# Springer Series in Statistics

Advisors :

D. Brillinger, S. Fienberg, J. Gani,

J. Hartigan, K. Krickeberg

### **Springer Series in Statistics**

L. A. Goodman and W. H. Kruskal, Measures of Association for Cross Classifications. x, 146 pages, 1979.

J. O. Berger, Statistical Decision Theory: Foundations, Concepts, and Methods. xiv, 425 pages, 1980.

R. G. Miller, Jr., Simultaneous Statistical Inference, 2nd edition. xvi, 299 pages, 1981.

P. Brémaud, Point Processes and Queues: Martingale Dynamics. xviii, 354 pages, 1981.

E. Seneta, Non-Negative Matrices and Markov Chains. xv, 279 pages, 1981.

F. J. Anscombe, Computing in Statistical Science through APL. xvi, 426 pages, 1981.

J. W. Pratt and J. D. Gibbons, Concepts of Nonparametric Theory. xvi, 462 pages, 1981.

V. Vapnik, Estimation of Dependences based on Empirical Data. xvi, 399 pages, 1982.

H. Heyer, Theory of Statistical Experiments. x, 289 pages, 1982.

L. Sachs, Applied Statistics: A Handbook of Techniques, 2nd edition. xxviii, 707 pages, 1984.

M. R. Leadbetter, G. Lindgren and H. Rootzen, Extremes and Related Properties of Random Sequences and Processes. xii, 336 pages, 1983.

H. Kres, Statistical Tables for Multivariate Analysis. xxii, 504 pages, 1983.

J. A. Hartigan, Bayes Theory. xii, 145 pages, 1983.

F. Mosteller, D.S. Wallace, Applied Bayesian and Classical Inference: The Case of *The Federalist* Papers. xxxv, 301 pages, 1984.

D. Pollard, Convergence of Stochastic Processes. xiv, 215 pages, 1984.

Frederick Mosteller David L. Wallace

# Applied Bayesian and Classical Inference

## The Case of The Federalist Papers

2nd Edition of Inference and Disputed Authorship: The Federalist



Springer-Verlag New York Berlin Heidelberg Tokyo Frederick Mosteller Department of Statistics Harvard University Cambridge, MA 02138 U.S.A. David L. Wallace Department of Statistics University of Chicago Chicago, IL 60637 U.S.A.

AMS Classification: 6201, 62F15

Library of Congress Cataloging in Publication Data Mosteller, Frederick Applied Bayesian and classical inference. (Springer series in statistics) Earlier ed. published in 1964. Bibliography: p. Includes index. 1. Federalist. 2. English language-Word frequency-Case studies. I. Wallace, David L. (David Lee), II. Mosteller, Frederick, 1916-1928 -Inference and disputed authorship, The Federalist. III. Title. IV. Series. JK155.M59 1984 342.73'029 84-5489

With 11 illustrations.

The first edition of this book, Inference and Disputed Authorship: The Federalist, was previously published by Addison-Wesley Publishing Company, Inc., Reading, MA, in 1964.

© 1964, 1984 by Springer-Verlag New York Inc.

Softcover reprint of the hardcover 2nd edition 1984

All rights reserved. No part of this book may be translated or reproduced in any form without written permission from Springer-Verlag, 175 Fifth Avenue, New York, New York 10010, U.S.A.

ISBN-13: 978-1-4612-9759-8 e-ISBN-13: 978-1-4612-5256-6 DOI: 10.1007/978-1-4612-5256-6

To the memory of Edwin Glenn Olds and Samuel Stanley Wilks

### **Preface to the Second Edition**

The new version has two additions. First, at the suggestion of Stephen Stigler, we have replaced the Table of Contents by what he calls an Analytic Table of Contents. Following the title of each section or subsection is a description of the content of the section. This material helps the reader in several ways, for example: by giving a synopsis of the book, by explaining where the various data tables are and what they deal with, by telling what theory is described where. We did several distinct full studies for the *Federalist* papers as well as many minor side studies. Some or all may offer information both to the applied and the theoretical reader. We therefore try to give in this Contents more than the few cryptic words in a section heading to speed readers in finding what they want.

Second, we have prepared an extra chapter dealing with authorship work published from about 1969 to 1983. Although a chapter cannot comprehensively cover a field where many books now appear, it can mention most of the book-length works and the main thread of authorship studies published in English. We found biblical authorship studies so extensive and complicated that we thought it worthwhile to indicate some papers that would bring out the controversies that are taking place. We hope we have given the flavor of developments over the 15 years mentioned.

