FREDERICK MOSTELLER
Harvard University
DAVID L. WALLACE
University of Chicago

Inference and Disputed Authorship: *The Federalist*

ADDISON-WESLEY PUBLISHING COMPANY, INC.
READING, MASSACHUSETTS · PALO ALTO · LONDON

MISCELLANY.

The FŒDERALIST. No. 62.

To the People of the State of New-York.

A FIFTH deſideratum illuſtrating the utility of a Senate, is the want of a due ſenſe of national character. Without a ſelect and ſtable member of the government, the eſteem of foreign powers will not only be forfeited by an unenlightened & variable policy, proceeding from the cauſes already mentioned; but the national councils will not poſſeſs that ſenſibility to the opinion of the world, which is perhaps not leſs neceſſary in order to merit, than it is to obtain, its reſpect and confidence.

An attention to the judgment of other nations is important to every government for two reaſons: The one is, that independently of the merits of any particular plan or meaſure, it is deſireable on various accounts, that it ſhould appear to other nations as the offspring of a wiſe and honorable policy: The ſecond is, that in doubtful caſes, particularly where the national councils may be warped by ſome ſtrong paſſion, or momentary intereſt, the preſumed or known opinion of the impartial world, may be the beſt guide that can be followed. What has not America loſt by her want of character with foreign nations? And how many errors and follies would ſhe not have avoided, if the juſtice and propriety of her meaſures had in every inſtance been previouſly tried by the light in which they would probably appear to the unbiaſſed part of mankind.

Yet however requiſite a ſenſe of national character may be, it is evident that it can never be ſufficiently poſſeſſed by a numerous and changeable body. It can only be found in a number ſo ſmall, that a ſenſible degree of the praiſe and blame of public meaſures may be the portion of each individual; or in an aſſembly ſo durably inveſted with public truſt, that the pride and conſequence of its members may be ſenſibly incorporated with the reputation and proſperity of the community. The half-yearly repreſentatives of Rhode-Iſland, would probably have been little affected in their deliberations on the iniquitous meaſures of that State, by arguments drawn from the light in which ſuch meaſures would be viewed by foreign nations, or even by the ſiſter States: whilſt it can ſcarcely be doubted, that if the [illegible] and ſtable body had been which will exempt the people of America from ſome of the dangers incident to leſſer republics, will expoſe them to the inconveniency of remaining for a longer time, under the influence of thoſe miſrepreſentations, which the combined induſtry of intereſted men may ſucceed in diſtributing among them.

It adds no ſmall weight to all theſe conſiderations, to recollect, that hiſtory informs us of no long lived republic which had not a Senate. Sparta, Rome and Carthage are in fact the only States to whom that character can be applied. In each of the two firſt there was a Senate for life. The Conſtitution of the Senate in the laſt, is leſs known. Circumſtantial evidence makes it probable that it was not different in this particular from the two others. It is at leaſt certain that it had ſome quality or other which rendered it an anchor againſt popular fluctuations; and that a ſmaller council drawn out of the Senate was appointed not only for life; but filled up vacancies itſelf. Theſe examples, though as unfit for the imitation, as they are repugnant to the genius of America, are notwithſtanding, when compared with the fugitive and turbulent exiſtence of other antient republics, very inſtructive proofs of the neceſſity of ſome inſtitution that will blend ſtability with liberty. I am not unaware of the circumſtances which diſtinguiſh the American from other popular governments, as well antient as modern; and which render extreme circumſpection neceſſary in reaſoning from the one caſe to the other. But after allowing due weight to this conſideration, it may ſtill be maintained that there are many points of ſimilitude which render theſe examples not unworthy of our attention. Many of the defects as we have ſeen, which can only be ſupplied by a ſenatorial inſtitution, are common to a numerous aſſembly frequently elected by the people, and to the people themſelves. There are others peculiar to the former, which require the controul of ſuch an inſtitution. The people can never wilfully betray their own intereſts: But they may poſſibly be betrayed by the repreſentatives of the people; and the danger will be evidently greater where the whole legiſlative truſt is lodged in the hands of one body of men, than where the concurrence of ſeparate and diſſimilar bodies is required in every public act.

The difference moſt relied on between the American and other republics, conſiſts in the principle of repreſentation, which is the [illegible] former move, and which [illegible] unknown to the

Detail from *The Federalist* No. 63 (numbered 62 in the original numbering system) as it appeared in *The New-York Packet*, March 4, 1788. This essay, one of twelve *Federalist* papers whose authorship has been in dispute between Alexander Hamilton and James Madison, also appeared in *The Independent Journal*, March 1, 1788.

CHAPTER 1

The Federalist Papers As a Case Study

1.1. PURPOSE

When two statisticians, both flanks unguarded, blunder into an historical and literary controversy, merciless slaughter is imminent. Our persistence needs explanation.

From the point of view of statistical methods, authorship problems fall into a general area called discrimination or classification problems. In these problems the task is to assign a category to an object or individual whose true category is uncertain. In our authorship problem the objects are essays written by either Hamilton or Madison. We reduce our uncertainty about the authorship of an "unknown" essay by comparing its properties with information obtained from essays whose authorship is known. Classifying plants in biology, skulls in anthropology, candidates for parole in criminology, and subjects according to personality in psychology are related operations that sometimes employ similar methods, even though the properties that aid the classification vary drastically from one area to another. The methods used to study one problem in discrimination can sometimes be extended to other areas of research. We are concerned with the methodology of discrimination studies, and we especially wish to compare a number of methods of discrimination all based on much the same data.

As explained later, for accidental personal reasons we began to study the authorship of the disputed *Federalist* papers. Because standard methods of historical research have not firmly settled this authorship problem, we felt justified in pursuing it with statistical methods. As the work progressed, we became dissatisfied with our rather catch-as-catch-can methods, although they may have been adequate for the immediate purpose of deciding authorship, and we realized that this problem presented an opportunity for a systematic comparison of two general methods of attack. One of these is the classical method of discrimination as devised by R. A. Fisher (1936). The other flows from the work of Thomas Bayes (1763) on statistical inference. Critics of the Fisherian approach complain that the method does not incorporate some important

practical information; critics of the Bayesian plan usually agree that the inclusion of the information would be an asset, but they regard the proper assessment of this information as a hopeless task. Harold Jeffreys (1939) had attempted to popularize the Bayesian approach. But it seems fair to say that it was not until about 1955 that statisticians began to think seriously of using Bayesian analysis directly in many practical problems, and even then it was mainly thinking. Even by 1963, very few life-sized problems have employed Bayesian methods for their solution, and far fewer involve substantial analyses of data.

To us, this lack of experience with the new method is unfortunate. So often in science and engineering, one finds that the difficulties anticipated from the armchair scarcely overlap those that confound one in the field or laboratory. With all this in mind, we decided to enlarge our effort with *The Federalist* papers to produce a case study of the use of Bayesian and other methods of discrimination. The main value of our work does not depend much upon one's view of the authorship question used as a vehicle for this case study, although we do add information there. Rather the value resides in the illustrative use of the various techniques and in the generalizations that emerge from their study. In retrospect, the methodological analysis could have been restricted to sets of papers whose authors are known. Still, the responsibility of making judgments about authorship in disputed cases adds a little hard realism and promotes additional care that might otherwise have been omitted.

Later, we explain the ideas and controversies about the methodology, but now let us turn to the subject of our case study—the authorship of the disputed *Federalist* papers.

1.2. *THE FEDERALIST* PAPERS*

The Federalist papers were written 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. As was common in those days, these short essays, about 900 to 3500 words in length, appeared in newspapers signed with a pseudonym, in this instance, "Publius." They covered nearly every phase of the proposed Constitution. Seventy-seven essays first appeared in several different newspapers, and then Hamilton wrote an additional eight essays designed to complete the job. All the papers then were published in book form and have been republished repeatedly both here and abroad.

---

* We are not historians, and we would be chagrined but not astonished to find some of the description in this section mildly in error, but it should be adequate to lay the setting for this report. A reader who wants more information and references will find the Adair (1944a, 1944b) articles both useful and fascinating. As for ourselves, we abbreviate Adair's description drastically and shamelessly. An additional good source is the preface to the Cooke (1961) edition.

Of the first 77 papers, it is generally agreed that Jay wrote five, and these are identified; drafts of Nos. 3, 5, and 64 exist, although that for No. 64 was mislaid for a while and then rediscovered.* Hamilton and Madison, as well as historians, seem agreed upon the authorship of an additional 57 papers, 43 by Hamilton and 14 by Madison. The authorship of another twelve is in flat dispute between Hamilton and Madison, and these are referred to below as "the disputed papers." An additional three are usually referred to as "Hamilton and Madison." In a nutshell, Hamilton said they were joint papers, and Madison said that Hamilton's material "was left with Mr. M on its appearing that the latter was engaged in it, with larger materials, and with a view to a more precise delineation; and from the pen of the latter, the several papers went to the Press." This gives us a definite "maybe" as far as the inclusion of Hamilton's material is concerned.

Why is authorship hard to settle? For one thing, after Hamilton and Madison had written the papers they each took political positions on important issues that differed from some of the views presented in their own papers. Old writings can be embarrassing. The papers, after all, were meant as propaganda, so there is no guarantee that the views expressed were invariably held by an author even at the time he wrote them. Definitely some positions expressed were not held at later times.

For concreteness, we mention two matters. Hamilton actually worked for a strong central government and against strong state governments, but he wrote in paper 28 that "the State governments will, in all possible contingencies, afford complete security against invasions of the public liberty by the national authority." Furthermore the party opposing Hamilton wanted strong state governments. As a similar difficulty, Madison in paper 44 took the position that once an end was sought by the law, the means were authorized. Later he adopted a "strict construction" theory of the Constitution. Hamilton used the means-end idea to support the creation of the National Bank, and Jefferson and Madison the "strict construction" to try to strike it down.

While it was generally known who had written *The Federalist*, no public assignment of specific papers to authors occurred until 1807, three years after Hamilton's death, when a Philadelphia publication printed a list, in a letter† to the editor signed M., claiming it to be based on Hamilton's own writing in his own copy. The letter announced that the executors of Hamilton's will deposited the copy in the Publick Library of New York. That copy has been missing for over 150 years. The list coincides with another list, the "Benson list," not made

---

* The *New York Times*, p. 1, November 19, 1959, reports the rediscovery as a major historical event. Number 64 has a special significance, as we see in the discussion of the Benson list.

† *The Port Folio*, November 14, 1807, p. 318. On p. 316 a small notice says "The information from M, relative to the respective shares which Gen. Hamilton, Mr. Jay, and Mr. Madison had in the composition of that imperishable collection of political essays, 'THE FEDERALIST,' is curious and valuable."

"... on the page opposite the memorandum quoted above he [Kent] pasted a copy of the article from the *City of Washington Gazette* which stated that Madison had written Numbers 10, 14, 17, 18, 19, 21, 37–58, 62–63, and that Jay was the author of Numbers 2, 3, 4, 5, 64. Underneath this clipping Kent wrote:

> I have no doubt that *Mr. Jay* wrote No 64 on the Treaty Power—He made a speech on that subject in the NY Convention, & I am told he says he wrote it. I suspect therefore from internal Evidence the above to be the correct List & not the one on the opposite page."

The article Kent clipped out was said to have been "furnished by Madison himself." A letter signed Corrector in the *Daily National Intelligencer* in the spring of 1817 gives to Madison 10, 14, 18–20, 37–58, 62–64, to Jay 2–5, the rest to Hamilton. The author says these numbers were copied from a "memorandum in the hand of Madison." Richard Rush's copy of *The Federalist* papers contains a list said to be in Madison's own handwriting: to Madison 10, 14, 18–20, 37–58, 62, 63, to Jay 2–5, and 64, and the rest to Hamilton. Two lists come from Jefferson! In one, said to be in Jefferson's handwriting: to Madison 10, 14, 17–19, 21, 37–58, 62 and 63, to Jay 2–5, 64, the rest to Hamilton. The other list is said to be in the handwriting of Gideon Granger, a member of Jefferson's Cabinet, who says that the "information derived from Jefferson": to Madison 10, 14, 37–48, joint 18–20, to Jay 2–5, 54, the rest to Hamilton. While there may be other lists, the flavor of the situation is well contained in these.

John Church Hamilton (1864), son of Alexander Hamilton, compared the disputed papers with other writings by Alexander Hamilton and decided that the latter wrote all the disputed papers. Cooke (1961, p. xxviii) feels Hamilton "produces some evidence" for Nos. 55–58. Nevertheless, Cooke gives more credence to Madison's claim.

In 1888, the historian and, later, senator, Henry Cabot Lodge reanalyzed the position and decided largely on the basis of a credibility analysis to return in the main to Benson's list.

About 1896, the Yale historian, E. G. Bourne (1901), gradually was sucked into the whirlpool of the *Federalists* while studying the use of history by framers of the Constitution. Bourne's approach was to match in parallel columns many lines from Madison's extensive notes of the Constitutional Convention and other writings with lines in the disputed papers. Madison's Convention notes seem to have been written up after *The Federalist*. Bourne felt that papers 51, 53, 62, 63 were definitely Madison's and that 49 and 50 were very likely Madison's. Similarly he felt the "joint" papers 18, 19, 20 could be fairly assigned to Madison. On balance, he concluded Madison's Gideon list to be correct.

Paul Leicester Ford (1898) took up the fight and also analyzed the disputed papers, giving 49–51 to Madison, and 52–58 and 62 and 63 to Hamilton. He also awarded 47 and 48 to Hamilton, clouding matters a bit more.

public until 1817, but said* to have been left with Egbert Benson shortly before Hamilton's fatal duel with Aaron Burr over remarks attributed to Hamilton concerning Burr's fitness for public office during a New York gubernatorial campaign. The Benson list reads "Nos. 2, 3, 4, 5, 54 by J. Nos. 10, 14, 37 to 48 inclusive, M. Nos. 18, 19, 20, M. & H. jointly. All the others by H." (In *The Port Folio* the names are spelled out, and it closes "all the rest by Mr. Hamilton.") Number 64 rather than 54 was actually written by Jay, so 54 is apparently an error. If 48 were similarly an error for 58, all but two of the papers in the controversy would be accounted for. Translating from Roman numerals helps promote such errors, especially when the original numerals have misprints. In the 1788 McLean edition, essay 48 is numbered LXVIII, essay 49 is LXIX, and essay 70 is numbered LXXX at the head of the essays.

Madison did not reply to the 1807 newspaper list. Of course, he may not have been aware of the article, and even if he were, he may not have been eager to enter his claim. Not until his retirement from the presidency did Madison's claim come forward. He claimed full authorship of numbers 49 through 58 and 62 and 63, and, based on Madison's corrected copy of *The Federalist*, the 1818 Gideon edition of *The Federalist* so appeared. Apparently, his count was accepted for nearly half a century.

Before we go further, the reader may as well know that there are more than two lists, and that they are not all alike.

The Kent (Chancellor James Kent) list is well described by Cooke† who says that

> "Because of differences in the ink and pen he used, Kent's statement may be divided into three parts, each of which was written at a different time. In the following copy of Kent's notes the three parts are indicated by Roman numerals:
>
> I. I am assured that Numbers 2. 3. 4. 5. & 54 [the number "6" was later written over the number "5"] were written by *Jay*‡ *Jay*. Numbers 10. 14. 37 to 48 [the number "9" was later written over the number "8"] both inclusive & 53 by *James Madison Jun.* Numbers 18. 19. 20. by Messrs *Madison & Hamilton* jointly—all the rest by Mr. *Hamilton.*
> II. (Mr. *Hamilton* told me that Mr. *Madison* wrote No. 68 [the number "4" was later written over the number "6"] & 69 [the number "4" was later written over the number "6"] or from pa. 101 to 112 of Vol 2d)
> III. NB. I showed the above mem. to General Hamilton in my office in Albany & he said it was correct saving the correction above made—See Hall's Law Journal Vol 6 p 461.

---

* Cooke (1961, p. xxiii) points out that the Benson list is suspect because no one authenticating it claims to have seen it. The list disappeared in 1818.

† Cooke (1961, pp. xxiv and xxvi) reproduced here with permission. See Preface.

‡ Misprint for John.

Adair, cited earlier, decided on the basis of all the historical evidence he could find, including his own original researches, that Madison wrote the disputed 12 papers, and was largely responsible for the joint papers 18, 19, and 20.

Nevertheless, he pursued the matter further and has encouraged us to do so.

Besides Cooke's edition, two other unabridged American editions (Rossiter, 1961, and Wright, 1961) have appeared since we began our study of the authorship problem. Among the 12 disputed papers, Cooke assigns to Madison Nos. 49 and 53, and the rest have "been attributed to Madison, but to indicate Hamilton's claim his name has been placed in brackets underneath that of Madison." Rossiter assigns 49–58 to Madison, "and probably 62, 63." He owes his confidence in this attribution ". . . chiefly to the scholarly labors of Professor Douglass Adair . . ." (p. xi) In Wright's edition, Adair's "assignment of authorship is followed, though question marks have been inserted after Madison's name at the head of Numbers 62 and 63, where there appears to be reason for doubt." (p. 10)

Recently Irving Brant (1961) reviewed again the data on 62 and 63 and presses the Madison claim to these largely on the basis of the Kent list. Broadus Mitchell (1957, p. 419) in a Hamilton biography gives all but Nos. 62 and 63 to Madison.

By and large the available historical evidence today is much the same as it was when Lodge attacked Madison's claim. Adair has noted that the preference shown for each man's claim has, over the years, swung with the popularity of the man's views, and with the evidence of the kind it is, if Hamilton's star should wax and Madison's wane, perhaps the recent thoughtful decisions can be reversed. Our own view is that the historical evidence is modest enough that a reasonable but stubborn skeptic could retain the Scotch verdict "Not proven," and that others with special doubts or beliefs could sensibly maintain the opposite of the current viewpoint favorable to Madison.

Later we develop our evidence, evidence of a different sort from that so far discussed. Using data internal to *The Federalist*, but not depending on its intellectual content, we provide new evidence that adds to the historical evidence and permits a more nearly definitive assignment. The point, of course, is not just to make an assignment of authorship, but to provide solid communicable evidence about the value of one's assignment, and that we are able to supply in plenty.

1.3. EARLY WORK

Frederick Williams introduced Frederick Mosteller to this authorship problem in 1941 when the latter was a graduate student at Princeton University. Influenced by Yule's (1938) and C. B. Williams' (1939) work on word counts and sentence length, Williams and Mosteller independently counted the number of

words in all the sentences of *The Federalist*,* and their first frustration was the discovery of an important empirical principle—people cannot count, at least not very high (Miller, 1956; Wundt, 1912). But they created checks to get accurate results. The second difficulty, common to nearly every investigation, is that special problems must be settled by judgment. How should quotations from other authors be handled? How should numbers written out or in numerals be counted in words, and so on? These matters took a great deal of time, but finally, the task was done. When Williams and Mosteller emerged from their bout with the desk calculators, their second frustration was:

Average sentence length: Hamilton—34.55 (words),
Madison—34.59 (words).

Still all was not lost: perhaps the variability in sentence length was the key to the style. They computed the standard deviation (a measure of variability) of the sentence lengths for each paper, then found the average of these for the two men:

Average standard deviation: Hamilton—19.2 (words),
Madison—20.3 (words).

Needless to say, these measures would be hopeless for discriminating between the two authors. You will observe that the sentence length is long, on the average, and that the large standard deviation means that some sentences were very long. Both Hamilton and Madison had developed a style of writing much admired in their period, rather in imitation of the *Spectator* papers. To a modern reader the style is oratorical and somewhat overwhelming. It sounds important and convincing when read aloud, but the listener may find it difficult to recall the ideas in a passage, though they are plentiful. The style is formal and complicated, the words are long, and the sentences are crowded with qualifications.

---

* The text used by Williams and Mosteller was: *The Federalist*, Sesquicentennial Edition, National Home Library Foundation, Washington, D.C., 1937.

From the "Bibliographical Note" (p. xxii):

"The text here given follows closely the original McLean edition of 1788, which is generally accepted as authentic. But as Hamilton's table of contents is a mere skeleton, we have used, with the permission of the publishers, G. P. Putnam's Sons, the more inclusive table of contents from the edition of Henry Cabot Lodge (1886)."

In preparing text for the high-speed computers, we used copies of The Modern Library edition of *The Federalist*, Random House, New York. This text appears to be identical with that of the Sesquicentennial Edition; indeed on p. iv, we read:

"The publishers are indebted to the National Home Library Foundation for making this edition possible and wish to thank them for the courtesy extended in the use of their plates."

The Modern Library edition has an additional four pages (619–622) that give The Declaration of Independence of the United States.

TABLE 1.3–1

NOUNS

Per cent	H	M
21	13	6
22	13	7
23	10	1
24	5	
25	1	
Totals	42*	14

TABLE 1.3–2

ADJECTIVES

Per cent	H	M
9		1
10	15	
11	20	1
12	4	6
13	2	5
14	1	
15		1
Totals	42*	14

TABLE 1.3–3

ONE- AND TWO-LETTER WORDS

Per cent	H	M
21		2
22	2	3
23	1	6
24	11	1
25	12	1
26	9	1
27	6	
28	1	
Totals	42*	14

TABLE 1.3–4

the's

Per cent	H	M
7	2	
8	12	2
9	12	6
10	10	1
11	6	2
12		2
13		1
Totals	42*	14

TABLE 1.3–5

DISCRIMINANT FUNCTION ANALYSIS

Class intervals	Hamilton	Madison	Disputed
.55–.60	–	1	–
.60–.65	–	2	–
.65–.70	1	4	3
.70–.75	5	5	6
.75–.80	23	2	3
.80–.85	12	–	–
.85–.90	1	–	–
	42*	14	12

* Two papers were pooled for this analysis; thus the total is 42 instead of 43.

Cooke (1961, p.xxviii) says of the remarkably similar prose styles of Hamilton and Madison:*

> "This was no unique phenomenon, for most educated Americans of the late eighteenth century—a few particularly gifted writers, like Jefferson, perhaps excepted—employed the same stylistic devices, the same standard phrases, and remarkably similar sentence structure."

An additional way to use the word counts was to make counts of additional variables. Williams and Mosteller computed for each paper the per cents of nouns, of adjectives, of one- and two-letter words, and of *the*'s.

None of the short words caused much trouble, but the workers were appalled by the number of special cases that seemed to arise in classifying words as nouns and as adjectives. Although they had been taught in school that these classes were clearly definable, they accumulated a great deal of evidence that would support the convictions of modern linguists that the categories of Latin grammar are often ill fitted to describe English (Roberts, 1958, pp. 131–150). Before finishing, they assembled quite a little book of decisions (not all of which look wise in retrospect, although they did serve to create consistency in the counting).

On the basis of these data, they constructed a statistic that was intended to separate Hamilton's writings from those of Madison by giving high scores to Hamilton's writings and low to Madison's. The statistic, which produces a weighted sum of the rates of occurrence of the four variables for each piece of writing, is called a *linear discriminant function.*

The frequency distributions for the percentages are shown in Tables 1.3–1, 1.3–2, 1.3–3, and 1.3–4, where H stands for Hamilton and M for Madison. The entries are the numbers of papers of known authorship falling between $p - .5$ and $p + .5$, where p is a whole per cent.

The frequency table for the discriminant function is shown in Table 1.3–5, grouped by class intervals of .05. The rightmost column gives the results for the 12 disputed papers.

Some casual calculations in this table are suggestive, but not definitive. Ratios of proportions or of probabilities sometimes give odds. For example, in the interval .65–.70 Hamilton has the fraction $\frac{1}{42}$ of his known papers, while Madison has the fraction $\frac{4}{14}$. The ratio of these two fractions would give odds of about 12 to 1 for Madison. Similarly, in the interval .70–.75 the odds are 3 to 1 (15 to 5) for Madison, and in the interval .75–.80 about 4 to 1 (23 to 6) for Hamilton. These calculations are rough in two ways. First, they have considerable uncertainty because, for this purpose, the number of papers we have available by either author is small; second, the use of these odds implies that each author has an equal chance of having written a disputed paper, an assumption the reader may not wish to accept and one whose consideration and revision we defer until Chapter 3.

---

* Quoted with permission. See Preface.

As often happens in discriminant analysis, the items needing to be sorted lie balefully in the middle, between the two criterion distributions. Had the problem been to assign the 12 papers in a block to one man, this evidence would support Madison. But that is not the problem. Each paper is to be settled separately. True odds factors of three or four to one are not compelling in decisions of this sort, and estimated odds of this magnitude are even less so.

Events drew Mosteller and Williams apart, and they never succeeded in continuing their work together. But a few years ago Mosteller had a number of inquiries about these calculations, and these inquiries led to the present investigation. The original decisions on *The Federalist* as to what material to include and what to omit (for example, quotations) are retained in the new studies.

