JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

This study has four purposes: to provide a comparison of discrimination methods; to explore the problems presented by techniques based strongly on Bayes' theorem when they are used in a data analysis of large scale; to solve the authorship question of *The Federalist* papers; and to propose routine methods for solving other authorship problems.

Word counts are the variables used for discrimination. Since the topic written about heavily influences the rate with which a word is used, care in selection of words is necessary. The filler words of the language such as *an*, *of*, and *upon*, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates, whereas more meaningful words like *war*, *executive*, and *legislature* do not.

After an investigation of the distribution of these counts, the authors execute an analysis employing the usual discriminant function and an analysis based on Bayesian methods. The conclusions about the authorship problem are that Madison rather than Hamilton wrote all 12 of the disputed papers.

The findings about methods are presented in the closing section on conclusions.

This report, summarizing and abbreviating a forthcoming monograph [8], gives some of the results but very little of their empirical and theoretical foundation. It treats two of the four main studies presented in the monograph, and none of the side studies.

[1] This work has been faciliated by grants from The Ford Foundation, the Rockefeller Foundation, and from the National Science Foundation NSF G-13040 and G-10368, contracts with the Office of Naval Research Nonr 1866(37) and 2121(09), and the Laboratory of Social Relations, Harvard University. The work was done in part at the Massachusetts Institute of Technology Computation Center, Cambridge, Massachusetts, and at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. Permission is granted for reproduction in whole or in part for purposes of the United States Government.

[2] Presented at a session of Special Papers Invited by the Presidents of The American Statistical Association, The Biometric Society (ENAR), and The Institute of Mathematical Statistics at the statistical meetings in Minneapolis, Minnesota, September 9, 1962.

1. THE FEDERALIST PAPERS AS A CASE STUDY IN DISCRIMINATION

THE *Federalist* papers were published anonymously in 1787–1788 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of the State of New York to ratify the Constitution. Of the 77 essays, 900 to 3500 words in length, that appeared in newspapers, it is generally agreed that Jay wrote five: Nos. 2, 3, 4, 5, and 64, leaving no further problem about Jay's share. Hamilton is identified as the author of 43 papers, Madison of 14. The authorship of 12 papers (Nos. 49–58, 62, and 63) is in dispute between Hamilton and Madison; finally, there are also three joint papers, Nos. 18, 19, and 20, where the issue is the extent of each man's contribution.

The controversy over the authorship of the disputed papers is thick with lists. Among these, the *Port Folio* list, the *Benson* list, and the *Kent* list purport to give Hamilton's allocation of the authorship of *The Federalist* papers among Jay, Madison, and himself. Other lists are attributed to Madison; among these his authorship allocation in the *Gideon* edition is the most important, because Madison, himself, gave the publishers the data and, after the list was published, confirmed its accuracy. Our information comes largely from secondary sources, and more complete historical data are contained in Adair's fascinating account [1, 2] of the controversy, as well as in Cooke's [4]. We cannot ourselves do justice in limited space to both the controversy and our contribution toward its resolution. No doubt the only reason the dispute exists is that Hamilton and Madison did not hurry to enter their claims. By the time those lists attributed to Hamilton were published, he was dead, and thereafter Madison waited a decade to make his own claim. Probably Hamilton and Madison were slow to claim their share in *The Federalist* because the arguments each gave there sometimes opposed political positions each adopted later.

Between 1885 and 1900 Henry Cabot Lodge and Paul Leicester Ford attacked Madison's claim, whilst E. G. Bourne supported it. Adair [1, 2] concluded that Madison wrote all of the disputed papers, but has continued to gather new evidence.

Recently three unabridged editions of *The Federalist* [4, 5, 12] have attributed Nos. 49–58 to Madison—two wholeheartedly, one with mild reservations. And Nos. 62 and 63 are assessed with more doubt. The bases of the attributions are mainly Adair's historical research. A major point is the mature consideration Madison gave his claim. He had no duel to jog his elbow, and his list is given paper by paper in a published book, rather than in clumps. And he had an opportunity to check the accuracy again after the list appeared. In a sense, however, the historical data we now have are quite similar to those that were available to Lodge and Ford who took a tack opposite to the presently favored direction. Perhaps the historian will welcome new evidence on the matter.

The writings of Hamilton and Madison are difficult to tell apart because both authors were masters of the popular *Spectator* style of writing—complicated and oratorical. To illustrate, in 1941 Frederick Williams and Frederick Mosteller counted sentence lengths for the undisputed papers and got means of 34.55 and 34.59 words respectively for Hamilton and Madison, and average standard deviations for papers of 19.2 and 20.3. These results show that for some measures the authors are practically twins.

As statisticians, our interest in this controversy is largely methodological. Problems of discrimination are widespread, and we wished a case study that would give us an opportunity to compare the more usual methods of discrimination with an approach via Bayes' theorem. Though much has been written recently about statistical methods using Bayes' theorem, we are not aware of any large-scale analyses of data that have been publicly presented. We think that this lack of experience is unfortunate because the difficulties one finds in the armchair often do not overlap with those that confound one in the field or laboratory. And so we use this problem in authorship as a case study.

(The classical approach to discrimination problems, flowing from Fisher's "linear discriminant function" and from the Neyman-Pearson fundamental lemma, is well exposited by Rao [10, Chapter 8]. Important recent writings on Bayesian statistics include the books by Jeffries [6], by Raiffa and Schlaifer [9], and by Savage and others [11].)

Adair in correspondence with one of the authors about early counts on *The Federalist* explained that he, Adair, had found that the words *while* and *whilst* discriminated Hamilton from Madison quite well. Adair encouraged us to pursue the matter further, and we did.

TABLE 2.1. FREQUENCY DISTRIBUTION OF RATE PER THOUSAND WORDS FOR THE 48 HAMILTON AND 50 MADISON PAPERS FOR *by*, *from*, AND *to*. THE UPPER LIMIT OF A CLASS INTERVAL IS NOT INCLUDED IN THE CLASS

Rate	*by*		Rate	*from*		Rate	*to*	
	H	M		H	M		H	M
1– 3	2		1– 3	3	3	20–25		3
3– 5	7		3– 5	15	19	25–30	2	5
5– 7	12	5	5– 7	21	17	30–35	6	19
7– 9	18	7	7– 9	9	6	35–40	14	12
9–11	4	8	9–11		1	40–45	15	9
11–13	5	16	11–13		3	45–50	8	2
13–15		6	13–15		1	50–55	2	
15–17		5		—	—	55–60	1	
17–19		3	Totals	48	50		—	—
	—	—				Totals	48	50
Totals	48	50						

2. WORDS AS DISCRIMINATORS

Words offer a great many opportunities for discrimination; there are so many of them. Some vary considerably in their rates of use from one paper to another by the same author; others show remarkable stability within an author. (A recent example by Brinegar [3] uses frequencies of word lengths for discrimination.)

Let us look in Table 2.1 at frequency distributions for rates of the high-frequency words *by*, *from*, and *to* in 48 Hamilton and 50 Madison papers (some exterior to *The Federalist*[3]). High rates for *by* usually favor Madison, low favor

[3] We do not give here the detailed references to all the exterior papers used. Papers numbered from 1 to 77 are from *The Federalist*. For the rest, consecutive code numbers imply related materials. The most important of these

TABLE 2.2. FREQUENCY DISTRIBUTION FOR *war*

Rate/1000	H	M
0 (exactly)	23	15
0+–2	16	13
2– 4	4	5
4– 6	2	4
6– 8	1	3
8–10	1	3
10–12	—	3
12–14	—	2
14–16	1	2
Totals	48	50

Hamilton; for *to* the reverse holds. Low rates for *from* tell little, but high rates favor Madison. In Table 2.2, we see that rates for the word *war* vary considerably for both authors. The meaning of *war* automatically suggests that the rate of use of this word depends on the topic under discussion. In discussions of the armed forces the rate is expected to be high; in a discussion of voting, low. We call words with such variable rates "contextual," and we regard them as dangerous for discrimination.

The best single discriminator we have ever discovered is *upon*, whose rate is about 3 per thousand for Hamilton and about 1/6 per thousand for Madison. We show its distribution in Table 2.3.

Nearly all words have low rates per thousand words; thus the occurrence of any one word at one spot in a text is a rare event. In many statistical problems distributions of rare events are governed by the Poisson law, which gives the probability of a count of x as a function of the mean count, λ, as follows:

$$P(x) = e^{-\lambda}\lambda^x/x! \qquad x = 0, 1, 2, \cdots.$$

For this distribution, the variance is also λ.

TABLE 2.3. FREQUENCY DISTRIBUTION FOR *upon*

Rate/1000	H	M
0 (exactly)	—	41
0+–1	1	7
1 –2	10	2
2 –3	11	
3 –4	11	
4 –5	10	
5 –6	3	
6 –7	1	
7 –8	1	
Totals	48	50

is Madison's 40,000 word paper on *Neutral Trade* which we chopped into 20 pieces of approximately 2000 words each and gave the code numbers 201–220.

To find out how well the Poisson fits the distribution of Hamilton and Madison word counts, we broke up a collection of *Federalist* papers into 247 blocks of approximately 200 words, and tabulated the frequency distribution of counts for each of many different words. For a few function words (filler words of the language such as articles, conjunctions, and prepositions) we give distributions for Hamilton in Table 2.4. The distributions for *an*, *any*, and *upon* are fitted rather well by the Poisson, but even a motherly eye sees disparities for *may* and *his*. The tail of the Poisson is not thick enough to fit the distribution for *his*, and the sizes of the fitted counts for $x=0$ and $x=1$ are far from the observed counts.

Another family of distributions, the negative binomial, possesses features that aid the fitting of *may* and *his*. Many probabilistic mechanisms lead to the negative binomial, and we mention one example for our problem. Suppose for each piece of writing an author uses a given word with Poisson frequencies but that he changes its rate of use from one piece of writing to another according to a gamma distribution. Then the negative binomial describes the distribution of counts over many pieces of writing. One way to write the negative binomial is

$$P(x) = \frac{\Gamma(x+\kappa)}{x!\Gamma(\kappa)}\left(\frac{\lambda/\kappa}{1+\lambda/\kappa}\right)^x\left(\frac{1}{1+\lambda/\kappa}\right)^\kappa \qquad x = 0, 1, 2, \cdots.$$

This distribution has mean λ and variance $\lambda(1+\lambda/\kappa)$. Since the variance of the Poisson was λ, and since the negative binomial has the Poisson as a limit for

TABLE 2.4. OBSERVED AND FITTED POISSON AND NEGATIVE BINOMIAL DISTRIBUTIONS FOR SELECTED WORDS—HAMILTON

	Occurrences:							
	0	1	2	3	4	5	6	7 or more
an								
obs.	77	89	46	21	9	4	1	
Poisson	71.6	88.6	54.9	22.7	7.0	1.7	.4	.1
*N.B. (6.41)	81.0	82.7	49.2	22.0	8.2	2.7	1.0	.2
any								
obs.	125	88	26	7	0	1		
Poisson	126.3	84.6	28.5	6.4	1.1	.2		
N.B. (∞)	same as Poisson							
may								
obs.	128	67	32	14	4	1	1	
Poisson	109.9	88.9	36.0	9.7	2.0	.3	.1	
N.B. (1.64)	128.2	69.4	30.1	12.1	4.6	1.7	.6	.3
upon								
obs.	129	83	20	9	5	1	6 or	
Poisson	121.6	86.1	30.6	7.3	1.3	.2	more	
N.B. (4.20)	131.1	77.1	27.9	8.2	2.1	.5	.1	
his								
obs.	192	18	17	7	3	2	4	3
Poisson	131.7	82.7	26.2	5.5	.9	.1		
N.B. (.154)	192.2	23.8	11.0	6.4	4.0	2.7	1.9	5.0

* Estimated values of κ follow N.B. in parenthesis.

