To clear up:
- lateness to class?
- late on assignments?
- not showing missed class? → tell me what’s going on.
- missed assignment? → mini-cloze test

To day: question genre: "what makes two types of lang. different?"
- altruistic vs. not, funny v. not, in event v. after event
dem. vs. reg.
- many approaches; today, an example statistical approach from a Bayesian perspective (as opposed to a classification perspective)

Desiderata:
- refinement: we want significant differences:
- where diffs (probably) not due to chance?

Zipf’s law (gloss): rare words make up surprisingly large fraction of data
- pub of x pub of (rank of x)\(^{-1}\). So tail is heavier than an exp. decay
- main idea of Monroe, Lee, Qu et al

 contrasts:
- logistic regression coefficients (e.g., Austin, "Gilbert")
- statistical significance flipping for bivariate detection (Klimberg, "oz")

Ref: FAB, how do I integrate into rel. res in logistic regression
(c) in real research: use hold-out set for exploring shifts, if want to be proper
and plan to do classification experiments
(True, it's said that you lose even more data this way)

Setting: two samples, one from one language and the other, from the other, e.g. Dem. Rep. on topic of abortion
like to be able to see, that, say, a word like 'pro-life' has diff. prob. of occ. w.r.t. the two generation processes

Assume words are drawn i.i.d. from a multinomial w. params
\[ \Theta = (\Theta_D, \ldots, \Theta_{ui}) \]
\[ \Theta_R = (\Theta', \ldots, \Theta_{ui}) \]

consider the count histograms: the observed evidence:
\[ (c_1^D, c_2^D, \ldots, c_i^D) \text{ vs. } (c_1^R, c_2^R, \ldots, c_i^R) \]

and establish notation

Following MCQ: 1st get intuition by looking at simple statistics

[3.2.1: diff. in frequencies: ] rank vi by \( c_i^D - c_i^R \) give away
- what if one side speaks more? \\
- be carefully about cherry-picking (p. 3+6)
  ranking makes it more clear: 'the', 'of', 'so'

3.2.2 diff. in proportions: normalize
\[ \text{rf}_i^D = \frac{c_i^D}{\sum c_i^D} \]
\[ \text{rf}_i^R = \frac{c_i^R}{\sum c_i^R} \]

Great visualization idea by MCQ!

ranking of top 20 D, top 20 P

> most freq. words getting most weight

most weight words not accounted for:

low freq. diffs: \( \frac{2}{N_0} - \frac{1}{N} \)
high freq. diffs: \( \frac{2000 - 900}{N} \)

\( \frac{10}{N} >> 1 \)

< note: restricted range of intuition (for review)
We know we're gonna want to get to a log-odds ratio:
<br>(Preceded: research drama — this won't go well, but it is leading to stug good)

Let's consider:

\[
\hat{rf}_i^R = \frac{rf_i^R}{1 - rf_i^R}
\]

Scale is \((0, 1)\) — more extreme gives bigger nice number.

Consider: 4:1. 

Log-odds-ratio:

\[
\log \left( \frac{rf_i^R/(1-rf_i^R)}{rf_i^O/(1-rf_i^O)} \right)
\]

- If equal, settle for 0.
- Informative sign.
- Problem of for unseen words.

See Fig. 2 — highest words are all low-freq, b/c ratios are more extreme for low counts (same 'reading' as before).

Let's go back to model-based approach:

--- Use notation from previous page that was deferred.

--- Should we keep trying ad hoc fixes? Or try a more 'principled' approach?

--- Let's try to explicitly take variance into account: what is the variance of the log-odds ratio?

--- What is its distribution?

Model: multinomial question (see notation that was deferred from previous page)

- Natural for independent rolls of dice or draws of words
- Altho' not true (Norvig's, Ken Church)

What do we know about \(\Theta_0^O, \Theta_R^R\)?

Express prior: Dirichlet is very convenient: conjugate prior for multinomial.

- Not that we know any other priors on multinomials.

Parameters \(\alpha_1, \alpha_2, \ldots, \alpha_n\); call \(\alpha_i = \Sigma \alpha_i\); require \(\alpha_i > 0\)

\[\theta_i \sim \text{Beta}(\alpha_i, 1)\]

- Mode: \(\theta_i \sim \frac{\alpha_i}{\sum \alpha_i}\) — so, all multinomials lie above pncts.

\[\text{Var} \approx \frac{\alpha_i}{\sum \alpha_i^2}\]

\[\text{Mean} \approx \frac{\alpha_i}{\sum \alpha_i}\]

- Probably better to have just talked about the mean.

Described as the "normalization trick".

- "Concentration" relates to spiked expected values.

Posterior \(P(\theta | D) \propto P(D | \theta)P(\theta) \approx \text{Dir}(\alpha_i + \xi_i, \ldots, \alpha_n + \xi_n)\)

- Very definitely made an imposition.

Showed Fig 4 — uninformative prior.

Fig 5 — informative prior.

Matthews et al.