CS/IS 6742: NLP and Social Interaction, Fall 2018
Oct. 30, 2018: Reference sheet for "What makes two sub-languages different?"

---

1 A key fact "we" know beforehand

There is a simple count-based approximation of the variance of the log-odds ratio between two posterior multinomial distributions based on Dirichlet priors, and the log-odds ratio is also known to be distributed approximately normal.

2 Notation

The color-based notation below was picked because it was easier at the time to change colors in powerpoint than to add subscripts or superscripts.

Suppose we have two language samples, S and S, drawn from the same vocabulary $V = v_1, v_2, \ldots v_{|V|}$.
We use i to index into the vocabulary.
We write $S_\bullet$ and $S_\bullet$ for the number of *tokens* in each of the two samples.[1]

Example: if S = "great great great", $S_\bullet$ = 3; there are three different tokens.[2]

We define

$$p(v_i) := \frac{\text{count}(v_i)}{S_\bullet} \tag{1}$$

and similarly for $p(v_i)$.

3 Log-odds

For a given i, the log-odds according to $p(v_i)$ is

$$\text{odds}_i := \frac{p(v_i)}{1 - p(v_i)} \tag{2}$$

and similarly for $p(v_i)$. And the *log-odds ratio* for v_i is $\log(\text{odds}_i/\text{odds}_i)$. What can this quantity range over?

4 Multinomial

A multinomial distribution for our choice of vocabulary has two (types of) parameters:

- $\overrightarrow{\phi} \in \Re^{|V|}$, where $\sum_i \phi_i = 1$ and for all i, $\phi_i \geq 0$. These are the probabilities on the sides of the "die" whose sides are labeled with the vocabulary items v_i.
- n, the number of draws (the sample size)

We'd like to find v_is where ϕ_i is really different from ϕ_i.

5 Re-estimated distribution

Suppose we have a Dirichlet prior on $\overrightarrow{\phi}$ parametrized by $\overrightarrow{\alpha} \in \Re^{|V|}$, where for all i, $\alpha_i \geq 0$; similarly for $\overrightarrow{\alpha}$. One can consider these vectors to represent *pseudocounts*.

Given a prior parametrized by $\overrightarrow{\alpha}$ and a sample S, we can have a re-estimated distribution over words, which we denote $\hat{p}(v_i)$, and similarly $\hat{p}(v_i)$. This gives us a new log-odds ratio, whose distribution under the hypothesis that $\phi_i = \phi_i$ is known according to §1. So we can test the corresponding z-score for significance.

---

[1] The "dot" notation is borrowed from statistics, to make one think of summing over all the values of the index variable replaced with the "dot".

[2] The number of *types* in S, on the other hand, is 1.