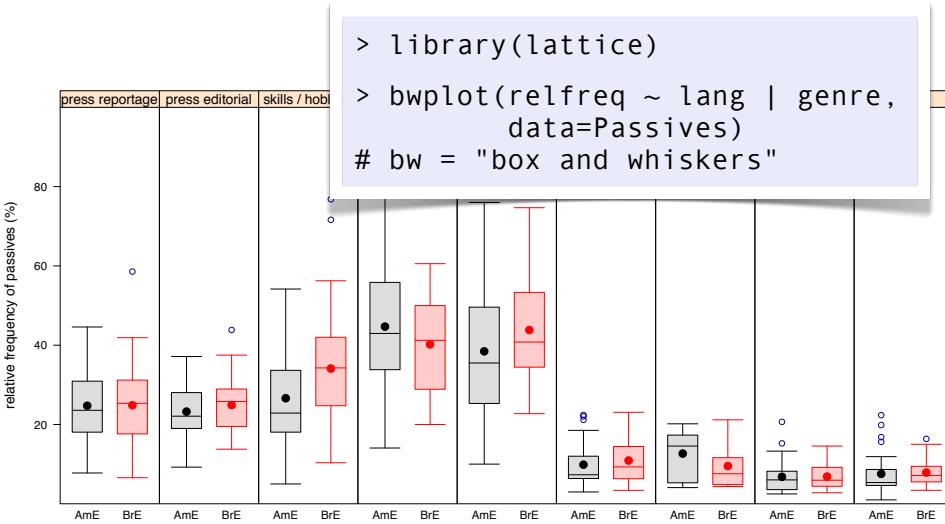
What are we really testing?

- ◆ 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: $\pi = 6.8\%$

33

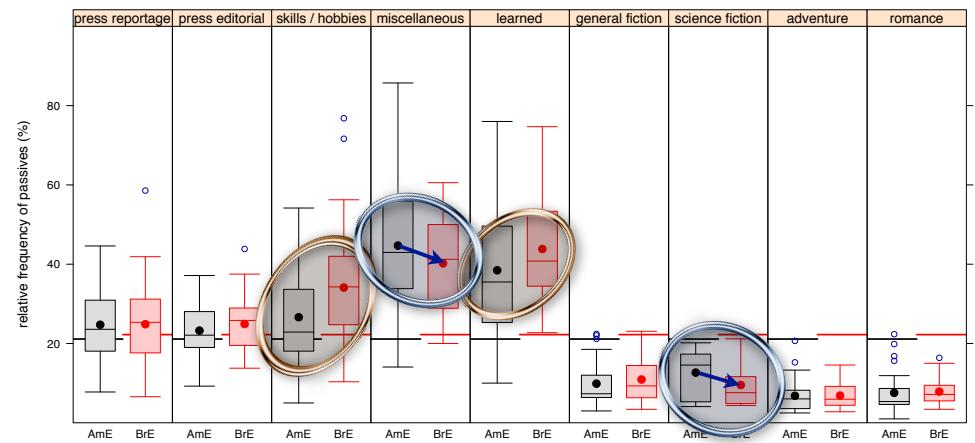
34

Average relative frequency?



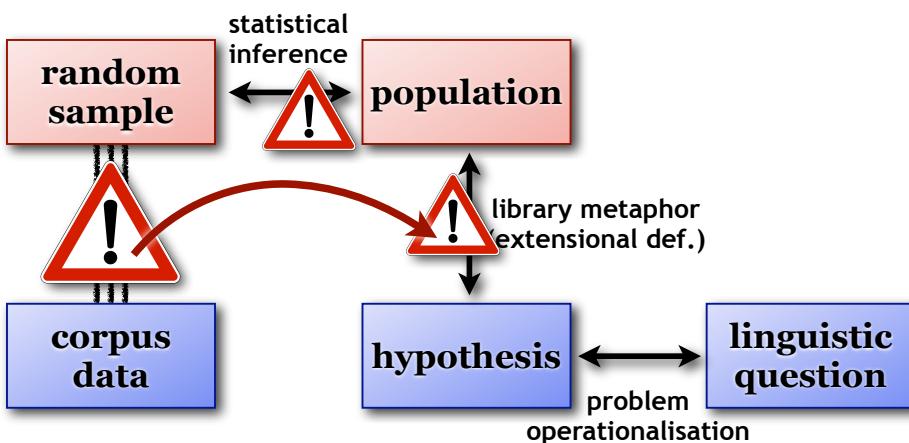
35

Average relative frequency?