TABLE 2.5. FUNCTION WORDS AND THEIR CODE NUMBERS FOR THE FEDERALIST STUDY

1 a	8 as	15 do	22 has	29 is	36 no	43 or	50 than	57 this	64 when
2 all	9 at	16 down	23 have	30 it	37 not	44 our	51 that	58 to	65 which
3 also	10 be	17 even	24 her	31 its	38 now	45 shall	52 the	59 up	66 who
4 an	11 been	18 every	25 his	32 may	39 of	46 should	53 their	60 upon	67 will
5 and	12 but	19 for	26 if	33 more	40 on	47 so	54 then	61 was	68 with
6 any	13 by	20 from	27 in	34 must	41 one	48 some	55 there	62 were	69 would
7 are	14 can	21 had	28 into	35 my	42 only	49 such	56 thing	63 what	70 your

$\kappa \to \infty$, λ/κ can be taken as a measure of the non-Poissonness of a negative binomial.

In our study of classical discrimination, the form of the distribution of frequencies for words scarcely plays a role. But in the main study, where Bayes' theorem is used, the form of the distribution matters, and we investigate the usefulness of both the Poisson and the negative binomial.

As we implied in discussing the word *war*, the words we want to use are noncontextual ones, words whose rate of use is nearly invariant under change of topic. For this reason, the little filler words, called function words (see Table 2.5), are especially attractive for discrimination purposes. Miller, Newman and Friedman [7] give a list of 363 function words and their frequencies in about 35,000 words of text assembled from the King James *Bible*, from William James, and from *The Atlantic* (1957). This created one pool of words for our use. From it we chose the 70 highest-frequency words on their list, and a random set of 20 from their low-frequency words.

Two additional pools of words were used. An alphabetical word index with frequencies for 70,000 words of text, half Hamilton, half Madison became available. From the 6700 different words of this index, we eliminated words whose low frequency (five occurrences or fewer) make them nearly useless for discrimination, and words that might be contextual: words like *law*, *executive*, *war*, and also many other words for which the contextuality was less evident; most concrete nouns and verbs, for example, *body*, *condemn*, *doctrine*, *friends*, *land*, *opened*. The remainder were tested for ability to discriminate by plotting their frequencies for Hamilton and Madison on binomial probability paper. We kept as potential discriminators those for which the authors' rates differed by at least three standard deviations. This created a set of 103 words.

In a third approach, used to select low-frequency words early in our work, we reviewed approximately equal numbers of papers, about five for each author, and kept or eliminated words on the basis of incidence. Then those words that

TABLE 2.6. ADDITIONAL WORDS AND CODE NUMBERS FOR THE FEDERALIST STUDY

*71 affect	*79 city	*87 direction	*94 innovation	102 perhaps	*110 vigor
*72 again	*80 commonly	*88 disgracing	*95 join	*103 rapid	*111 violate
*73 although	*81 consequently	89 either	*96 language	104 same	*112 violence
74 among	*82 considerable	*90 enough (and in sample of 20)	97 most	105 second	*113 voice
75 another	*83 contribute		98 nor	106 still	114 where
76 because	*84 defensive	*91 fortune	*99 offensive	107 those	115 whether
77 between	*85 destruction	*92 function	100 often	*108 throughout	*116 while
78 both	86 did	93 himself	*101 pass	109 under	*117 whilst

* Means word emerged from screening study, the rest from random sample of low frequency function words.

TABLE 2.7. NEW WORDS FROM THE WORD INDEX STUDY TOGETHER WITH THEIR CODE NUMBERS

118 about	130 choice	142 intrust+s+ed+ing	154 proper
119 according	131 common	143 kind	155 propriety
120 adversaries	132 danger	144 large	156 provision+s
121 after	133 decide+s+ed+ing	145 likely	157 requisite
122 aid	134 degree	146 matter+s	158 substance
123 always	135 during	147 moreover	159 they
124 apt	136 expence+s	148 necessary	160 though
125 asserted	137 expense+s	149 necessity+ies	161 truth+s
126 before	138 extent	150 others	162 us
127 being	139 follow+s+ed+ing	151 particularly	163 usage+s
128 better	140 I	152 principle	164 we
129 care	141 imagine+s+ed+ing	153 probability	165 work+s

were kept were tried on another set of papers, and cumulative totals of Hamilton and Madison incidences in papers recorded for each word. This process continued through four sets of papers for each author, and the result was a selection of 28 low-frequency words.

All told then for the three pools of words, (1) the Miller-Newman-Friedman list yielded 70 high-frequency words and the random sample of 20 low-frequency function words, (2) the word index study yielded 103 words selected from about 6700 for ability to discriminate, and (3) the succession of sets of papers yielded 28 words (screening set) selected for ability to discriminate from among about 3000 words. Because the lists overlap, the number of different words among the sets of 70, 20, 103, and 28 is 165. We do not give a Venn diagram, but Table 2.5 gives the 70 high-frequency function words, Table 2.6 gives the random sample and the 28 obtained from the successive waves, and Table 2.7 gives the additional words obtained from the word index.

3. WEIGHT-RATE ANALYSIS

A. The study, its strengths and weaknesses

While our main study is Bayesian in character, we want also to see how a more usual approach handles this problem of discrimination. We present it first because the reader may be more familiar with this approach.

The plan is to construct a linear discriminant function, a weighted sum of the rates for words, using about half the data for each author, called the *screening set*, to select the discriminating words and to get weights to apply to their observed rates. Once the words are chosen and weighted, the function is tried out on the other half of the material of known authorship, not only as a test of the chosen function, but also to calibrate the weighted sum on material uncontaminated by the effects of selection and weighting. Finally, the discriminant function is applied to the disputed papers, and the values obtained are compared with those obtained from the *calibrating set* for each author by methods that are described later.

Why do we need a calibrating set, why not use all the data for selection and weighting? First, when for some purpose we choose the best looking few from a large set of variables they may not perform as well as they appeared to in the material that was used to select them—this loss is called the regression effect. Second, usually weights must be chosen to optimize results if new data are very

like those in the material used for weighting. When they are not, there is an additional slump. Third, we do not know what the selectivity may have done to the relations among the chosen variables, and we need estimates of location and variability that cannot be obtained from data used for selection and weighting. Fourth, since we have weighted the words as if they are uncorrelated, and they are not quite, the calibrating set is required to adjust for correlation in the components leading to the total score. Thus we avoid having the *final* analysis lean upon the assumption of zero correlations. It is our observation that studies of discrimination rarely employ calibrating sets, even though they may be gravely needed. The magnitude of the selection and weighting effects may not be adequately appreciated. We illustrate their sizes for this problem below.

An analysis of rates is the basis for this study, and rates, themselves, lead to one weakness that we have done nothing about. Low-frequency words often have zero frequencies and therefore zero rates in the papers. A zero rate for a 1000-word paper is quite different information from a zero rate for a 3000-word paper, and a proper analysis should take some account of that fact.

A major weakness of this study is that it does not have the serious protection against contextual words provided for the main study discussed later. Partly this is a lapse in our work, but partly it comes from the limited amount of material available when we use a calibrating set as well as a selection and weighting set of papers.

B. Materials and techniques

As a pool of words for this study, we use those of the screening study and the two sets of function words (section 2), in all, the 117 words with code numbers 1 to 117 (see Tables 2.5 and 2.6).

We use 23 Hamilton and 25 Madison papers to choose words and find weights for them. Next, these weights are applied to the remaining papers of known authorship, which form a calibrating set of 25 Hamilton and 25 Madison papers uncontaminated by the process of selection and weighting. Finally the weights are applied to the rates for the disputed and joint papers and these are assessed.

If x_i denotes the observed rate for the ith word in a given paper, and if W_i is a weight for that word, a linear discriminant function takes the value y for the given paper, obtained by summing over words:

$$y = \Sigma W_i x_i.$$

The weights are chosen so that when Hamilton is the author and the rates apply to him, the value of y tends to be large (say), and when Madison is the author, the value small. In addition to assigning direction, the weights give better discriminating words more importance.

The set of weights for words that maximizes the standardized difference between the means[4] of the discriminant function for the two sets of papers is proportional to

$$\frac{\mu_{iH} - \mu_{iM}}{\sigma^2_{i\,H} + \sigma^2_{i\,M}},$$

[4] Difference divided by the square root of the sum of the variances of scores for Hamilton and Madison.

TABLE 3.1. WEIGHT-RATE ANALYSIS: WORDS, WEIGHTS, AND IMPORTANCES (TIMES 10^4)

	Weight	Importance		Weight	Importance		Weight	Importance
Group 1			*Group 3*			*Group 5*		
upon	1394	3847	as	−0140	0339	innovation	−1681	0336
Group 2			at	0247	0318	language	−1448	0304
although	−1754	0351	by	−0146	0542	vigor	2174	0543
commonly	1333	0267	of	0037	0281	voice	−2159	0410
consequently	−1311	0459	on	−0271	0796	*Group 6*		
considerable	0784	0251	there	0463	0972	destruction	1709	0342
enough	0683	0403	*Group 4*					
while	2708	0704	would	0085	0428			
whilst	−2206	0993						

where μ_{ij} and σ_{ij}^2 are the mean and variance of the rate for word i for author j and where words are assumed uncorrelated.

Since we do not know the theoretical means and variances, we need to estimate them, or something proportional to them.

For the μ's we use median rates in the screening set for each author unless one or both of the medians is zero, and then average rate for each author is used. To replace the variances we used squares of ranges (range is largest minus smallest rate).

Our reasons for using ranges were to cut down on total amount of calculation and to defend against contextual words. Recall that ranges are somewhat more affected by outliers than standard deviations, and the more contextual the word, the more wildly its rates vary. Thus the weight chosen, dropping the subscript i, is

$$W = \frac{\bar{x}_H - \bar{x}_M}{R_H^2 + R_M^2}$$

where R stands for range and $\bar{x}$ stands for median or average as appropriate.

Naturally we preferred not to put the entire 117 words into the discriminant, because 90 are unselected and offer little hope of helping. Suppose one rate is exactly at Hamilton's average and another at Madison's; then the difference in score, $W\bar{x}_H - W\bar{x}_M = W(\bar{x}_H - \bar{x}_M)$, is one measure of the importance of the variable, for it gives the contribution of the particular word to the difference in score between an average Hamilton paper and an average Madison paper. We arbitrarily chose an importance of .025 as the cut-off, and this left us with the 20 words displayed in Table 3.1.

C. Results for the screening and calibrating sets[5]

In Table 3.1, we see that *upon* is the most important word by a factor of 4. Other leaders are *whilst*, *there*, and *on*. The words are grouped because we used the word index as a source to help classify words more or less as to their degree of apparent danger from contextuality for the study. Thus all of the 117 words in the pool fall into 6 classes—Group 1 for *upon* because it is so outstanding and because it cannot easily be classified either as a high- or as a low-frequency

[5] The calculations in this part of the study were largely supervised by Charles L. Odoroff.