1.4. RECENT WORK—PILOT STUDY

About 1959, after previous correspondence, Douglass Adair informed Mosteller that he had found a pair of words (which we call marker words) that distinguished Hamilton and Madison quite well when the words occurred. Hamilton uses *while* and Madison in a corresponding situation uses *whilst*. Adair has pursued his investigation of these words, including verification in the original newspapers. He has also found an earlier writer, Bailey (1916), who noticed the *while-whilst* distinction. On the other hand, both words have low rates per thousand words of text, so that not all papers can be discriminated by them. Furthermore, since there are a few exceptional reversals of use in writings of known authorship, a skeptic might well feel that one occurrence of such a marker word would not be enough to justify classification of the whole paper. For example, he might attribute an occurrence to a change by a typesetter. Indeed Cooke (1961, pp. xxviii–xxix) says*

> "To attempt to find in any of the disputed essays words which either man used and which the other never employed is futile, if only because the enormous amount which each wrote allows the assiduous searcher to discover almost any word in the earlier or subsequent writings of both."

And Cooke goes on (p.606) to explain the further difficulties of using marker words. He notes that *whilst* in No. 51 was put there by Hamilton in the McLean edition as a change from *and* in the original newspapers.

We do not attempt to use single words on an all-or-none basis to assist in the identification—rather we try to get so many words and clues that the total evidence is overwhelming, although no one clue is. And to this end, we use rates and other methods for weighting evidence.

Adair's findings encouraged us to look for additional marker words and to begin a large study. We knew we wanted the counts done on high-speed com-

---

* Quoted with permission. See Preface.

TABLE 1.4–1

INCIDENCE: NUMBERS OF PAPERS IN WHICH WORD OCCURRED AT LEAST ONCE

	enough	*while*	*whilst*	*upon*	Number of papers examined
Hamilton	14	10	0	23	23
Madison	0	0	13	4	19
Disputed	0	0	5	1	12
Joint	1	0	2	2	3

TABLE 1.4–2

RATES PER 1000 WORDS

	enough	*while*	*whilst*	*upon*	Total words in 1000's
Hamilton	0.59	0.26	0	2.93	45.7
Madison	0	0	0.47	0.16	51.0
Disputed	0	0	0.34	0.08	23.9
Joint	0.18	0	0.36	0.36	5.5
					126.1

puters, but while the programmers worked, we decided to have word counts made on a few papers by hand, as described in Section 2.6.

We then made a rough screening plan that involved looking at incidences of each word in 6 Hamilton and 5 Madison papers (we call this set wave A); this screening led to the elimination or choice of marker word candidates. Straightforward objective rules were used, except that contextual words (see Section 2.1) were discarded freely. We then tried the words that looked promising on a set of papers on similar topics (5 by Hamilton and 5 by Madison), labeled the *Exterior set*, not in *The Federalist* (there is danger in the screening process of wasting the known Madison *Federalist* papers). This led to further elimination of marker word candidates. By now, Adair's *while-whilst* pair had emerged as important candidates (as we already knew), thereby suggesting that the screening method was not hopeless. Another word, *upon*, emerged as a Hamilton marker. These three words were tried on a second set of 10 *Federalist* papers (6 Hamilton and 4 Madison), wave B, and *upon* by itself separated the 10 papers correctly and emphatically because Hamilton always had a rate of use higher than a prechosen number, about 1.5 per 1000 words, and Madison always a lower rate, usually zero. We decided that *upon* was definitely a marker word.

Later the computer programming was completed and a few more *Federalist* papers became available (6 Hamilton and 5 Madison), wave C. The word *enough* emerged as another Hamilton marker. Summary data on the four strong marker words are given in Tables 1.4–1 and 1.4–2.

TABLE 1.4–3

HAMILTON KNOWN PAPERS

Paper	Wave	Total words	Absolute counts				Rate per 1000 words
			enough	*while*	*whilst*	*upon*	*upon*
Federalist							
1	A	1,570	1	–	–	6	3.82
6	A	1,893	1	–	–	4	2.11
7	A	2,245	–	–	–	11	4.90
8	A	1,986	2	–	–	3	1.51
9	A	1,629	1	1	–	4	2.46
11	B	2,495	–	–	–	6	2.40
12	B	2,140	1	–	–	7	3.27
24	B	1,805	–	1	–	7	3.88
26	C	2,379	7	–	–	6	2.52
36	B	2,721	–	–	–	6	2.21
59	C	1,847	–	–	–	3	1.62
65	B	2,013	1	1	–	10	4.97
66	A	2,215	1	1	–	11	4.97
70	B	3,029	3	2	–	6	1.98
71	C	1,702	3	1	–	3	1.76
72	C	2,032	2	2	–	5	2.46
74	C	906	–	–	–	2	2.21
77	C	1,970	1	1	–	10	5.08
		36,577					
Exterior (identified in Section 2.2)							
Cont. 1		1,188	–	–	–	2	1.68
Cont. 2		1,380	2	–	–	5	3.62
Pac. 1		2,945	–	1	–	6	2.04
Pac. 2		2,470	1	1	–	10	4.05
Pac. 3		1,178	–	–	–	1	.85
		9,161					
Total		45,738	27	12	0	134	
Rate per 1000 words:			.59	.26	0	2.93	

Looking at Table 1.4–2 for the disputed papers, note the absence of *enough* and *while*, the low rate for *upon* and the presence of *whilst*. The pattern of the whole line gives strong evidence for Madison's authorship.

These data show rather clearly that the disputed papers as a whole are Madisonian, but in this form they cannot settle the papers singly. Note that *enough* occurs among the joint papers, which otherwise look Madisonian. What is wanted from these and other data is, for each paper, a good measure of the weight of evidence toward Madison or Hamilton.

TABLE 1.4–4

MADISON KNOWN PAPERS

Paper	Wave	Total words	Absolute counts				Rate per 1000 words
			enough	*while*	*whilst*	*upon*	*upon*
Federalist							
10	A	2,987	–	–	–	–	–
14	A	2,149	–	–	1	–	–
37	A	2,709	–	–	1	1	.37
38	A	3,314	–	–	2	4	1.21
39	C	2,585	–	–	–	–	–
40	B	2,718	–	–	–	–	–
41	B	3,506	–	–	1	–	–
42	C	2,711	–	–	–	2	.74
43	C	3,029	–	–	1	–	–
44	B	2,561	–	–	2	–	–
45	C	2,117	–	–	2	–	–
46	B	2,601	–	–	2	–	–
47	A	2,547	–	–	–	–	–
48	C	1,561	–	–	–	–	–
		37,095					
Exterior (identified in Section 2.2)							
Helv. 1		2,945	–	–	1	–	–
Helv. 2		2,704	–	–	3	1	.37
Helv. 3		2,458	–	–	1	–	–
N. Am. 1		3,048	–	–	3	–	–
N. Am. 2		2,798	–	–	4	–	–
		13,953					
Total		51,048	–	–	24	8	
Rate per 1000 words:			0	0	.47	.16	

By breaking down the data of this preliminary study, we can give the reader some idea of the state of the evidence at this point, before we turn to more systematic approaches. Tables 1.4–3, 1.4–4, and 1.4–5 show the data for Hamilton, Madison, and the Disputed and Joint papers. They show all four words as strong discriminators, though *upon* is far and away the best.

On the basis of Table 1.4–5 the evidence seems to be that all 12 disputed papers are by Madison and that most of the material in Nos. 18–20 is too. For us the problem is to extend this evidence and to measure its strength. We are not faced with a black-or-white situation, and we are not going to provide an absolutely conclusive settlement. Outside those parts of logic and mathematics that are divorced from the empirical world, strong confidence in conclusions is the most an investigation can be expected to offer.

TABLE 1.4–5

JOINT AND DISPUTED PAPERS

Paper	Total words	Absolute counts				Rate per 1000 words
		enough	*while*	*whilst*	*upon*	*upon*
Joint						
18	2,084	1	–	1	1	.48
19	2,019	–	–	1	–	–
20	1,438	–	–	–	1	.70
Total	5,541	1	–	2	2	
Rate per 1000 words:		.18	0	.36	.36	
Disputed						
49	1,594	–	–	1	–	–
50	1,103	–	–	–	–	–
51	1,911	–	–	2	–	–
52	1,841	–	–	–	–	–
53	2,160	–	–	1	–	–
54	1,996	–	–	–	2	1.00
55	2,034	–	–	–	–	–
56	1,560	–	–	1	–	–
57	2,200	–	–	3	–	–
58	2,082	–	–	–	–	–
62	2,380	–	–	–	–	–
63	3,020	–	–	–	–	–
Total	23,881	0	0	8	2	
Rate per 1000 words:		0	0	.34	.08	

1.5. PLOTS AND HONESTY

Most writers on this disputed authorship problem tumble over themselves to assure the reader that there is no question of honesty or integrity on the part of either Hamilton or Madison, that the whole matter is one of innocent mistakes, that there has been no attempt at deceit, that it is a question of memory, not veracity. Madison said much the same thing, attributing Hamilton's disagreement, essentially, to carelessness. From the point of view of the historical evidence, the assumption of honesty is fairly important. Current historical attribution leans on Madison's mature consideration for his claim and on the Gideon edition which agrees with and is that claim. Further, it soft-pedals the Benson list because the authenticators did not say they saw it. The Kent list, with its contradictory notes written over a period of years, corrected in who knows whose handwriting, is one basis for allowing Hamilton to yield some of the disputed papers to Madison. If Hamilton ever had a claim, it seems to pass

into limbo largely by default because of the vanishing of lists and of his own copy of *The Federalist*, and the general doubt surrounding the authenticity of all claims attributed to him. Except then for Nos. 62 and 63, current historical evidence largely relies on Madison's attribution in the Gideon edition. We are not trying to slight the excellent historical work; we are trying to get to the critical assumptions underlying the arguments.

Thus, we feel that the historical evidence is strong, but not enormously strong, because it leans so heavily upon the integrity and memory of one man.

Our own evidence has a different weakness. It depends upon the extent of the editing done by one or the other man. Of course, since the papers were written quickly, we suppose they were not heavily edited except by their authors, but we cannot guarantee that. If Hamilton wrote and Madison edited Nos. 49 through 58, say, but not other Hamilton papers, then Madison might remove *upon*'s and *enough*'s, change *while*'s to *whilst*'s, and so on, to confuse our methods, but there is a lot against this, as we see later. Although we cannot overcome this criticism completely, we can mitigate it by producing a great deal more evidence, based on many marker words, and even more, on high-frequency words, such as *by*, *of*, and *to*. With the frequently used words, the differences between these authors' rates of use are relatively slight, and the chance much greater that neither author would be aware of the differences. As the reader will see later, we are encouraged to find considerable consistency among the papers generally agreed to be Hamilton's and among those generally agreed to be Madison's. If the editing had been heavy, we might hope that it would partially reveal itself through inconsistency in papers other than the disputed ones. Because this trouble does not arise, we have some evidence for the lack of cross-editing, and some for the accuracy of the Gideon edition, but we find the amount hard to evaluate. We pursue this further in Section 3.7F, where we can give the reader a better notion of the large quantity and cunning quality of editing required to match the facts.

1.6. THE PLAN OF THE BOOK

In all we give four parallel studies, each with somewhat different methodology: first, in Chapter 3, the main study based on Bayes' theorem, with Chapter 4 as its technical appendix; second, in Chapter 5, a classical linear discrimination study; third, in Chapter 6, a "robust" Bayes analysis that is cheaper and less subject to troubles from distributional assumptions than the main study, but also less sensitive; fourth, in Chapter 7, a simplified rate study along somewhat classical lines. Later chapters attend to a variety of matters.

Chapter 2 lays the foundation for the word counts on which all these studies are based.

CHAPTER 2

Words and Their Distributions

2.1. WHY WORDS?

When we leave general style as a poor bet and pay attention to words, we find that Hamilton and Madison use certain words at quite different rates. Douglass Adair brought this spectacularly to our attention by pointing out their uses of *while* and *whilst*. In our work, we have used individual words as the principal basis for measuring likelihood of authorship. Early investigations convinced us that most single variables, carefully selected or not, have little discriminating value, and that a large pool of variables provides the greatest hope of success. Sentence length is a good example of a stylistic variable which had even been used effectively elsewhere, yet failed miserably here. Since the rate for each word can be regarded as a variable, words supply a pool of thousands of variables. Furthermore, words are easily recognized and effective for discrimination.

Our "word" is a composite of all words of the same spelling, capitalization neglected. To put our worst foot forward at once, we do not distinguish *abuse* as noun from *abuse* as verb, the *great* of *great plan* from that of *Great Britain*, nor the personal pronoun *I* from the Roman numeral *I*. Distinguishing between words of the same spelling is rejected almost solely because it cannot be done routinely, and certainly not yet by a high-speed computer (except in experimental programs). Actually, the words used as discriminators in most of our work rarely admit these ambiguities.

Variables involving grammatical concepts have attractions. One fine class consists of single words, each split into its different uses (see Section 8.5). Such variables seem more likely to discriminate between authors, and indeed to underlie the discriminatory ability of words as we use them. The difficulties of this proposed improvement are huge. Even if we assume that one could list the (word, use) pairs to be considered, the task of classification looks unfeasible. That it could be made acceptably objective is doubtful, and the difficult decisions, if nature showed its usual malice for the investigator, would occur not frequently but just where discrimination exists.

TABLE 2.1–1

FREQUENCY DISTRIBUTION OF RATE PER THOUSAND WORDS FOR THE 48 HAMILTON AND 50 MADISON PAPERS FOR *by*, *from*, AND *to*.

by			*from*			*to*		
Rate per 1000 words	H	M	Rate per 1000 words	H	M	Rate per 1000 words	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	Totals	48	50	55–60	1	
17–19		3				Totals	48	50
Totals	48	50						

Let us examine a few words for their ability to discriminate and for their consistency of rate. For this purpose, we discuss some results on 98 items of writing: 48 by Hamilton, 50 by Madison. We call each item a "paper" (the precise sources of the writings are described in Section 2.2). One class of words that we use has been called *function* words—the filler words of the language, such as *a*, *an*, *by*, *to*, and *that*. Generally they include prepositions, conjunctions, pronouns, and certain adverbs, adjectives, and auxiliary verbs.

Table 2.1–1 gives frequency distributions for the rates of use per thousand words of text for the function words *by*, *from*, and *to*. We employ rate here rather than frequency because the papers vary in length—from 906 words to 3551. A length of 2000 is somewhat typical.

Casual inspection of Table 2.1–1 suggests that low rates for *by* are favorable to Hamilton's authorship, and high rates to Madison's. Rates for *to* are in the opposite direction. Very high rates for *from* suggest Madison, but low rates give practically no information. It appears that *by* discriminates better than *to*, which in turn is superior to *from*.

We like the function words rather well because many of them are not much influenced by the context of the writing, but other sorts of more meaningful words also seem relatively free from context, for example, *commonly*, *innovation*, *fortune*, *vigor*, and *voice*. In Table 2.1–2, we give the rates for *commonly* and *innovation*.

In contrast with *by*, *from*, and *to*, which are high-frequency words, *commonly* and *innovation* have low frequencies. Zero is the most frequent rate for both words. Our longest paper has 3551 words, so that the lowest possible nonzero

TABLE 2.1–2

FREQUENCY DISTRIBUTIONS FOR *commonly* AND *innovation*

commonly			*innovation*		
Rate per 1000 words	H	M	Rate per 1000 words	H	M
0 (exactly)	31	49	0 (exactly)	47	34
$0^{+}-$ .2	(cannot occur)		$0^{+}-$ .2	(cannot occur)	
.2– .4	3	1	.2– .4		6
.4– .6	6		.4– .6	1	6
.6– .8	3		.6– .8		1
.8–1.0	2		.8–1.0		2
1.0–1.2	2		1.0–1.2		1
1.2–1.4	1		Totals	48	50
Totals	48	50			

rate is $1000 \times 1/3551 \approx .28$.* Generally speaking, occurrences of *commonly* are favorable to Hamilton's authorship, of *innovation* to Madison's. Nonoccurrence of either word says relatively little about the authorship.

Words such as *law*, *executive*, *liberty*, *money*, *trade*, *war*, and *states* vary greatly in their rate with the context of a paper. Since *The Federalist* papers deal with specific topics in the proposed Constitution, variation is to be expected. Suppose that an investigator allowed such words among his potential discriminators and found one, say, *trade*, to discriminate well among the known papers. It is still quite possible that a peculiar assignment of tasks between Hamilton and Madison might find the one who had ordinarily not discussed trade writing on that topic in a disputed paper and thus throwing the analysis off. The assumptions underlying our later analyses seem inappropriate for contextual words because of this dependence on external information, such as who would be likely to write about trade. While such words provide evidence, we cannot evaluate it to our own satisfaction, let alone to that of one who believes that, say, Hamilton was so expert on trade that he would never have agreed to let Madison write on it. Consequently, we decided to eliminate these meaningful, contextual words. Unfortunately, recognition of words that should be so discarded is difficult. In some of our studies, contextual words were eliminated on an *ad hoc*, intuitive basis; in others the problem was met by constructing lists of noncontextual words to be used as candidates for discriminators.

To give an example of a contextual word, we display the distribution of rates for *war* in Table 2.1–3.

For both authors, the rates vary from 0 to 15 per thousand. While occurrences of the word look somewhat favorable to Madison's authorship, Hamilton

* We use the symbol "$\approx$" for "is approximately equal to" or "approximately equals."

TABLE 2.1–3

FREQUENCY DISTRIBUTION FOR *war*

Rate per 1000 words	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

TABLE 2.1–4

FREQUENCY DISTRIBUTION FOR *upon*

Rate per 1000 words	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

occasionally has a startlingly high rate. Incidentally, it is amusing to find that Madison uses it more, while Hamilton is spoken of by some biographers as having strong military ambitions.

Our best single word for discrimination is *upon*, and its distribution is shown in Table 2.1–4.

Low rates for *upon* go with Madison, high rates with Hamilton. The spread or variability of the distributions seems appropriate for both authors: Hamilton's consistent with his 3/1000 rate and Madison's consistent with his .18/1000.

To summarize, our discussion of the distributions in Tables 2.1–1 through 2.1–4 gives the reader some notion of the sorts of variables used in this study and alerts him to the special worries we have about contextual words.

2.2. VARIATION WITH TIME

In our work, we find it necessary to use papers not included in *The Federalist* as well as those that are. Since Madison had but 14 known papers in *The Federalist*, we felt forced to enlarge our sample of his writings from other material. This means in turn that we have analyzed parts of his writings from the years 1780 through 1806, and thus arose the opportunity to see how his rates varied through the years. Since we rarely went outside *The Federalist* for Hamilton's writings, we could not make a corresponding study.

Before proceeding, the reader may wish to review the list of writings of the two authors, exterior to *The Federalist*, that we employ. In the Appendix a special list of such references is given together with the code numbers we use to identify the materials.

The Madison papers that we studied fall into time periods as follows:

		Year
(a)	1 paper (about 2900 words) An essay on money in *Freneau's Magazine*, not published until 1791.	1780
(b)	2 papers (about 5800 words total) The North American No. I and The North American No. II, anonymous political writings.	1783
(c)	14 papers (about 37,000 words total) The 14 known papers from *The Federalist*.	1787–8
(d)	1 paper (about 3600 words) (Oct.) Observations on a draft Constitution for Virginia.	1788
(e)	7 papers (about 11,000 words total) Further Freneau essays.	1791–2
(f)	5 papers (about 13,000 words total) The five Helvidius papers, written in reply to a series by Hamilton (signed Pacificus) on powers of the Executive. Helvidius's remarks on war were in accord with his questioning of the powers of the Executive branch.	1793
(g)	20 papers (about 40,000 words total) The long *Neutral Trade* paper, which we have broken into twenty pieces of about 2000 words each.	1806

Since groups (a) and (d) contain only one paper each, we combined group (a) with group (b) and group (d) with (c). The five resulting groups are listed in Table 2.2–1, together with the code designation of each paper and the word count for each group.

At the outset of the study, groups II and V were noted to have particular interest because of the time lapse between the papers, the contextual change, and the large number of words in each. Group II represents the *Federalist* group, and the papers of group V are primarily concerned with questions of neutral trade in time of war. These two groups are especially convenient to compare since the total number of words is about the same in each. Within each period, we have pooled all the papers, and Table 2.2–2 shows the rates for the five periods for every seventh function word on one of our basic lists (see Section 2.5A), plus one function word, *her*, that we thought would display contextuality.

In studying these rates, one has to face squarely the impossibility of clearly separating temporal effects from contextual ones. Generally, the absence of a positive reason for attributing a change to contextuality leaves us with time as a residual cause. We supposed that in the *Neutral Trade* papers, group V, discussions of countries would lead to the frequent use of *her*, and so it did.

In Table 2.2–2, *by*, *from*, *must*, *one*, *some*, and *where* appear fairly stable. The only word suggesting a trend with time is *in* (and possibly *from*), and the

TABLE 2.2–1

MADISON'S PAPERS, GROUPED BY DATE OF WRITING

Group	Number of papers	Date	Code number	Description	Length
I	3	1780 1783	302 121, 122	M–4 NA–1, NA–2	8,725 words
II	15	1787–88	10, 14, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 141	14 known *Federalists*, N9	40,646
III	7	1791–92	301, 311, 312, 313, 314, 315, 316	M–1 to M–3, M–5 to M–9, N–1 to N–8	10,889
IV	5	1793	131, 132, 133, 134, 135	Helv. 1–5	12,779
V	20	1806	201 through 220	*Neutral Trade*	40,561

TABLE 2.2–2

MADISON'S RATES PER 1000 WORDS OVER A 25-YEAR PERIOD

Group	I	II	III	IV	V	Total
Length in thousands of words	8.725	40.646	10.889	12.779	40.561	113.600
Years	1780–83	1787–88	1791–92	1793	1806	
Word						
any	1.38	1.92	.83	3.36	3.23	2.40
by	7.45	11.91	12.95	10.80	11.54	10.29
from	4.47	4.99	5.14	5.32	6.80	5.65
her	2.06	.66	.28	1.10	5.46	2.49
in	15.47	21.03	21.39	24.49	26.89	23.11
must	2.18	2.36	2.48	2.43	1.46	2.04
one	3.90	3.08	4.22	2.19	2.20	2.83
some	.57	1.53	.72	1.10	1.19	1.21
there	1.26	.84	1.84	3.13	1.14	1.33
where	1.95	1.97	1.56	1.09	2.27	1.93
would	5.39	4.43	2.85	5.24	1.95	3.56

words *any*, *her*, *there*, and *would* vary considerably from one group to another. Our general conclusions based on this table and on similar studies with other words are that pronouns and auxiliary-verb forms are potentially contextual and offer risky discrimination. Furthermore, we concluded that for the main study special pains should be taken to examine all candidates for the final list of discriminators for evidence of large variability between groups of writings.

Essentially we had three choices: to revise our model for word distributions in such a way as to handle excessive variability between groups of writings within the model, or to restrict our source of words to *The Federalist* papers alone in hopes that rates are homogenous there, or, finally, to try to choose words whose rates are not much affected by the group of writings. We chose the third of these alternatives, partly for expediency, because we regard the first as an ideal choice that we could not afford, but in definite preference to the second. Since 10 of the disputed papers come in a clump, the group as a whole might differ substantially in its rates from those of its author or authors in other *Federalist* writings. Therefore the second choice could easily lay the study open to the danger we wish to avoid. Quite aside from these considerations, persons interested in such studies will prefer methods that spot authorship over a period of years and a variety of topics to those that are tightly restricted.

2.3. HOW FREQUENCY OF USE VARIES

A mathematical model is a description of a process in mathematical terms. For example, the parabolas that describe the behavior of bodies moving through a vacuum near the surface of the earth, together with the rules for generating them, form a mathematical model for bodies in flight. Similarly, the mathematical descriptions of the behavior of light in systems of mirrors and lenses are mathematical models of the real thing. One advantage of the mathematical model is that it can be manipulated in the absence of the structure it represents; another is that consequences can sometimes be derived from the mathematics that are not obvious from inspection of apparatus or data. The models for moving bodies and light do not ordinarily have uncertainty or variability built into them. We need a model to describe the changing rates with which words are used. One reason is that we do not have infinite amounts of data associated with each paper, and we need to be able to appraise our uncertainty. We need to be able to assemble information from different words to strengthen our inference, and without a model one is hard put to know how to do this.

Even the finest mathematical models do not represent a physical process perfectly or completely, but if we can represent a process fairly closely, we expect the consequences from the model to be fairly close to the truth. This section and the next are devoted to considering what would be an adequate model for describing the variation in the use of words.