36

Problems with statistical inference



37

5

Rethinking corpus frequencies

38

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

39

Studying variation in language

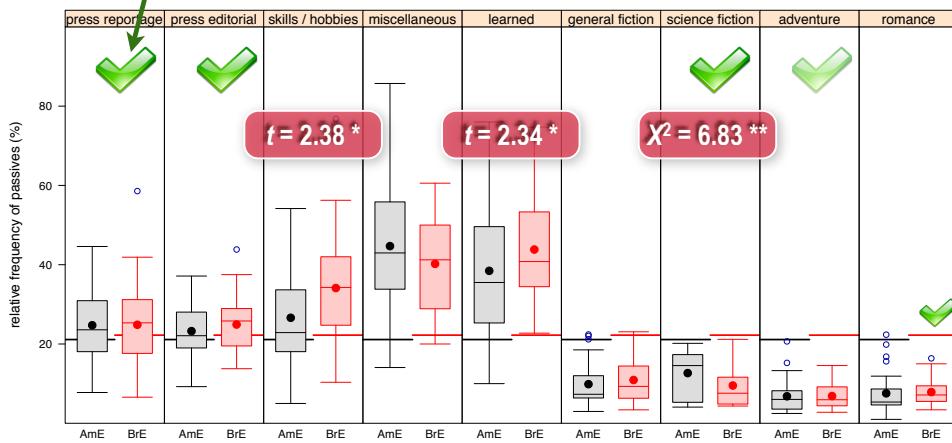
- ◆ **Approach 1:** restrict study to sublanguage in order to eliminate non-randomness
 - data from this sublanguage (= single section in library) can be pooled into large random sample
- ◆ **Approach 2:** goal of quantitative corpus analysis is to **explain variation** between texts in terms of
 - random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

research question

40

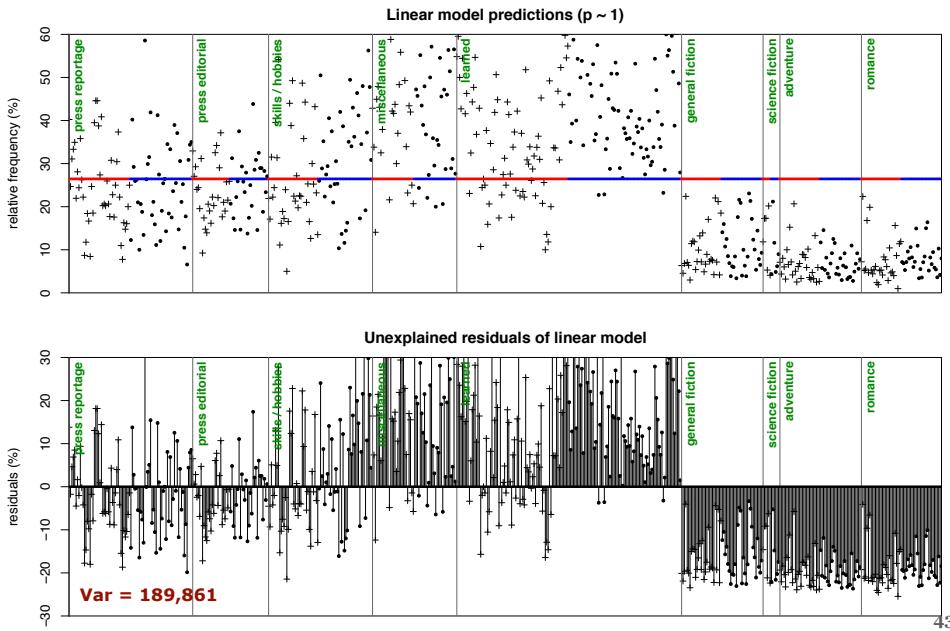
Eliminating non-randomness

binomial distribution
across texts (almost)



41

Linear model for passives



43

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:

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

I'm just an ANOVA ...