TABLE 3.2. WEIGHT-RATE ANALYSIS: RESULTS FOR CALIBRATING SET

Paper	Hamilton						Total y	Paper	Madison						Total y
	Word group								Word group						
	1	2	3	4	5	6			1	2	3	4	5	6	
13	.29	.07	.40	.12	0	0	.88	134	0	−.09	.09	.04	0	0	.05
15	.45	.09	.39	.04	−.12	0	.85	135	0	−.15	−.23	.08	−.20	0	−.50
16	.44	.29	.22	.16	0	0	1.11	201	0	.01	−.36	.02	.11	0	−.22
17	.54	.17	.34	.06	.14	0	1.26	202	.07	−.06	−.28	.00	0	0	−.28
21	.42	.41	.14	.05	0	.08	1.11	203	0	−.06	−.16	.02	−.07	0	−.27
22	.52	.10	.23	.05	.06	0	.96	204	0	−.63	−.19	.03	0	0	−.79
23	.54	0	.06	.02	.00	0	.62	205	0	−.28	−.06	.00	.04	0	−.29
25	.14	.04	−.07	.09	.11	0	.31	206	.07	−.49	−.25	.03	0	0	−.64
27	.39	.15	.24	.02	0	0	.80	207	0	−.46	−.23	.00	−.17	0	−.86
28	.26	0	.37	.06	0	0	.70	208	0	−.09	−.21	.02	−.11	0	−.38
29	.64	.28	.39	.07	0	0	1.39	209	.13	−.42	−.20	.01	−.16	0	−.65
30	.92	.07	.03	.10	.11	.09	1.32	210	0	−.21	−.36	.00	−.08	0	−.65
31	1.05	.08	.14	.02	0	.10	1.38	216	0	−.20	.01	.03	0	0	−.16
32	.20	−.12	.12	.14	.15	0	.49	217	0	−.08	−.30	.03	−.07	0	−.42
33	.72	0	−.05	.08	0	0	.75	218	0	−.19	−.23	.01	0	0	−.42
34	.63	.16	.19	.08	0	0	1.06	219	0	−.29	.09	.01	−.15	0	−.53
35	.56	.12	.06	.06	0	0	.80	220	0	.02	−.23	.02	0	0	−.19
60	.50	.04	.15	.11	−.06	0	.73	301	0	−.18	−.01	.03	0	.09	−.07
61	.28	0	.27	.10	0	0	.64	302	0	−.11	−.14	.08	−.05	0	−.22
67	.60	0	.11	.03	−.10	0	.63	311	0	−.38	−.12	.05	.00	0	−.44
68	.19	.10	.01	.04	0	0	.34	312	0	0	−.09	.00	0	0	−.08
69	.64	.16	.29	.09	0	0	1.18	313	0	.10	.05	0	0	0	.15
73	.79	.07	.12	.11	.10	0	1.19	314	0	−.15	−.17	.03	−.10	0	−.38
75	.37	.75	.07	.12	0	0	1.31	315	0	−.24	−.10	.04	−.29	0	−.59
76	.64	.32	.10	.11	0	0	1.17	316	.09	−.51	−.02	0	−.14	0	−.57
$\bar{y}$	.51	.13	.17	.08	.02	.01	.92	$\bar{y}$	.01	−.20	−.16	.02	−.06	.00	−.38
s	.22	.18	.14	.04	.06	.03	.32	s	.04	.19	.12	.02	.09	.02	.25

word. We gave groups we thought contextual higher numbers. For convenience of exposition, we take the rest up out of numerical order.

Group 4 consisted of function words from the Miller-Newman-Friedman list (see section 2) which were personal pronouns or auxiliary verbs.

Group 3 consists of the high-frequency function words (code number 1–70, excluding pronouns and auxiliaries) and the members of the random sample of low-frequency words. Aside from *upon*, the words obtained in the screening study were placed in groups 2, 5, or 6. Group 2 includes those that are function words and those on a list of words we called "well-liked" words. From the original word index, without any evidence of differential use by Hamilton and Madison, we made up two lists of words; the "well-liked" list contains mostly abstract adverbs that we assessed as largely free from contextuality, and a larger list of more meaningful words, that did not seem so contextual that we should delete them. The latter words fall into group 5. Group 6 includes what is left: words we regarded as possibly highly contextual.

As in the main study, the reason we went to all this trouble was to get some quantitative idea of what such intuitive classifications might be worth. We were encouraged to see that groups 4 and 6 were practically annihilated by the weights and that groups 2 and 3 had a number of members. But the calibrating set is the ultimate test.

Now that the words are selected and the weights readied, the weights are applied to the papers in the Hamilton and Madison screening sets. We expect the two screening sets to be well separated since the words and weights are chosen to do just that. The Hamilton mean is .87, the Madison one −.41, the midpoint .23. Paper 113, Pacificus III, has the lowest Hamilton score, .40; paper 10 with −.19 the highest Madison score. So the observed separation between the extremes is about 4 Madison standard deviations and somewhat less than 3 Hamilton standard deviations. Observe that the score 0 plays no special role in this analysis. All these papers can be classified perfectly by the discriminant function.

Next we test the discriminant function on the calibrating set, with the results shown in Table 3.2. Hamilton's average is .92, Madison's −.38, with midpoint .27; these three numbers are rather similar to those obtained from the screening set, an encouraging sign. For both sets the standard deviations are larger. One expects the means of the two groups to move toward one another, but here, instead, the standard deviations have grown. The smallest Hamilton score is .31, the largest Madison .15, so there is still no overlap between the papers of known authorship. On the other hand, the distance between these closest papers is only .5 Hamilton standard deviations or .6 Madison's. Still it is quite cheering to get no overlap, though it does not imply that we will make no errors in future classifications.

D. Regression effects

Before going on, let us see how much the various word groups have deteriorated between the screening set and the calibrating set. As a measure of discrimination we choose $(\bar{y}_H - \bar{y}_M)/[(s_H + s_M)/2]$ because it is easy to compute and has the rough interpretation "number of standard deviations the means

TABLE 3.3. WEIGHT-RATE ANALYSIS: DISCRIMINATION INDICES

Word group	Screening set	Calibration set
Group 1	3.3	3.8
Group 2	2.5	1.9
Group 3	3.5	2.5
Group 4	1.0	1.7
Group 5	2.7	.9
Group 6	1.3	.3
Total	6.9	4.5

are apart." The symbols in this formula refer to means and standard deviations in the screening or calibrating sets for groups or totals, whichever are appropriate. In Table 3.3 the results are shown for both screening set and the calibrating set.

The excellence of group 1, *upon*, actually an improvement from screening to calibration, illustrates an important point: when a few variables are selected from many, if these few really are head and shoulders better than the rest, there need be no regression effect owing to selection, though there may be a loss owing to choice of weights. The loss from 2.5 to 1.9 in group 2 is in the expected direction, but not spectacular. The loss in group 3 may be partly selectivity and partly the result of reduced paper length.

We shall not try to account for the change in group 4, which may be due to the change from *Federalist* material in Madison's screening set to the *Neutral Trade* paper in his calibrating set. Both group 5 and group 6 fell flat on their faces. The discrimination in the total dropped from nearly 7 standard deviations in the screening set to 4.5 in the calibrating set—a drop is certainly expected, but the size of the drop cannot be anticipated in this method. On the positive side, we have left 4.5 standard deviations' worth of discrimination as measured in an independent set of papers. And we are glad we provided for both a screening set and a calibrating set so as not to be misled by the original 7 standard deviations.

E. Results for the disputed papers

Scores for the joint and the disputed papers are shown in Table 3.4. The average score for the 12 disputed papers is −.31, slightly higher than Madison's −.38 for the calibrating set. Except for paper 55, all have negative scores and fall well below the midpoint .27 for the two calibrating sets. Paper 55 has a score that is higher than the most Madisonian of the Hamilton calibrating papers, and falls slightly on Hamilton's side. Later we try to take a more quantitative view. Superficially, then, all but No. 55 look Madisonian, and No. 55 is hard to judge because it is near the middle.

Among the joint papers, all three fall below the midpoint .27, though only No. 19 falls very far below it, and No. 20 is quite near the middle.

There are many ways to try to assess these results, and we present two here. As a first method we compare, using t-statistics, each disputed paper with the Hamilton calibrating set and with the Madison set. Let y be the total score for a disputed paper, compute

TABLE 3.4. RESULTS FOR JOINT AND DISPUTED PAPERS

Joint papers	Word group						Total
	1	2	3	4	5	6	
18	.07	−.04	−.17	.02	0	.08	−.03
19	0	−.04	−.15	.02	0	0	−.18
20	.10	.05	−.09	.01	.15	0	.22
$\bar{x}$	.05	−.01	−.14	.02	.05	.03	.00
Disputed papers							
49	0	−.22	−.22	.12	−.14	0	−.46
50	0	−.16	−.18	.08	0	0	−.26
51	0	−.37	−.31	.04	−.11	0	−.76
52	0	0	−.27	.04	0	0	−.23
53	0	−.18	−.17	.02	−.16	0	−.48
54	.14	−.18	−.44	.02	0	0	−.46
55	0	.04	.24	.04	0	0	.32
56	0	−.09	.05	.02	0	0	−.01
57	0	−.30	−.15	.02	−.08	0	−.50
58	0	−.15	−.24	.05	0	0	−.33
62	0	−.13	−.21	.02	−.09	0	−.41
63	0	−.03	−.13	.03	0	0	−.13
$\bar{x}$	.01	−.15	−.17	.04	−.05	0	−.31

$$t_H = \frac{y - \bar{y}_H}{s_H\sqrt{1 + \frac{1}{n_H}}}, \qquad t_M = \frac{y - \bar{y}_M}{s_M\sqrt{1 + \frac{1}{n_M}}}.$$

Here $n_H = n_M = 25$, so the numbers of degrees of freedom for the t's are 24. We use these t's to compute the area P_H to the left of t_H and P_M, that to the right of t_M, for these t-distributions. Here P_H evaluates the chance of getting a more Madisonian result than the one observed *if Hamilton wrote the paper*, and P_M is a similar quantity for Madison.

If both P_H and P_M were large, say .4, the meaning is that the discrimination is poor to start with and that the paper itself is difficult to discriminate on the basis of the data being examined. This does not happen in our problem because $\bar{y}_H$ and $\bar{y}_M$ are far apart. If P_H is large and P_M is small we incline toward Hamilton as the author, and vice versa. If both P_H and P_M are small, say .01 and .02, the potential discrimination is strong, but the paper in question is difficult to discriminate on these data.

In Table 3.5 the t's and P's are given for the joint and the disputed papers. Paper No. 55 still seems quite up in the air. The rest of the disputed papers except possibly 56 seem to be Madison's.

As a second method of appraising the results, a method more nearly comparable with the main study, we study approximate confidence limits for the likelihood ratio. Assume that the score $\bar{y}$ is approximately normally distributed for each author. Then a new piece of writing of length similar to those pieces previously studied gives the ratio

TABLE 3.5. OUTCOMES FOR THE JOINT AND DISPUTED PAPERS: t-VALUES, P-VALUES, 90% CONFIDENCE LIMITS FOR LOG LIKELIHOOD, AND ESTIMATED LOG ODDS

Joint papers	t_H	t_M	P_H	P_M	Lower confidence limit	Upper confidence limit	Estimated log odds
18	−2.924	1.326	.00372	.09866	− 9.4	.3	−3.2
19	−3.358	.779	.00131	.22180	−12.5	−1.5	−4.7
20	−2.146	2.305	.02110	.01506	− 4.5	5.0	.1
Disputed papers							
49	−4.247	− .341	.00014	.63196	−19.5	−3.0	−7.2
50	−3.614	.457	.00070	.32589	−14.5	−2.0	−5.6
51	−5.138	−1.463	.00002	.92178	−26.8	−3.8	−8.6
52	−3.529	.563	.00086	.28933	−14.8	−1.9	−5.3
53	−4.309	− .419	.00012	.66053	−20.0	−3.1	−7.3
54	−4.219	− .306	.00016	.61888	−19.3	−3.0	−7.1
55	−1.850	2.678	.03833	.00658	− 2.9	7.0	1.4
56	−2.864	1.401	.00428	.08701	− 9.0	.7	−2.9
57	−4.362	− .486	.00011	.68431	−20.4	−3.2	−7.4
58	−3.847	.163	.00039	.43594	−16.4	−2.5	−6.2
62	−4.092	− .145	.00021	.55704	−18.3	−3.0	−6.8
63	−3.212	.964	.00187	.17233	−11.5	−1.0	−4.2

$$K = \frac{\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-(y-\mu_1)^2/2\sigma_1^2}}{\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-(y-\mu_2)^2/2\sigma_2^2}}.$$

Unfortunately we do not know the μ's and σ's. We could, however, set confidence limits on K, or what is equivalent, on its natural logarithm λ. A method for obtaining conservative 90 per cent confidence limits for λ has been worked out by Ann Mitchell and results for the disputed papers are shown in Table 3.5.

For the 10 disputed papers excluding Nos. 55 and 56 the evidence seems very strong for Madison. For No. 55 the confidence is at least .90 that the odds are between 18 to 1 for Madison and 1100 to 1 for Hamilton. This evidence does not seem so strong to us, nor do the somewhat stronger odds in Madison's favor for paper No. 56.

