Rethinking Corpus Frequencies

Stefan Evert
stefan.evert@uos.de | purl.org/stefan.evert
Institute of Cognitive Science
University of Osnabrück

How often is *kick the bucket* really used?

What are the characteristics of “translationese”?

Do Americans use more split infinitives than Britons? What about British teenagers?

What are the typical collocates of *cat*?

Can the next word in a sentence be predicted?

Do native speakers prefer constructions that are grammatical according to some linguistic theory?

Evidence from frequency comparisons / estimates

A toy research question

*How many passives are there in English?*

☞ Slightly more interesting version:

*Is the passive used more often in British English than in American English?*

Standard methodology
(e.g. McEnery & Wilson 2001, Sec. 3.4)

- Obtain representative corpora of AmE and BrE
- Count passives in both corpora
  - AmE: 54 passives | BrE: 50 passives
  - contrary to implicit expectation of research question
- Different sample sizes ☞ use relative frequency
  - unit of measurement? — here: sentences (simple)
  - AmE: 54 passives out of 300 sentences = 18%
  - BrE: 50 passives out of 200 sentences = 25%
Standard methodology
(e.g. McEnery & Wilson 2001, Sec. 3.4)

- Must test significance to make sure observed differences are not just a coincidence
  - main cause: “random” choice of corpus material
- Standard tests for frequency comparison based on contingency table = cross-classification
  - null hypothesis $H_0$: same proportion of passives

<table>
<thead>
<tr>
<th></th>
<th>AmE</th>
<th>BrE</th>
</tr>
</thead>
<tbody>
<tr>
<td>passive</td>
<td>$k_1$</td>
<td>$k_2$</td>
</tr>
<tr>
<td>active</td>
<td>$n_1-k_1$</td>
<td>$n_2-k_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AmE</th>
<th>BrE</th>
</tr>
</thead>
<tbody>
<tr>
<td>passive</td>
<td>54</td>
<td>50</td>
</tr>
<tr>
<td>active</td>
<td>246</td>
<td>150</td>
</tr>
</tbody>
</table>

Standard methodology
(e.g. McEnery & Wilson 2001, Sec. 3.4)

- Traditional: Pearson’s chi-squared test
  - test statistic: $X^2 = 3.1571$, df = 1
  - p-value has meaningful interpretation: $p = 0.0756$
- Recommended: Fisher’s exact test
  - better accuracy & robustness than chi-squared test
  - similar p-value here: $p = 0.0717$
- Observed difference is not significant
  - customary “significance level”: $p < .05$
  - more conservative levels are $p < .01$ and $p < .001$

What does “not significant” mean?
- either: there is no difference between AmE and BrE
- or: selected corpora did not provide enough evidence

Significance vs. effect size
- if the “true” difference between AmE and BrE is small, it is more difficult to detect with sufficient confidence

Possible solution: use larger corpora
- larger sample size increases “power” of tests
- even small effects become significant

Criticism

- Recent criticism of significance testing:

Fundamental problem:
all significance tests assume random samples
- even a “representative” or “balanced” corpus is not a perfectly random sample of language
Goals for today

- The logic of statistical hypothesis testing
  - illustrated by the library metaphor
- Why a corpus isn’t a random sample
- Measuring non-randomness
- Consequences for significance tests
- Rethinking corpus frequencies

Back to our toy problem

How many passives are there in English?

- American English style guide claims that “In an average English text, no more than 15% of the sentences are in passive voice. So use the passive sparingly, prefer sentences in active voice.”
- This Web page actually states that only 10% of English sentences are in passive voice (as of January 2009)!
- We have doubts and want to verify this claim
The library metaphor
first seen at ICAME 25 (2004)

♦ Extensional definition of a language:
“All utterances made by speakers of the language under appropriate conditions, plus all utterances they could have made”

♦ Imagine a huge library that contains anything that has ever been written or said in the language → library metaphor (Evert 2006)

• necessarily a hypothetical library. The ICAME 30 Conference was awful and boring, is part of the English language, even though no-one has an occasion to say it!