Setting up a complete statistical model for the frequency of use for even a single word is a very difficult task, and fortunately, for our purposes, it is un-

necessary. We examine word frequencies only in good-sized blocks of text, mostly of 1000 words or more, and of at least 100 words in all intended studies. Over pieces longer than 100 words, details of local dependence introduced by grammatical structure and style, such as avoidance of repetitions, ought not to be especially important when we study total counts. In what follows, we talk in terms of models without any allowance for local dependence, with the intent that the models be judged by and used for their behavior in pieces of text of 100 or more words.

How does frequency of use vary among and within chunks of writing? Two counter-pressures that we all recognize are those for splendid isolation and for clubbiness. In writing, we avoid using the same word over and over, and this isolation tends to space occurrences and make the rate quite constant. But for emphasis, parallelism, or clarity, we may repeat a word several times in a brief passage. One concrete way to raise the distributional problem is this: if a word has just been used, is it more or less likely to be used in the next, say, 200 words than if it has not been?

2.3A. Independence of words from one block of text to another. One could scarcely hope that the chance of occurring soon does not depend on whether or not the word just occurred. The absence of such dependence implies statistical independence. We often use such an assumption as statistical independence as a base line to measure departures against. Let us discuss the idea first for pairs of adjacent blocks of 200 words. Suppose that in the long run $\frac{1}{2}$ the blocks contain the word at least once and that we do have independence. Then among a great many pairs of adjacent blocks $\frac{1}{4}$ of the pairs would not contain the word at all, $\frac{1}{2}$ would contain it in exactly one block, and $\frac{1}{4}$ would contain it in both blocks. An empirical test of independence can show deviations from this simple theory. Perhaps many more than $\frac{1}{2}$ the pairs have occurrences in just one of the two adjacent blocks—just as in baseball, for example, too many double-headers are split for independence to hold.

The classical theory of the binomial distribution (see, for example, Mosteller, Rourke, and Thomas, 1961, Chapter 7) tells how the counts of occurrences should behave for probabilities of incidence more general than $\frac{1}{2}$, and for more than two blocks. In what follows we examine results for four adjacent blocks. We use four instead of two blocks to magnify the effects of dependence. We now present the special theory that assumes independence for four blocks.

THEORY. Let p be the fraction of blocks in which a word occurs; then the formula for the desired binomial probabilities can be shown to be as follows:

Number of blocks with word present	0	1	2	3	4
Probability of that number	$(1 - p)^4$	$4p(1 - p)^3$	$6p^2(1 - p)^2$	$4p^3(1 - p)$	p^4

We can estimate p as the observed fraction of all blocks in which the word occurs. Then the estimated probabilities in the above array can be multiplied by the number of sets of four blocks, 39 in our problem, to make a direct comparison of frequencies. Tables of the binomial distribution (National Bureau of Standards, 1949) facilitate the calculation.

Application. To get appropriate data, we broke a sequence of Hamilton papers (*Federalist* 13, 16, 17, 23, 25–35, 59–61, 68, 69, 71–77) into 247 blocks of about 200 words each. Then we formed 39 sets of four adjacent blocks, each set contained entirely in a single paper. For any set, a given word can occur in 0, 1, 2, 3, or 4 blocks. Table 2.3–1 shows the observed frequencies for 51 words, selected as discussed later, and beneath them frequencies computed on the basis of the binomial distribution, where p is again estimated as the fraction of blocks in which the word occurred.

To the eye, most of the observed and fitted distributions agree rather well. Some exceptions are *his, one, only, than.* We thought it wise to compute for each word the binomial dispersion index (Hoel, 1954, pp. 175–177):

$$\chi^2 = \frac{\sum(x - \bar{x})^2}{\bar{x}[1 - (\bar{x}/4)]},$$

where the sum is over the 39 sets, the 4 is the number of blocks in a set, x is the count for a single set of four blocks, and $\bar{x}$ is the average number of blocks per set with occurrences.

The purpose of the index is to measure the agreement between observed set-to-set variation and the theoretical variability supplied by the binomial distribution. The denominator provides the theoretical estimate

$$\sigma^2 = npq \approx n\left(\frac{\bar{x}}{n}\right)\left(\frac{n - \bar{x}}{n}\right),$$

and the numerator provides the sum of squares that makes the distribution of the index, as an approximation, belong to a standard family of distributions much used in the statistical literature, the family of χ^2-distributions. The particular member of that family which is appropriate for this problem is the one designated by the expression "38 degrees of freedom," one degree of freedom less than the number of squares, 39, that are added. Fortunately, tables of the percentage points of these distributions are widely available, and we make use of them in what follows.

Instead of computing one χ^2 for one word, we examine the observed χ^2-values for the 51 words and compare the resulting distribution with the theoretical distribution for χ^2 which is broken into tenths in Table 2.3–2.

The theoretical distribution therefore has $\frac{51}{10} = 5.1$ values expected per interval. The observed distribution is reasonably consistent with this uniformity except for the upper tail, where quite a few words yield extremely high values of χ^2. Possibly, serious nonbinomiality—block-to-block dependence—is restricted

TABLE 2.3–1

HAMILTON'S INCIDENCE DISTRIBUTION FOR SETS OF 4 BLOCKS

Word		Blocks 0	1	2	3	4
all	observed	4	8	15	10	2
	binomial	2.6	10.1	14.6	9.4	2.3
also	observed	31	6	2	0	0
	binomial	29.9	8.2	.8	.0	.0
an	observed	1	2	13	13	10
	binomial	.4	3.2	10.7	15.9	8.8
any	observed	2	9	17	9	2
	binomial	2.4	9.8	14.6	9.8	2.4
at	observed	5	7	16	6	5
	binomial	2.6	10.1	14.6	9.4	2.2
been	observed	5	15	15	4	0
	binomial	6.5	14.7	12.4	4.7	.7
can	observed	9	12	5	12	1
	binomial	5.0	13.5	13.5	6.0	1.0
do	observed	27	12	0	0	0
	binomial	28.4	9.3	1.2	.1	.0
down	observed	35	4	0	0	0
	binomial	35.0	3.7	.2	.0	.0
even	observed	19	14	5	1	0
	binomial	18.5	15.2	4.7	.6	.0
every	observed	10	18	8	2	1
	binomial	10.5	16.3	9.5	2.5	.2
from	observed	0	3	14	17	5
	binomial	.6	4.4	12.1	15.0	7.0
had	observed	21	15	2	1	0
	binomial	21.3	13.9	3.4	.4	.0
has	observed	9	14	11	5	0
	binomial	7.9	15.5	11.4	3.8	.5
her	observed	35	3	1	0	0
	binomial	34.2	4.5	.2	.0	.0
his	observed	19	11	5	1	3
	binomial	13.7	16.4	7.3	1.5	.1

(*cont.*)

TABLE 2.3–1 (*cont.*)

Word		Blocks 0	1	2	3	4
if	observed	1	13	13	7	5
	binomial	2.2	9.4	14.6	10.1	2.6
into	observed	13	13	12	1	0
	binomial	11.7	16.4	8.7	2.0	.2
its	observed	5	7	12	9	6
	binomial	1.9	8.6	14.5	10.9	3.1
may	observed	2	10	13	12	2
	binomial	2.2	9.4	14.6	10.1	2.6
more	observed	8	11	11	8	1
	binomial	5.4	13.8	13.2	5.6	.9
must	observed	9	16	9	5	0
	binomial	8.8	15.9	10.7	3.2	.4
my	observed	37	2	0	0	0
	binomial	37.8	1.2	.0	.0	.0
no	observed	11	14	9	5	0
	binomial	9.9	16.2	9.9	2.7	.3
now	observed	32	5	2	0	0
	binomial	30.4	7.8	.7	.0	.0
on	observed	3	12	13	9	2
	binomial	4.7	13.1	13.7	6.3	1.1
one	observed	1	22	10	5	1
	binomial	5.4	13.8	13.2	5.6	.9
only	observed	18	10	8	3	0
	binomial	14.4	16.3	6.9	1.3	.1
or	observed	0	4	15	17	3
	binomial	.8	5.3	13.0	14.1	5.8
our	observed	22	12	4	1	0
	binomial	20.4	14.4	3.8	.4	.0
should	observed	7	14	13	4	1
	binomial	6.5	14.7	12.4	4.7	.6
so	observed	9	16	9	5	0
	binomial	8.8	15.9	10.7	3.2	.4
some	observed	18	13	8	0	0
	binomial	16.8	15.8	5.5	.9	.0

TABLE 2.3–1 (*cont.*)

Word		Blocks 0	1	2	3	4
such	observed	7	17	14	1	0
	binomial	8.8	15.9	10.7	3.2	.4
than	observed	7	14	5	10	3
	binomial	4.4	12.8	13.9	6.7	1.2
their	observed	0	9	19	8	3
	binomial	1.9	8.6	14.5	10.9	3.1
then	observed	31	6	2	0	0
	binomial	29.9	8.2	.8	.0	.0
there	observed	7	12	9	10	1
	binomial	4.7	13.1	13.7	6.3	1.1
things	observed	34	5	0	0	0
	binomial	34.2	4.5	.2	.0	.0
this	observed	1	1	5	15	17
	binomial	.1	1.1	6.4	16.2	15.2
unto	observed	39	0	0	0	0
	binomial	39	0	0	0	0
up	observed	31	6	2	0	0
	binomial	29.9	8.2	.8	.0	.0
upon	observed	3	11	14	11	0
	binomial	3.3	11.3	14.4	8.2	1.8
was	observed	18	14	6	0	1
	binomial	16.8	15.8	5.5	.9	.0
were	observed	19	14	5	1	0
	binomial	18.5	15.2	4.7	.6	.0
what	observed	15	16	6	2	0
	binomial	14.4	16.3	6.9	1.3	.1
when	observed	14	18	6	1	0
	binomial	15.2	16.2	6.4	1.1	.1
who	observed	10	14	9	6	0
	binomial	8.3	15.7	11.1	3.5	.4
will	observed	2	7	12	10	8
	binomial	1.9	8.6	14.5	10.9	3.1
would	observed	2	4	10	12	11
	binomial	.5	3.8	11.4	15.5	7.9
your	observed	39	0	0	0	0
	binomial	39	0	0	0	0

TABLE 2.3–2

χ^2-DISTRIBUTION IN SETS OF FOUR BLOCKS

$\chi^2_{(38)}$	Expected frequency	Observed frequency
0 –27.32	5.1	2
27.32–30.51	5.1	5
30.51–33.00	5.1	0
33.00–35.19	5.1	4
35.19–37.34	5.1	5
37.34–39.57	5.1	3
39.57–42.03	5.1	5
42.03–45.12	5.1	6
45.12–49.56	5.1	7
49.56–∞	5.1	12
		49 (Two words had no occurrences)

to a modest fraction of the words, perhaps 20 per cent. To make a more serious assessment would require building a model for nonbinomiality, an unnecessary step at this point. Here it is enough to see that for most words the binomial distribution gives a fair picture, but that it is not entirely adequate. There is some dependence, at least for some of the words. Results for Madison are similar, and we do not present them.

2.3B. Frequency of occurrence. Although the study of distributions for blocks is instructive, our main studies depend upon the counts of occurrences of words. By treating all blocks in all papers alike, we can get some notion of the appropriate family of distributions. It is pleasant to work with these blocks which are approximately equal in length, because the varied length of the papers has been a nagging backache in most of our analyses.

For our study of distributions of words we have chosen function words from the Miller-Newman-Friedman list (see Section 2.5A); we take the words with total frequencies between 45 and 180 in the 35,000 words of text that they counted.

The distribution of the number of occurrences for the 51 words chosen is shown in Table 2.3–3. For example, in the Hamilton text there are 45 blocks in which *this* did not occur, 80 in which *this* occurred exactly once, and so on. The order of display in the table was determined by the count in the cell for zero occurrences. With the notable exceptions of *her* and *his*, the frequency distributions appear orderly to the eye because the counts in the cells for high numbers of occurrences shrink as the count in the zero cell increases. For more detailed study, we select a few representative and unrepresentative words from this list: *an, from, any, may, upon, can, every, his, do, my.* These 10 words exemplify for Hamilton all the distributions, and include an exceptional word *his.*

Just as the binomial distribution set a base line in the study of incidence, the Poisson distribution corresponds to independence for counts of occurrences.

TABLE 2.3–3

DISTRIBUTION OF OCCURRENCES FOR FUNCTION WORDS OF CLASSES 4 AND 5

	Hamilton										
	0	1	2	3	4	5	6	7	8	9	14
this	45	80	71	39	8	4					
*an**	77	89	46	21	9	4	1				
or	86	69	49	21	9	6	6	1			
would	90	47	29	28	22	14	5	7	4	1	
*from**	93	82	51	13	5	2	1				
will	105	60	31	26	12	8	3	2			
its	116	82	29	11	4	3	2				
their	118	66	34	16	9	2	1	1			
if	118	87	31	7	3	1					
*any**	125	88	26	7		1					
*may**	128	67	32	14	4	1	1				
*upon**	129	83	20	9	5	1					
at	129	70	42	4	2						
all	132	67	32	13	2	1					
there	138	75	23	8	2	1					
been	138	65	24	14	3	3					
than	143	66	29	6	2	1					
on	145	67	27	7	1						
one	149	77	16	3	2						
more	152	70	15	7	2	1					
*can**	157	60	20	5	2	2	1				
has	157	57	20	11	2						
should	161	58	26	2							
who	163	53	25	3	2	1					
no	167	60	11	8	1						
so	170	55	19	2	1						
such	173	56	13	5							
must	173	49	14	9	1		1				
into	183	50	12	2							
only	185	54	7	1							
*every**	186	46	14	1							
what	188	44	11	3			1				
was	192	42	7	2	3	1					
were	194	44	7	1	1						
when	195	40	9	3							
had	200	35	8	4							
some	200	38	9								
even	204	39	4								
*his**	192	18	17	7	3	2	4	1	2		1
our	212	23	9	1		1	1				
*do**	228	16	2	1							
then	230	17									
up	231	14	2								
also	232	15									
now	234	13									
things	236	11									
down	240	6	1								
*my**	241	6									
her	241	3	1	1						1	
unto	247										
your	247										

(cont.)

TABLE 2.3–3 (*cont.*)

	Madison									
	0	1	2	3	4	5	6	7	8	9	14
on	63	80	55	32	20	8	4				
this	80	83	71	21	6	1					
*from**	90	93	42	17	8	9	3				
or	103	74	51	16	12	3	2	1			
*an**	122	77	40	14	8		1				
was	129	55	32	16	20	5	3		2		
been	132	74	35	12	4	5					
at	133	86	31	8	1	3					
*any**	145	90	19	8							
all	146	68	28	16	4						
*may**	156	63	29	8	4	1	1				
than	157	84	19	2							
its	158	66	29	7	2						
more	158	68	30	4	2						
their	159	71	17	12	2		1				
no	165	71	16	7	2	1					
has	167	64	19	9	1	1	1				
would	167	49	31	9	1	3		2			
such	170	66	22	3	1						
if	171	60	22	6	3						
one	172	66	15	5	3			1			
will	172	55	19	8	3	3		2			
into	178	71	11	2							
were	179	58	18	5	1	1					
only	182	65	11	4							
so	182	59	17	3	1						
had	185	52	16	6		2	1				
her	200	26	10	9	4	4	3	2	1	2	1
must	202	47	10	2	1						
some	205	53	4								
there	208	47	4	2	1						
*every**	209	43	4	6							
*can**	211	44	6	1							
*his**	213	21	9	11	2	2	1	2	1		
should	218	32	12								
then	219	34	8	1							
what	220	33	8	1							
even	221	38	3								
also	222	36	4								
who	227	28	5	2							
now	236	24	2								
when	236	23	2	1							
*do**	244	15	2	1							
our	245	13	4								
down	246	15	1								
up	254	8									
*upon**	254	7	1								
things	254	7	1								
*my**	259	3									
unto	262										
your	262										

Note: Words marked * are studied further in Table 2.3–4.

This distribution is a fair approximation for many kinds of counts, such as the number of radioactive particles striking a Geiger counter in a short interval, accidents at a given place for comparable times, wire worms per plot, blood corpuscles per square, the number of men on base when a home run is hit, and the number of vacancies in the United States Supreme Court in a given year. Usually, an indefinitely large number of occurrences is possible when the Poisson distribution is applied. But when the average number of occurrences is small, this limitation may not matter, as in the last two examples above.

If we know the average number of occurrences, we know all about the Poisson distribution. For example, suppose that a word occurs, on the average, twice per block; then the Poisson distribution gives the following probability table of occurrences.

Number of occurrences	0	1	2	3	4	5	6 or more
Probability	.135	.271	.271	.180	.090	.036	.017

More generally, if λ is the average number of occurrences, then the probability of exactly x occurrences for a given block is

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots,$$

where $x! = x(x-1)\cdots 2 \cdot 1$ and $e(=2.718\ldots)$ is the base of the natural logarithms. We can estimate λ from the average number of occurrences per block for a given word, then we can compute the probabilities from tables (for example, Molina, 1942), and multiply the results by the number of blocks to make direct comparisons of frequencies.

In Table 2.3–4 we carry out this program for 10 words which typify the distributions shown in Table 2.3–3. (The third distribution is the fitted negative binomial discussed later.)

While the fitted Poisson and the observed distribution agree fairly well for some of the words, for example, *an, from, any, upon, can, every, do, my*, even the most motherly eye can scarcely make twins of the distributions for *may* or those for *his*. Thus, we are forced to revise our notions of the form of the distribution. The Poisson family is not rich enough to describe the variation in the frequencies.

Like Ulysses, statisticians are never at a loss, and naturally we have several other distributions as candidates. Some of these, called *contagious distributions*, are designed to produce clumping of occurrences. While we have explored some of these, we found few that are mathematically manageable.

We call the parts of a frequency distribution far from the mean "the tails of the distribution." In the Poisson distribution, the right-hand tail is long but

TABLE 2.3–4

OBSERVED AND FITTED POISSON AND NEGATIVE BINOMIAL DISTRIBUTIONS FOR SELECTED WORDS

		Hamilton Occurrences							
		0	1	2	3	4	5	6	7 or more
	observed	77	89	46	21	9	4	1	
an	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
	observed	93	82	51	13	5	2	1	
from	Poisson	90.5	90.9	45.6	15.3	3.8	.8	.1	
	N.B. (7)	93.2	84.7	44.2	17.2	5.6	1.6	.4	.1
	observed	125	88	26	7	0	1		
any	Poisson	126.3	84.6	28.5	6.4	1.1	.2		
	N.B. (∞)	same as Poisson							
	observed	128	67	32	14	4	1	1	
may	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
	observed	129	83	20	9	5	1		
upon	Poisson	121.6	86.1	30.6	7.3	1.3	.2		
	N.B. (4.20)	131.1	77.1	27.9	8.2	2.1	.5	.1	
	observed	157	60	20	5	2	2	1	
can	Poisson	141.0	78.9	22.3	4.2	.6	.1	.0	
	N.B. (1.06)	157.3	57.6	20.7	7.3	2.6	.9	.3	.3
	observed	186	46	14	1				
every	Poisson	180.9	56.2	8.8	.9	.1			
	N.B. (1.58)	185.5	48.1	10.4	2.2	.3	.5†		
	observed	192	18	17	7	3	2	4	4
his	Poisson	131.7	82.7	26.2	5.5	.9	.1	.0	.0
	N.B. (.154)	192.2	23.8	11.0	6.4	4.0	2.7	1.9	5.0
	observed	228	16	2	1				
do	Poisson	225.7	20.3	.9	.0				
	N.B. (.21)	228.7	14.8	2.7	.7				
	observed	241	6						
my	Poisson	241.1	5.8	.1					
	N.B. (∞)	same as Poisson							

Note: Estimated values of κ follow N.B. in parentheses; an * means κ was estimated from the variance. A † means sum of remaining frequencies.

TABLE 2.3–4 (*cont.*)

OBSERVED AND FITTED POISSON AND NEGATIVE BINOMIAL DISTRIBUTIONS FOR SELECTED WORDS

		Madison Occurrences							
		0	1	2	3	4	5	6	7 or more
	observed	122	77	40	14	8	0	1	
an	Poisson	106.1	95.9	43.4	13.1	2.9	.5	.1	
	*N.B. (2.45)	121.8	80.2	37.2	14.9	5.4	1.9	.6	
	observed	90	93	42	17	8	9	3	
from	Poisson	76.6	94.3	57.9	23.7	7.3	1.8	.4	
	N.B. (2.62)	95.9	79.9	45.8	22.5	10.0	4.2	1.7	2.0
	observed	145	90	19	8				
any	Poisson	146.8	84.9	24.7	4.8	.7	.1		
	N.B. (∞)	same as Poisson							
	observed	156	63	29	8	4	1	1	
may	Poisson	136.1	88.9	29.2	6.4	1.1	.2	.0	
	N.B. (1.15)	157.7	63.9	25.2	9.4	3.7	1.3	.5	.3
	observed	254	7	1					
upon	Poisson	253.2	8.6	.2					
	N.B. (.11)	254.4	6.6	.9	.1				
	observed	211	44	6	1				
can	Poisson	209.2	47.1	5.3	.4				
	N.B. (2.6)	210.9	43.9	6.6	.6				
	observed	209	43	4	6				
every	Poisson	201.4	53.0	7.0	.6				
	N.B. (.695)	209.6	39.8	9.2	2.3	1.1†			
	observed	213	21	9	11	2	2	1	3
his	Poisson	167.3	74.9	17.0	2.6	.3	.0	.0	.0
	N.B. (.149)	222.7	14.7	8.4	5.2	3.4	2.4	1.3	3.9
	observed	244	15	2	1				
do	Poisson	240.9	20.2	.9	.0				
	N.B. (.16)	245.0	13.5	2.7	.7				
	observed	259	3						
my	Poisson	259.0	3.0						
	N.B. (∞)	same as Poisson							

Note: Estimated values of κ follow N.B. in parentheses; an * means κ was estimated from the variance. A † means sum of remaining frequencies.

OBSERVE

observe
an Poisson
*N.B. (

observ
from Poisso
N.B. (

observ
any Poisso
N.B.

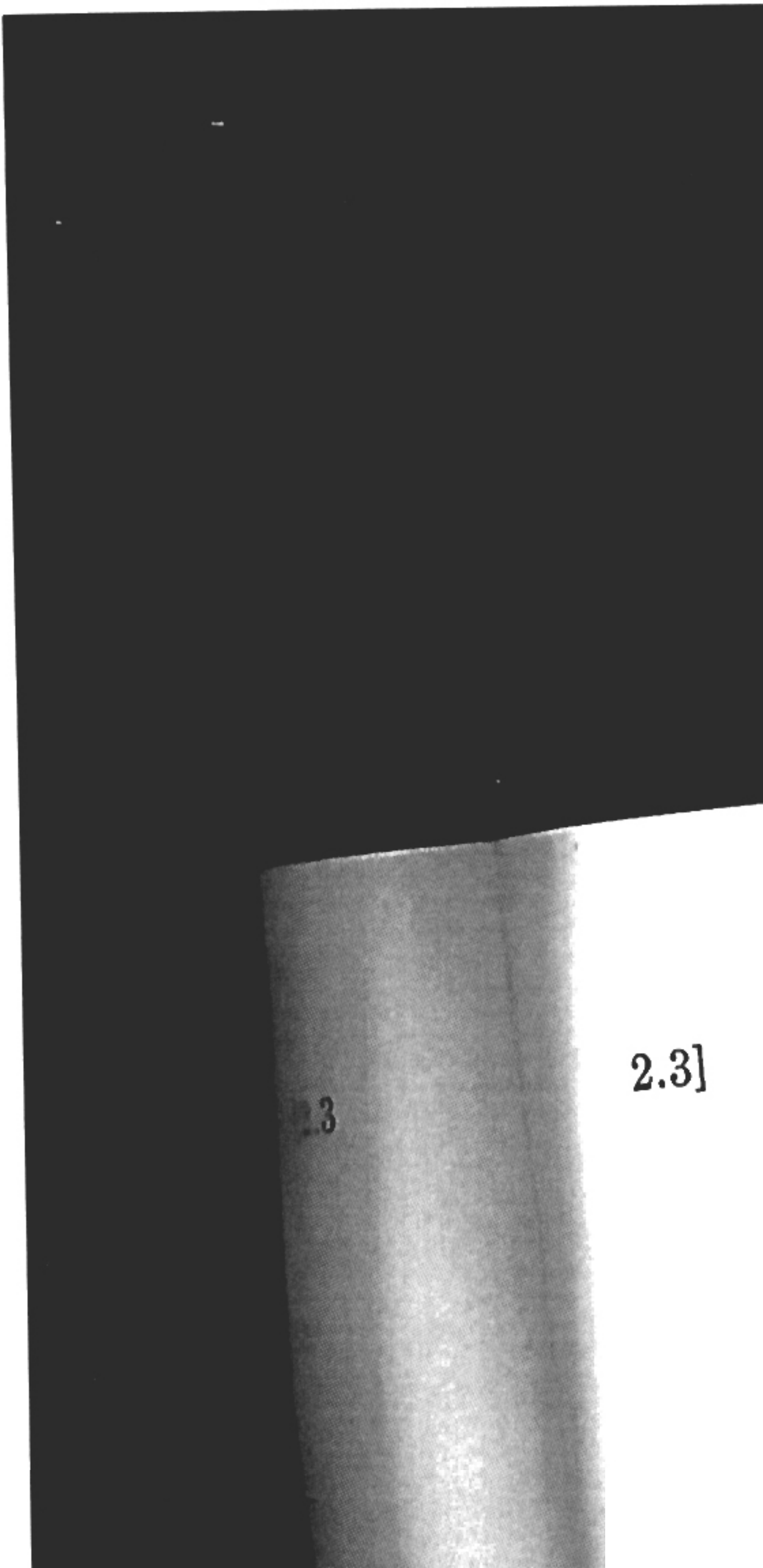

2.3]
2.3

TABLE 2.3–5

A COMPARISON OF NEGATIVE BINOMIAL AND POISSON DISTRIBUTIONS

Count x	Negative binomial probability $\lambda = \kappa = 2$	Poisson probability $\lambda = 2$
0	.250	.135
1	.250	.271
2	.188	.271
3	.125	.180
4	.078	.090
5	.047	.036
6	.027	.012
7	.016	.003
8	.009	.001
9 and beyond	.010	.000+

thin. What we need to describe words like *may* and *his* is a family that offers a fatter tail than the Poisson. One such family is the *negative binomial.* A variety of mechanisms can lead to the negative binomial, but one which we present informally here (and formally in Section 4.1) is a two-stage sampling operation. Suppose that, before writing a block, the author chooses a rate for that block, and then behaves according to the Poisson theory for that rate and block. For the next block he chooses another rate and so on. If he chooses his rates randomly in a way that we will explain later, he winds up with the negative binomial. Without being quite so fanciful, we can say that the negative binomial offers one way to introduce a greater block-to-block variation of rates than the Poisson can indicate.