Among the joint papers, the problem is one of extent and the log odds have been given for general interest.

Two objections to these confidence limits are: first, they are very likely too conservative for the answer they claim to give; second, they may be sensitive to the assumption of normality. These objections cancel, but one is uncertain where the net would be.

For those who wish a point estimate of the log odds from this weight-rate study, we give one obtained from t-distributions in the right-most column of Table 3.5. Among the 12 disputed papers, the odds are satisfyingly fat except

for Papers 55 and 56, and Madison is strongly indicated as the author. These conclusions are only for the weight-rate study and do not override the results of the more definitive main study. We feel the shortcomings mentioned at the close of section 3A make the results of the weight-rate study less satisfactory than those of the main study, and that no matter what school of statistics one prefers there is a suitable interpretation from the main study. On the other hand, we must emphasize that the weaknesses of the weight-rate study are not entirely inherent in the method, but reside partly in our execution of it, and partly in the limited materials of this problem.

4. INTRODUCTION TO BAYES' THEOREM AND ITS APPLICATIONS

How does an observation change our beliefs? In uncertain inference, Bayes' theorem offers one answer. We use numerical probabilities to express degrees of belief about propositions such as "Hamilton wrote paper No. 52" and use Bayes' theorem to adjust these probabilities for the evidence in hand. A sequence of examples illustrates both the ideas and the problems involved in our two different uses of Bayes' theorem and leads us to our actual application.

TABLE 4.1. POISSON PROBABILITIES WITH $w\mu_H = .5$, $w\mu_M = 1.0$

Frequency	Hamilton	Madison
0	.607	.368
1	.303	.368
2	.0758	.184
3	.0126	.0613
4	.00158	.0153
5	.000158	.00307
6	.0000132	.000511

A. An example of applying Bayes' theorem with initial odds and parameters known

Suppose Hamilton's and Madison's use of the word *also* are well represented by Poisson distributions with parameters $w\mu_H$ and $w\mu_M$, where w is paper length in thousands of words and the μ's are the rates per thousand. Suppose further that the rates are known to be $\mu_H = .25$ and $\mu_M = .50$. Then for an unknown paper of length $w = 2$ (2000 words), the probabilities for 0 to 6 usages by each author are shown in Table 4.1 to three significant figures.

Suppose *also* is used four times in a paper of 2000 words written either by Hamilton or by Madison; what are the odds that Hamilton wrote the paper? Naturally, the answer depends on our uncertainty. If we were nearly sure before the observation that Hamilton wrote it, the comparison of the probabilities, .00158 and .0153 would not much weaken our belief. But if initially we thought the authorship was a tossup, the evidence would give us new odds of 10 to 1 (.0153/.00158) in favor of Madison. Now we develop these ideas formally.

Let p_1 and $p_2 = 1 - p_1$ be the probabilities before the observations that Hypotheses 1 and 2 respectively are true. For example, Hypothesis 1 might be that Hamilton wrote the paper, Hypothesis 2 that Madison wrote it. Let $f_i(x)$, $i = 1$,

2 be the conditional probabilities of observing the result x, given that Hypothesis i is true. Assume that x is one of a discrete set of possible observations. The probability that the specific result x occurs is $p_1f_1(x)+p_2f_2(x)$. The conditional probability that Hypothesis 1 is true given observation x is

$$P(\text{Hypothesis } 1 \mid x) = \frac{p_1f_1(x)}{p_1f_1(x) + p_2f_2(x)}.$$

This result is a special case of Bayes' theorem. It can be extended to the continuous case.

Both computational and intuitive advantages accrue if we use odds instead of probabilities. The odds for Hypothesis 1 relative to Hypothesis 2 are given by the ratio of their probabilities. Thus the odds are defined to be

$$\text{Odds}(1, 2 \mid x) = \frac{P(\text{Hypothesis } 1 \mid x)}{P(\text{Hypothesis } 2 \mid x)} = \frac{p_1f_1(x)}{p_2f_2(x)} = \left(\frac{p_1}{p_2}\right)\left(\frac{f_1(x)}{f_2(x)}\right) \qquad (1)$$

$$= (\text{initial odds}) \times (\text{likelihood ratio}) = \text{final odds}.$$

Thus p_1/p_2 is the initial odds determined by our beliefs prior to the execution of the experiment leading to the result x, and $f_1(x)/f_2(x)$ is the likelihood ratio determined by the data of the experiment itself.

To return to our example, suppose initially we are very sure that Hamilton wrote the new paper, say $p_1=.999$ and $p_2=.001$, then the initial odds for Hamilton are 999/1 or 999. The observed count of 4 usages in the paper gives likelihood ratio $f_1(4)/f_2(4) \approx .00158/.0153 \approx .1$. Thus the final odds are

$$\text{Odds}(\text{Hamilton, Madison} \mid x = 4) \approx 999(0.1) = 99.9,$$

or about 100 to 1 for Hamilton.

B. Selecting words and weighting their evidence

Bayes' theorem, as described in equation (1), can help us to select words for the final discrimination, and to weight the evidence from each word, as we now illustrate.

Let us extend our previous example and suppose that independent Poissons with the known rates given in Table 4.2 apply to *also*, *an*, and *because*. Suppose a 2000-word paper by one of our two authors contains 4 *also*'s, 7 *an*'s, and 0 *because*'s. If Hamilton is the author, the probability of the triple (4, 7, 0) is the product of three Poisson probabilities $f_P(4 \mid .5)f_P(7 \mid 12)f_P(0 \mid .9)$, while if Madison is the author the probability is $f_P(4 \mid 1)f_P(7 \mid 9)f_P(0 \mid 1)$. Finally the likelihood ratio for the 3 words is the product of the likelihood ratios for the separate words, and the weighting problem is solved.

TABLE 4.2. RATES PER THOUSAND FOR *also*, *an*, AND *because*

Word	Hamilton rate	Madison rate
also	.25	.50
an	6.00	4.50
because	.45	.50

The logarithmic form of Bayes' theorem can be obtained by taking the logarithm of both sides of equation (1) to get

$$\text{Final log odds} = \text{Initial log odds} + \log \text{likelihood ratio},$$

and then evidence from independent measurements is additive instead of multiplicative.

For a single pair of Poisson distributions the likelihood ratio is

$$K = (\mu_N/\mu_M)^x e^{-w(\mu_H-\mu_M)}$$

and the log likelihood ratio is

$$\lambda(x) = x \log (\mu_H/\mu_M) - w(\mu_H - \mu_M).$$

For one observation on each of n independent words the log likelihood ratio is

$$\sum \lambda_i(x_i) = \sum [x_i \log (\mu_{iH}/\mu_{iM}) - w(\mu_{iH} - \mu_{iM})],$$

where the subscripts have their obvious meanings.

In choosing independent words for discriminators, the criterion is whether their contribution is worth the cost of including them. For known rates, no bias arises from selection. In the example, *because* cannot contribute much to the log odds for any plausible observation, indeed its likelihood ratios for $x=0$, 1, 2, 3 are, successively, 1.11, 1.00, .90, .81. If the cost of observing the frequency of *because* is high, or in more complicated problems, if the cost of computing is large, then we may prefer to discard the word. Of course, that decision should *not* be preceded by a peek at the frequencies of *because*'s in the unknown material.

C. *Initial odds*

All this seems perfectly straightforward; wherein lie the difficulties? The first is in the choice of the *initial odds*, here, the values of p_1 and p_2. Your final odds—posterior odds—will differ from mine, if we choose different values for p_1. In some problems the choice of p_1 might be made on the basis of objective frequencies. But in others, personal degree of belief may be involved. In our own problem you might regard the initial authorship odds as 1 for an unknown paper, regarding a 50-50 chance as appropriate to your degree of ignorance. Or if you are an historian with knowledge of the problem you may have quite strong beliefs that lead to the assignment of large odds in one or the other direction. Nevertheless, the analysis of the data, using more and better words than the three in the example, may provide a likelihood ratio that overwhelms most of the variation in initial odds. Then the final odds, though still variable from person to person, would be very large, or very small.

For example, a likelihood ratio of 10^{-6} would convert strong initial odds of 10^3 for Hamilton to final odds of 10^{-3} (1000 to 1 for Madison). In terms of probabilities, for the same example, it converts an initial probability of .999 to a final probability of .001. Naturally, for any strength of data, opinions can be sufficiently extreme that the direction of odds cannot be changed.

In summary, by the factorization of final odds into the product of initial odds and likelihood ratio, the difficulties in assessing the initial odds have been sepa-

rated from the statistical analysis needed to determine the likelihood ratio. Further, if the likelihood ratio is large enough, or small enough, the final probabilities (not odds) of authorship are changed only slightly by wide changes in the initial odds. The statistical analysis in our study is largely devoted to the likelihood ratio term.

D. Unknown parameters

The second difficulty arises from uncertainty in the data distribution. We do not know that it is Poisson, indeed we have seen that there is evidence for the negative binomial. But even if the form of the distribution were known exactly, we do not know its parameters exactly, a difficulty magnified by the selection of best words from many candidates. And, finally, can we be confident that the parameters remain constant from one sort of text to another? For now, let us look at the difficulty of the unknown parameter value.

The usual source for information on parameter values is observation of large amounts of data known to be sampled from each distribution. If these data provide precise point estimates $\hat{\mu}_H$, $\hat{\mu}_M$ of the parameters, we would, in the tradition of large sample theory, use the estimates in place of the known values to evaluate the likelihood ratios. We use this same approximation even when our estimates are not so precise.

In the actual problem, we have about 94,000 words of text known to be written by Hamilton, 114,000 words by Madison—seemingly vast amounts, yet surprisingly little for most purposes. If the word *also* had true rates .25 and .50 in Hamilton and Madison writings, we would expect only 24 and 57 occurrences, respectively, and our standard errors for the observed mean rates would be $\sqrt{.25/94} \approx .052$ and $\sqrt{.50/114} \approx .066$. These standard errors are not negligible, and selection effects are serious. The word *also* was chosen as one of the better words for discrimination, on the basis of the difference between observed rates, so the true rates are likely to be closer together than the observed rates.

One method of making allowances for the selection effects and for guiding the choice of point estimates for approximate determination of likelihoods makes use of Bayes' theorem, this time in application to many continuous parameters, rather than to two hypotheses.

We express our state of knowledge about each author's rates by probability distributions. These probability distributions are posterior distributions obtained from an analysis of papers of known authorship together with prior information about comparative and average rates of word use. This analysis, a major part of our study, is described in section 5, and yields as output for each word, the posterior joint density, $p(\mu_H, \mu_M)$, of Hamilton's and Madison's rates. Because the rates μ_H and μ_M are dependent a priori and a posteriori, the marginal posterior densities, $p(\mu_H)$ and $p(\mu_M)$, can only be obtained by numerical integrations from $p(\mu_H, \mu_M)$.

Let us return to the *also* example, and the problem of assessing the evidence from 4 occurrences in an unknown paper of 2000 words. What we need as the likelihood ratio factor for changing initial odds to final odds is the ratio of the probability of observing 4 occurrences of *also* if Hamilton is the author to the probability of 4 occurrences if Madison is the author. Our model does not specify these probabilities; it specifies the conditional probability of 4 occurrences

given that Hamilton is the author and given that his rate is μ_H; in symbols, $P\{x=4 \mid H, \mu_H\}$. The desired unconditional probability can be found by averaging the conditional probability over the posterior distribution of the rate $\tilde{\mu}_H$:

$$P\{x = 4 \mid H\} = \int P\{x = 4 \mid H, \mu_H\} p(\mu_H) d\mu_H = E(P\{x = 4 \mid H, \tilde{\mu}_H\}),$$

where the tilde $\sim$ over μ_H indicates that the expectation is taken over the distribution of μ_H. Similarly

$$P\{x = 4 \mid M\} = \int P\{x = 4 \mid M, \mu_M\} p(\mu_M) d\mu_M = E(P\{x = 4 \mid M, \tilde{\mu}_M\}),$$

so the unconditional likelihood ratio is the ratio of the two expectations or integrals.

