Federalist Paper Authorship

Let's write a program to decide whether or not James Madison or Alexander Hamilton wrote Federalist Paper 49. Both men have claimed to have written it, and hence the authorship is in dispute. First we used historical essays to estimate p_i, the probability that Hamilton generates the word i (independent of all previous and future choices of words). Similarly we estimated q_i, the probability that Madison generates the word i. For each word i we observe the number of times that word occurs in Federalist Paper 49 (we call that count c_i). We assume that, given no evidence, the paper is equally likely to be written by Madison or Hamilton.

Define three events: H is the event that Hamilton wrote the paper, M is the event that Madison wrote the paper, and D is the event that a paper has the collection of words observed in Federalist Paper 49. We would like to know whether $P(H|D)$ is larger than $P(M|D)$. This is equivalent to trying to decide if $P(H|D)/P(M|D)$ is larger than 1.

The event $D|H$ is a multinomial parameterized by the values p. The event $D|M$ is also a multinomial, this time parameterized by the values q.

Using Bayes Rule we can simplify the desired probability.

$$\frac{P(H|D)}{P(M|D)} = \frac{\frac{P(D|H)P(H)}{P(D)}}{\frac{P(D|M)P(M)}{P(D)}} = \frac{P(D|H)P(H)}{P(D|M)P(M)} = \frac{P(D|H)}{P(D|M)}$$
$$= \frac{\binom{n}{c_1,c_2,\ldots,c_m}\prod_i p_i^{c_i}}{\binom{n}{c_1,c_2,\ldots,c_m}\prod_i q_i^{c_i}} = \frac{\prod_i p_i^{c_i}}{\prod_i q_i^{c_i}}$$

This seems great! We have our desired probability statement expressed in terms of a product of values we have already estimated. However, when we plug this into a computer, both the numerator and denominator come out to be zero. The product of many numbers close to zero is too hard for a computer to represent. To fix this problem, we use a standard trick in computational probability: we apply a log to both sides and apply some basic rules of logs.

$$\begin{aligned}\log\Big(\frac{P(H|D)}{P(M|D)}\Big) &= \log\Big(\frac{\prod_i p_i^{c_i}}{\prod_i q_i^{c_i}}\Big) \\ &= \log(\prod_i p_i^{c_i}) - \log(\prod_i q_i^{c_i}) \\ &= \sum_i \log(p_i^{c_i}) - \sum_i \log(q_i^{c_i}) \\ &= \sum_i c_i \log(p_i) - \sum_i c_i \log(q_i)\end{aligned}$$

This expression is "numerically stable" and my computer returned that the answer was a negative number. We can use exponentiation to solve for $P(H|D)/P(M|D)$. Since the exponent of a negative number is a number smaller than 1, this implies that $P(H|D)/P(M|D)$ is smaller than 1. As a result, we conclude that Madison was more likely to have written Federalist Paper 49. That is the standing assumption currently made by historians!