Corpus Linguistics
(L415/L615)
Statistics for Corpus Linguistics

Markus Dickinson
Department of Linguistics, Indiana University
Fall 2015

Statistics for Corpus Linguistics
We will more or less follow the presentation in Gries (2009) ... with some pointers from Stephanie Dickinson

Point of emphasis: learning to quantitatively think about one's data

General breakdown at looking at distributional data:
- Frequencies of occurrence of linguistic elements
  - Frequency lists
  - Dispersion statistics
- Frequencies of co-occurrence (cf. collocations)

We’ll also break things down into:
- Descriptive statistics
- Inferential statistics: evaluate data from significance perspective

Observed frequencies

**Observed absolute frequencies**: basic counts
- e.g., in spoken part of ICE-GB, *give* occurs 297 times, *bring* occurs 128 times
- e.g., in written part of ICE-GB, *give* occurs 144 times, *bring* occurs 69 times

Sometimes logarithms are taken, to create a linear distribution

**Observed relative frequencies**: adjust for size of corpora, e.g., frequencies per 1,000,000 words:
- Spoken: 637,682 words, Written: 423,581 words
  - *give*: 465.75 words/million (spoken), 339.96 words/million (written)
  - *bring*: 200.73 words/million (spoken), 162.90 words/million (written)

Proportions & Relative frequency ratio

**Proportions** (=relative frequencies) are often used for hypothesis testing

**Relative frequency ratio**: the quotient of the relative frequencies of a word in two corpora, e.g.:
- *source* appears 14 times in Perl corpus of 6,065 words
- *source* appears 1 times in Python corpus of 5,596 words

Proportions: \( \frac{14}{6065} \approx 0.00231 \) and \( \frac{1}{5596} = 0.000179 \)

Relative frequency ratio: \( \frac{14}{6065} - \frac{1}{5596} \approx 12.92 \)

Entropy

One can also consider **entropy**: average uncertainty of a random variable:

\[ H = - \sum p(x) \cdot \log_2 p(x) \]

- what is the entropy of the different forms of *give* here?
- and how does the entropy of *sing* compare?

\( H(give) = 2.1, H(sing) = 1.4 \)

- In other words: the form of the lemma *sing* is more predictable (less uncertain)

Zero frequencies

Problem: zero frequencies for things which may be possible
- Smoothing/discounting techniques can adjust for this, e.g., Good-Turing smoothing
Dispersion

Consider the following 3-part “corpus”:

\[ q w e e r | q r r t t t | q y y y y y \]

Overall relative frequency of \( y \): 28.57% (6/21)
- Range from 0% to 85.71%
- \( q \): relative frequency is 14.29% across all subcorpora

One can measure degree of dispersion \( (DP_{\text{norm}}) \)
- Reporting standard deviations can also help

Interval/Ordinal-scaled data

Q: How to compare the relationship between two variables?
- Examine a scatter plot
- Calculate a correlation coefficient
  - E.g., Pearson’s \( r \), Kendall’s tau
  - Generally: 0: no correlation, 1: strong positive correlation, \(-1\): strong negative correlation

For more than two variables, linear modeling could be helpful
- Simplest linear models require interval-scaled variables
- Typically: build comprehensive model & remove non-significant predictors in stepwise fashion

Co-occurrence

We have already talked about collocations, so we’ll just mention a few pointers for consideration:
- Use type frequencies in addition to token frequencies?
- Use a window-based approach?
- Use collocations of more than two words?
- Use discontinuous n-grams?

Cross-Tabulations

Tests such as Pearson’s chi-squared can show how observed frequencies (and percents) compare between groups/conditions

Cross-tabulation of two variables (corpus & verb form):

<table>
<thead>
<tr>
<th></th>
<th>LOB</th>
<th>FLOB</th>
<th>BROWN</th>
<th>FROWN</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>pres perf</td>
<td>4,196</td>
<td>4,073</td>
<td>3,358</td>
<td>3,499</td>
<td>15,306</td>
</tr>
<tr>
<td>simp past</td>
<td>35,821</td>
<td>35,276</td>
<td>37,223</td>
<td>36,250</td>
<td>144,470</td>
</tr>
<tr>
<td></td>
<td>40,017</td>
<td>39,349</td>
<td>40,761</td>
<td>39,749</td>
<td>159,876</td>
</tr>
</tbody>
</table>

Relationship between tenses & corpus parts
\( (\chi^2 = 130.8; df = 3; p < 0.001) \), but ...

Thinking about data

Is the pattern really about variety/dialect?
- British (LOB, FLOB) corpora feature more present perfects than American (Brown, Frown) corpora

Instead of tense \( \times \) corpus:
- We might want: tense \( \times \) variety \( \times \) time
  - LOB, Brown < FLOB, Frown

Simple way: slice up the table on previous slide
- Alternative: generalized models (see paper)

Generalized linear models

Generalized linear model predicting the probability of a binary variable
- In what conditions (variety, time) are you relatively more likely to have pres. perf.?
General tips

Gries lists a few hints (in section 3.3) that are worth remembering:

- Plot your data
  - e.g., A linear correlation may actually be curvilinear
- Look at effect size
  - It’s not all about significance (which tends to happen with large corpora)
- Look at pairwise comparisons
  - Significance doesn’t mean that all pairs significantly differ from each other

Exploratory statistics

In addition to testing hypotheses, one can use statistical techniques to generate hypotheses

- e.g., hierarchical agglomerative cluster analysis
  - This produces tree diagrams which are relatively easily interpretable
  - n objects clustered into m < n groups: large within-group similarity, small between-group similarity
  - clustered on the basis of x characteristics (features)
- Example: clustering Russian verbs meaning ‘to try’

Number of clusters

Clustering varies by:

- Technique (e.g., more even-size or elongated clusters?)
- Similarity measurement (e.g., distance? curvature?)
- Number of clusters (often user-specified)