THEORY. As in the Poisson distribution, let λ be the mean frequency of use for blocks of fixed size, and let κ be another number ($\kappa > 0$); then for the negative binomial the probability of exactly x occurrences is

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

When $x = 0$, the coefficient on the right is to be taken as unity.

To illustrate the relation with the Poisson distribution, let $\kappa = 2$, $\lambda = 2$; then the formula reduces to

$$P(x) = (x+1)/2^{x+2}, \qquad x = 0, 1, 2, \ldots,$$

and we display the numerical result in Table 2.3–5.

The main point of Table 2.3–5 is that for large values of x (beyond 4), the negative binomial has substantially larger probabilities than the Poisson with the same mean. It can be shown that if κ is very large, the negative binomial closely approximates the Poisson and that when κ is small, it does not. We could take $1/\kappa$ as a measure of non-Poissonness, but we actually use λ/κ for reasons we give in Chapter 4.

We estimate λ from the average of the observed distribution, and κ by more grisly devices (from mode or variance) and then fit the distributions. Some are displayed in Table 2.3–4. We find it heartwarming to see the improvement in agreement, especially for *may* and *his* for Hamilton, and *an*, *from*, *may*, and *his* for Madison when the Poisson is replaced by the negative binomial.

2.4. CORRELATIONS BETWEEN RATES FOR DIFFERENT WORDS

Although we said earlier that we would set aside the fine structure of diction, we must attend to the gross relations among words. In our later work we largely behave as if words were independent, and then inquire what steps to take to correct for the failure of that assumption. How nearly independent or uncorrelated are rates of use of words of the kind we ultimately employ for discriminations?

If we make repeated counts of the makes of 100 successive automobiles seen on the road, we will find that when the count of *Ford*'s is high, that of *Chevrolet*'s is lower on the average, and vice versa. The restraint that the total counts be of fixed size creates a negative correlation between the counts in different categories. This correlation* reaches its extreme for two categories, where, for example, the correlation between the number of sons and the number of daughters in families of size five is exactly -1. That is, if you know the number of sons you can state exactly the number of daughters. This negative multinomial correlation is somewhat artificial and arises even if the category that occurs on one trial is utterly independent of that on another.

Given such independence of trials, we can compute the theoretical correlation coefficient from the true proportions belonging to the various categories. Let p_i $(= 1 - q_i)$, $i = 1, \ldots, k$ be the proportion associated with the ith category. Then the correlation between the observed counts in categories i and j is

$$\rho_{ij} = -\sqrt{p_i p_j / q_i q_j}.$$

* A coefficient of correlation is a way of measuring the degree of relation between two variables, or more properly, the degree of linear relation. When the coefficient is $+1$, the relation between the two is perfect—indeed a straight-line graph relates them. The same is true when the coefficient is -1, except that then when one variable is high, the other is low. Independence implies zero correlation, and often small correlations imply lack of predictability of the value of one variable from the other.

TABLE 2.4–1

OBSERVED DISTRIBUTION OF CORRELATION COEFFICIENTS

Midpoint of class interval	Frequency for Hamilton	Frequency for Madison
—.425	1	1
—.375	3	1
—.325	3	2
—.275	7	9
—.225	12	12
—.175	32	31
—.125	54	70
—.075	56	72
—.025	66	56
.025	39	53
.075	34	46
.125	34	30
.175	22	11
.225	19	20
.275	9	8
.325	3	5
.375	7	1
.425	3	3
.475	1	3
.525	0	1
.575	1	0
Total	406	435
$\bar{r}$	—.0026	—.0120
S.D.	.1547	.1450

When $p_i = p_j = q_i = q_j = \frac{1}{2}$, $\rho_{ij} = -1$ as mentioned earlier. Among the words we use, *of* and *to* have the largest proportions, about .06 and .04, which would create a correlation of about —.05 between them. Other correlations from this multinomial source are considerably smaller. Since correlations from the multinomial source seem negligible, the correlations that we attend to are those created by actual affiliations or substitutabilities between words. To assess the magnitudes of these correlations, the Pearson correlation coefficient was computed for rates of all pairs of the 30 words used in the main study of Chapter 3 for each author, using 48 Hamilton papers and 50 Madison papers as the basis for the rates. Miles Davis made the calculations and Ivor Francis provided an analysis of the results.

The frequency distributions for 406 Hamilton and 435 Madison correlations are displayed in Table 2.4–1 (29 correlations are missing from Hamilton because of his failure ever to use one of the words in this text). The mean correlation,

$\bar{r}$, for both authors is slightly negative (about as expected), and for each the observed standard deviation is about .15, which is close to $1/\sqrt{n} \approx \frac{1}{7} \approx .14$, the standard deviation that a theoretical correlation of zero would imply. Thus most of the true correlations are probably slightly negative but near zero, and the deviations that we observe are mainly sampling variation. For very low-frequency words, the correlations are likely not very meaningful. For words that have an average rate of over $\frac{1}{1000}$ for each author, the average correlations for the two authors that exceed .15 in absolute value (about 1.5 standard deviations from the average) are:

an–there,	.20,	*of–to*,	−.39,
an–this,	.19,	*on–to*,	−.19.
by–this,	.24,		

While some interpretations suggest themselves, the main point is that these numbers are quite modest. Later, in Sections 3.7 and 4.7, we use correlations to adjust our appraisals.

2.5. POOLS OF WORDS

The reader can abandon the notion that our study is a unified whole, conceived carefully in advance, and executed as planned. Instead, we have made many studies, some contributing to all the others, some utter failures, abandoned early on. Most dealt with the search for words that especially help us to discriminate. Here we summarize those studies that contributed the words finally used.

Three pools of words produced four sets of possible discriminators, and Table 2.5–1 may help the reader prepare for their discussion.

TABLE 2.5–1

POOLS OF WORDS AND THEIR OUTPUTS

Miller-Newman-Friedman list of 363 function words	Screening study based on all different words (about 3000) in 11 *Federalist* papers	Index with frequencies based on some Hamilton *Federalist* papers outside the screening study and some Madison non-*Federalist* writings
produced	produced	produced
70 "unselected" high-frequency words and 20 random low-frequency words	28 selected words	103 selected words
which total to 165 different words		

TABLE 2.5–2

FUNCTION WORDS AND THEIR CODE NUMBERS

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

TABLE 2.5–3

ADDITIONAL WORDS AND THEIR CODE NUMBERS

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

Note: An * means the word emerged from the screening study; the rest came from a random sample of low-frequency function words.

2.5A. The function words. We obtained a pool of words by using a list of "function" words, made up earlier, for a different purpose, by Miller, Newman, and Friedman (1958). Their list of 363 words is an expansion of a much smaller list compiled by Fries (1952). Their list gives frequency counts based upon 35,000 words of text taken from the King James *Bible*, William James, and *The Atlantic* (1957). Though not directly relevant to the *Federalist* period, the counts have been helpful, but many words are so rare that they are of no value for our study. Function words, as opposed to "content" words, include, in addition to the filler words already mentioned, ordinals, cardinals, and some others. The list is objective with respect to the *Federalist* problem; thus it relieves us of a large onus of choice and plays a central role in the present work.

From the Miller-Newman-Friedman list, we eliminated some biblical words, cardinals larger than one, and personal pronouns except in the possessive form. The main point is that the selection of the remaining words has nothing to do with their ability to decide authorship in our problem. In the present work, the 70 most frequent words in the Miller-Newman-Friedman text have been used. This is the first set of Table 2.5–1. The list, shown in Table 2.5–2, is not totally satisfactory because certain types of function words are potentially dangerous. Personal pronouns and auxiliary verbs, especially with respect to mood and tense, are likely to be related to external details, and inference from them is difficult.

We added 20 words for purposes of estimation: a random sample of 10 from words with counts 11 to 21 and another 10 from words with counts 22 to 44 in the Miller-Newman-Friedman list. These form a set of low-frequency function words unselected for ability to decide authorship. They are shown in Table 2.5–3.

2.5B. Initial screening study. The study has been described sketchily in Chapter 1, and some results of it are given there. Here we give the rules followed and the words selected. During this screening, we thought that contexts were important, and we eliminated words by the hundreds for their potential dependence on context. Still, our attitude toward this danger was not as serious and quantitative as it later became. Consequently, we believe that our studies, using the words selected from this pool, still suffer more from contexts than they would have with more care.

Our plan was to explore low-frequency words separately from middle- and high-frequency ones (definitions were vague). The present description applies to that low-frequency study.

General plan. Papers were studied in waves. A wave consisted of about 10 papers, nearly equally divided between the authors. Words were scored on a wave according to the number of Hamilton and Madison papers in which the word occurred. Thus, (3,2) means that the word occurred in three Hamilton and two Madison papers in the set. If a word scored (5,0) or (0,5) we would be most

TABLE 2.5–4

SURVIVORS OF THE LOW-FREQUENCY SCREENING STUDY

Hamilton markers (H, M)		Madison markers (H, M)	
(14, 0)	*enough*	(0, 13)	*whilst*
(10, 0)	*while*	(2, 13)	*consequently*
(8, 0)	*destruction*	(0, 8)	*although*
(8, 0)	*offensive*	(1, 9)	*violate* + *s* + *d* + *ing*
(10, 1)	*affect* + *ed*	(3, 12)	*pass* + *es* + *ed* + *ing*
(9, 1)	*commonly*	(1, 8)	*voice*
(9, 1)	*vigor* + *ous*	(1, 8)	*throughout*
(6, 0)	*city* + *cities*	(2, 10)	*language*
(6, 0)	*contribute*	(0, 5)	*fortune* + *s*
(6, 0)	*defensive*	(0, 5)	*join* + *ed*
(8, 1)	*direction*	(0, 5)	*violence*
(5, 0)	*disgracing*	(1, 7)	*again*
(5, 0)	*rapid*	(1, 7)	*function* + *s*
(13, 4)	*considerable* + *ly*	(1, 7)	*innovation* + *s*

Note: Incidences in Hamilton and Madison papers (H, M).

encouraged, but (3,2) or (3,4) would be discouraging. Strong marker words like *while* and *whilst* were desired.

At the end of a wave, a word was retained or discarded on the basis of the cumulative score to date. In all, there were four waves (including the Exterior set) as described in Chapter 1.

Details of study. More specifically, we began with 6 Hamilton and 5 Madison *Federalists*, and we retained words that scored (2,0), (0,2), (3,0), (0,3), (4,0), (0,4), (5,0), (0,5), (6,0), (4,1), (1,4), (6,1), (5,1), (1,5), (6,2), (2,5), and discarded all others. About 3000 different words were in the original pool of 11 papers, and 305 words survived the first screening (wave A).

Because Madison had so few known *Federalists*, we feared to use them up early, and so we obtained as our next wave an Exterior set of 10 papers. The 305 words that survived the wave-A screening were scored in the same manner in this Exterior set, and the cumulative score based on these first 21 papers was obtained for each word. For example, *argument*, pooled with its plural, scored (5,1) on wave A, and (1,3) on the Exterior set for a total of (6,4), but *upon* scored (6,2) and (5,1) for an (11,3) total. The latter word was earmarked for special study because of its high Hamilton rate. We did not eliminate any words at this point, but proceeded to wave B.

Wave B had 10 papers (6 Hamilton and 4 Madison). The cumulative score for each word described above was further augmented by adding on its score based on these ten papers. The cumulative score is now a number pair (x,y), where x can vary from 0 through 17, and y from 0 through 14.

TABLE 2.5–5

WORDS THAT SURVIVED UNTIL WAVE C, BUT FAILED THERE

Hamilton markers	Score before wave C	Wave C score	Final score
criterion	(4, 0)	(0, 1)	(4, 1)
finance + *s*	(4, 0)	(1, 1)	(5, 1)
hostility	(7, 1)	(0, 1)	(7, 2)
intimate + *d*	(4, 0)	(1, 1)	(5, 1)
occasional	(4, 0)	(0, 0)	(4, 0)
pervert	(4, 0)	(0, 1)	(4, 1)
probability	(6, 1)	(4, 2)	(10, 3)
utility	(7, 1)	(0, 1)	(7, 2)
utmost	(4, 0)	(0, 0)	(4, 0)
wide	(4, 0)	(0, 0)	(4, 0)
Madison markers			
commensurate	(0, 4)	(0, 0)	(0, 4)
death	(0, 4)	(0, 0)	(0, 4)
ensues + *ing*	(0, 4)	(0, 0)	(0, 4)
expression	(1, 6)	(2, 1)	(3, 7)
face	(0, 4)	(1, 1)	(1, 5)
fundamental	(0, 5)	(2, 0)	(2, 5)
probably	(2, 8)	(2, 3)	(4, 11)
work + *s*	(1, 8)	(1, 0)	(2, 8)

At this point, the index $z = (x - y)^2/(x + y)$ was introduced as a measure of the discrimination achieved by a word. We decided to retain only those words for which $z \geq 3.6$. For Hamilton markers, this means that $(x,0)$ words are kept if x is 4 or more, $(x,1)$ are kept if x is 6 or more, $(x,2)$ are kept if x is 8 or more, $(x,3)$ are kept if x is 10 or more. Similar cutoffs hold for y in Madison markers, with scores $(0,y)$, $(1,y)$, $(2,y)$, and $(3,y)$.

These machinations left us 46 words to test on wave C (6 Hamilton and 5 Madison papers). We retained those words with final z-values of at least 4.5. This is the third set of words mentioned in Table 2.5–1. These 28 words are shown in Table 2.5–3 together with their code numbers, and also in Table 2.5–4, ordered according to their final z-values. That there are equal numbers of Hamilton and Madison markers is quite accidental.

Words that lost out in the final screening are shown in Table 2.5–5, together with their scores. For us, it was a sad personal blow that *probability* and *probably* failed this final test.

Someone may well ask at this point whether the whole enterprise is a boondoggle. Is it not reasonable that none of the words has predictive power and that we are merely keeping a chance selection of words? Possibly all these words

are used with approximately equal incidences by the two authors, and we have chosen only those extreme words that would occur by chance in any such large collection. To answer such an inquiry, suppose a word has an equal chance for incidence for both authors. There are 6 Hamilton and 5 Madison papers in wave C. Then in the (x,y) pairs for these 11 papers of wave C, the quantities $(x/6) - (y/5)$, or equivalently $5x - 6y$, would be distributed about zero with no dependence on the forecasting power of the information in the previous waves. To test this, we made two frequency distributions for $5x - 6y$, one for 24 Hamilton markers, the other for 22 Madison markers entering wave C, with the following results.

Center of class interval: $5x - 6y$

	−25	−20	−15	−10	−5	0	5	10	15	20	25	Total
Hamilton					4	5	6	3	4	1	1	24
Madison	1	2	3	2	5	6	2	1				22

Visual inspection shows that the markers are discriminating. We conclude then that the discriminators have some power.

2.5C. Word index with frequencies. Late in our work we were able to construct an index of Hamilton and Madison words covering 18 Hamilton *Federalists* and 19 Madison papers. (Hamilton: 16, 23, 25, 27–35, 60, 61, 68, 69, 75, 76; Madison: Helvidius 4, 5, M1 through M9, N1 through N8, *Neutral Trade* 201–209.) In the index, five counts were given for each word: total frequency for the two authors, frequency for each author, and incidence in papers for each.

To obtain this large list of words became feasible only when the alphabetical lists and counts from many papers could be merged into a single listing. This merging was accomplished on papers totaling about 70,000 words, divided nearly equally between Hamilton and Madison. The merging could be done conveniently only for those papers whose initial alphabetization was available on a tape output. The only Madison papers included are external to the *Federalist*, while the Hamilton papers were *Federalist* papers, but not the wave-A, -B, or -C papers used in the screening study.

What was available, then, was an index restricted to words actually used by Hamilton and Madison in a substantial body of writings. The listings contained about 6700 different words, where varying forms (for example, plurals, verb conjugates) account for much of the total.

We analyzed the index to find out whether there were any outstanding words that we had missed in the screening study. Naturally the screening study can easily miss a fine discriminator that does not do well on the starting set (wave A). On binomial probability paper (Mosteller and Tukey, 1949) we set three standard-deviation confidence limits on the Hamilton-Madison split in frequencies.

TABLE 2.5–6

NEW WORDS FROM THE INDEX STUDY AND THEIR CODE NUMBERS

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

If a word fell outside these limits, it was noted as a possible discriminator. The technique ignores the negative binomiality of words. The resulting list had about 240 words, and a great many of them were quite contextual. After eliminating words that we regarded as contextual, we were left with 103 words. Forty-eight were new, and these are listed in Table 2.5–6, together with their code numbers. These new words are used in the main Bayesian study (Chapter 3) and in the robust Bayesian study (Chapter 6).

2.6. WORD COUNTS AND THEIR ACCURACIES

Our counts have been done partly on high-speed computers (machine counts) and partly by hand. In this section we describe that work.

Machine counts. Certain *Federalist* papers were typed on punched cards by personnel at the Littauer Statistical Center of Harvard University. A program for the counting of frequencies of single words was prepared by Wayne Wiitanen, C. Harvey Willson, and Robert A. Hoodes, under the direction of Albert E. Bea-

```
    THE SUPPOSITION THAT EACH CONFEDERACY INTO WHICH THE STATES WOULD   13   037
BE LIKELY TO BE DIVIDED WOULD REQUIRE A GOVERNMENT NOT LESS             13   038
COMPREHENSIVE THAN THE ONE PROPOSED, WILL BE STRENGTHENED BY ANOTHER    13   039
SUPPOSITION, MORE PROBABLE THAN THAT WHICH PRESENTS US WITH THREE       13   040
CONFEDERACIES AS THE ALTERNATIVE TO A GENERAL UNION.  IF WE ATTEND      13   041

LESS        5  ( 3, 60, 63), (19, 43, 46), (38, 56, 59), (71, 22, 25), (84, 5, 8),
LIBERTY     1  (85, 14, 20),
LIGHT       1  (74, 59, 63),
LIKE        1  (31, 26, 29),
LIKELIHOOD  1  (67, 62, 71),
LIKELY      1  (38,  4,  9),
LINKS       1  (46, 26, 30),
```

FIG. 2.6–1. Portion of machine output for Hamilton's *Federalist* No. 13.

ton. The first run was proofed and corrected. Another run on each paper was made and a second proofing done. This corrected run is discussed here.

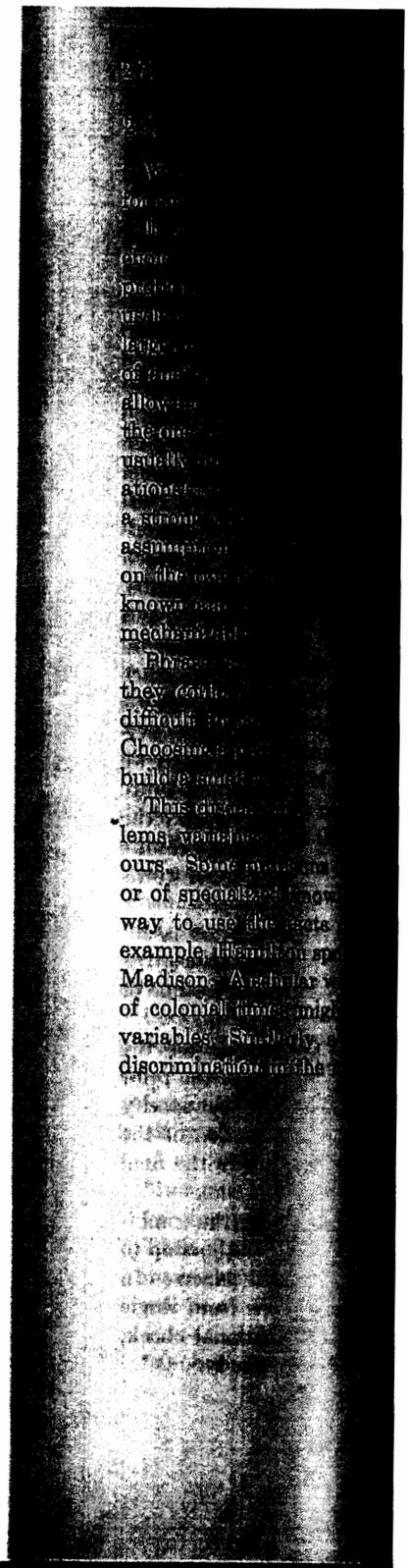
For each separate paper in the final run, the total frequency of every word was tabulated separately, but a count for the entire paper was summed on hand machines.

Figure 2.6–1 illustrates parts of a page of machine output for Hamilton's paper No. 13. Each line at the top corresponds to a single card input. The text also illustrates the style of writing.

Following each line of text are the paper number and the line number (essentially punched card number). After the words are alphabetized they are counted, and then each appearance is listed as a number triple (x, y, z), where x is the line number and y and z are the position of first and last letters of the word on that line. For example, *likely* appears in line 038 from characters 4 through 9.

Hand counts. Certain *Federalist* papers were typed on roll paper (adding-machine paper, one word to a line) and proofed. At the same time, the page number and line number were written opposite each word (for some but not all papers). The words on the roll paper were then cut and sorted into alphabetical order. (During this operation a deep breath created a storm of confetti and a permanent enemy.) The count of each word was tabulated by hand and typed or written on a master sheet. A complete count for the total number of words in the particular paper was also tabulated and recorded. This total count was later checked and rechecked (but only a few words were recounted).

The disputed papers were counted both by hand and by machine. This gave an opportunity to compare the results. A reconciliation of the hand counts and of the machine counts for the disputed papers was carried out by Theodore S. Ingalls, Ralph A. Stewart, Jr., and Cleo Youtz. It has been reported in Mosteller and Wallace (1962).



2.7. CONCLUDING REMARKS

We close with some remarks about other sorts of variables and considerations for similar studies.

It would be ridiculous to claim that words are the only or even the best choice for a set of variables. But for hope of success in a difficult discrimination problem, requirements like the following are needed, though no rules can make useless variables discriminate. The pool of potential discriminators should be large enough, say, 50 to 1000, to offer a good chance of success. For some kinds of analysis, the pool needs to be delimited by systematic rules so that proper allowance can be made for selection. For when many variables are involved, the ones that appear to be best in a sample may not actually be the best; they usually do not work so well later as they appeared to initially. These deteriorations are called selection and regression effects. Enough data are needed to allow a strong grip on the distribution theory for the variables used. Distributional assumptions on each variable should be acceptable to those of divergent views on the question of authorship. Finally, measurement of the variables on the known and disputed papers should be objective and routine, and, preferably, mechanizable.

Phrases are attractive, and, were it possible to choose a moderately large list, they could be handled much as words. The measurement problem is more difficult to carry out, and the available data would be much less adequate. Choosing a pool of phrases is not easy. Possibly, the index could be used to help build a small study of phrases, though the problems are substantial.