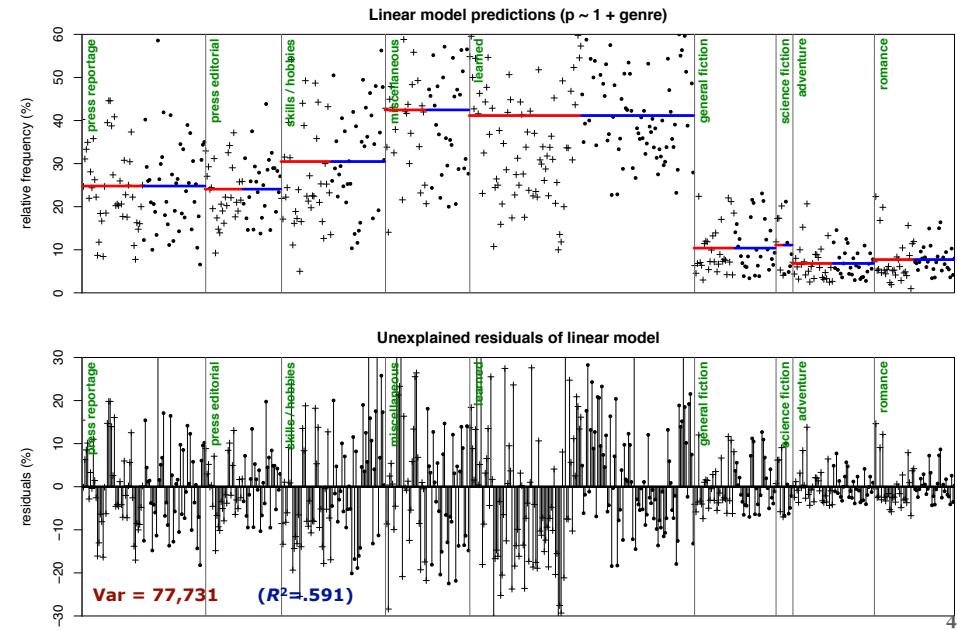
relative frequency in text i

overall average "intercept"

unexplained "residuals" + sampling variation

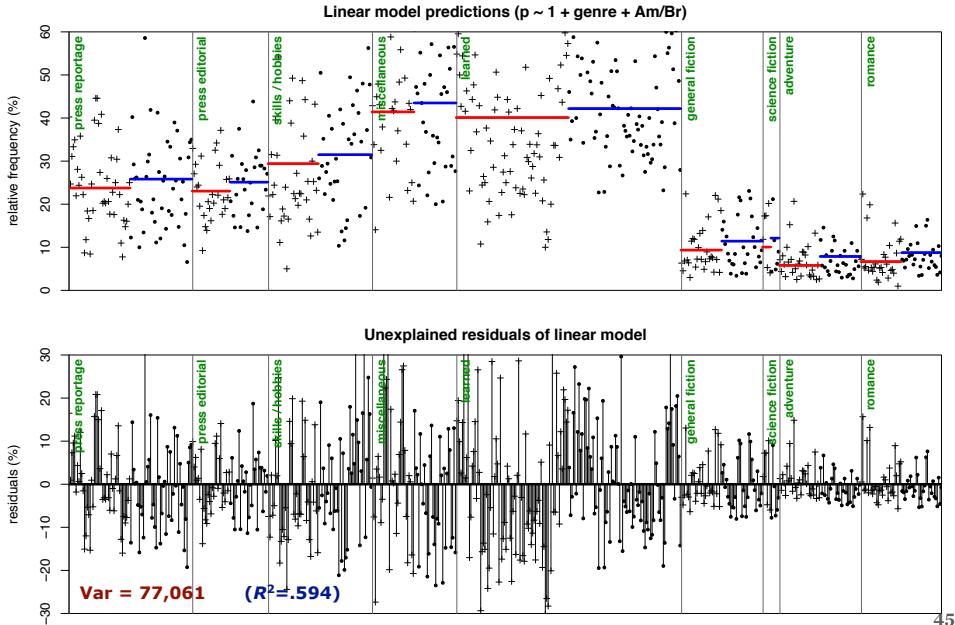
42

Linear model for passives



44

Linear model for passives



Linear models in R

```
# linear model "formula": response ~ explanatory factors
> LM <- lm(relfreq ~ genre + lang, data=Passives)

# analysis of variance shows which factors are significant
> anova(LM)      # see ?anova.lm for details

# individual coefficients + standard errors
> summary(LM)    # difficult to interpret for genre (why?)
> confint(LM)     # corresponding confidence intervals
```

Linear model for passives

- ◆ Goodness-of-fit (analysis of variance, F test)
 - 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%)

46

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 = ...