The library metaphor

♦ Random sampling in the library metaphor (for a random sample of sentences = tokens)
  • walk to a random shelf ...
  ... select a random book ...
  ... open it on a random page ...
  ... and pick a random sentence from the page
  → 1 item for our sample
  • Repeat \( n \) times for sample size \( n \)

From research question to statistical analysis

random sample \( \xrightarrow{\text{statistical inference}} \) population (library)

linguistic hypothesis \( \xleftarrow{\text{problem operationalisation}} \) hypothesis

Statistics

Inference from random sample

♦ Take a sample of, say, 100 sentences
  • we observe \( k = 19 \) passives
  • if style guide is right, we expect \( k_0 = 15 \) passives
  • \( k > k_0 \) → reject claim of style guide?

♦ Principle of statistical inference
  • if sample is picked at random, sample frequency \( k \) should roughly be equal to expected frequency \( k_0 \)

♦ Take another sample, just to be sure
  • observe \( k = 13 \) passives
  • \( k < k_0 \) → claim of style guide confirmed?
Sampling variation

- Random sampling ensures that, on average, relative frequency equals population proportion.
- But for every individual sample we will get a different value because of chance effects → sampling variation.
- The main purpose of statistical methods is to estimate & correct for sampling variation.
  - that’s all there is to statistics, really 😊

Quantifying sampling variation

- Many corpus linguists set out to test $H_0$.
  - each one draws a random sample of size $n = 100$.
  - assume $H_0$ to be true → rejection is a mistake!
  - how many of the samples have the expected number $k = 15$ of passives, how many have $k = 19$, etc.?
- We don’t need an infinite number of monkeys (or corpus linguists) to answer these questions.
  - randomly picking sentences from our metaphorical library is like drawing balls from an infinite urn.
- Described by binomial distribution

The binomial distribution

$$Pr(k) = \left( {n \atop k} \right) (\pi_0)^k (1 - \pi_0)^{n-k}$$

percentage of samples = probability

Binomial sampling distribution

- risk of false rejection = p-value = 26.2%
Why a corpus isn’t a random sample

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
  - Brown: 500 text fragments @ 2,000 words
    - total size = 1 million running words
  - BNC: 4,048 documents ranging from 25 to 428,300 words
    - total size = 100 million running words

The key problem

- unit of sampling (text or text fragment)
  ≠ unit of measurement (e.g. sentence)
  - recall sampling procedure in the library metaphor ...
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

- 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

- 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?

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
A sample of random samples is a random sample

- 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

Measuring non-randomness

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)
- 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

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 \( \Rightarrow \) best possible fit
- Non-randomness \( \Rightarrow \) larger sampling variation

Passives in the Brown corpus

Passives in the LOB corpus
Consequences of non-randomness

- 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**

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)
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 \( \rightarrow \) large amount of evidence

- **t-test: not significant**
  - p-value: \( .1340 > .05 \) \( (t=1.50, \text{df}=619.96) \)
  - confidence interval for difference: \( -0.6\% - +4.9\% \)
  - \( H_0 \): same average relative frequency in AmE and BrE

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: \( \pi = 6.8\% \)

Average relative frequency?
Average relative frequency?

Problems of the statistical analysis of corpus data

Studying variation in language

- It seems absurd now to measure & compare relative frequencies in “language” (= library)
  - proportion \( \pi \) 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 \( \rightarrow \) one factor behind this variation

Rethinking corpus frequencies
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

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:

\[ p_i = \beta_0 + \beta_1 \text{genre} + \beta_2 \text{AmE/BrE} + \epsilon_i \]
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%)

- 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 = ...
Don’t try this at home, kids!

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 ($\pi_i$)

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

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

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

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%

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
Model diagnostics

Linear Model

Generalised Linear Model

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

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!)

References (1)

References (2)


References (3)