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 $S_*$ and $S'_*$ for the number of tokens in each of the two samples.\footnote{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”.}

Example: if $S =$ “great great great”, $S_* = 3$; there are three different tokens.\footnote{The number of types in $S$, on the other hand, is 1.}

We define

$$p(v_i) := \frac{\text{count}(v_i)}{S_*}$$

and similarly for $p(v_i)$.

3 Log-odds

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

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

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:

- $\phi \in \mathbb{R}^{|V|}$, 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 $v_i$s where $\phi_i$ is really different from $\phi_{i'}$.

5 Re-estimated distribution

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

Given a prior parametrized by $\alpha$ 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 the Dirichlet distribution. So we can test the corresponding $z$-score for significance.