47

48

Linear models in R

```
# more intuitive than coefficients: model predictions for each
# genre and language variety; based on "dummy" data frame with
# all possible genre/language combinations (ordered by genre)
> Predictions <- unique(Passives[, c("genre", "lang")])
> Predictions <- Predictions[
  order(Predictions$genre, Predictions$lang), ]
# predicted average relative frequency of passives in each category
> transform(Predictions,
  predicted=predict(LM, newdata=Predictions))
# confidence and prediction intervals (see Unit #6 for details)
> cbind(Predictions, predict(LM,
  newdata=Predictions, interval="confidence"))
> cbind(Predictions, predict(LM,
  newdata=Predictions, interval="prediction"))
```

49

Linear models in R

	genre	lang	fit	lwr	upr
1	press	reportage	AmE 23.76	21.25	26.27
311	press	reportage	BrE 25.83	23.32	28.34
45	press	editorial	AmE 23.05	19.92	26.17
355	press	editorial	BrE 25.12	22.00	28.25
72	skills / hobbies		AmE 29.40	26.69	32.12
382	skills / hobbies		BrE 31.48	28.78	34.18
108	miscellaneous		AmE 41.42	38.44	44.40
420	miscellaneous		BrE 43.50	40.52	46.48
138		learned	AmE 40.11	38.15	42.06
450		learned	BrE 42.18	40.23	44.14
218	general	fiction	AmE 9.35	6.33	12.38
530	general	fiction	BrE 11.43	8.41	14.46
247	science	fiction	AmE 10.06	3.63	16.48
559	science	fiction	BrE 12.13	5.71	18.56
253		adventure	AmE 5.80	2.77	8.82
565		adventure	BrE 7.87	4.85	10.90
282		romance	AmE 6.67	3.65	9.70
594		romance	BrE 8.75	5.73	11.78

50

Linear models in R

```
# alternative: compute partial effects with effects package
> library(effects)
> eff <- Effect(c("genre", "lang"), LM)
> print(eff) # predicted values for each category

# visualization of partial effects
> plot(eff, multiline=TRUE, rotx=30)

# some display options
> topR <- list(x=0.95, y=0.95, corner=c(1,1))
> plot(eff, multiline=TRUE, rotx=30,
  ci.style="bars", key.args=topR)

# partial main effects average over other predictors
> plot(Effect("genre",LM), rotx=30, ylim=c(0,50))
```

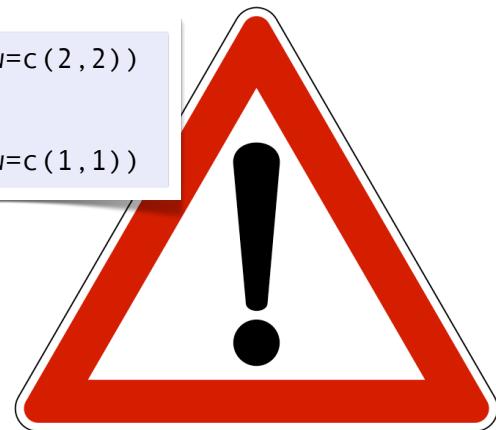
51

Interactions

- ◆ Our linear model assumes that the difference AmE/BrE is the same across all genres
 - both predictors are independent **main effects**
- ◆ Include **interaction term** in model to account for genre-specific differences
 - > LM <- lm(relfreq ~ genre + lang + genre:lang, data=Passives)
 - > anova(LM)
- ◆ Interaction is **not significant** here ($p = .055$)
 - data give no evidence for genre-specific differences