This discussion is not meant to discourage others from using, in such problems, variables that they feel have deeper or more appropriate significance than ours. Some problems of authorship can take advantage of facts about authors, or of specialized knowledge. Indeed, we are disappointed not to have found a way to use the facts of Hamilton's and Madison's differing childhoods. For example, Hamilton spoke French fluently as a youth, but we are not clear about Madison. A scholar who is well equipped in the study of the regional languages of colonial times might know how to look for, and might find, some striking variables. Similarly, a psychiatrist, clinical psychologist, or poet might discover discrimination in the imagery of the language.

CHAPTER 3

The Main Study

In this chapter, we present the methods and results of the main study. To simplify the exposition, we describe the methods only for the simpler model based on Poisson distributions of word frequencies but we give the numerical results for both the Poisson and the negative binomial. The technical development for the full model based on negative binomial distributions is postponed to Chapter 4, along with detailed mathematical treatments of special problems.

Section 3.1 begins with an overview of the methods of the main study, and describes the principal sources of difficulties and how we go about meeting them. Starting in Section 3.1A, we develop the procedures for assessing the evidence on authorship of a disputed paper; first, we derive Bayes' theorem in its simplest form for assessing the evidence of a single word, then use it to combine evidence from several words and to guide in the selection of words. Second, we treat the relation of the statistical evidence and prior evidence on authorship. With the completion of this much of the development in Section 3.1C, the reader without interest in the statistical methodology can turn to the results for *The Federalist* in Section 3.4, although a glance at the final selection of words described in Section 3.3 would be desirable.

With Section 3.1D, we begin to treat one of the central methodological issues —unknown parameters. Here we indicate the final method, leaving most of the intermediate steps for Section 3.2. We end Section 3.1 with a formal outline of the logic of our two kinds of uses of Bayes' theorem.

Section 3.2 treats the problem of getting posterior distributions for the parameters, with the major effort spent on the nature and source of the prior distribution. We introduce here the notion of underlying constants that specify which prior distribution is being used. We complete the presentation of the method in Section 3.3 with the choice of the 30 words that are used in the final analysis of *The Federalist*.

We present the results, in the form of log odds of authorship for the combination of all words, in Section 3.4, first for the papers of known authorship to provide a check on the method, then for the disputed *Federalist* papers. In Section 3.5, the log odds are broken down into the contributions from single words, and from groups of words. Section 3.6 presents an additional check, resorting to some papers by Hamilton not previously used in the study.

Section 3.7 serves several important functions. It describes the general magnitudes of adjustments in the final log odds needed because of imperfect assumptions and approximations. The range of adjusted log odds is given in Section 3.7E. In Section 3.7F, we discuss the question, "Can the odds be believed?," both within the mathematical model and with respect to difficulties external to the model.

Probability has several interpretations. We must mention two: probability as degree of belief, and probability as relative frequency. In the relative frequency interpretation, probability may be applied only to events that can be repeated over and over under much the same conditions. The degree-of-belief interpretation is more widely applicable, but specifying the probabilities is often difficult in the absence of conditions of symmetry or long-run relative frequencies. When the two interpretations are simultaneously applicable, the same numerical values would normally be assigned in each interpretation.

Much of what we do in the main study concerns propositions that would not ordinarily be assigned frequency probabilities; for example, the proposition "Hamilton wrote paper No. 52." To be able to use a concept of probability for such a proposition greatly facilitates the analysis of data and the interpretation of evidence, because we can use techniques derived from Bayes' theorem, which is discussed below. To achieve a notion of probability that can also be given a frequency interpretation is highly desirable, for it strengthens the interpretation and widens the acceptability of the results. In ways that are explained much later, our work comes close to merging the two interpretations.

Right here, we need to discuss these ideas a bit further. In the course of this discussion, Bayes' theorem is often mentioned. One need only know that it is a mathematical device that combines evidence from data with prior information. The reservations that people have had about using Bayes' theorem concern the meaningfulness and source of some of the probabilities needed for its application. Its correctness has never been in dispute. However, we warn readers who are well informed in probability that our approach to a Bayesian analysis may differ considerably from their preconceptions.

Coolidge (1925) illustrates conditions required to use Bayes' theorem and have probabilities that are relative frequencies with the example of repeated drawings from an urn containing white and black balls in an unknown ratio. He points out that Bayes' theorem was originally derived on the assumption that initially all compositions of urns were equally likely.*

> "Imagine an immense number of urns containing black and white balls in varying proportions, but with a fixed number of urns with each mixture. Then if an urn be drawn at random and n drawings, with replacement, be made therefrom, showing just r white balls, the probability that the next ball will be white is accurately given by [a formula derived from Bayes' theorem].

* Quoted with permission. See Preface.

It is only when we can give a really precise statement of this sort that Bayes' principle can be used with perfect confidence, and the cases are rare.

"Why not, then, reject the formula outright? Because, defective as it is, Bayes' formula is the only thing we have to answer certain important questions which do arise in the calculus of probability. The question as to the likelihood that a coin which showed a given succession of heads and tails should be bad is real and insistent. To say what might reasonably have been expected from a good coin under the circumstances does not, by any means, cover the case. Therefore we use Bayes' formula with a sigh, as the only thing available under the circumstances. . . ." (p. 100)

In striving for a match between theory and fact, one need not suppose that all compositions of urns are equally likely, and Bayes' theorem has too often been cast aside because that has not been understood. One approach to the problem of choosing distributions over "compositions of urns" might be to investigate a large class of events with a view to finding what distributions occur. Although this old idea has rarely been tried, one large investigation was carried out by Egon Pearson (1925) "with a view to making more clear the use and the limitations of Bayes' Theorem in the field of practical statistics." (p. 388)

He made counts for many variables that might be expected to obey a binomial distribution with a view to finding the prior distribution of the probability p of success. Examples: fraction of men seen smoking a pipe; fraction of vehicles drawn by horses (p's do change with time!); fraction of chestnut colts born of bay mares; fraction of verbs in Carlyle. Pearson found that a uniform prior distribution was inconsistent with his data, but that a U- or V-shaped density produced a better fit.

Berkson (1930), apparently commenting on this study, says:*

"There are writers, however, who, admitting that the assumption is to be questioned, believe it may be subjected to experimental test, and have essayed to actually sample at random the probabilities that characterize the universes of our experience. It would be impertinent to assert that an experimental investigation is bound to be futile, but the utility of this sort of procedure seems to us exceedingly dubious. We doubt indeed that any clear meaning can be assigned to the concept of 'the universes of our experience,' of which random samples are to be obtained. But granting the existence of such a distribution of *a priori* probabilities we doubt the relevancy of its estimation to any practical problem. In any actual investigation, we deal with a definite slice of possible experience; an anthropologist is not concerned with the universes dealt with in the investigation of an economist or an epidemiologist. If *a priori* probabilities are of interest to him, they are those that obtain in his peculiar world of observation. It appears to us quite as

* Quoted with permission. See Preface.

wide of the mark aimed at, to call in a formula which obtains its *a priori* probability from experience in general, as to obtain it from the unique experience at hand, and indeed it may be argued that, as between the two, the latter is the more reasonable." (pp. 54–55)

While we find Pearson's study most educational, we admit that Berkson's criticism is a telling one. Why, indeed, should a man studying the authorship of *The Federalist* papers use stud books as a basis for prior distributions?

Pearson concluded,* in part (although we do not know that these remarks would now represent his views):

"The exact form of the distribution . . . seems to be something which each statistician can only determine for himself *a posteriori* by an examination of his own statistical experience, and Bayes' Theorem [with a uniform prior] can only be accepted as providing a valuable working rule for prediction, on the assumption that among the problems with which most statisticians are confronted there is in fact a distribution . . . whose difference from [uniformity] is of the same order as that observed in the experiments . . . in this paper." (p. 433)

Coolidge implies: lacking better, we do the best we can. We do have better. We can meet Berkson's criticism head on by getting data appropriate to the problem, and this is what we do. Berkson and Pearson are both agreed that the most we can hope for is a good approximation. What is good enough for all the rest of applied mathematics is going to have to be good enough for statistical inference. The sort of approximation we have in mind is the one that the statistician ordinarily makes when he chooses a data distribution and proceeds to use it freely without further question, even though he knows that it is not, and cannot be, exactly right. We do not suggest careless acceptance of either prior or data distributions, and in the main study, we pay more attention than usual to the effects of distributions.

3.1. INTRODUCTION TO BAYES' THEOREM AND ITS APPLICATIONS

How shall we use new observations to change our beliefs? In the realm of uncertain inference, Bayes' theorem offers one answer. Our problem is to assess our beliefs about the authorship of the disputed *Federalist* papers in the light of the evidence provided by the word frequencies. In the main study, we express our beliefs in terms of probabilities of authorship, and we use the standard form of inference based on Bayes' theorem. The object, then, is to determine the posterior, or conditional, or final probabilities of authorship, so named because they are conditional on or posterior to the evidence.

---

* Quoted with permission. See Preface.

Neither using Bayes' theorem in practical problems nor formulating probabilities as degree of belief would have won many popularity prizes over the last couple of centuries, so the reader may properly expect difficulties. We want to see what some of these difficulties are in a large analysis of data.

In this main study, we use probabilities freely to express uncertainty, not only about authorship, but also about imperfectly known quantities (parameters) that describe how the frequencies of words vary. We think these uses of probability set the main study off a bit from the objective or relative-frequency approach to statistics in the direction of what is sometimes called "Bayesian" inference. Just how large the offset is, and just how practical the methods are, the reader will judge for himself, although he may have to wade upstream through our personal views to do it.

Before proceeding to the technical development of the main study, we want to give a general description of our approach—foretell the sources of difficulties, and state roughly how we deal with each.

One essential feature of our approach is the use of a range of values rather than a single value for any final probability, say, that Hamilton wrote paper No. 52. We can thus handle some forms of uncertainty without having to make the specific critical assumptions and computations needed to lead to a single "right" value. Of course, a range of values will be satisfactory only if the same general conclusion about authorship is appropriate to all values in the range.

Consider briefly the basic components of Bayesian inference as applied to a single disputed paper. Suppose a set of word frequencies has been observed for the paper. To determine the final odds of authorship, we need to know the initial odds, as well as the respective sampling distributions of the set if Hamilton or if Madison wrote the paper. Given this information, Bayes' theorem is remarkably simple, but to get each of these pieces of information requires work, and even their meaning is troublesome. An additional problem—the choice of words to be used in the inference—is intertwined with the other problems.

Consider first the initial odds of authorship. Here, as in many applications of Bayesian inference, the initial or prior probabilities are present because they are required for Bayes' theorem, and not because of any desire to make some profit from prior information. Initial odds are a major source of uncertainty. They vary from person to person according to what prior evidence—here, mainly historical evidence—is available, and how the available evidence is assessed. There is no "correct" assessment, and we have no desire nor any special competence to attempt one.

Fortunately, the analysis of the statistical evidence can be carried out separately from the determination of initial odds and the two combined at the very end. We concentrate on the former, and allow each reader to put in the initial odds as he chooses. We hope to produce such strong statistical evidence as to overwhelm any moderate assessment of initial odds.

The first of several restrictions on words arises in the separation of the assessment of the statistical evidence from the choice of initial probabilities, because the statistical evidence must be independent of the historical evidence on which the initial probabilities are based. To accomplish this, we eliminate most meaningful words, i.e., what we call contextual words. The contextual words undoubtedly provide evidence, but the assessment of this evidence and of the fraction of it that is new is more a historical than a statistical problem. We hope to produce enough evidence without the contextual words.

In Chapter 2, we studied the form of distributions of word frequencies in pieces of writing by a single author, and we found that for most noncontextual words, the negative binomial family of distributions fits well. In the main study, we assume that word frequencies are independently distributed according to negative binomial distributions with parameters that depend on the word, author, and length of paper. In Bayesian inference, the exact form of the distribution plays an important role, and as L. J. Savage says, the probability model that is chosen should be big enough to "fit an elephant." In our judgment, the negative binomial is adequate for *The Federalist*. As we proceed through the main study, we shall describe several more studies that support this judgment.

The Poisson family of distributions performed distinctly less well in the studies in Chapter 2. Nevertheless, we have carried both families through the main study. in part to see where the choice of distributions is or is not important, and in part to allow motivation and exposition of our procedures for the simpler Poisson model.

To specify a sampling distribution requires choosing not only the form but also the values of the parameters. Our knowledge of the parameters is imperfect, and more seriously so because we select words to be used from a large pool. The unknown parameters and the associated problem of selectivity cause the chief conceptual and computational difficulties in the main study. In facing these difficulties we make a second, more serious use of Bayes' theorem.

If we knew the parameters, we could then calculate the probability of x occurrences of a word in a disputed paper if Hamilton wrote the paper, and the corresponding probability if Madison wrote the paper, the two necessary numbers for assessing the evidence from this word. The source of information about the parameters is the papers of known authorship. From these, the parameters can be estimated within a range of uncertainty. The probability of x occurrences in a disputed paper if Hamilton is the author will vary according to what values of the parameters are chosen within this range. What is required is an average of these probabilities. We represent our state of uncertainty about these parameters by a probability distribution and use the weighted average with respect to this probability distribution.

We form a conceptual pool of all the words that might be used. The general rate of use of a word is a source of recognizable inhomogeneity in the pool. To take account of this stratification is easy and of no import here. Suppose that,

apart from this, the pool can be regarded as homogeneous. Suppose also, for the moment, that the distribution of the parameters over the words in the pool be known. This distribution is then the correct prior distribution for use in Bayes' theorem in combination with the observed word rates in the papers of known authorship. Then, for any word chosen from the pool, we obtain a posterior probability distribution for its parameters. That distribution is then used to carry out the averaging to eliminate the unknown parameters from the sampling distributions for the unknown paper.

The allowance for selectivity in this process is automatic, and works roughly as follows. Consider some measure of discriminating ability of words, say, a relative difference between mean rates. The distribution of this measure in the pool will be concentrated about "no discriminating ability." The posterior distribution for any word will reflect this prior concentration at "no ability" and the evidence for apparent discrimination in the papers of known authorship. For rare words, the evidence in the known papers is relatively slight, and the apparent discriminating ability will be heavily discounted. For high-frequency words, the balance reverses, and the evidence in the known papers will be discounted only slightly.

We do not know the distribution over the pool of words, but we can estimate it sufficiently well for our needs. Using again the feature of providing a range of final answers, we carry through the entire analysis for each of about a dozen distributions in the pool. Actually, we have here a family of distributions, and are choosing several members of the family by choosing parameters of the family. We call these parameters "underlying constants" to avoid confusion with the word-rate parameters. Our final results, log odds of authorship, are given for each of several sets of underlying constants. Each set completely specifies the distribution in the pool.

How, finally, do we get any information on the distribution in the pool of words? Quite apart from questions of authorship, what we have now is a straightforward, but messy, estimation problem, akin to the estimation of the "between" variance component in Model II analysis of variance. Using the word frequencies for a sample of words that are *not* preselected for discriminating ability, we can estimate the parameters for each word separately; then combining all words, we can estimate the parameters of the distribution of the word-rate parameters. There is nothing especially Bayesian in the way we handle this final problem.

We turn now to the technical development of this program. A sequence of examples introduces both the ideas and problems involved.

3.1A. An example applying Bayes' theorem with both initial odds and parameters known. Suppose Hamilton's and Madison's use of the word *also* is well represented by Poisson distributions whose parameters are the expected numbers of occurrences $w\mu_H$ 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

TABLE 3.1–1

POISSON PROBABILITIES WITH EXPECTED COUNTS

$w\mu_H = .62,\ w\mu_M = 1.34$

Frequency	Hamilton	Madison
0	.538	.262
1	.334	.351
2	.103	.235
3	.0214	.105
4	.00331	.0352
5	.000411	.00943
6	.0000424	.00211

rates are known to be $\mu_H = .31$ and $\mu_M = .67$. Then for an unknown paper of length $w = 2$ (2000 words), the probabilities for 0 to 6 usages by each author are shown to three significant figures in Table 3.1–1.

Suppose *also* is used four times in a 2000-word paper written either by Hamilton or by Madison; what are the odds that Hamilton wrote the paper? Naturally, the answer depends on our uncertainty. If initially we thought the authorship was a tossup, the comparison of the probabilities .00331 and .0352 would give us new odds of about 10 to 1 (.0352/.00331) in favor of Madison. But if we were nearly certain before the observation that Hamilton wrote the paper, even though the evidence would reduce the odds in favor of Hamilton by a factor of 10, we might still be left with a probability nearly 1 that Hamilton wrote it. Now we develop these ideas more formally.

Let p_1 and $p_2 = 1 - p_1$ be the probabilities before the observation 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. For simplicity, 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 probability is obtained by adding its two similar components; the first is the probability that Hypothesis 1 is true and that x occurs, or $p_1f_1(x)$.) 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.

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

$$(1)\quad \text{Odds } (1, 2 \mid x) = \frac{P \text{ (Hypothesis 1} \mid x)}{P \text{ (Hypothesis 2} \mid x)} = \frac{p_1 f_1(x)}{p_2 f_2(x)} = \left(\frac{p_1}{p_2}\right)\left(\frac{f_1(x)}{f_2(x)}\right)$$
$$= \text{(initial odds)} \quad \times \quad \text{(likelihood ratio)} = \text{final odds.}$$

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 that 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 four usages in the paper gives the likelihood ratio $f_1(4)/f_2(4) \approx .00331/.0352 \approx .1$. Thus the final odds are

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

or about 100 to 1 for Hamilton.

3.1B. Selecting words and weighting their evidence. Bayes' theorem, as described in Eq. (1), provides one rationale for selecting the words that we use to determine authorship, and it provides automatically for weighting the evidence from each word, as we now illustrate.

TABLE 3.1–2

RATES PER THOUSAND WORDS FOR *also*, *an*, AND *because*, AND A MEASURE OF IMPORTANCE FOR DISCRIMINATION

Word	Hamilton rate	Madison rate	Importance ($w = 2$)
also	.31	.67	.55
an	6.00	4.50	.86
because	.45	.50	.01

Let us extend our previous example and suppose that the counts for *also*, *an*, and *because* are independent Poisson variables, with the known rates given in Table 3.1–2. Suppose that 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

$$f_P(4 \mid .62) f_P(7 \mid 12) f_P(0 \mid .9),$$

and if Madison is the author, the probability is

$$f_P(4 \mid 1.34) f_P(7 \mid 9) f_P(0 \mid 1).$$

Finally, the likelihood ratio for the three words is the product of the likelihood ratios for the separate words, and the weighting problem is solved.

The numerical values of the Poisson probabilities instruct us further. They give the likelihood ratio as

$$\frac{.00331}{.0352} \times \frac{.0437}{.117} \times \frac{.407}{.368} \approx .0940 \times .374 \times 1.11 \approx .039.$$

The contribution of *because*, the third factor, 1.11, is near unity since the two authors use the word with nearly identical rates.

The logarithmic form of Bayes' theorem can be obtained by taking the logarithm of both sides of Eq. (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 observation x drawn from a Poisson distribution, the likelihood ratio, Hamilton to Madison, is

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

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

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

where the subscript i stands for the ith word and the other subscripts have their obvious meanings.

In choosing independent words as discriminators, the criterion is whether their contribution is worth the cost of including them. When rates are known, no bias arises from selection. (Selection effects arise because chosen variables are not as good at discriminating as they appear to be.) 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* in the unknown material, for medical and psychological experiments are not the only ones that require controls to guard against the frailty of an investigator's objectivity. Objective decisions come easily when one has no knowledge of the unknowns, but once one has the information, trying to ignore it puts one in the position of the small boy who is trying not to think of a rhinoceros that is standing on top of a flagpole.

One useful measure for the importance of a word is the difference between expected log likelihood ratio given Hamilton's authorship and that given Madison's. For a 2000-word paper ($w = 2.0$), this measure is

$$2(\mu_H - \mu_M) \log(\mu_H/\mu_M).$$

In Table 3.1–2, we list importances for *also*, *an*, and *because*, and the value for the latter is negligible compared with the others. For the Poisson, this measure of importance of a word can be described as the difference between the log odds for a paper whose observed rate for this word is at the Hamilton mean and the log odds for a paper whose observed rate for this word is at the Madison mean. An importance of zero implies identical means, and an importance of, say, log 2 implies that a paper of unknown authorship with observed rate at the Hamilton mean would be assigned odds for Hamilton against Madison twice as large as the odds that would be assigned if the observed rate were at the Madison mean. This discussion concerns only the contribution of this one word. Among words whose summed mean rate for Hamilton and Madison exceeds 1.0, we have discarded those words for which this swing in log odds between the mean rates for a 2000-word paper was less than .34 = log 1.4. For low-frequency words, a less stringent criterion seems reasonable. We computed the swing in log odds between a paper with no occurrences and one with one occurrence, and discarded the word if this swing in log odds was less than .48 = log 1.62. (The actual rules were made on the basis of the negative binomial distribution and a 3000-word length of paper, but the essentials of the decisions are contained in the above remarks.)

3.1C. Initial odds. All this seems perfectly straightforward. Wherein lie the difficulties? The first place 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 differing 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 odds as 1 for an unknown paper, thus setting a 50-50 chance as appropriate to your degree of ignorance. Or you might notice that Hamilton wrote 43 of the known papers to Madison's 14, and then assign prior odds of $43/14 \approx 3$. 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, an analysis 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 (1000 to 1 for Hamilton) to final odds of 10^{-3} (1000 to 1 for Madison). In terms of probabilities, the same likelihood ratio converts an initial probability

of .999 to a final probability of .001, and any initial probability less than .999 to a final probability of less than .001. Naturally, for any strength of data, opinions can be so 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—what might be called the evaluation of the historical evidence—have been separated 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) are changed only slightly by wide changes in the initial odds. The statistical analysis in our study is devoted largely to the likelihood ratio. What we call "odds" in the bulk of this monograph are these likelihood ratios or, in more complex situations, the analogous factors for converting initial odds to final odds. Thus our odds apply without adjustment only for even initial odds, while a simple adjustment is needed for any other initial odds.

By this factorization and the resulting factorization of the problem, we have largely avoided having to specify the initial probabilities, often considered the major obstacle to using Bayes' theorem. Consequently, our approach coincides with much of the work on classification problems (cf. Anderson, 1958, Chapter 6; or Rao, 1952, Chapter 8). In our handling of nuisance parameters, we make more critical use of Bayes' theorem, and to that problem we now turn.

3.1D. Unknown parameters. A second and more serious difficulty arises from uncertainty in the data distribution. We do not know that it is Poisson; indeed we have seen evidence in favor of the negative binomial. But even if the form were known exactly, we would not know the parameters of the distribution exactly, nor could we be confident that the parameters remain constant from one sort of text to another. We mention the latter problem briefly in Section 3.3 and treat it in detail in Section 4.10.

Let us look at the difficulty of the unknown parameter values. For purposes of exposition, assume that Poisson distributions adequately represent the distribution of the word frequencies. Then for each word the unknown parameters are the rates per 1000 words of text in Hamilton's and in Madison's writings. Denote these rates by μ_H and μ_M.

The usual way to get information on parameter values is to observe large amounts of data known to be sampled from each desired distribution. If these data provided precise point estimates of the parameters, we would, in the tradition of large-sample statistics, use the estimates in place of the known values to evaluate the likelihood ratios.

In the actual problem, we have 94,000 words of text known to be written by Hamilton and 114,000 words by Madison, seemingly vast amounts, yet surprisingly little for handling any but words of the highest frequency. For example, if the word *also* had the rates .31 and .67 in Hamilton's and Madison's writings, we would expect only 29 and 76 occurrences, respectively,

Similarly,

$$P\{x = 4 \mid \mathrm{M}\} = \int P\{x = 4 \mid \mathrm{M}, \mu_{\mathrm{M}}\} p(\mu_{\mathrm{M}})\, d\mu_{\mathrm{M}} = E(P\{x = 4 \mid \mathrm{M}, \tilde{\mu}_{\mathrm{M}}\}),$$

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

If the posterior densities of the rates are known for each word and each author, we need now, in principle, only carry out the integrations and obtain the two probabilities required for each word and thence the likelihood ratio. But even for the Poisson family of distributions, carrying out the integration requires bivariate numerical integrations. The integrals look univariate, but the marginal density of μ_{H} appearing in the integrand must itself be obtained by an integration from the joint density $p(\mu_{\mathrm{H}}, \mu_{\mathrm{M}})$ that Bayes' theorem yields. (The rates cannot be treated independently, because they are *a priori* dependent.) In the extension to the negative binomial family, the integrals become four-dimensional.