Even for the Poisson family of distributions, carrying out the integrations requires, except for some special cases, bivariate numerical integrations. The integrals look univariate, but the marginal densities of μ_H must each be obtained by an integration from the joint distribution of μ_H, μ_M that Bayes' theorem yields, for example:

$$P\{x = 4 \mid H\} = \int\int P\{x = 4 \mid H, \mu_H\} p(\mu_H, \mu_M) d\mu_H d\mu_M.$$

In the extension to the negative binomial family, the integrals become four-dimensional.

A natural approximation to an expectation is:

$$E(g(\tilde{\mu})) = \int g(\mu) p(\mu) d\mu = g(\hat{\mu})$$

in which the function is evaluated at some central value of the distribution $p(\mu)$. The mean might be the preferred measure but it can be determined only by integrations of the type being avoided. The mode is a more feasible choice well articulated with the use of Bayes' theorem. The result of using this approximation is identical with that obtained when the estimated rates $\hat{\mu}_H$, $\hat{\mu}_M$ are used as the known rates.

5. HANDLING UNKNOWN PARAMETERS OF DATA DISTRIBUTIONS

When parameters of a data distribution are unknown, Bayes' theorem requires a prior distribution for these quantities, and the parameters are treated as random variables rather than constants. For the Poisson, the rate parameters μ_H and μ_M for each word would be assigned a prior distribution, and the prior would allow for effects of selection of apparently good discriminators from large pools of words. The data on words is sufficiently limited that a flat prior is inappropriate.

A. Evaluating the prior

We like to have prior distributions based on data, even feebly. To get a hold on the prior distributions we use the observed data on a pool of words unselected

for their ability to discriminate; in particular we use the 90 function words from the Miller-Newman-Friedman list (see section 2).

To get average frequency of occurrence separated from differential frequency by the two authors we introduce a pair of parameters for each word:

$$\sigma = \mu_H + \mu_M, \qquad \tau = \frac{\mu_H}{\mu_H + \mu_M}.$$

Clearly σ measures average frequency, and τ ability to discriminate. For fixed $\tau(\neq 1/2)$ the bigger σ, the better the discrimination. When $\tau = 1/2$, the authors have equal rates. For authors writing together on the same topic at the same period, we suppose that the prior distribution for τ for any word would be nearly symmetric and unimodal. The spread may depend on σ. Our plan is to get a rough estimate of the distribution of τ over our pool of 90 words, and to use that estimate as a prior for the τ for any word. We cannot expect to determine the distribution of τ exactly, but we can hope to estimate it within a range adequate for our uses.

A graphical treatment is instructive. Let $s = m_H + m_M$ estimate σ, and $t = m_H/s$ estimate τ, where the m's are observed rates for a given word. For our material s has a small standard error, so for practical purpose we take $s \approx \sigma$. The standard error of t depends on both τ and σ.

In Figure 5.1 the (s, t) pairs are plotted for the 90 unselected function words. (The square root scale is used for technical reasons. The circled points have been moved to the left.) Curves B and C in the figure give two-standard-error bands around the null value $\tau = .5$. Curve A is located 2 standard errors above the value $\tau = .55$, and Curve D the same distance below $\tau = .45$.

The variation vertically in Figure 5.1 is more than Poisson variation can stand. But if we want to estimate the fraction of words with τ outside the interval .3 to .7, the estimate should be less than the observed fraction .12 of t's outside, because sampling variation alone could put some t's outside even if no τ's are. Similarly the true proportion outside .4 to .6 is less than the observed .28.

If the distribution of τ values were a beta $\tau^{\gamma-1}(1-\tau)^{\gamma-1}/\mathrm{B}(\gamma, \gamma)$ with equal arguments γ, we could investigate how the probabilities vary with γ. We chose the form $\gamma = \beta_1 + \beta_2\sigma$ to allow decreased variability for τ with increased σ. After reviewing the behavior of γ we decided that a range of γ between 5 and 20 was plausible. Therefore we selected the pairs of values for β_1 and β_2, shown in Table 5.1, that would more than handle the range.

B. The interpretation of the prior

More formally, we suppose there are underlying linguistic quantities β_1, β_2 that determine the general similarity of word occurrences by Hamilton and

TABLE 5.1. VALUES OF β_1, β_2 CHOSEN FOR THE INITIAL CALCULATIONS

β_1	5	2	10	20	5	2	20
β_2	1	1	1	1	5	10	10

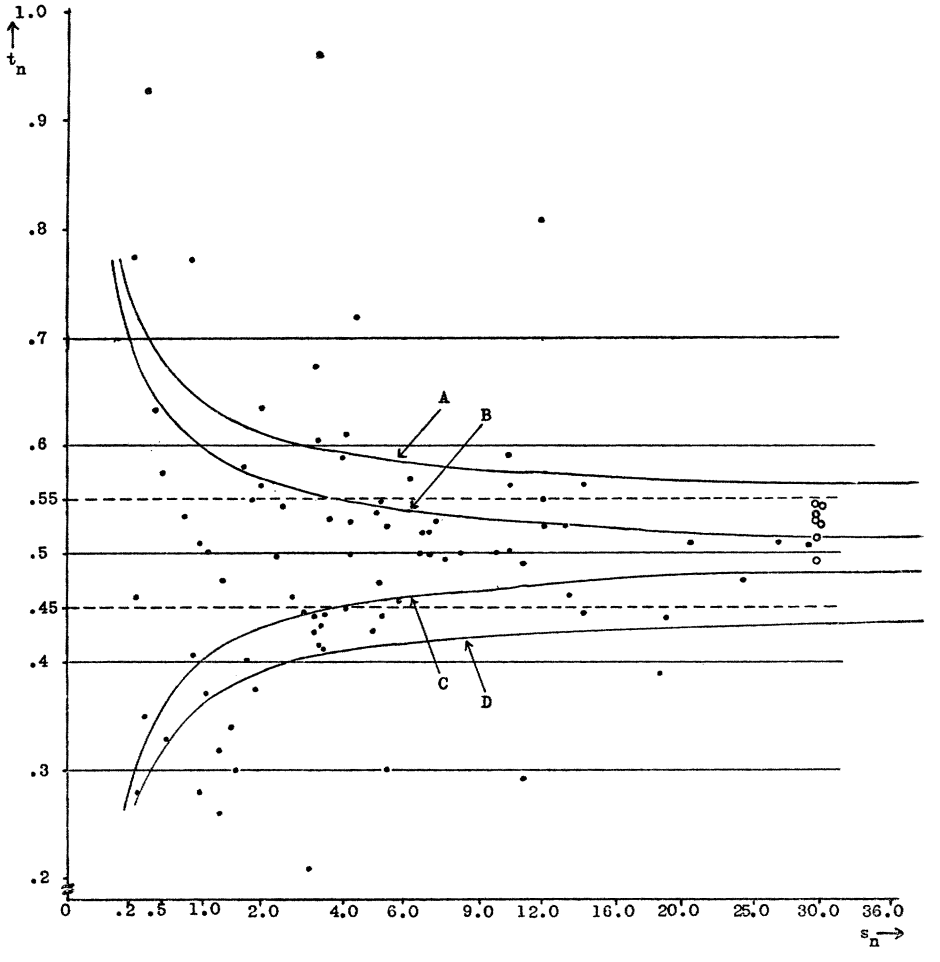

FIG. 5.1. Sample estimates (s_n, t_n) of the parameters (σ_n, τ_n) for 90 function words.
Curves B and C show 2 standard error bands for t_n, if $\tau_n = .5$
Curve A shows 2 standard error band above $\tau_n = .55$
Curve D shows 2 standard error band below $\tau_n = .45$

Madison for a large pool of words—not just the 90 unselected ones. We have further assumed that given β_1, β_2, the prior distributions of the τ for each word in the pool, given its σ, are adequately represented by independent, symmetric beta distributions, both of whose arguments are $\gamma = \beta_1 + \beta_2\sigma$.

C. Effects of varying the prior

Logically the β's are parameters of the distribution of the differential rate parameter τ, but this terminology is intolerable, so we call them "underlying constants." To try next to introduce a prior distribution for β_1 and β_2 is to invite an infinite regression. Instead we repeatedly carry out the analysis assuming known β_1, β_2, for several sets shown in Table 5.1. Naturally after the initial evaluations further refinements can be introduced. The essential feature

is the many analyses, with their fluctuating log odds. Nevertheless, the variation may not be enough to change the final assessments of authorship.

D. The posterior distribution of σ, τ

For any pair of underlying constants β_1, β_2, the posterior density of (σ, τ) given the data x_H, x_M on the papers of known authorship, is, by Bayes' theorem:

$$p(\sigma, \tau \mid x_H, x_M) = C(X)p(\sigma, \tau)p(x_H, x_M) \mid \sigma, \tau).$$

Return to the *also* example. We now use the actual observed rates instead of the simplified ones presented earlier. For the Poisson, the likelihood of observing 26 counts in 94,000 words of Hamilton text, and 80 in 114,000 of Madison text, with rates $\mu_H = \sigma\tau$, $\mu_M = \sigma(1-\tau)$, has logarithm:

$$\begin{aligned}\log p(x_H, x_M) \mid \sigma, \tau) = &- 94\sigma\tau + 26 \log [94\sigma\tau] - \log 26! - 114\sigma(1 - \tau) \\ &+ 80 \log [114\sigma(1 - \tau)] - \log 80!.\end{aligned}$$

The prior density with $\beta_1 = 10$, $\beta_2 = 0$ (our preferred choice) has logarithm:

$$\log p(\sigma, \tau) = \text{const.} + (10 - 1) \log [\tau(1 - \tau)],$$

where the constant includes log B(10, 10) and the constant prior assigned to σ. Then by Bayes' theorem the posterior density of (σ, τ) for *also* has logarithm

$$\begin{aligned}\log p(\sigma, \tau \mid x_H, x_M) = \text{const.} &- \frac{94 + 114}{2} + (80 + 26)\log \sigma \\ &+ (114 - 94)\sigma(\tau - 1/2) + (26 + 10 - 1) \log \tau \\ &+ (80 - 10 - 1) \log (1 - \tau).\end{aligned}$$

Approximate methods tell us that the mode of the posterior is near $\hat{\sigma} = .99$, $\hat{\tau} = .316$. The effect of this prior is to give an estimate $\hat{\tau} = .316$, somewhat nearer to 1/2 than the estimate $t = .282$ based on the observed rate.

Solve for $\hat{\mu}_H = .31$, $\hat{\mu}_M = .67$. At last we are ready to estimate odds for this set of underlying constants. If a paper of 2000 words has 4 *also*'s we multiply both rates by 2.000, and use the Poisson tables to get $f_P(4 \mid .62)/f_P(4 \mid 1.34)$, just as described at the start of section 4.

E. Negative binomial

The entire study was carried out in parallel for the negative binomial and Poisson data distributions. The negative binomial introduces many complications that strongly influenced our allocation of effort, but few new ideas. This brief statement of the treatment of prior distributions is given for completeness.

For each word, four parameters are needed: the mean rate μ and non-Poissonness $\delta = \mu/\kappa$ for each author. The mean rates were handled exactly as for the Poisson. Study of moments estimates, in the spirit of Figure 5.1, suggested that non-Poissonness δ was nearly independent of the rate μ. A tail reducing transformation from δ to $\zeta = \log(1+\delta)$ was made for each author, and then transformed to $\xi = \zeta_H + \zeta_M$, $\eta = \zeta_H/\xi$ analogously to σ and τ.

