

## 1 Key facts about log-odds ratios

Monroe, Colaresi, and Quinn seem to have known ahead of time that:

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

## 2 (Extremely non-nuanced) expectations on political discussion of abortion in the 1999-2000 U.S. Senate

- Democrats (“left”, “blue”): language such as “pro-choice”, “women’s rights”
- Republicans (“right”, “red”): language such as “pro-life”, “murder”, “child”

## 3 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  $V = v_1, v_2, \dots, v_{|V|}$ .

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.<sup>1</sup>

Example: if  $S = \text{“great great great”}$ ,  $S_\bullet = 3$ ; there are three different tokens.<sup>2</sup>

We define

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

and similarly for  $p(v_i)$ .

## 4 Log-odds

Here is where we do the finesse of moving straight to log-odds ratio rather than just the expected ratio of  $p(v_i)$  and  $p(v_i)$ .

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

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

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?

<sup>1</sup>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”.

<sup>2</sup>The number of *types* in  $S$ , on the other hand, is 1.

## 5 Multinomial

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

- $\vec{\phi} \in \mathbb{R}^{|V|}$ , 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  $v_i$ .
- $L$ , the number of draws (the sample length)

We’d like to find  $v_i$ s where  $\phi_i$  is really different from  $\hat{\phi}_i$ .

## 6 Re-estimated distribution

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

Given a prior parametrized by  $\vec{\alpha}$  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  $\phi_i = \hat{\phi}_i$  is known according to §1. So we can test the corresponding  $z$ -score for significance.