In the presence of a substantial amount of data, the posterior distributions of the rates are rather sharply concentrated and a natural approximation for the expectation of any function, say g, over this distribution is the function evaluated at some central value of the rate that we label $\hat{\mu}$:

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

The mean might be the preferred choice for $\hat{\mu}$, but it can be determined only by integrations of the type being avoided. The mode, the position of the highest point of the density, is a more feasible choice and one well articulated with the use of Bayes' theorem. We do use the modes of the posterior distributions in this way, discussing necessary adjustments later. The unadjusted result of using this approximation is identical with that obtained if the modal estimated rates were used as the known rates. Except for very high-frequency words like *the* or *of*, the modal estimated rates are not the same as the observed mean rates in the Hamilton and Madison papers, because the prior distributions have the effect of moving the estimated rates closer together, thereby making allowances for selection of the apparently best words.

3.2. HANDLING UNKNOWN PARAMETERS OF DATA DISTRIBUTIONS

When parameters of a data distribution are unknown, Bayes' theorem requires both data sampled from the data distributions and a prior distribution for the parameters, treated now as random variables rather than as constants. Bayes' theorem then yields a posterior distribution for the parameters.

For the Poisson family, the rate parameters μ_{H} and μ_{M} for each word must be assigned a prior distribution. In our problem, Bayes' theorem uses the data on papers of known authorship to obtain a posterior distribution for the word

rates. Finally, we complete the classification of unknown papers as described in Section 3.1D.

How to choose prior distributions is the main topic of this section and our approach introduces two essential features of our study: we use more than one choice of prior distribution, and we base our choices of priors on data, even if feebly.

3.2A. Choosing prior distributions. The data in the papers of known authorship are sufficient to dominate any prior information on the rate of use of a word averaged over the two authors, but not large enough to dominate prior information on the comparative rates for the same word. We expect both authors to have nearly identical rates for almost any word. We choose the prior distributions of rates to represent this prior expectation. The effect of these prior distributions is to reduce the apparent quality of discriminators, and thereby we automatically allow for effects of selection of apparently good discriminators from large pools of words.

To separate average rate of use from a comparison of the rates for the two authors, we introduce a pair of parameters for each word:

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

(Our σ has nothing to do with standard deviation.) Clearly, $\frac{1}{2}\sigma$ measures average frequency, and τ measures the ability to discriminate. For fixed $\tau(\neq \frac{1}{2})$, the bigger σ, the better the discrimination. When $\tau = \frac{1}{2}$, the authors have equal rates.

Prior information about σ arises from a few studies of word rates in texts written more than one hundred years before or after *The Federalist*. The net effect of this information is almost negligible in comparison to the observed rates in the 94,000 words of Hamilton text and the 114,000 words of Madison text. We use a flat prior for σ for each word.

For authors writing together on the same topic at the same period, we suppose that the prior distribution of τ for any word would be very nearly symmetric and unimodal with much probability near $\frac{1}{2}$. The spread may depend on the combined rate σ of the word, but otherwise the same distribution might reasonably be expected to apply, independently, to almost any word free from severe contextuality. The distribution's concentration around $\frac{1}{2}$ is critical, yet hard to specify without reference to any data. Our plan is to get rough estimates of the distribution of τ over a group of words, unselected for ability to discriminate, and representative of the pools of words from which all words were selected; in particular, we use the 90 function words from the Miller-Newman-Friedman list (see Section 2.5A). 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 m_H and m_M be Hamilton's and Madison's observed rates for a given word. Then $s = m_H + m_M$ estimates σ

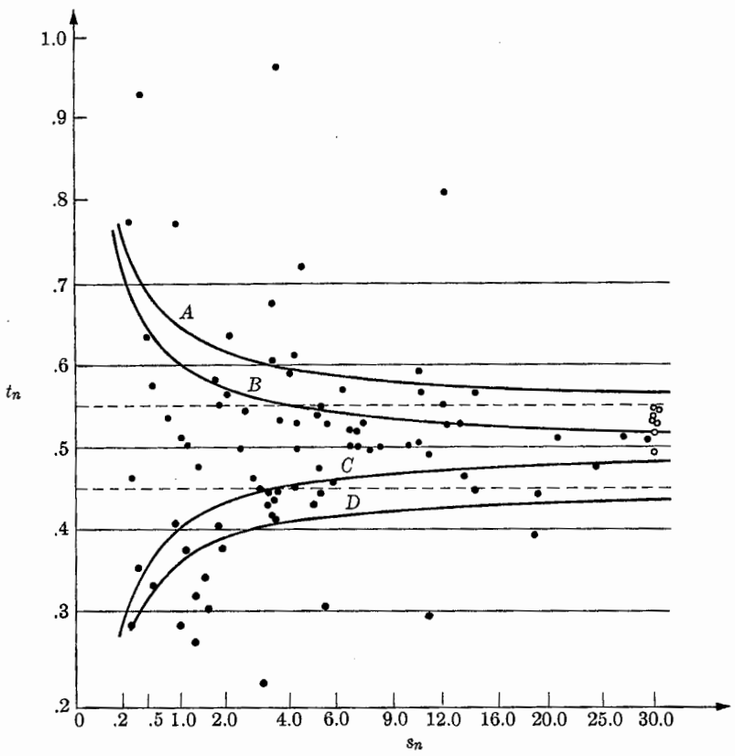


FIG. 3.2–1. Sample estimates (s_n, t_n) of the parameters (σ_n, τ_n) for 90 function words. Curves B and C show two-standard-error bands for t_n if $\tau_n = .5$. Curve A shows a two-standard-error band above $\tau_n = .55$. Curve D shows a two-standard-error band below $\tau_n = .45$.

and $t = m_H/s$ estimates τ. For our material, s has a small standard error, and so, for practical purposes, we take $s \approx \sigma$. The standard error of t depends upon both τ and σ.

In Fig. 3.2–1 the (s, t) pairs are plotted for the 90 unselected function words. (The square-root scale on the horizontal axis is used to stabilize the variance of s. The circled points have been moved to the left to bring them into the figure.) Curves B and C in the figure give two-standard-error bands around the null value $\tau = .5$. Curve A is located two standard errors above the value $\tau = .55$, and curve D the same distance below $\tau = .45$.

The variation vertically in Fig. 3.2–1 is much larger than can be expected with $\tau = .5$ for every word and Poisson variation in the rates. 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

TABLE 3.2–1

PAIRS OF 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

variation alone could put some t's outside even if no τ's are. Similarly the true proportion outside .4 to .6 is almost surely less than the observed .28.

If the distribution of τ values were a beta distribution with probability density function $\tau^{\gamma-1}(1-\tau)^{\gamma-1}/\mathrm{B}(\gamma,\gamma)$ 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 effects of changing γ on the tails of the distribution, 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 3.2–1, that would more than handle the range.

The analysis of Fig. 3.2–1 for the Poisson model is an oversimplification because part of the variation, in excess of Poisson variation, is due to non-Poisson variation in the counts for each author and not to variation in rates between the authors. Our final choices of β's for bracketing the true prior distribution are carried out for the negative binomial distribution. The detailed discussion of the choices of parametrization and of families of distributions and initial choices of β's is given here in a relatively simple situation, not only for exposition of the method but also because our initial choices for the negative binomial model are based in large part on direct or analogous use of this Poisson development.

3.2B. The interpretation of the prior distributions. More formally, we suppose there are underlying linguistic quantities β_1, β_2 that determine the general similarity of word occurrences by Hamilton and Madison for a large pool of words—including the pools of words from which we selected our 165. We further assume that given β_1, β_2, the prior distribution of the τ for each word in the pool, given its σ, is adequately represented by a symmetric beta distribution, both of whose arguments are $\gamma = \beta_1 + \beta_2\sigma$. We assume that the distributions of the τ's are independent across words.

3.2C. Effect of varying the prior. Logically, the β's are parameters of the distribution of the differential-rate parameter τ, but because this terminology is intolerable, 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 3.2–1. 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 matter in the final assessments of authorship.

3.2D. The posterior distribution of (σ, τ). For any pair of underlying constants β_1, β_2, the posterior density of the parameters (σ, τ), given the vectors of data x_H, x_M on the papers of known authorship, is, by Bayes' theorem:

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

where $C(X)$ is a constant, $p(\sigma, \tau)$ is the prior density of the parameters $(\tilde{\sigma}, \tilde{\tau})$, and $p(x_H, x_M \mid \sigma, \tau)$ is the density for the vectors of data, given the parameters.

Return to the example of *also*. For the Poisson, the likelihood of observing 26 counts in 94,000 words of Hamilton, and 80 in 114,000 of Madison, 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

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

Approximate methods tell us that the mode of the posterior density is near $\hat{\sigma} = .99$, $\hat{\tau} = .316$. The effect of this prior is to give an estimate $\hat{\tau} = .316$ instead of the estimate $t = .282$ based on the observed rates. The allowance for selection effects is thus moderate for *also*, but stronger for low-frequency words, as seems appropriate.

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 four *also*'s, we multiply both rates by 2.000 and use the Poisson tables to get

$$\frac{f_P(4 \mid .62)}{f_P(4 \mid 1.34)} \approx .1,$$

described at the start of Section 3.1.

TABLE 3.2-2

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

3.2E. 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 influence our allocation of effort, but few new ideas. The following 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. A study of estimates based on moments, in the spirit of Fig. 3.2–1, suggested that non-Poissonness δ was nearly independent of the rate μ, and this independence was the reason for using the measure δ. A tail-reducing transformation from δ to $\zeta = \log(1 + \delta)$ was made for each author, and then transformed to $\xi = \zeta_H + \zeta_M$, $\eta = \zeta_H/\xi$, so that ξ and η are analogous to σ and τ.

We introduced five underlying constants $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$, and assumed that, given the β's, $(\sigma, \tau, \xi, \eta)$ 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, the parameter τ has the symmetric beta density:

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

η has the symmetric beta density:

$$\frac{[\eta(1 - \eta)]^{\beta_3 - 1}}{B(\beta_3, \beta_3)};$$

ξ has the gamma density with mean β_4, argument β_5:

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

Each quintuple of β's is called a set of underlying constants and is assigned a code number. Table 3.2–2 shows the values of the underlying constants for the six sets used in the displayed log odds in Section 3.4. Altogether, 21 different sets were used in some phase of the study.

TABLE 3.2–3

FINAL WORDS AND WORD GROUPS
ESTIMATED NEGATIVE BINOMIAL PARAMETERS

Code	Word	Set 31 μ_1	μ_2	σ	τ	δ_1	δ_2	Set 22 δ_1	δ_2
	B3A								
60	*upon*	3.24	.23	3.47	.932	.25	.39	.29	.48
	B3B								
3	*also*	.32	.67	.99	.327	.09	.10	.13	.14
4	*an*	5.95	4.58	10.53	.565	.02	.02	.07	.07
13	*by*	7.32	11.43	18.75	.390	.35	.40	.37	.43
39	*of*	64.51	57.89	122.40	.527	.24	.25	.26	.28
40	*on*	3.38	7.75	11.12	.304	.34	.42	.37	.44
55	*there*	3.20	1.33	4.53	.706	.23	.24	.25	.27
57	*this*	7.77	6.00	13.77	.564	.21	.21	.25	.23
58	*to*	40.79	35.21	76.00	.537	.39	.45	.41	.48
	B3G								
73	*although*	.06	.17	.23	.267	.11	.11	.21	.18
78	*both*	.52	1.04	1.56	.334	.12	.14	.15	.18
90	*enough*	.25	.10	.35	.727	.47	.52	.54	.64
116	*while*	.21	.07	.28	.744	.23	.25	.29	.35
117	*whilst*	.08	.42	.50	.153	.15	.13	.27	.20
123	*always*	.58	.20	.78	.742	.07	.07	.13	.13
160	*though*	.91	.51	1.42	.639	.08	.08	.11	.12
	B3E								
80	*commonly*	.17	.05	.23	.763	.05	.05	.14	.16
81	*consequently*	.10	.42	.52	.189	.16	.14	.25	.20
82	*considerable(ly)*	.37	.17	.54	.684	.07	.08	.13	.15
119	*according*	.17	.54	.71	.238	.30	.30	.36	.34
124	*apt*	.27	.08	.35	.770	.06	.07	.13	.16
	B3Z								
87	*direction*	.17	.08	.25	.693	.31	.32	.39	.43
94	*innovation(s)*	.06	.15	.20	.278	.06	.06	.15	.13
96	*language*	.08	.18	.26	.316	.05	.05	.11	.10
110	*vigor(ous)*	.18	.08	.26	.680	.02	.02	.08	.09
143	*kind*	.69	.17	.86	.799	.25	.22	.29	.27
146	*matter(s)*	.36	.09	.45	.790	.05	.05	.12	.13
151	*particularly*	.15	.37	.51	.282	.14	.16	.20	.21
153	*probability*	.27	.09	.36	.757	.02	.02	.07	.08
165	*work(s)*	.13	.27	.40	.326	.46	.42	.55	.47

Note: Estimates are based on set 31 of underlying constants. Estimates of δ_1 and δ_2 for set 22 are also shown in final columns.

PS
AMETERS

		Set 22	
δ_1	δ_2	δ_1	δ_2
.25	.39	.29	.48
.09	.10	.13	.14
.02	.02	.07	.07
.35	.40	.37	.43
.24	.25	.26	.28
.34	.42	.37	.44
.23	.24	.25	.27
.21	.21	.25	.23
.39	.45	.41	.48
.11	.11	.21	.18
.12	.14	.15	.18
.47	.52	.54	.64
.23	.25	.29	.35
.15	.13	.27	.20
.07	.07	.13	.13
.08	.08	.11	.12
.05	.05	.14	.16
.16	.14	.25	.20
.07	.08	.13	.15
.30	.30	.36	.34
.06	.07	.13	.16
.31	.32	.39	.43
.06	.06	.15	.13
.05	.05	.11	.10
.02	.02	.08	.09
.25	.22	.29	.27
.05	.05	.12	.13
.14	.16	.20	.21
.02	.02	.07	.08
.46	.42	.55	.47

ants. Estimates of δ_1 and

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† were used in the calculation of likelihood ratios for the unknown papers as if they were the known parameters. Table 3.2–3 shows the parameters estimated in this way from set 31 of underlying constants for each of the 30 words finally chosen for the study. The final columns give the non-Poissonness parameters estimated from set 22. The estimated rates from set 22 differ from those from set 31 by .01 or less. Set 22 is our preferred choice, but many side studies use the estimates from set 31.

3.2F. Final choices of underlying constants. For both the Poisson and negative binomial families, a variety of sets of β's were used, each giving a different prior distribution. The pool of 90 unselected words provides evidence on plausible ranges for the β's as indicated in Section 3.2A. The methods for estimating the β's are presented in Section 4.5. The 6 sets shown in Table 3.2–2 and used in the displayed log odds in Section 3.4 are spread over the estimated range. The one set that most closely represents the available evidence is set 22. For the Poisson, β_3, β_4, β_5 are irrelevant.

How β_1 and β_2 influence the posterior mode of τ can most easily be seen by means of an approximate sampling equivalent of the prior information. The posterior mode of τ, using β_1 and β_2, is approximately equal to the sample estimate t from a modified sample in which the number $\beta_1 + \beta_2\hat{\sigma}$ of extra occurrences has been added to the occurrences of the word in both the Hamilton and the Madison texts. Here, $\hat{\sigma}$ is the estimated combined rate, estimated from the original data. The net discriminating power of the word decreases as $\beta_1 + \beta_2\sigma$ increases; for low-frequency words, β_1 is important; for high-frequency words, β_2 is important.

The final three β's concern the non-Poissonness parameters δ_1, δ_2 of the negative binomial. Large values of β_3 require δ_1, δ_2 to be nearly equal; small values of β_3 allow them to be nearly independently determined. The final two, β_4 and β_5, influence the determination of the average non-Poissonness in a way that is neither easily described nor very important.

3.3. SELECTION OF WORDS

The pools of words and the 165 words chosen for inclusion in the study were described in Chapter 2. The analysis of the disputed papers in the main study is based on 30 words: "the final 30 words." Here we describe how these words were chosen.

An essential feature of the main study is that allowance for selection and regression effects is made entirely through the prior distributions. We assume

* The computation was programmed for digital computer by Miles Davis.
† Subscripts 1 and 2 refer to Hamilton and Madison, respectively.

that the prior distributions apply to any and every word chosen from some large pool of words. Then, according to the model, the prior distribution will reduce the apparent discriminating ability to the appropriate extent.

With this selection feature, we may choose words for inclusion by whatever methods are convenient, so long as they are independent of the unknown papers. Thus the words obtained from the screening and the index studies may be included. What matters more is that we may continue to eliminate words without changing the treatment or validity of the remaining words.

Two restrictions are important. First, to estimate the underlying constants of the prior distribution, we must not select the words on the basis of discriminating ability. For this purpose we use the 70 words with highest frequency in the Miller-Newman-Friedman list and the random sample of 20 words of low frequency from that list. Second, we assume that the variation in parameters measuring word rates for these 90 unselected words is representative of that in the entire pool from which all 165 words were chosen.

We grouped words according to categories of use and source that we felt might bear on contextuality. The groups were made up of the 165 words in the study, though the definitions were made to apply to pools from which the 165 were chosen. Groups A, B, G (for gamma), and D contain words occurring on the Miller-Newman-Friedman list: D, the pronouns and auxiliary verbs; A, *upon*; B, the remaining words among the 70 high-frequency function words; G, the remaining low-frequency function words. Our 90 unselected words are directly representative of the pools corresponding to A, B, G, D. Group E consists of words that occur on the list of "well-liked" words (mostly adjectives and adverbs) made up, without regard to discriminating ability, from the index. Group Z (for zeta) consists of words on a larger, incompletely defined list, made up from the index and including mostly abstract nouns. We regarded the remaining words, group H (for eta) as severely contextual and did not use them further in the main study. Of the groups we retained, D and Z seemed to us most likely to have strong effects of contextuality. This division of the 165 words gave group sizes:

A, 1; B, 46; G, 30; D, 31; E, 9; Z, 31; H, 17.

Most of the 165 words show no appreciable discriminating ability, and we kept only those words that had sufficient "importance" as described in Section 3.1. After this assessment, groups A, B, G, D, E, Z had 1, 11, 9, 5, 6, 13 words, respectively. In this step we eliminated a handful of words for which the computation of the posterior mode was not successful.

We evaluated log odds for the known papers and examined them 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 present in Section 4.10, we retained only those words that are homogeneous across the Madison papers. In addition, we eliminated all words in the group containing personal pronouns and auxiliary verbs because we feared that they

were cont
B3E, B3Z,
negative
to retain th
errors.

3.4. LOG

In this sec
reserving adj
log odds from
and 30 sets o
for the final 30 wo
groups and indiv

As explained in
the authorship o
odds, our log odd
final odds.

3.4A. Checkin
from the 30 words
treating each pape
log odds obtained b
the indicated set of
eter values.

Looking first at
stants, we see that
paper, negative log
for No. 113 with log
is for No. 134 with log
lying constants, has
though, the log odds
consistently and for
are can be better app

Log odds	
0	
.1	
.5	
1	e
2	
3	

* The computation o
Davis.

were contextual. The net was 30 words in five groups, called B3A, B3B, B3G, B3E, B3Z, and displayed in Table 3.2–3 with the estimated parameters for the negative binomial analysis for set 31 of underlying constants. We have decided to retain the cumbersome code for groups of words, rather than risk introducing errors.

3.4. LOG ODDS

In this section and the next, we give the results of our log odds computations, reserving adjustments for later sections. What we present is a sprinkling of the log odds from among the $113 \times 30^2 \approx 10^5$ numbers for 113 papers, 30 words, and 30 sets of underlying constants.* First we examine the log odds in total for the final 30 words: then in Section 3.5, we examine log odds for separate word groups and individual words.

As explained in Section 3.1C, our log odds are logarithms of the final odds on the authorship of a paper when the initial odds are 1 to 1. For any other initial odds, our log odds are the logarithm of the factor that changes initial odds to final odds.

3.4A. Checking the method. In Table 3.4–1 we present the total log odds from the 30 words for each of the 98 papers of known authorship, obtained by treating each paper as if it were a disputed paper. Each entry is the approximate log odds obtained by using, for each word, the estimated parameters based on the indicated set of underlying constants as if the estimates were the true parameter values.

Looking first at the negative binomial model for set 22 of underlying constants, we see that every Hamilton paper has positive log odds; every Madison paper, negative log odds. The weakest result among the Hamilton papers is for No. 113 with log odds of 3.0; the weakest result among the Madison papers is for No. 134 with log odds of −.8. Paper 134, for all four sets of Poisson underlying constants, has log odds pointing mildly in the wrong direction. As a whole, though, the log odds for all sets of underlying constants and both models 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
0	1 to 1	4	55 to 1
.1	1.1 to 1	5	150 to 1
.5	1.6 to 1	10	22,000 to 1
1	$e \approx 2.7$ to 1	15	3.3×10^6 to 1
2	7 to 1	20	480×10^6 to 1
3	20 to 1	25	7.1×10^9 to 1

* The computation of log odds was programmed for digital computer by Miles Davis.

TABLE 3.4–1

TOTAL NATURAL LOG ODDS FOR 98 PAPERS OF KNOWN AUTHORSHIP
TOTAL FOR 30 FINAL WORDS
6 SETS OF UNDERLYING CONSTANTS, 2 DISTRIBUTIONS
MODAL APPROXIMATION

Hamilton Paper number	Paper length (thousands of words)	Negative binomial Set of underlying constants 22	33	38	21	11	31	Poisson Set of underlying constants 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
15	3.1	20.3	18.5	18.4	21.5	22.1	21.7	48.9	45.4	45.2	51.4
16	1.9	24.9	23.1	25.2	27.5	28.2	26.8	46.6	42.5	45.6	51.0
17	1.6	19.8	17.9	19.1	21.5	19.8	21.2	32.0	29.1	31.1	35.1
21	2.0	15.6	14.0	15.4	17.3	18.1	16.9	35.6	32.2	34.5	39.0
22	3.5	30.2	27.6	29.1	33.0	32.8	32.5	68.4	63.2	64.4	72.8
23	1.8	16.2	15.1	15.0	16.9	14.1	18.0	35.0	32.1	33.6	38.0
24	1.8	15.0	13.1	13.7	16.3	14.9	16.0	27.1	24.7	25.6	29.3
25	2.0	11.6	10.3	15.1	14.4	15.2	12.5	16.7	14.2	18.0	20.0
26	2.4	16.5	15.5	16.8	17.9	16.4	18.0	36.1	32.5	35.7	40.2
27	1.4	17.2	15.9	18.3	19.4	18.1	18.7	25.6	23.1	25.4	28.5
28	1.6	13.3	12.8	12.2	13.2	12.6	13.7	19.4	18.6	17.4	19.6
29	2.2	27.8	25.6	24.1	28.7	26.0	29.5	56.9	53.3	53.2	59.9
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12.9	13.2	43.4	40.1	39.4	45.8
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	53.6	49.9	49.1	56.2

28	1.6	13.3	12.8	12.2	13.2	12.6	13.7	19.4	18.6	17.4	19.6
29	2.2	27.8	25.6	24.1	28.7	26.0	29.5	56.9	53.3	53.2	59.9
30	2.0	16.7	13.9	11.7	17.2	12.9	18.2	43.4	40.1	39.4	45.8
31	1.7	24.8	21.6	18.6	25.4	20.4	26.5	53.6	49.9	49.1	56.2
32	1.4	3.3	3.4	3.9	3.2	2.7	3.7	4.8	4.5	3.9	4.7
33	1.6	9.7	8.3	5.7	9.1	6.3	10.4	22.5	21.1	19.4	22.9
34	2.2	17.1	15.6	13.0	16.8	14.7	18.2	38.1	36.1	34.1	39.1
35	2.2	14.1	13.2	12.5	14.4	12.5	16.0	33.0	30.7	30.4	34.9
36	2.7	22.1	20.9	21.7	23.6	24.0	23.5	45.3	42.5	42.6	47.6
59	1.8	14.0	13.7	14.5	14.7	14.8	15.0	18.9	17.9	17.5	19.7
60	2.2	14.9	13.4	12.8	15.3	14.1	15.9	29.4	27.3	26.6	30.7
61	1.5	10.0	9.3	10.1	10.7	10.5	10.5	15.1	14.0	14.1	16.0
65	2.0	16.5	14.3	12.7	17.0	15.2	17.4	35.5	33.0	32.2	37.2
66	2.2	18.8	16.5	15.1	19.5	17.2	20.0	41.5	38.6	37.9	43.6
67	1.4	11.7	10.9	7.8	10.6	8.5	12.3	22.8	22.0	19.6	22.6
68	1.5	9.7	8.9	10.3	10.6	10.9	9.9	12.7	11.7	12.2	13.7
69	2.6	23.3	21.1	19.6	24.0	21.4	24.9	52.0	48.7	47.6	54.3
70	3.0	16.6	14.9	18.0	19.2	19.5	17.6	32.8	29.5	32.7	36.8
71	1.7	12.0	10.6	15.1	14.6	13.8	13.2	19.8	16.7	21.5	24.0
72	2.0	17.1	15.6	17.5	18.9	17.9	18.2	30.3	27.5	29.8	33.5
73	2.3	14.9	12.6	10.9	15.2	10.4	16.5	40.2	36.9	36.5	42.6
74	0.9	12.4	11.4	12.8	13.6	13.9	12.8	17.7	16.2	17.6	19.6
75	1.9	16.5	15.3	15.9	17.4	14.5	17.5	31.3	28.6	30.5	34.2
76	1.8	18.7	16.5	16.3	19.9	18.4	19.8	33.6	31.0	31.3	35.8
77	2.0	18.9	16.8	15.1	19.4	17.1	20.1	38.1	35.5	34.7	39.9
101	1.2	6.5	5.9	7.0	7.3	7.5	6.7	8.9	8.1	8.4	9.6
102	1.4	15.7	13.9	15.4	17.5	16.0	16.7	26.0	23.3	25.4	28.8
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

(cont.)