We introduced 5 underlying constants β_1, β_2, β_3, β_4, β_5, and assumed that, given the β's,

TABLE 5.2. FINAL CHOICES OF SETS OF UNDERLYING CONSTANTS

Set	β_1	β_2	β_3	β_4	β_5
22	10	0	12	1.25	2.0
31	10	0	12	.83	1.2
33	15	0	12	.83	1.2
38	5	5	6	.83	1.2
21	5	1	6	1.25	2.0
11	5	1	1.5	1.25	2.0

$(\sigma, \tau, \xi, \eta)$ are independent across words;
(σ, τ), ξ, η are independent of each other for each word;
σ has a distribution that can be adequately approximated by a constant density;
conditional on σ, τ has the symmetric beta density

$$[\tau(1-\tau)]^{\beta_1-\beta_2\sigma-1}/\mathrm{B}(\beta_1+\beta_2\sigma,\beta_1+\beta_2\sigma);$$

η has the symmetric beta density $[\eta(1-\eta)]^{\beta_3-1}/\mathrm{B}(\beta_3, \beta_3)$;
ξ has the gamma density with mean β_4, argument β_5:

$$(\beta_5/\beta_4)^{\beta_5}\xi^{\beta_5-1}e^{-\beta_5\xi/\beta_4}/\Gamma(\beta_5).$$

Each quintuple of β's is called a set of underlying constants and is assigned a code number. Table 5.2 shows the values of the underlying constants for six sets used in the displayed log odds in section 6.

For each set of underlying constants, the mode of the posterior distribution of $(\sigma, \tau, \xi, \eta)$ was determined for each word. The corresponding values of the parameters μ_1, δ_1 and μ_2, δ_2 (subscripts 1 and 2 refer to Hamilton and Madison, respectively) were used in the calculation of likelihood ratios for the unknown papers as if they were the known parameters. Table 5.3 shows the parameters estimated in this way from set 31 of underlying constants for each of the 30 words finally chosen for the study.

F. Final choices of underlying constants

For both the Poisson and negative binomial families, a variety of β's were used, each giving a different prior distribution. From the pool of 90 unselected words, ranges of β's were estimated. The six choices shown in Table 5.2 are spread over the estimated range. For the Poisson, β_3, β_4, β_5 are irrelevant.

The effect of β_1 and β_2 on the estimation of the differential-use parameter τ is essentially to add to the occurrences of the word in both the Hamilton and in the Madison texts, $\beta_1+\beta_2\hat{\sigma}$ extra occurrences, where $\hat{\sigma}$ estimates the combined rate σ (the extra occurrences are not used to estimate σ). The net discriminating power of the word decreases as $\beta_1+\beta_2\hat{\sigma}$ increases; for low-frequency words, β_1 is important; for high-frequency words, β_2 is important.

The final 3 β's concern the non-Poissonness parameters δ_1, δ_2 of the negative binomial.

G. Choice and grouping of words for use on the unknown papers

The 165 words that entered the main study were reduced to 30 words classed

into 5 groups before the disputed papers were analyzed. In the first step, 120 words were eliminated, because their discriminating power was low, because computation of posterior mode of parameters had failed, or because the word was regarded as very likely contextual corresponding to Group 6 in the weight-rate study of section 3.

Log odds for the remaining 45 words were evaluated for the known papers and examined for troubles. A major trouble was a systematic deviation between the Madison word rates in *The Federalist* and those in the long *Neutral Trade* paper. By methods that we do not present here, we retained only those words that were homogeneous across the Madison papers. In addition, we eliminated all words in the group containing personal pronouns and auxiliary verbs as potentially contextual. The net was 30 words in 5 groups, called B3A, B3B, B3G, B3E, B3Z, and displayed in Table 5.3, with the estimated

TABLE 5.3. FINAL WORDS AND WORD GROUPS: ESTIMATED NEGATIVE BINOMIAL PARAMETERS BASED ON UNDERLYING CONSTANTS SET 31

Code No.	Word	μ_1	μ_2	σ	τ	δ_1	δ_2
B3A							
60	upon	3.24	.23	3.47	.932	.25	.39
B3B							
3	also	.32	.67	.99	.327	.09	.10
4	an	5.95	4.58	10.53	.565	.02	.02
13	by	7.32	11.43	18.75	.390	.35	.40
39	of	64.51	57.89	122.40	.527	.24	.25
40	on	3.38	7.75	11.12	.304	.34	.42
55	there	3.20	1.33	4.53	.706	.23	.24
57	this	7.77	6.00	13.77	.564	.21	.21
58	to	40.79	35.21	76.00	.537	.39	.45
B3G							
73	although	.06	.17	.23	.267	.11	.11
78	both	.52	1.04	1.56	.334	.12	.14
90	enough	.25	.10	.35	.727	.47	.52
116	while	.21	.07	.28	.744	.23	.25
117	whilst	.08	.42	.50	.153	.15	.13
123	always	.58	.20	.78	.742	.07	.07
160	though	.91	.51	1.42	.639	.08	.08
B3E							
80	commonly	.17	.05	.23	.763	.05	.05
81	consequently	.10	.42	.52	.189	.16	.14
82	considerable(ly)	.37	.17	.54	.684	.07	.08
119	according	.17	.54	.71	.238	.30	.30
124	apt	.27	.08	.35	.770	.06	.07
B3Z							
87	direction	.17	.08	.25	.693	.31	.32
94	innovation(s)	.06	.15	.20	.278	.06	.06
96	language	.08	.18	.26	.316	.05	.05
110	vigor(ous)	.18	.08	.26	.680	.02	.02
143	kind	.69	.17	.86	.799	.25	.22
146	matter(s)	.36	.09	.45	.790	.05	.05
151	particularly	.15	.37	.51	.282	.14	.16
153	probability	.27	.09	.36	.757	.02	.02
165	work(s)	.13	.27	.40	.326	.46	.42

parameters for the negative binomial analysis for set 31 of underlying constants.

Groups B3A, B3B, B3G contain function words from the Miller-Newman-Friedman list (see section 2), *upon* in A, high-frequency words (by their count) in B, and low-frequency words in G. Groups B3E and B3Z contain words obtained from the index, as described in section 2; the "well-liked" words in E, the rest in Z. The low-frequency words among the final 30 are sometimes called "markers"—Hamilton markers or Madison markers—because even one occurrence is indicative of, say, Hamilton.

6. THE RESULTS OF THE MAIN STUDY

In this section we give our results, reserving adjustments for section 7. Our presentation is a sprinkling of the log odds from from among the $113 \times 30^2 \approx 10^5$ numbers for 113 papers, 30 words, and 30 sets of underlying constants.

A. Checking the method

We present the total log odds for 11 papers by each author in Table 6.1. For each author, we give his first 8 papers from *The Federalist*, and, in addition, for Hamilton, Pacificus I, II, and III (code number 111, 112, 113), for Madison, Helvidius II, III, and IV (code number 132, 133, 134). Papers exterior to *The Federalist* are displayed to give a notion of what happens when the method

TABLE 6.1. TOTAL NATURAL LOG ODDS FOR PAPERS OF KNOWN AUTHORSHIP, 11 BY HAMILTON AND 11 BY MADISON; TOTAL FOR 30 FINAL WORDS; 6 SETS OF UNDERLYING CONSTANTS, 2 DISTRIBUTIONS

Paper number	Paper length (thousands)	Hamilton									
		Negative binomial						Poisson**			
		Set of underlying constants						Set of underlying constants			
		22	33	38	21	11	31	22, 31	33	38	11, 21
1	1.6	13.9	11.9	13.5	15.7	13.7	14.8	22.9	20.5	22.4	25.7
6	1.9	16.8	15.7	16.4	17.7	17.4	17.5	27.4	25.5	26.1	29.2
7	2.2	16.6	14.3	13.8	17.5	14.4	18.0	36.6	33.7	33.7	39.0
8	2.0	14.0	13.0	16.5	16.5	16.3	15.1	20.7	18.3	21.4	23.8
9	1.6	11.6	10.2	11.2	12.6	12.5	11.8	16.7	15.3	15.8	18.0
11	2.5	16.3	15.4	15.6	17.2	16.6	17.4	30.7	28.7	28.4	32.2
12	2.1	13.0	11.3	12.0	14.1	13.4	13.8	25.2	23.0	23.7	27.3
13	1.0	7.8	7.3	7.4	8.0	7.7	8.0	12.0	11.4	11.1	12.4
111	2.9	11.9	12.0	9.9	10.9	9.7	12.6	27.0	26.6	23.4	26.5
112	2.5	9.3	7.5	7.6	10.0	8.8	10.2	23.4	21.1	21.2	25.3
113*	1.2	3.0	2.2	4.6	4.0	4.7	2.9	2.8	2.0	3.2	3.8
		Madison									
10	3.0	−17.5	−17.2	−16.6	−18.2	−19.8	−18.3	−30.5	−29.5	−28.5	−31.0
14	2.1	−20.0	−18.5	−20.6	−22.6	−24.2	−20.6	−28.7	−26.5	−28.8	−31.5
37	2.7	−20.2	−18.9	−19.8	−23.4	−25.5	−21.2	−32.7	−30.4	−33.0	−35.8
38	3.3	−16.5	−15.3	−17.8	−19.6	−21.8	−17.5	−25.4	−22.9	−27.5	−29.2
39	2.6	−24.6	−23.6	−24.5	−26.6	−28.9	−25.8	−45.1	−42.5	−43.8	−47.7
40	2.7	−19.2	−18.5	−19.7	−20.9	−21.9	−20.3	−30.1	−28.6	−29.3	−31.7
41	3.5	−15.6	−15.1	−15.4	−17.5	−18.9	−16.5	−27.6	−26.5	−26.3	−28.7
42	2.7	−11.9	−11.1	−11.0	−13.4	−15.2	−12.4	−21.1	−20.0	−21.2	−22.6
132	2.7	−20.3	−19.3	−20.5	−22.9	−26.7	−21.7	−31.9	−29.3	−32.1	−35.1
133	2.5	−13.3	−11.7	−14.0	−15.6	−16.9	−13.6	−20.4	−18.7	−20.6	−22.7
134*	1.7	−0.8	−0.1	−1.7	−1.9	−2.4	−0.8	0.9	1.4	0.3	0.1

* The paper by each author that is most poorly identified is marked by a star.

** For the Poisson, set 31 is the same as set 22, set 11 is the same as set 21.

is applied to a larger variety of writings, and these particular papers are chosen because they contain the one paper most poorly identified by the log odds for each author. Looking first at the negative binomial distribution for set 22 of underlying constants, we see that every Hamilton paper has positive and every Madison negative log odds. Since these papers contain the worst outcomes, the entire 48 Hamilton and 50 Madison papers are assigned log odds in the proper direction by this distribution and set of underlying constants. Paper No. 134 for four Poisson sets of underlying constants has log odds pointing mildly in the wrong direction. As a whole, though, the log odds for all sets point consistently and forcibly in the right direction. Just how strong these odds are can be better appreciated by consulting the following brief table of antilogs.

Log odds	Odds	Log odds	Odds	Log odds	Odds	Log odds	Odds
0	1 to 1	1	$e \approx 2.7$ to 1	4	55 to 1	15	3.3×10^6 to 1
.1	1.1 to 1	2	7 to 1	5	150 to 1	20	480×10^6 to 1
.5	1.6 to 1	3	20 to 1	10	22,000 to 1	25	7.1×10^9 to 1

Since we know who wrote each of these papers, the log odds in Table 6.1 offer a check on the method. Essentially, each paper has been treated as if it were a disputed paper, and the log odds computed. And the evidence is that the method works well.

B. Effects of prior distributions and of data distributions

Beyond this, we can examine the effect of varying the underlying constants. Visual inspection will assure the reader that the variation in log odds from one set to another is modest compared to the variation from one paper to another. But the change in log odds from the negative binomial distribution to the Poisson is huge. For the moment the reader may wish to be conservative and attend to the negative binomial odds—if median odds of 3 million to 1 can ever be called conservative. Wasn't it a Damon Runyon character who announced that nothing in life is better than 3 to 1?

C. Who wrote the disputed papers?

Next, the *pièce de resistance*, Table 6.2, presents total log odds for the joint and disputed papers. Attending to the 12 disputed papers, we see that every set of underlying constants gives odds for all papers strongly in favor of Madison. The weakest of these are papers 55 and 56, and the lowest odds for No. 55 are 240 to 1 ($e^{5.5}$) in favor of Madison, not absolutely overwhelming, in the language of section 4. Essentially, No. 55 does not have its share of marker words, no matter who wrote the paper, and the high-frequency words produced no information.