52

```
> par(mfrow=c(2, 2))
> plot(LM)
> par(mfrow=c(1, 1))
```



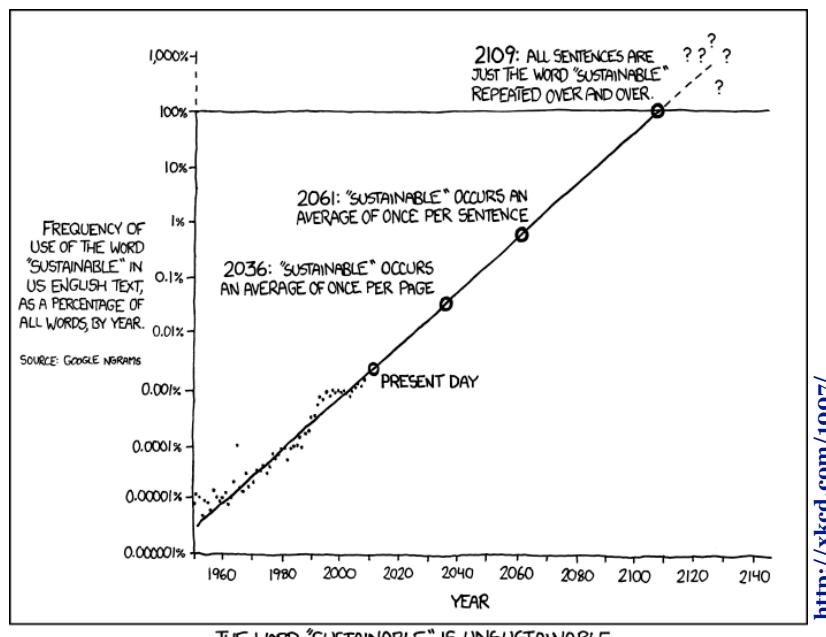
Linear models are not appropriate!

53

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%

54



55

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)

$$f_i \sim B(n_i, \pi_i)$$

binomial sampling ("family")

$$\pi_i = \frac{1}{1 + e^{-\theta_i}}$$

"link" function

"response"

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

linear predictor

56

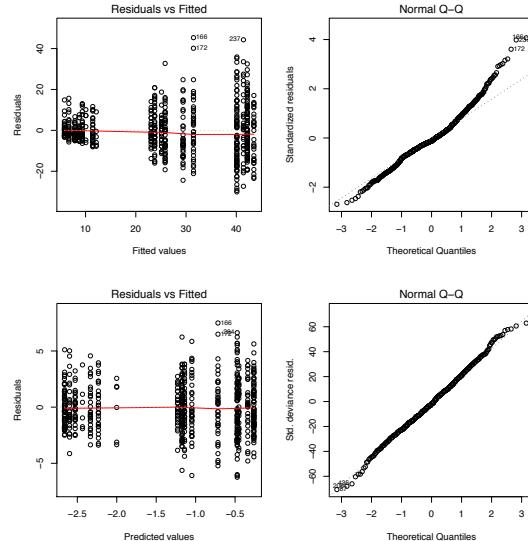
GLM for passives

- ◆ 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

57

Model diagnostics comparison

Linear Model



Generalised Linear Model

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

GLM in R

(note the extra options needed!)

```
# for GLM with binomial family, responses are pairs of
# passive / active counts ( $f_k, n_k - f_k$ ) = “successes” / “failures”
> response.matrix <- cbind(Passives$passive,
   Passives$n_s - Passives$passive)

# genre * lang is shorthand for main effects + all interactions
> GLM <- glm(response.matrix ~ genre * lang,
   family="binomial", data=Passives)

# individual coefficients + standard errors
> anova(GLM, test="Chisq") # interaction significant now
> summary(GLM) # even more difficult to interpret than for LM
> confint(GLM)

# diagnostics plot ( ; separate multiple commands in single line)
> par(mfrow=c(2,2)); plot(GLM); par(mfrow=c(1,1))
```

58

GLM in R

(note the extra options needed!)

```
# predictions for each genre and language variety
> transform(Predictions, predicted =
  100 * predict(GLM, type="response",
  newdata=Predictions))

# calculate confidence intervals from standard errors
> res <- predict(GLM, type="response",
  newdata=Predictions, se.fit=TRUE)
> transform(Predictions,
  predicted = 100*res$fit,
  lwr=100*(res$fit - 1.96*res$se.fit),
  upr=100*(res$fit + 1.96*res$se.fit))

# we can't compute prediction intervals for new texts — why?
```

59

60

Partial effects for GLM

```
# the effects package also supports GLM
> eff <- Effect(c("genre", "lang"), GLM)
> print(eff) # predicted values for each category

# visualization of partial effects (note the interaction)
> plot(eff, multiline=TRUE, rotx=30,
       ci.style="bars", key.args=topR)

# partial main effects average over interactions in this case
> plot(Effect("genre", GLM), rotx=30,
       ylim=c(-3,0)) # ylim refers to linear predictor
> plot(Effect("lang", GLM), rotx=30,
       ylim=c(-3,0))
```



61

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!)

62

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 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.

63

64

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.

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.