TABLE 3.4–1 (*cont.*)

TOTAL NATURAL LOG ODDS FOR 98 PAPERS OF KNOWN AUTHORSHIP
TOTAL FOR 30 FINAL WORDS
6 SETS OF UNDERLYING CONSTANTS, 2 DISTRIBUTIONS
MODAL APPROXIMATION

Madison Paper number	Paper length (thousands of words)	Negative binomial Set of underlying constants 22	33	38	21	11	31	Poisson Set of underlying constants 22, 31	33	38	11, 21
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
43	3.0	−28.5	−27.2	−28.2	−31.2	−35.6	−29.6	−48.8	−46.7	−46.7	−50.8
44	2.6	−20.8	−19.7	−21.0	−23.0	−26.5	−21.8	−34.3	−32.3	−33.4	−36.6
45	2.1	−14.2	−12.9	−15.7	−16.5	−17.6	−15.1	−19.6	−17.3	−20.5	−22.8
46	2.6	−27.0	−25.7	−27.0	−29.5	−32.3	−28.2	−44.4	−42.1	−43.0	−47.0
47	2.5	−21.9	−21.2	−22.0	−23.9	−27.6	−22.9	−36.9	−35.2	−35.3	−38.3
48	1.6	−13.1	−12.5	−12.6	−13.9	−15.1	−13.4	−18.3	−17.6	−17.5	−19.0
121	3.0	−14.9	−14.2	−15.1	−16.1	−18.4	−15.8	−25.5	−23.9	−24.9	−27.5
122	2.8	−24.7	−23.8	−25.9	−27.2	−29.8	−26.7	−45.2	−41.9	−45.1	−49.7
131	2.9	−16.6	−14.9	−18.7	−19.8	−22.7	−17.6	−26.7	−23.6	−28.2	−31.0
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

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
135	3.0	−20.7	−19.5	−22.0	−24.1	−27.0	−22.0	−34.7	−32.2	−34.7	−38.0
141	3.6	−18.5	−17.4	−19.9	−20.9	−23.9	−19.7	−28.5	−26.3	−28.5	−31.1
201	2.0	−21.1	−20.5	−19.8	−21.9	−23.6	−21.7	−33.7	−32.5	−31.7	−34.5
202	2.0	−13.1	−12.6	−11.5	−14.1	−15.3	−13.7	−24.6	−23.5	−23.8	−25.7
203	2.0	−17.2	−15.9	−17.8	−19.4	−21.2	−17.9	−26.6	−24.5	−26.6	−28.9
204	2.0	−22.8	−21.3	−24.2	−25.4	−26.8	−24.4	−36.9	−33.3	−37.6	−41.6
205	1.9	−14.5	−14.1	−14.6	−15.3	−14.0	−15.4	−21.0	−19.6	−20.5	−22.5
206	1.9	−18.3	−17.5	−17.7	−20.1	−18.9	−19.5	−32.6	−30.4	−32.7	−35.5
207	1.9	−19.0	−18.0	−20.3	−21.3	−23.7	−20.3	−28.0	−25.8	−28.4	−31.3
208	2.0	−21.4	−20.7	−21.4	−23.1	−26.3	−22.4	−34.6	−33.0	−33.4	−36.2
209	2.1	−13.7	−13.0	−14.1	−16.0	−14.8	−14.7	−25.1	−23.1	−26.3	−28.4
210	2.1	−26.4	−25.2	−26.2	−28.7	−31.5	−27.3	−40.8	−38.9	−39.4	−42.9
211	2.0	−21.7	−20.5	−22.2	−23.9	−25.5	−22.8	−30.8	−28.6	−30.4	−33.4
212	2.2	−14.7	−14.0	−14.9	−17.0	−17.0	−15.4	−25.7	−24.0	−26.4	−28.2
213	2.0	−16.9	−16.0	−16.4	−18.3	−21.1	−17.3	−24.6	−23.6	−23.5	−25.6
214	2.0	−12.7	−11.2	−13.8	−14.9	−16.0	−13.0	−13.0	−11.4	−13.6	−15.0
215	2.0	−19.3	−17.9	−20.3	−22.0	−25.4	−20.0	−27.1	−25.1	−27.2	−29.6
216	2.0	−5.7	−5.2	−6.2	−6.5	−7.4	−6.3	−9.6	−8.5	−9.8	−11.1
217	2.2	−20.6	−19.9	−19.7	−22.0	−24.2	−21.3	−35.8	−34.6	−33.8	−36.7
218	2.0	−18.9	−17.6	−19.7	−21.4	−23.3	−19.8	−34.2	−31.4	−34.2	−37.5
219	2.1	−16.8	−15.6	−17.7	−19.1	−19.8	−17.6	−23.5	−21.5	−23.8	−26.1
220	2.0	−15.3	−14.6	−14.5	−16.3	−16.8	−15.7	−26.4	−25.1	−25.2	−27.5
301	2.0	−14.1	−13.3	−14.1	−15.5	−17.3	−14.5	−19.9	−18.8	−19.2	−21.1
302	2.9	−17.6	−16.6	−17.4	−19.5	−21.4	−18.3	−32.9	−31.1	−31.6	−34.7
311	1.5	−14.8	−13.9	−14.8	−16.1	−17.5	−15.3	−19.6	−18.3	−19.3	−21.3
312	1.4	−9.9	−9.4	−9.3	−10.2	−10.4	−10.1	−14.0	−13.4	−13.2	−14.4
313	1.5	−4.0	−3.8	−3.6	−4.0	−4.0	−4.1	−5.3	−5.3	−4.8	−5.2
314	1.5	−15.1	−14.2	−15.2	−16.6	−17.1	−15.5	−20.3	−19.0	−20.0	−21.8
315	1.5	−15.3	−14.2	−15.8	−17.2	−18.7	−15.7	−20.0	−18.5	−19.9	−21.8
316	1.5	−9.8	−9.2	−9.6	−11.2	−12.6	−10.6	−15.9	−14.3	−16.5	−18.2

TABLE 3.4–2

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
MODAL APPROXIMATION

Paper number	Paper length (thousands of words)	Negative binomial Set of underlying constants 22	33	38	21	11	31	Poisson Set of underlying constants 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

Since we know who wrote each of these papers, the log odds in Table 3.4–1 offer a check on the method. We have used these papers in estimating parameters and have used the log odds for words in the choice of final words, and so we might expect the method to work a bit better here than elsewhere; and we discuss how much better in Section 4.8. But the results here seem very strong and satisfactory.

We can examine the effect of varying the underlying constants. The values of the β's for each of these sets are shown in Table 3.2–2. As a rough rule, the changes run to 10 per cent of the log odds. The changes for these and other sets of underlying constants are studied in Section 4.5E, and summarized in Section 3.7B.

Visual inspection will assure the reader that the variation in log odds from one set of underlying constants to another is modest compared to the variation from one paper to another. Different paper lengths explain only part of this variation; the rest is primarily natural random variation. Examine, for example, the log odds for the Madison papers Nos. 201–220, all of almost equal length.

The changes in log odds from the negative binomial distribution to the Poisson are huge, running between 50 and 150 per cent increases! This shows at once that the choice of data distribution does matter immensely. For the moment, the reader may wish to be conservative and attend to the negative binomial odds—if median odds of three million to one can ever be called conservative.

Readers of Damon Runyon may remember that ". . . nothing between human beings is 1 to 3 . . . ."* How can anyone so believe his model as to accept as remotely reasonable odds of millions to one, when only 48 papers by Hamilton and 50 papers by Madison are in hand? The question is indeed appropriate, and the Poisson log odds are not defensible. But we believe that the negative binomial log odds are defensible, at least after we make the adjustments presented in Section 3.7. The basis is primarily the moderate contributions from each of many nearly independent words. We postpone further discussion to Sections 3.5B, 3.7E, and 3.7F where we treat possibilities outside the framework of the mathematical model.

3.4B. The disputed papers. Next, the *pièce de résistance*, Table 3.4–2, 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 for papers Nos. 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 3.4A. Essentially, No. 55 does not have its share of marker words, no matter who wrote the paper, and the high-frequency words produce no information.

---

* We are indebted to Frank Anscombe, who recalled that the passage occurs in the story "A Nice Price" in the book *Money From Home*.

TABLE 3.5–1

LOG ODDS BY WORD GROUPS FOR SET 22 OF UNDERLYING CONSTANTS

		Negative binomial						Poisson				
		Word group										
Authorship	Paper number	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
	11	3.2	8.2	−.2	1.5	3.6	8.8	14.2	−.3	2.2	5.9
	12	4.5	3.4	.2	1.9	3.1	12.6	5.8	−.1	2.7	4.2
	111	2.6	11.9	1.9	−3.0	−1.4	7.4	22.7	2.3	−3.4	−1.8
	112	5.6	−1.3	2.2	−.7	3.4	19.8	−3.6	2.5	−1.2	5.9
	113	−.3	−1.2	.8	.9	2.8	−.9	−1.4	.8	1.1	3.2
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
	40	−6.0	−8.5	−.3	−.9	−3.5	−8.3	−16.0	−.5	−.7	−4.7
	41	−7.3	−6.9	−1.3	1.6	−1.7	−10.7	−14.3	−1.9	2.4	−3.0
	132	−3.2	−7.0	−5.4	−2.4	−2.3	−5.5	−10.0	−10.8	−2.2	−3.5
	133	−5.6	−1.4	−1.4	−1.9	−2.9	−7.5	−4.8	−1.6	−2.4	−4.1
	134	−4.2	5.0	−.3	.2	−1.7	−5.2	7.8	−.1	.4	−2.0
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

Among the joint papers, No. 20 looks especially mixed, but the small log odds are confounded with the brevity of the paper. The matter is further complicated because Madison's notes suggest that he borrowed much of it from Felice and Sir William Temple. We discuss this further in Section 8.4.

3.5. LOG ODDS BY WORDS AND WORD GROUPS

3.5A. Word groups. The total log odds makes a strong prediction in the right direction for almost all of the papers of known authorship, and strong predictions for Madison for each disputed paper. To show how consistently the different word groups behave and to what extent each contributes to the total, we present, in Table 3.5–1, the log odds for each of the five word groups, for set 22 of underlying constants, and both distributions. All disputed and joint papers are included, but only a selection from the papers of known authorship. For each author, we give his first 7 papers from *The Federalist*, and, in addition, for Hamilton, Pacificus I, II, and III (code numbers 111, 112, 113); for Madison, Helvidius II, III, and IV (code numbers 132, 133, 134). Some exterior papers are included to illustrate the consistency of behavior of word groups over changes in source of writing. The papers include numbers 113 and 134, the papers of Hamilton and Madison most poorly identified in total log odds.

All groups look remarkably consistent, considering their different strengths, for a weak set should point in the wrong direction occasionally. This general consistency is a further sign in support of the method.

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 the marker words. Although this does not prove that cleverness in selecting variables fails to pay, it does show that routine can pay.

3.5B. Single words. The contribution of each separate word to the log odds is shown in Table 3.5–2 for the negative binomial distribution having set 22 as underlying constants for all disputed and joint papers and for some of the known papers that were used in Section 3.5A. The log odds for *upon* have been given already in Table 3.5–1 as word group B3A and are not repeated. For a few words, the Poisson log odds for set 22 are shown alongside those for the negative binomial.

The words are divided into three groups: in Table 3.5–2A the 9 words of highest frequency, in Table 3.5–2B the 11 Hamilton marker words, and in Table 3.5–2C the 9 Madison marker words. For the Hamilton markers, every negative log odds means that the word did not occur. For the Madison markers, every positive log odds corresponds to a nonoccurrence of the word.

TABLE 3.5–2A

LOG ODDS FOR SINGLE WORDS: 9 HIGH-FREQUENCY WORDS

Authorship	Word: / Number:	*an* 4	*of* 39	*there* 55	*this* 57	*to* 58	*both* 78	*though* 160	*by* 13	*by** 13	*on* 40	*on** 40
	Paper number											
Hamilton	1	.6	.5	−.8	.6	.9	.6	−.5	.0	.2	−.4	−.6
	6	.2	1.7	2.1	−.3	−1.5	.1	.7	1.9	3.2	4.3	6.7
	7	.7	1.7	2.1	1.0	−.3	.9	−.8	−1.2	−2.2	−.2	−.1
	8	.6	.7	−1.3	.4	.1	.8	.7	1.8	3.2	.5	1.2
	9	.3	1.9	−.1	−.3	−.4	.7	.0	1.6	2.6	.7	1.3
	11	.4	1.5	1.4	1.0	−.8	1.0	−.8	.6	1.3	3.7	6.8
Madison	10	−.3	−1.7	.0	−1.5	−.8	−1.2	−1.0	−2.2	−4.9	−.9	−1.9
	14	−.5	−.7	−3.2	−.2	−.7	.9	−.1	.1	.3	−2.4	−4.8
	37	−.9	−.7	−2.1	−.2	−1.3	.4	−.3	−1.3	−2.6	−1.5	−3.2
	38	.1	−.9	−2.1	.2	−.6	−1.4	−.5	−1.2	−2.7	.7	2.0
	39	−1.7	−.1	−3.7	.5	−.5	−.5	.1	−1.9	−3.9	−4.1	−9.7
	40	−.7	−1.0	−2.1	.0	.1	.4	−.3	−4.5	−10.4	.8	1.9
Joint	18	−1.3	.3	−.7	.3	−2.3	.2	1.1	−3.1	−6.0	−1.8	−3.4
	19	−.3	.0	−2.1	−.2	−1.6	.8	.3	−.8	−1.4	−2.7	−5.4
	20	−.5	−1.0	2.2	−.5	−1.4	.0	.0	−1.0	−1.6	.1	.4
Disputed	49	.6	−.2	−.8	−1.1	−.2	.7	−.5	.1	.3	−3.5	−6.5
	50	−.6	−.7	−1.8	.0	−1.6	.5	−.4	−.3	−.3	−3.3	−5.2
	51	−1.1	−.3	.0	.0	−2.1	−.3	−.6	−1.3	−2.3	−4.1	−8.5
	52	−1.3	−.9	−2.8	.4	.1	.1	−.6	−1.3	−2.1	−3.6	−7.1
	53	−1.2	−.3	−1.5	.0	−.7	−.7	−.7	−2.5	−4.8	1.4	2.8
	54	−.3	−.5	−2.1	1.2	−1.3	−.3	.3	−1.8	−3.3	−3.6	−7.2
	55	−1.8	−.2	.5	−.3	.0	.2	−.1	1.1	2.1	.6	1.4
	56	−1.2	1.0	.0	.4	−2.1	−.5	−.5	1.2	1.9	−1.4	−2.4
	57	−1.2	.9	−.2	−.3	−.7	.2	−.7	−1.1	−2.0	−3.0	−6.3
	58	−.4	−.9	−1.4	.0	−1.5	.2	−.1	−.7	−1.2	−3.3	−6.8
	62	.1	−.2	−3.4	.3	−.8	−.6	−.2	−1.4	−2.6	−2.6	−5.5
	63	.4	−.5	.8	−.5	−2.0	−1.2	−.4	−4.1	−10.1	−1.6	−3.5

Note: Set 22 of underlying constants; columns marked * give Poisson log odds; all others are negative binomial.

TABLE 3.5–2B

LOG ODDS FOR SINGLE WORDS: 11 HAMILTON MARKERS

62	.1	—.2	—3.4	.3	—.8	—.6	—.2	—1.2	—2.6	—2.6	—6.6	3.5
63	.4	—.5	.8	—.5	—2.0	—1.2	—.4	—4.1	—10.1	—1.6	—3.5	

Note: Set 22 of underlying constants; columns marked * give Poisson log odds; all others are negative binomial.

TABLE 3.5–2B

LOG ODDS FOR SINGLE WORDS: 11 HAMILTON MARKERS

Authorship	Word:	*enough*	*while*	*always*	*commonly*	*considerable*	*apt*	*direction*	*vigor*	*kind*	*matter*	*probability*
	Number:	90	116	123	80	82	124	87	110	143	146	153
	Paper number											
Hamilton	1	.8	—.1	.5	.9	—.2	1.4	— 1	.5	—.6	.9	.8
	6	.8	—.2	—.6	—.2	.9	—.3	.8	—.1	1.3	—.4	.7
	7	—.2	—.2	—.7	.9	—.3	.8	—.1	—.1	.4	.7	1.4
	8	.9	—.2	1.7	.9	.4	—.3	.6	.9	—.8	—.4	1.4
	9	.8	.9	—.5	—.1	.4	—.2	.7	—.1	.6	.8	—.2
	11	—.2	—.2	—.8	—.2	.8	—.4	—.1	.9	1.6	—.5	.6
Madison	10	—.2	—.3	.8	—.3	—.5	.7	—.1	—.2	—1.0	—.6	.6
	14	—.2	—.2	—.7	—.2	—.3	—.3	—.1	—.1	—.8	—.4	—.3
	37	—.2	—.2	—.8	—.2	—.4	.3	—.1	—.2	—1.0	.6	—.4
	38	—.3	—.3	—1.0	—.3	—.5	—.5	—.2	—.2	—1.1	.5	—.5
	39	—.2	—.2	—.8	—.2	—.4	—.4	—.1	—.2	—.9	—.5	—.4
	40	—.2	—.2	—.8	—.2	—.4	—.4	—.1	—.2	—1.0	—.6	—.4
Joint	18	.8	—.2	.3	—.2	.4	—.3	—.1	—.1	—.8	—.4	—.3
	19	—.2	—.2	—.6	.9	—.3	—.3	.6	—.1	—.8	—.4	—.3
	20	—.1	—.1	.5	—.1	.5	—.2	—.1	.5	—.6	1.6	—.2
Disputed	49	—.1	—.1	.4	—.1	—.2	—.2	—.1	—.1	—.6	—.3	—.2
	50	—.1	—.1	—.3	—.1	—.2	—.1	—.1	—.1	—.4	—.2	—.1
	51	—.2	—.2	—.6	—.2	—.3	—.3	—.1	—.1	—.7	—.4	—.3
	52	—.2	—.2	.4	—.1	—.3	—.3	—.1	—.1	—.7	—.4	—.3
	53	—.2	—.2	—.7	—.2	—.3	.8	—.1	—.1	—.8	1.5	—.3
	54	—.2	—.2	—.6	—.2	.4	—.3	—.1	—.1	—.8	—.4	—.3
	55	—.2	—.2	.3	—.2	.4	—.3	—.1	—.1	—.8	—.4	—.3
	56	—.1	—.1	—.5	—.1	.5	—.2	—.1	—.1	.6	—.3	—.2
	57	—.2	—.2	.3	—.2	—.3	—.3	—.1	—.1	—.8	—.5	—.3
	58	—.2	—.2	—.6	—.2	—.3	—.3	—.1	—.1	—.8	—.4	—.3
	62	—.2	—.2	.2	—.2	.3	—.3	—.1	—.2	—.9	—.5	—.3
	63	—.2	—.3	—.9	.8	—.5	—.4	—.2	—.2	—1.1	—.6	—.4

Note: Set 22 of underlying constants; all entries are negative binomial log odds.

TABLE 3.5–2C

LOG ODDS FOR SINGLE WORDS: 9 MADISON MARKERS

Authorship	Word: / Number:	*also* 3	*although* 73	*consequently* 81	*innovation* 94	*language* 96	*particularly* 151	*work* 165	*whilst* 117	*whilst** 117	*according* 119	*according** 119
	Paper number											
Hamilton	1	.4	.1	.4	.1	.1	.3	.1	.4	.5	.4	.6
	6	.5	.1	.5	.1	.1	.3	.1	.5	.6	−.6	−.5
	7	.6	.2	.6	.1	.1	.4	.2	.6	.8	.6	.9
	8	−.1	.1	.5	.1	.1	.3	.2	.5	.7	.5	.8
	9	−.1	.1	.4	.1	.1	.3	−.7	.4	.5	.4	.6
	11	.0	.2	.6	.1	.2	.4	.2	.7	.8	.6	1.0
Madison	10	.8	.2	−.7	.2	.2	−.3	.2	.8	1.0	−1.2	−2.7
	14	.0	.2	−1.5	−.7	−.5	.3	−.8	−1.1	−1.0	.6	.8
	37	−.9	−.7	−1.4	−.7	−1.1	.4	−.7	−1.0	−.8	−.4	−.1
	38	−.7	−1.0	−.6	.2	−.4	−.3	−.6	−1.5	−2.4	−.8	−1.2
	39	.0	.2	.6	−.7	.2	−.9	.2	.7	.9	−2.2	−9.5
	40	−.9	.2	.7	−1.0	.2	−.4	.2	.7	.9	−.4	−.1
Joint	18	.6	.1	.5	.1	.1	.3	.2	−1.1	−1.0	−1.4	−3.1
	19	.5	.1	.5	.1	.1	.3	.2	−1.1	−1.0	.5	.8
	20	−.8	.1	.4	.1	.1	.2	.1	.4	.5	.4	.6
Disputed	49	−.2	.1	−1.0	.1	.1	.3	−.6	−1.2	−1.2	.4	.6
	50	−.3	−.9	.3	.1	.1	−.6	.1	.3	.3	.3	.4
	51	−.1	.1	−1.6	.1	.1	−1.0	.1	−1.8	−2.9	.5	.8
	52	−.6	.1	.5	.1	.1	−.5	.1	.5	.6	.5	.7
	53	−1.4	−.8	.5	−1.0	.1	−.4	.2	−1.1	−1.0	.6	.9
	54	.0	−.8	−1.5	.1	.1	−.5	.2	.5	.7	−.5	−.4
	55	.0	.1	.5	.1	.1	.3	.2	.5	.7	−1.1	−1.7
	56	−.2	.1	.4	.1	.1	−.5	.1	−1.2	−1.2	.4	.6
	57	.0	.2	.6	−.7	.1	.3	.2	−2.1	−4.6	−.5	−.4
	58	.0	−.8	−.9	.1	.1	−.5	.2	.6	.7	.5	.8
	62	.0	.2	−1.9	.1	.1	.4	.2	.6	.8	.6	.9
	63	−.8	.2	.7	.2	.2	−.3	.2	−.9	−.7	−.3	−.1

Note: Set 22 of underlying constants; columns marked * give Poisson log odds; all others are negative binomial.

From the log odds for single words we see that, with the exception of the two words *upon* and *on*, the total log odds are composed of moderate contributions from many words. Taking mean log odds over the 12 disputed papers, *upon* gives −4.3, *on* gives −2.3, the total for 30 words gives −15.0. Thus the total log odds has as its components about 30 per cent from *upon*, 15 per cent from *on*, and 55 per cent from the remaining 28 words. In Section 3.7E, the magnitude of these contributions is examined in more detail. If *on* and *upon* are excepted, only three instances of odds higher than 20 to 1 remain for single words in the disputed and joint papers. All in all, this slight dependence on huge contributions by single words makes the final odds much more acceptable.

The contribution from low-frequency words comes mostly from an occurrence of the word. "No occurrence" usually has a small effect, rising for the strongest low-frequency marker *whilst* to between .4 and .8, depending on paper length. Most failures of low-frequency word groups to discriminate correctly are caused by the occurrence of too few words.

The log odds for Madison paper No. 39 based on *according* illustrates how the negative binomial stamps down the odds compared to the Poisson, when a low-frequency marker word has an unusually high rate. In paper 39, 9 occurrences of *according* give a Poisson log odds of −9.5 which is reduced to −2.2 for the negative binomial! It comes to this: the negative binomial provides automatic damping for low-frequency words and thus prevents words from getting badly out of hand. We eliminated most words with severe outliers, but we still need something like this negative binomial effect to protect us against the ever-present possibilities of an unexpectedly high rate.