Among the joint papers, No. 20 looks especially mixed, but the small log odds are confounded with shortness of paper. We have started an investigation of this writing, for the matter is complicated because Madison's notes suggest that he borrowed much of it from Sir William Temple or Felice. Is it that Hamilton's or Temple's words make the log odds low?

TABLE 6.2. TOTAL NATURAL LOG ODDS FOR THE PAPERS OF JOINT AND DISPUTED AUTHORSHIP. TOTAL FOR THE 30 FINAL WORDS. 6 SETS OF UNDERLYING CONSTANTS, 2 DISTRIBUTIONS

Paper number	Paper length (thousands)	Negative binomial						Poisson**			
		Set of underlying constants						Set of underlying constants			
		22	33	38	21	11	31	22, 31	33	38	11, 21
Joint											
18	2.1	−11.0	−10.8	−9.0	−11.4	−11.2	−11.4	−20.1	−19.5	−18.9	−20.5
19	2.0	−12.1	−12.0	−10.8	−12.2	−12.9	−12.5	−18.6	−18.4	−16.7	−18.3
20	1.4	−4.6	−5.0	−1.9	−3.6	−3.3	−4.6	−7.0	−7.6	−5.8	−6.0
Disputed											
49	1.6	−13.2	−12.2	−12.9	−14.6	−15.8	−13.4	−18.1	−17.1	−17.6	−19.3
50	1.1	−14.3	−13.7	−13.7	−15.1	−15.9	−14.5	−18.2	−17.5	−17.4	−18.9
51	1.9	−21.9	−20.9	−22.1	−24.0	−25.4	−23.0	−33.4	−31.3	−32.7	−35.9
52	1.8	−16.0	−15.7	−15.0	−16.5	−17.1	−16.6	−23.1	−22.5	−21.6	−23.4
53	2.2	−15.8	−15.0	−16.2	−17.4	−18.5	−16.4	−22.0	−20.7	−21.7	−23.6
54	2.0	−14.3	−13.6	−13.2	−15.7	−16.1	−14.8	−22.9	−21.7	−22.7	−24.3
55	2.0	−5.8	−5.5	−5.9	−6.2	−6.4	−6.1	−7.1	−6.6	−6.9	−7.6
56	1.6	−8.7	−8.2	−8.8	−9.6	−9.9	−9.0	−10.6	−10.0	−10.4	−11.4
57	2.2	−16.7	−15.7	−17.2	−18.4	−20.8	−17.6	−26.1	−24.2	−25.9	−28.6
58	2.1	−18.0	−17.1	−17.6	−19.4	−21.5	−18.5	−26.3	−25.1	−25.2	−27.4
62	2.4	−16.5	−16.0	−16.0	−17.3	−17.5	−17.3	−26.9	−25.6	−25.6	−28.0
63	3.0	−18.5	−17.7	−17.7	−19.6	−21.1	−19.1	−32.2	−31.2	−30.2	−32.9

** For the Poisson, set 22 is the same as set 31, set 11 is the same as set 21.

D. The behavior of word groups

To show how consistently the different word groups behave and to what extent each contributes to the total, we present the log odds by word groups for set 22 in Table 6.3. All groups look quite consistent, considering their differing strengths—a weak set must have negative log odds occasionally. This general consistency is a further sign of good discrimination.

The set B3B is stronger than B3A (*upon*) which in turn looks nearly as strong as the other three groups put together. Recall that B3B contains the high-frequency function words: *to, this, there, on, of, by, an, also.* So in the end the high-frequency words outshone all marker words. While this does not prove the cleverness in selecting variables fails to pay, it does show that routine can pay.

E. Illustrative single words

The few words selected for Table 6.4 are chosen to exemplify log odds for a high-frequency word, *of*, for low-frequency Madison markers *whilst* and *innovation*, and for low-frequency Hamilton markers, *enough* and *probability*. But *according* was chosen to illustrate a special point, the way the negative binomial stamps down the odds compared to the Poisson when a low-frequency marker word has an unusually high rate. See, in Madison's paper 39, log odds of −9.52 changed to −2.48! It comes to this, the negative binomial provides automatic damping for low-frequency words and thus prevents words from getting badly out of hand.

TABLE 6.3. LOG ODDS BY WORD GROUP FOR SET 22 OF UNDERLYING CONSTANTS

Paper number	Negative binomial					Poisson				
	Word group									
	B3A	B3B	B3G	B3E	B3Z	B3A	B3B	B3G	B3E	B3Z
	Hamilton									
1	4.7	2.0	1.9	3.1	2.3	11.6	2.8	2.2	4.0	2.3
6	2.5	9.2	1.7	.3	3.2	5.2	14.3	2.1	.7	5.2
7	6.4	4.6	−.2	2.6	3.3	23.3	6.9	−.3	3.0	3.8
8	1.2	3.0	4.9	2.2	2.8	2.1	5.1	7.4	2.4	3.6
9	2.9	3.5	2.6	1.0	1.6	6.0	5.4	3.0	1.2	1.2
	Madison									
10	−6.5	−6.8	−.9	−2.1	−1.2	−9.1	−14.1	−1.9	−3.5	−1.8
14	−5.0	−7.9	−1.4	−1.9	−3.8	−6.6	−12.3	−1.5	−2.5	−5.8
37	−3.2	−9.2	−3.1	−1.3	−3.4	−5.5	−16.1	−3.4	−1.2	−6.5
38	.3	−4.6	−6.4	−3.0	−2.9	.8	−8.4	−10.1	−3.4	−4.4
39	−5.8	−11.7	−.8	−2.7	−3.6	−7.9	−21.1	−1.4	−10.0	−4.7
	Joint									
18	−2.1	−8.1	1.3	−1.0	−1.1	−3.6	−14.4	1.9	−2.7	−1.3
19	−4.8	−7.6	−.9	1.4	−.2	−6.2	−13.0	−.9	1.7	−.2
20	−.9	−7.6	.8	1.0	2.0	−1.6	−9.9	.7	1.1	2.6
	Disputed									
49	−4.0	−5.5	−.8	−1.3	−1.6	−4.9	−9.4	−.9	−1.1	−1.9
50	−2.9	−9.0	−1.1	.2	−1.5	−3.4	−12.2	−1.2	.3	−1.8
51	−4.6	−9.3	−3.8	−1.9	−2.4	−5.8	−16.4	−5.4	−2.6	−3.3
52	−4.4	−10.2	.2	.2	−1.8	−5.6	−15.9	.1	.4	−2.2
53	−5.1	−6.4	−4.6	1.4	−1.2	−6.6	−10.1	−5.4	1.7	−1.7
54	−.2	−8.6	−1.3	−2.3	−1.9	−.6	−15.2	−1.7	−3.1	−2.3
55	−4.8	−.1	.8	−.7	−1.0	−6.2	1.1	.6	−1.4	−1.3
56	−3.9	−2.4	−3.1	1.0	−.4	−4.8	−3.1	−3.5	1.2	−.4
57	−5.1	−5.9	−2.6	−.9	−2.1	−6.7	−10.9	−5.4	−.8	−2.4
58	−4.9	−8.6	−1.3	−1.3	−2.0	−6.4	−15.1	−1.5	−1.0	−2.4
62	−5.5	−8.1	−.2	−1.5	−1.2	−7.3	−14.1	−.9	−3.2	−1.5
63	−6.6	−8.4	−1.5	.2	−2.3	−9.2	−19.1	−1.6	.6	−2.9

F. Summing up

In summary, the following points are clear:

1) Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers. Weakest support is given for No. 55. Support for Nos. 62 and 63, most in doubt by current historians, is tremendous.

2) Prior distributions are not of major importance. While choice of underlying constants (choice of prior distributions) matters, it doesn't matter very much, once one is in the neighborhood of a distribution suggested by a fair body of data. We conclude from this that the emphasis on the difficulty, even impossibility, of choosing prior distributions as a criticism of the use of Bayes' theorem is not well placed. Here again is the old story in statistics that modest variation in the weights used does not much change the inference. And, of

TABLE 6.4. LOG ODDS FOR SINGLE WORDS FOR 5 EACH OF HAMILTON, MADISON, AND DISPUTED PAPERS. SET 31. STAR MEANS WORD DID NOT OCCUR

	Negative binomial						Poisson			
	whilst	enough	of	innovation	probability	according	whilst	of	probability	according
Hamilton										
1	.49*	.83	.56	.13*	.85	.48*	.56*	.78	.85	.65*
6	.58*	.80	1.79	.16*	.79	−.60	.67*	2.82	.80	−.53
7	.68*	−.24*	1.80	.19*	1.74	.64*	.80*	2.98	1.88	.94*
8	.61*	1.01	.77	.16*	1.78	.58*	.71*	1.17	1.93	.83*
9	.51*	.82	2.02	.14*	−.29*	.49*	.58*	3.06	−.29*	.68*
Madison										
10	.87*	−.30*	−1.79	.24*	.60	−1.34	1.06*	−3.09	.59	−2.72
14	−1.08	−.23*	−.80	−.77	−.38*	.61*	−1.05	−1.27	−.39*	.89*
37	−.94	−.28*	−.74	−.73	−.48*	−.42	−.85	−1.30	−.49*	−.19
38	−1.72	−.32*	−.99	.27*	−.58*	−.88	−2.45	−1.86	−.60*	−1.26
39	.77*	−.27*	−.17	−.74	−.46*	−2.48	.92	−.34	−.47*	−9.52
Disputed										
49	−1.23	−.19*	−.25	.14*	−.28*	.48*	−1.24	−.39	−.29*	.66*
50	.35*	−.14*	−.80	.10*	−.20*	.35*	.39*	−1.05	−.20*	.46*
51	−2.06	−.21*	−.41	.17*	−.37*	.56*	−2.95	−.65	−.35*	.80*
52	.57*	−.21*	−1.02	.17*	−.33*	.54*	.65	−1.52	−.33*	.77*
53	−1.08	−.23*	−.41	−1.78	−.38*	.62*	−1.04	−.68	−.39*	.90*

course, the prior distributions for the parameters are a method of weighting. On the other hand, the choice of the prior can matter. It is the agreement with data that puts us in a particular region of the space of prior distributions.

Some people ask "Who is to be the official guesser of priors?" and "Won't different people get different results?." The answers are "No one," and "Yes." The position is similar to that of the choice of the data distribution in any sort of statistical analysis. There is no official chooser of data distributions, and people who choose different ones do get different answers. Furthermore, no wise chooser ever feels he has chosen just exactly the correct data distribution. Naturally, the chooser uses a variety of more or less objective methods to help him choose, as we have done in studying word distributions. Our study indicates that in some problems at least and in ours in particular there are ways of getting priors that are not blind guesses, but that have a foundation in data, a foundation whose firmness is like that of the choice of data distributions.

Some personalists may feel we go too far in pressing the correspondence between choosing prior distributions and the fitting of distributions by considerations of real data. In the end, they feel that it is degree of belief that is being adjusted rather than objective frequencies, and that our emphasis is misplaced. We feel that where real data is available, it would be used, and, where it is not, expertise in the subject matter area would be important, and agreement hard to find. In such cases the observed data may need to be very substantial if it is to swamp the variation in priors owing to ignorance.

3) Data distributions are important and need investigation. Looking again at Table 6.1 for Paper No. 111 for instance, we find that the log odds more than doubled when we went from the negative binomial to the Poisson. A 100

per cent change in log odds, or a factor of 22,000 in the odds, is not to be ignored. While we have made modest progress on the question of the assessment of data distributions, we have not marred the surface of the problem. The fact is clear, in Bayesian studies the data distribution can be enormously important, and in this study it clearly is. If Bayesian methods are to be widely used we need some new ways of choosing and assessing data distributions. And we presume that the investigation will require a new empirical and theoretical effort comparable in magnitude to that extended on the theory of distributions until 1935.

4) Routine pays off. We were surprised that in the end, it was the utterly mundane high-frequency function words that did the best job. Though we love them for their lack of contextuality, their final strength was as unexpected as it was welcome. The result is reminiscent of a hard fact perennially being discovered in the social sciences. For the forecasting of a great many facets of a man's behavior, it is hard to beat the usual tired old socio-economic variables and the standard personal background data.

5) The main study is robust against changes in type of essay. The method performed satisfactorily on both *Federalist* material and on essays exterior to *The Federalist*. Such robustness is encouraging in considering the method for other studies and in assessing the generality of the present inference. The disputed papers could be as contextually disparate as the Pacificus or Helvidius papers for all we know.

6) Two of the joint papers are mainly the work of Madison, the third presents a puzzle. The joint papers Nos. 18 and 19 seem mainly to be the work of Madison. No. 20 presents an interesting new problem—can we sort out the contribution of a possible, unwitting, third party, Sir William Temple, so as to assess properly Hamilton's share?

7. ADJUSTMENTS TO THE LOG ODDS

The log odds presented depend on many assumptions and approximations. We have studied the effects of several important ones and briefly describe the results, but not the theory or method.

A. Effect of modal estimation

The effect of using modal estimates is to exaggerate the odds. A very rough rule is: log odds should be deflated by about 15 per cent.

B. Effect of correlation

Treating words as independent may either inflate or deflate the estimated odds. Our investigation suggested that the observed odds should be deflated by about 9 per cent.

C. Effect of regression

A study of regression compares the mean log odds in the papers of known authorship and in the disputed papers with the expected log odds if the estimated parameters are true. For both distributions, the expected log odds show that the prior distributions forecast that in fresh papers the mean log

odds will be about 3 less than the log odds already observed for the papers of known authorship. The disputed papers are fresh and rather clearly by Madison, and they show less regression than forecast for Madison papers. Indeed, for the negative binomial they are about 1.6 standard deviations above the forecast, and for the Poisson about .6 standard deviations above.

D. Effect of changes in word counts

The recent Cooke edition [4] of *The Federalist* calls attention to changes between the original newspaper edition and various book editions. This valuable list came too late to affect our calculations, but only two changes are at all important and these affect two important marker words in two disputed papers.

Of the 2 *whilst*'s in paper No. 51, one was added by Hamilton in preparing the McLean edition on which our work is largely based. Removing it changes the log odds in set 22 by one unit from -21.9 to -21.2.

In paper No. 58, an *upon* in the newspaper edition was deleted for the McLean edition, increasing the final log odds by 2.8 from -18.0 to -15.2. No other changes of single words could have caused so great a shift as these for *upon* and *whilst*, yet the odds are still satisfyingly fat.

E. Are the odds to be believed?

Even after the deflations discussed above, the odds cannot be taken entirely at face value. The possibility of some sorts of what we call outrageous events that we regard as outside the framework of the model must be allowed for. Everyone worries whether Hamilton wrote and Madison heavily revised the disputed papers. We think this unlikely not only for historical reasons, but because the task of making the set of rates and log odds match well with Madison's criterion set implies enormous revisions from Hamilton's original copy. We take this up at greater length elsewhere [8]. The possibility of blunders or errors in calculation is always present. Such possibilities reduce extreme odds very considerably, but would, if the reader thought highly of the work, reduce modest odds relatively little. Thus in the framework of the model of papers independently produced by one or the other author subject to much the same writing habits as usual, odds of millions to one are understandable. But when outrageous events are given their rightful role, thousands to one may be still a bit extreme.

To the statistician who asks how we can get odds of even thousands to one when we have only about 100 papers altogether, the answer is that the distribution theory, which is rather well founded in data, supplies some of this strength, but the major part comes from the independent fairly modest contributions from many variables.

8. CONCLUSIONS

Here we present the conclusions of our study without argument. We may go too far in these conclusions, considering that we have only one large data analysis to base them on, but they should serve to stimulate thought, even if they later require revision.

Results on *The Federalist* as a case study.

(1) Our data independently supplement that of the historian. On the basis of our data alone, Madison is extremely likely, in the sense of degree of belief, to have written the disputed *Federalists*, with the possible exception of No. 55, and there our evidence is weak, suitable deflated odds are 80 to 1 for Madison. No. 56, next weakest, is a strong 800 to 1 for Madison. The data are strong for all the rest, including the two papers historians feel weakest about, Nos. 62 and 63.

(2) Among the joint papers, Nos. 18 and 19 look as if Madison wrote the lion's share. No. 20 requires a more subtle analysis of its possible contamination by third parties before Hamilton's share can be assessed.

(3) Choice of prior distribution mattered little compared to other sources of variation.

(4) Changes in the data distribution had enormous effects on the output, but both Poisson and negative binomial performed better in the disputed papers than the theory of these models forecasts.

(5) The main study shows stable discrimination for essays on various subjects, the writing spread over a quarter of a century.

(6) The classical weight-rate study gives results similar to the main study, except that the results appear weaker. No. 55 slightly favors Hamilton, and No. 56 slightly favors Madison. Sampling variation could reverse either one.

Remarks on authorship problems

(1) The function words of the language appear to be a fertile source of discriminators, and luckily the high-frequency words are the strongest. Use of our rate table (not presented here) for high-frequency words may help an investigator quickly form a pattern of rates for a new author.

(2) Contextuality is a source of risk. For this reason it is important to have a variety of sources of material, to allow "between writings" variability to emerge and to give a basis for the elimination of words that show substantial heterogeneity.

(3) Pronouns and auxiliary verbs appear to be dangerously contextual; other function words are not entirely safe, and should not be taken for granted, but should be investigated for contextuality.

(4) Paired authorship problems should not ordinarily be as difficult as this one, but there is plenty of room for original ideas on the problem of selecting one from among many authors, or of factoring a collection of essays into component groups of similar authors.

Remarks on discrimination problems generally.

(1) Where important and germane variables are available they should, of course, be tried.

(2) If obvious methods fail, the systematic exploration of a very large pool of variables may strike oil. Routine drilling paid off for us more than clever thoughts about words and other variables that ought to discriminate.

(3) In the presence of large numbers of variables, preparing for selectivity is necessary—in classical studies through calibrating or validating sets of data

whose origin is known but that are uncontaminated by the selection and weighting processes; in Bayesian studies through realistic priors that can be based on pools of variables.

(4) In other studies, our notion of contextuality has its counterparts in the usual sources of variation arising in the analysis of variance such as locations, regions, times, and laboratories. The discriminator needs to be prepared to handle such variation by the elimination or proper weighting (usually nonlinear) of evidence subject to such variation—and what evidence is not? Including sufficiently varied background data lessens the risk of being misled by unsuspected contextuality.

(5) Do use intuition to group variables, to help with selection, and to escape contextuality.

(6) If flexible high speed computing is available, one method of validating is to treat each piece of known material as an unknown and to obtain the variables and weights from all the rest exclusive of this known piece. Then apply the results to the known piece to get a validating observation. Then repeat for other knowns. This is expensive because it means repeating the entire study for each piece used as a validator, but as computation becomes inexpensive, this approach becomes less unreasonable.

Remarks on Bayesian studies

(1) Study of variation of results with different priors is recommended. Bracketing the prior is often easy. When founded in data, the choice of the prior has a status like that of the data distribution—subjectivity tempered with empiricism.

(2) Where possible, priors should be empirically oriented. Planning ahead for the collection of data suitable for estimating the form and underlying constants of prior distributions is useful and important, and very likely not hard to do once one gets in the frame of mind of preparing to use Bayes' theorem. The remark is all the more germane in the field of repetitive studies, where there has never been much excuse for the failure to collect such data.

(3) In any method of inference, data distributions matter enormously, and their study and choice requires the development of new and systematic methods.

(4) Statisticians need to provide richer sets of priors and data distributions than we now have, with a view to their manageability as well as their utility in studies of data. For example, mixture distributions can give satisfactory tail properties, but are grisly mathematically.

(5) The moment one leaves the simplest application of Bayes' theorem, the applier finds himself involved in a welter of makeshifts and approximations. This trouble and its cures require systematic developments. The prospect of useful exact treatments can be neglected.

(6) Simple Bayesian methods are needed that can be applied without appeal to high speed computers.

(7) Users of Bayes' theorem will find that many statistical jobs are readily handled by standard devices, but not by available Bayesian techniques. To avoid such devices for the sake of consistency reduces the quality of the re-

search. On the other hand, time spent developing Bayesian procedures for a standard problem can be a profitable pursuit.

(8) In summary, for large scale data analysis, Bayesian methods require new studies of theoretical and empirical distributions and their approximation and estimation comparable in extent to those provided by statisticians up to 1935.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge with thanks the many who have helped us. These include the following.

Mr. Miles Davis has handled the programming of the high-speed calculations. Wayne Wiitanen, C. Harvey Willson, and Robert A. Hoodes, under the direction of Albert E. Beaton programmed the word counts. Roger Carlson, Robert M. Elashoff, Ivor Francis, Robert Kleyle, Charles Odoroff, P. S. R. S. Rao, and Marie Yeager have assisted with many parts of this work. Mrs. Cleo Youtz has also supervised a number of the studies. Mrs. Virginia Mosteller cooperated in the screening study.

We acknowledge with thanks the many helpful discussions and suggestions received from colleagues: Frank J. Anscombe, Carl Christ, William G. Cochran, Arthur P. Dempster, Martin Diamond, John Gilbert, Morris Halle, William Kruskal, David Levin, P. J. McCarthy, Ann Mitchell, John W. Pratt, Howard Raiffa, L. J. Savage, Robert Schlaifer, Maurice M. Tatsuoka, and John W. Tukey.

We appreciate the careful calculations and other work of Linda Alger, Eleanor Beissel, Mary Blyman, John Burnham, Helen Canning, Adelle Crowne, Roy D'Andrade, Abraham Davidson, Sara Dustin, Jane Hallowell, Joanna Handlin, Elizabeth Ann Hecht, Ann Hilken, Helen V. Jensen, Kathryn Karrasik, Vincent Kruskal, Christine Lyman, Nancy McCarthy, William Mosteller, Loneta Newburgh, Eva Pahnke, Eleanor Rosenberg, Astrid Salvesen, Mary Skillman, Lucy Steig, Elizabeth Thorndike, Henry Tibery, Bruce B. Venrick, Druscilla Wendoloski, and Richard Wendoloski.

For work on the manuscripts leading to this paper we thank Mrs. Angela Klein, Miss Janet Mendell, Mrs. Cleo Youtz, and Miss Rosemarie Stemphel.

REFERENCES

[1] Adair, Douglass, "The authorship of the disputed Federalist papers. Part I.," *The William and Mary Quarterly*, Volume 1, No. 2 (1944), 97–122.

[2] Adair, Douglass, "The authorship of the disputed Federalist papers. Part II.," *The William and Mary Quarterly*, Volume 1, No. 3 (1944), 235–64.

[3] Brinegar, Claude, "Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship," *Journal of the American Statistical Association*, 58 (1962), 85-96.

[4] Cooke, Jacob E., Editor, *The Federalist*. Cleveland, Ohio: World Publishing Company, Meridian Books, 1961.

[5] *The Federalist Papers*. Introduction by Clinton Rossiter. New York: The New American Library, Mentor Books, 1961.

[6] Jeffries, Harold, *Theory of Probability*, Third Edition. London: Oxford at the Clarendon Press, 1961.

[7] Miller, G. A., Newman, E. B., and Friedman, E. A., "Length-frequency statistics for written English," *Information and Control*, Volume 1, No. 4 (1958), 370–89.
[8] Mosteller, F., and Wallace, D. L., *Methods of Inference Applied to The Federalist.* Reading, Massachusetts: Addison-Wesley Publishing Company, in press.
[9] Raiffa, Howard, and Schlaifer, Robert, *Applied Statistical Decision Theory.* Boston, Massachusetts: Harvard Business School, 1961.
[10] Rao, C. Radhakrishna, *Advanced Statistical Methods in Biometric Research.* New York: John Wiley and Sons, Inc., 1952.
[11] Savage, L. J., and other contributors, *The Foundations of Statistical Inference.* London: Methuen, 1962.
[12] Wright, Benjamin Fletcher, Editor, *The Federalist.* Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1961.