3.5C. Contributions of marker and high-frequency words. What log odds would be assigned to a paper containing each Hamilton marker once, no Madison markers, and each high-frequency word at the Hamilton mean rate? Table 3.5–3 presents the answer to this and similar questions. Our purpose is to see whether either author is being treated unfairly by our words, or put another way, whether there are delicate matters that need our attention.

We divided the 30 final words into 10 high-frequency words, 11 Hamilton markers, and 9 Madison markers corresponding to sections A, B, C of Table 3.5–2, with *upon* added to the high-frequency words. Then, we obtained log odds for twelve artificial 2000-word papers which are formed to give the 12 ($2 \times 2 \times 3$) combinations of 1 or 0 occurrences of every Hamilton marker, 1 or 0 occurrences of every Madison marker, and high-frequency words at rates nearest to the Hamilton mean rate, the Madison mean rate, or the average mean rate. Table 3.5–3 shows for each combination the log odds under the negative binomial having set 31 as underlying constants, and it shows the contribution of each of the three component groups of words.

The log odds in such artificial papers show no gross weighting toward one author or the other, and what asymmetry there is seems to favor Hamilton. Thus, in total, the paper with Hamilton usage in all three categories has log

TABLE 3.5–3

CONTRIBUTIONS OF MARKER AND HIGH-FREQUENCY WORDS, 2000-WORD PAPERS
SET 31 OF UNDERLYING CONSTANTS, NEGATIVE BINOMIAL

11 Hamilton markers (word numbers 80, 82, 87, 90, 110, 116, 123, 124, 143, 146, 153)
Expected number of occurrences for Hamilton: 7.0; for Madison: 2.4
0 occurrences of each Hamilton marker give log odds —4.4
1 occurrence of each Hamilton marker gives log odds 7.6

9 Madison markers (word numbers 3, 73, 81, 94, 96, 117, 119, 151, 165)
Expected number of occurrences for Hamilton: 2.3; for Madison: 6.4
0 occurrences of each Madison marker give log odds 3.5
1 occurrence of each Madison marker gives log odds —6.0

10 high-frequency words (numbers 4, 13, 39, 40, 55, 57, 58, 60, 78, 160)
Occurrences of each at the Hamilton mean rate give log odds 12.3
Occurrences of each at the Madison mean rate give log odds —11.1
Occurrences of each at the average mean rate give log odds 1.8

Total log odds for 30 words

Hamilton markers		None	1 each	None	1 each
Madison markers		None	None	1 each	1 each
High-frequency word rate	Hamilton mean	11.4	23.4	2.0	14.0
	Madison mean	—11.9	.1	—21.4	—9.4
	Average	1.0	13.0	—8.5	3.5

odds 23.4, the corresponding Madison paper, —21.4, and the "in-between" paper with no markers, 1.0.

Among the high-frequency words, the excess of 12.3 over —11.1 is due mostly to *upon*. The 11 Hamilton markers give log odds 7.6, stronger than the —6.0 obtained from 9 Madison markers. In partial compensation, the expected number (7.0) of Hamilton markers in a Hamilton paper is not quite as large a fraction (7.0/11) of the 11 Hamilton markers as the expected number (6.4) of Madison markers in a Madison paper is of the 9 Madison markers. Expected log odds are studied in more detail in Section 4.8.

In the main study, we assess the evidence of each word without regard to possible imbalance of Hamilton and Madison markers or of the potential log odds that can be attained. Nevertheless, we are pleased that possible log odds seem well in balance, as one protection against ill effects of violations of assumptions. Further, the slight weighting in favor of Hamilton means that any im-

balance has not tended to increase the observed evidence for Madison in the disputed papers. Not that we have anything against Madison, but as the data have turned out, their evidence is just a bit easier to defend than equivalent results in the direction of Hamilton would be.

3.6. LATE HAMILTON PAPERS

As a separate validation of the main study, we checked the method on four late Hamilton *Federalist* papers, numbers 79, 80, 82, and 85, which Hamilton wrote especially for the McLean edition. From the point of view of the main study these are fresh untouched papers that did not contribute to the determination of the constants. We chose the four shortest papers among the papers numbered 78 through 85.

Considering their lengths, all four papers give strong log odds for Hamilton, as they should. Table 3.6–1 shows the total log odds and its breakdown by word groups, and at the top, the lengths of the papers.

TABLE 3.6–1

LOG ODDS FOR FOUR LATE HAMILTON *Federalist* PAPERS
NEGATIVE BINOMIAL DISTRIBUTION
SET 22 OF UNDERLYING CONSTANTS

Paper number:	79	80	82	85
Paper length (thousands of words):	1.0	2.3	1.5	2.5
Word group				
B3A	1.8	3.5	3.2	6.5
B3B	1.9	5.7	9.0	6.1
B3G	—.7	1.9	.3	2.6
B3E	.9	.2	.2	—1.0
B3Z	.2	2.2	—.3	1.2
Total (30 words)*	4.0	13.5	12.4	15.4

* Total log odds may disagree with sum of log odds by word groups because of rounding.

The short paper No. 79 has the weakest log odds, while the other three are satisfactorily high.

The log odds may be compared with expectations if the model and prior distributions used were correct, as is done for the papers of disputed and of known authorship in Section 4.8. The log odds for papers 82 and 85 are slightly above expectation; for paper 80, slightly below; for paper 79, about $1\frac{1}{2}$ standard errors below expectation. Results for individual words indicate that too few Hamilton and too many Madison marker words occurred. The high-frequency words and *upon* behaved according to expectations.

3.7. ADJUSTMENTS TO THE LOG ODDS

When we presented the log odds in earlier sections, we put off an examination of the many underlying assumptions and approximations. In this section, we plug up this hole through studies that estimate corrections to the log odds for the disputed papers. Although most of these studies are specialized, they suggest that per cent reduction in log odds gives a good way to extrapolate to our problem. Consequently, the adjustments reduce extreme odds most heavily.

The reader must understand our intent. The assumptions and approximations were introduced to make the entire study feasible. We merely try to establish the general order of magnitude of adjustments needed to allow for the errors introduced. To get accurate corrections would, we believe, be as difficult as carrying out the analysis without the assumptions and approximations.

3.7A. Correlation. Treating words as independent in the calculation of log odds may either inflate or deflate the estimated odds. In Section 4.7, an investigation of the effect of correlation on 11 high-frequency words suggests that the log odds based on these words should, on the average, be adjusted toward zero by an amount variously estimated between .6 and 1.2. The adjustment amounts to 6 per cent to 12 per cent of the average log odds for these 11 words, and we apply the same range of per cent reduction to the total log odds.

An alternative way of viewing the correlation adjustment is instructive. Among the final 30 words, *on* and *upon* form the one pair where a correlation would be clearly expected. Since *upon* and *on* are the two strongest words, the need to adjust for their correlation was evident before we undertook any general exploration of correlations among words. We expected that the log odds based on independence would need to be deflated, and found for this pair of words that the log odds should be shifted toward zero by about .9, or by about 13 per cent of the combined log odds for these two words. This .9 reduction in log odds is in the middle of the range appropriate as the total adjustment for the 11 high-frequency words, so that the correlations beyond that between *upon* and *on* produce very nearly no net adjustment. The result of the study of correlations among the 11 high-frequency words can then be given the alternative summarization: the log odds from *on* and *upon* should be reduced by about 13 per cent; the net effect of other correlations tends neither to reduce nor to increase the log odds.

3.7B. Effects of varying the underlying constants that determine the prior distributions. Set 22 of underlying constants,

$$\beta_1 = 10, \beta_2 = 0, \beta_3 = 12, \beta_4 = 1.25, \beta_5 = 2,$$

determines the prior distribution that seems best by the estimates made on pools of words. These underlying constants are not precisely determined, and over a range of values that seem very likely to bracket the true prior, the log odds might vary up to about 12 per cent above or below the log odds for set 22.

Most of the allowance is due to the choice of β_1, β_2. An increase of one unit in β_1 decreases log odds by about two per cent; an increase of one unit in β_2 is nearly equivalent to an increase of 1.5 in β_1. The range $5 \leq \beta_1 + 1.5\beta_2 \leq 15$ is an adequate bracket. The effect of changes in β_3, β_4, β_5 is much less, and the suggested total allowance of ± 12 per cent is generous.

3.7C. Accuracy of the approximate log odds calculation. How accurate is the modal approximation to log odds? For five of the final 30 words, we here describe some answers obtained by the approximate method described in Section 4.6. The five words 40 (*on*), 58 (*to*), 60 (*upon*), 90 (*enough*), 117 (*whilst*), include the two strongest words (60, 40), the strongest rare word (117), the word with the highest non-Poissonness (90), and a word (58) with very high rate and considerable non-Poissonness.

The modal estimates $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}$, $\hat{\tau}$, $\hat{\delta}_1$, $\hat{\delta}_2$ are given in Table 3.2–3 for all 30 words. For these five words, the modes $\hat{\sigma}$, $\hat{\tau}$, $\hat{\delta}_1$, $\hat{\delta}_2$ and approximate standard deviations of the posterior distributions of these four parameters are shown in Table 3.7–1. If the posterior distribution of $\hat{\sigma}$, $\hat{\tau}$, $\hat{\delta}_1$, $\hat{\delta}_2$ could be regarded as approximately normal, the correction to the modal approximation would need take account only of the variances and covariances and the second derivatives of the negative binomial probabilities for the unknown papers.

TABLE 3.7–1

MODES AND APPROXIMATE STANDARD DEVIATIONS OF POSTERIOR DISTRIBUTIONS OF σ, τ, δ_1, δ_2

	σ	τ	δ_1	δ_2
40 *on*				
Mode	11.13	.304	.37	.44
S.D.	.45	.018	.14	.15
58 *to*				
Mode	76.01	.537	.41	.48
S.D.	1.20	.008	.15	.16
60 *upon*				
Mode	3.49	.930	.29	.48
S.D.	.25	.018	.13	.32
90 *enough*				
Mode	.36	.721	.54	.64
S.D.	.09	.073	.26	.46
117 *whilst*				
Mode	.51	.161	.27	.20
S.D.	.08	.051	.26	.13

Note: Set 22 of underlying constants.

But the non-Poissonness rates are non-negative quantities, and such large standard deviations relative to the modes, as exhibited in Table 3.7–1, indicate that the distributions must be severely skewed with long tails to the right.

Most of the main study has been based, not on δ_1 and δ_2, but on transformed parameters ξ and η, chosen (in Section 3.2E) to reduce the long tails of the distributions of the δ's. To study the accuracy of the modal approximation, we transformed again, from ξ to $\log \xi$, to reduce still further the asymmetry and provide better approximations.

After transforming, we estimated the adjustments required in the log odds for the range of occurrences actually found. Generally, but not always, the calculations implied a reduction in log odds toward zero. For four of the five words investigated, the changes, word by word, averaged roughly from 5 per cent to 15 per cent of the estimated log odds. The largest adjustment is for the word *upon*—for which reductions of 25 per cent and 15 per cent seem appropriate for Hamilton and Madison, respectively. As an overall net correction, a reduction of 15 per cent from this source might be applied to the negative binomial log odds, perhaps 18 per cent for Hamilton and 12 per cent for Madison, because *upon* is such an important word and its correction has a larger impact on Hamilton than on Madison.

3.7D. Changes in word counts. The recent Cooke edition (1961) of *The Federalist* calls attention to many changes between the original newspaper edition and various book editions. This valuable list came too late to affect our calculations, but two changes are worth noting here.

Cooke says that one of the two *whilst*'s in paper No. 51 was added by Hamilton in preparing the McLean edition, on which our work is largely based. If, as some have argued, Hamilton made use of newspaper text with Madison's own changes in preparing papers by Madison for the McLean edition, then Hamilton's anomalous use of *whilst* would be explained, and we might even be justified in using the McLean text. We are not especially competent to decide this issue, and so we rely on the original newspaper edition as the standard. This path of least resistance is the easier because the paper itself is very strongly decided without the further *whilst*.

Removing the extra *whilst* is straightforward, and it reduces the log odds for paper No. 51 based on *whilst* from −1.9 to −1.2, and the unadjusted total log odds from −21.9 to −21.2, all based on set 22 of underlying constants.

In paper No. 58, an *upon* in the newspaper text was deleted from the McLean edition. Putting it back reduces the log odds given by *upon* from −4.9 to −2.1, and the unadjusted total log odds from −18.0 to −15.2.

These two changes show that shifts of one marker word can have a substantial effect on the log odds. Nevertheless, the log odds that remain are still enormous, and we emphasize that no other changes of single words could have caused so large a change in the log odds as did either of these.

The counts for individual words are subject to error, notwithstanding the checks we have built into the counting and recording processes. A final check

on the word counts for the 30 words used in the calculation of log odds showed some minor errors, and we here indicate the nature of the changes in log odds that are induced. The log odds presented do not incorporate these changes; rather, we treat the counting errors as an additional source of error in the analysis and make a rough allowance for its effect on the log odds.

From nine errors among the joint and disputed papers, the two largest changes are one of .5 in favor of Hamilton in the log odds for paper No. 56 from an extra occurrence of *considerable*, and a change of .4 in favor of Madison in paper No. 51 from an extra occurrence of *particularly*. No other change exceeds .25 in log odds, and the changes are nearly balanced in direction.

Some larger errors occurred among the papers of known authorship where fewer checks were earlier employed. The effect of these errors on the estimation of parameters would be negligible. The effect on the log odds assigned each paper when treated as an unknown would be larger, but these log odds were only illustrative and no essential changes occur.

The direction of changes induced by the errors in counts appears unrelated to author, and has no consistent strengthening or weakening effect on the log odds. An allowance of plus or minus three per cent in the log odds seems adequate for this source of variation.

TABLE 3.7–2

APPROXIMATE ADJUSTED TOTAL NATURAL LOG ODDS TOTAL FOR THE FINAL 30 WORDS

Paper number	Adjusted for word changes only	Adjusted log odds combined estimated reductions		
		Median (23%)	Maximum (40%)	Minimum (3%)
Joint				
18	−11.0	−8.5	−6.6	−10.7
19	−12.1	−9.3	−7.3	−11.7
20	−4.6	−3.5	−2.8	−4.5
Disputed				
49	−13.2	−10.2	−7.9	−12.8
50	−14.3	−11.0	−8.6	−13.9
51	−21.2	−16.3	−12.7	−20.6
52	−16.0	−12.3	−9.6	−15.5
53	−15.8	−12.2	−9.5	−15.3
54	−14.3	−11.0	−8.6	−13.9
55	−5.8	−4.5	−3.5	−5.6
56	−8.7	−6.7	−5.2	−8.4
57	−16.7	−12.9	−10.0	−16.2
58	−15.2	−11.7	−9.1	−14.7
62	−16.5	−12.7	−9.9	−16.0
63	−18.5	−14.2	−11.1	−17.9

3.7E. Approximate adjusted log odds for the disputed papers. The adjustments discussed in the preceding sections may be combined to give, besides the specific adjustments for word changes given in Section 3.7D, an estimated 23 per cent reduction of the total log odds for each paper, based on the log odds for set 22 of underlying constants. Taking the extremes of each adjustment range leads to a range of reductions between 3 per cent and 40 per cent. These values arise from independently applying the midpoints or the ends of the ranges of each reduction: 9 ±3 per cent for correlation, 0 ±12 per cent for choice of prior distribution, 15 ±5 per cent for modal approximation, 0 ±3 per cent for errors in word count.

Table 3.7–2 shows, for each disputed paper and joint paper, the log odds adjusted for specific word changes, and the final log odds based on the median estimated reduction, and on each of the two extremes of estimated reductions.

3.7F. Can the odds be believed? Even after adjustment, some of the odds in the preceding section exceed a million to one. The average of the log odds for the "median" adjustments comes to about −11 or odds of about 60,000 to 1. Such odds seem large, indeed scary, and in common sense one shies from them. To understand them in their proper perspective requires both a fair appraisal of the model's strength and a sensible attitude toward what we shall call outrageous events. We feel that the calculated odds after adjustment are believable in the absence of outrageous events, but the possibility of such events might require serious reduction of the more extreme odds.

Strength of the model. To justify the odds is to justify the choice of data distribution. The general study of such choices and assessments of the reliance to be placed on the choices is a vastly important, yet too little investigated, subject. In our own problem, we assume that the allowances for selection effects, for length of paper, and for estimation of parameters are roughly satisfactory; thus we assume that the reliability of data distributions is the critical issue. For simplicity, assume that we have for each author observations on 50 papers of equal lengths.

Consider first the odds achievable from a single dichotomous variable, say the presence or absence of some marker word. Suppose the word is present in 49 Hamilton papers and in 1 Madison paper. With just this information, an odds factor more extreme than 49 or $\frac{1}{49}$ is unjustifiable, and one less extreme is appropriate when one remembers errors in estimation. Results from several such marker words cannot be compounded to improve the odds without outside evidence for their independence. On the other hand, for markers of lower quality than 49 to 1, odds may be built up to a total near 49 to 1 by near independence that can be checked by the available data. A look at the log odds for the disputed papers by single words in Table 3.5–2 shows that, except for *upon*, *on*, *by*, and *there*, single words never contribute above 12 to 1, and marker words rarely above 5 to 1 odds, and then only on the basis of double occurrences. We feel, on balance, that the odds from marker words are not in serious excess,

even before the adjustments for correlation and errors in the modal approximations.

For words of moderate to high frequency, larger odds can be achieved, but the distributional problems are harder. For these words, we have internal evidence for modest dependence, and we have studied the amount of adjustment needed to be applied to the odds compounded under independence. For each single word, we have used a negative binomial distribution. Studies of the distributions for large numbers of words make the negative binomial appear satisfactory for all but a few words, and these have long since been eliminated.

To illustrate the distributional problem, consider *upon*, the one word that gives a huge contribution to the odds. Our discussion is based on the frequency distribution for *upon* in the 48 Hamilton and 50 Madison papers as given in Table 2.1–4 or Table 4.8–3. Let us neglect variation of paper length.

Consider an observed rate of 4.5 in an unknown paper. The likelihood ratio is the quotient of a small Hamilton probability, fairly well determined, and a tiny Madison probability. The determination of the latter depends critically on the distribution chosen to extrapolate the Madison data to rates far above those observed, and consequently the likelihood ratio may be very large but is also sensitive to the choice. For observed rates above zero and below two, the likelihood ratio is more firmly determined, and moderate odds are justified.

Finally, consider zero occurrences. For Madison the probability is well estimated as about .8, while for Hamilton, the estimate is small for any reasonable family of distributions we might choose, say $\frac{1}{50}$ or less, which implies a likelihood ratio of 40 or more. The support for a reasonable family is that the observed distribution for *upon* is quite regular, and well fit by a negative binomial. So too are the distributions for many other words. There is not, among these words, evidence of frequent aberrant observations. In fact, after the 15 per cent reduction for modal approximation, the odds contributed by no occurrence of *upon* in a 2000-word paper are just about 60 to 1, strong but not much beyond what sense easily allows. Fortunately, none of the disputed papers show high rates of *upon*. The believability of the huge odds based on *upon* for papers with Hamilton-like rates would be much more suspect than the more modest ones for papers with Madison-like rates.

For other high-frequency words, much the same analysis is appropriate, except that the authors' distributions overlap more, so that the observed results never fall so far into the tail of either distribution. Only *on* gives consistently strong odds, but from the alternative explanation of the correlation effect, we may allocate the bulk of the correlation adjustment to the *on-upon* correlation, and take the adjustment away from the odds for *on*. The effect is roughly to halve the log odds for *on*, and then the odds rarely exceed 10 to 1.

As one final indication of acceptability of the log odds based on the negative binomial distribution, the study of Section 4.9 shows that probability predictions under the negative binomial model are not excessive when judged by a criterion that strongly penalizes strong incorrect predictions by single words.

The unadjusted log odds for the negative binomial (words: *by*, *from*, and *to*) perform remarkably well without deflation. In contrast, the Poisson log odds for these words are best halved.

So much for justification of the odds given by our model.

Outrageous events. One could say: "Without judging the matter, let us suppose that you have handled these technical details with surpassing skill; are you not, thus far, neglecting some earthy possibilities?" We must answer "Yes," and try to indicate a few kinds of things that might seriously deflate the model's odds.

Although the chances of a blunder in large-scale calculations are fairly high, we do have parallel work in the other three studies that is largely independent, and that offers some checks on the work—not precise checks, but checks on reasonableness that reduce the chances of grave blunders. Other sorts of poor workmanship may recommend themselves to the reader.

The reader has to ask himself whether we have somehow falsified the evidence. Deliberate fraud or hoax is not impossible in scientific work; at a guess, the chance is well over one in a million. In this problem, such activities could be expensive to detect if we managed to suppress words that were favorable to Hamilton in the disputed papers, but left good discrimination elsewhere.

Certainly every reader wonders whether Hamilton wrote the disputed papers and Madison edited them so fiercely that for the variables we use Hamilton's traces have been washed away. Historical evidence would be important here, but we think it unlikely that these authors would mark up one another's manuscripts in that manner. If one or two marker words were at issue, we could imagine the possibility of such changes' or even of slight printer's errors' mattering, but with so many small words the possibility seems remote.

How could Madison change a Hamilton paper of 2000 words, so that our 30 final words appear with average Madison rates instead of average Hamilton rates? He would have had to remove about 50 occurrences of Hamilton words among the final 30, and to add about 20 occurrences of Madison words.

If instead of thinking in terms of a single paper written by Hamilton and heavily revised by Madison, we were to think of this possibility for all 12 disputed papers, the results of the regression study of Section 4.8 would be relevant. That study shows that word rates in the 12 disputed papers are very near to the Madison rates, and the log odds are close to predictions from theory, given Madison as author. Thus, judged by our 30 words, the disputed papers are much like typical Madison papers. For Madison to have edited more than a few Hamilton papers and not have some of the resulting papers appear like a Hamilton-Madison mixture or a Madison parody instead of a fresh Madison paper would be remarkable indeed.

Still the chance of this does not seem so remote that odds of millions to one are appropriate. If you thought the odds were 1 to 1000 that Hamilton wrote a disputed paper but that it was thoroughly edited by Madison, then 1000-to-1 odds are the most that Madison can have for the paper.

More generally, let $P(\mathrm{M})$ be the probability that Madison wrote the paper, let $1 - a$ be the probability that no outrageous event occurred, let $1 - b$ be

the probability that Madison wrote the paper, given that no outrageous event occurred, and let A be the probability that Madison wrote the paper, given that an outrageous event did occur. Then

$$P(\mathrm{M}) = (1 - a)(1 - b) + aA,$$

and we suppose that a, b are small. Note that $P(\mathrm{M})$ is least when $A = 0$, so that the minimum odds in favor of Madison are

$$(1 - a)(1 - b)/[1 - (1 - a)(1 - b)].$$

Roughly this is $1/(a + b)$. If a is much smaller than b, then the odds are roughly $1/b$, which is the point we have been making. But going to the other extreme, if a is large and b small, then the odds are roughly $1/a$. So, for example, for $a \approx 10^{-3}$, $b \approx 10^{-6}$, the odds go down from 10^6 to 10^3 but for $a \approx 10^{-3}$, $b \approx 10^{-2}$, the odds only go down from 100 to 1 to about 90 to 1.

Clearly, frauds, blunders, poor workmanship, and coincidences snip away at very high odds, and the reader has to supply the adjustment for these. The chance of being brought to what we call "roguish ruin" by one or another sort of outrageous event is sufficiently large that final odds of millions to one cannot be supported, but these long odds can be understood in their own place within the mathematical model. Whenever strong discrimination or precise estimation exists, something akin to these odds of millions to one arises, no matter what approach to statistics is being used. For example, a difference between means of 10 standard deviations may be reported. When we see such a report, we all go through the same process of having reservations, of wanting to verify the correctness or incorrectness of the results, and of continuing to hold reservations if verification is not feasible. (We ask: Has someone multiplied by an extra $\sqrt{n}$? Isn't the wrong measure of variance being used?) But when the results are the usual one to three standard deviations, we are more relaxed. A good many people on first hearing about odds of millions to one are shocked and feel that the methods are automatically discredited. But outrageous events have an impact on all approaches to inference, not just on the Bayesian approach. We have tried to explain both the sense in which long odds or huge differences can occur and the need for reservations in making final appraisals. We do not, however, wish to be whipped by the backlash of the question, "Don't you trust yourselves?" Yes, we do, but not millions to one.

Finally, we do not regard the need for discussion of outrageous events as a shortcoming of our analysis. Rather it shows its strength. Had we obtained modest odds of say five to one, the chances of outrageous events would scarcely modify the results, and the emphasis of the discussion would have been on the uncertainty of the attribution, and whether five to one is of much help to historians. Strong results force the discussion of outrageous events.