We have also corrected a few typographical errors.

As usual, many have helped us. Erich Lehmann and Walter Kaufmann-Bühler suggested that we prepare this new edition. Many have advised us about the writing or contributed material—Persi Diaconis, Bradley Efron, Alvar Ellegård, John D. Emerson, Irene Fairley, Katherine Godfrey, Katherine Taylor Halvorsen, David C. Hoaglin, Peter J. Kempthorne, Erich L. Lehmann, Colin Mallows, Lincoln E. Moses, Marjorie Olson, Stephen L. Portnoy, Stephen M. Stigler, and Wing H. Wong. Augustine Kong carried out an important computer search of the literature. Cleo Youtz edited and re-edited the manuscript of the new material and checked it for accuracy. The work was partly facilitated by National Science Foundation Grant SES 8023644 to Harvard University.

### **Preface to the First Edition**

We apply a 200-year-old mathematical theorem to a 175-year-old historical problem, more to advance statistics than history. Though problems of disputed authorship are common in history, literature, and politics, scholars regard their solutions as minor advances. For us the question of whether Hamilton or Madison wrote the disputed *Federalist* papers has served as a laboratory and demonstration problem for developing and comparing statistical methods. While we help solve this historical problem, our practical application of Bayes' theorem to a large analysis of data is a step in testing the feasibility of a method of inference that has been heavily criticized in the past, but which is currently being explored with fresh attitudes and fresh mathematics. Furthermore, large practical applications have until now been few, and our work helps fill that gap.

Historians will find that our results strengthen their general trend of opinion in favor of Madison's authorship of the disputed papers by providing a different sort of evidence from that they have ordinarily used. They can add these results to evidence of other kinds.

Methods like ours can be used for other authorship studies, and we anticipate that the cost will become relatively cheap in the future. Preparing text for a high-speed computer is currently responsible for a major part of the cost. Savings should come when an electronic reader becomes available that can read text directly from printed material into the computer.

Some may feel that we should make more definite pronouncements about the comparative value of different methods of inference—especially Bayesian versus classical. Obviously we are favorably disposed toward the Bayesian approach, but in the field of data analysis at least, the techniques are in their infancy. As these techniques evolve, their merits and demerits may be easier to judge.

Even though individual statisticians may claim generally to follow the Bayesian school or the classical school, no one has an adequate rule for deciding what school is being represented at a given moment. When we have thought we were at our most Bayesian, classicists have told us that we were utterly classical; and when we have thought ourselves to be giving a classical treatment, Bayesians have told us that the ideas are not in the classical lexicon. So we cannot pretend to speak for anyone but ourselves, and we do this mainly in Chapter 9 by pointing out some developments that we think statistical inference needs. After an introduction to the historical problem and the data to be used (Chapters 1 and 2), we launch four parallel studies. They are the main Bayesian study (Chapters 3 and 4), a standard approach to discrimination (Chapter 5), and two other simplified approaches: one Bayesian (Chapter 6) and one using Bayesian ideas but not a Bayesian interpretation (Chapter 7). Chapter 8 presents ancillary studies and describes a simplified approach to an authorship problem. Finally, Chapter 9 summarizes our conclusions. A reader who wants to know quickly our main results could read Chapters 1, 2, 3, 8, and 9. He would skip the parallel developments of Chapters 5, 6, and 7, and the extensive and mathematical Chapter 4. Though a reading of Chapter 4 might well come last and then only by the mathematically inclined, statisticians may find the most professional meat there and in Chapter 9.

The main study gives results in terms of odds. These odds tend to be very large for most of the papers, and they are not to be taken at face value. To appreciate them or, perhaps better, to depreciate them properly, the reader should study Section 3.7F with care.

A word about our exposition of mathematics outside of Chapter 4 may help bring order out of chaos. When we found we could in a reasonable space explain a mathematical development in detail for those without much mathematical preparation, we have tried to do this. But where the needed mathematical preparation was substantial, we have not tried to simplify the exposition. We felt that the unprepared reader would have to skip through these more difficult developments in any case and that he would have no trouble in recognizing these spots. We have, of course, tried to make the principal results understandable for all.

In discussing our work on The Federalist at the Minnesota meetings reported below, Jerzy Neyman suggested that categorizing statistical methods as Bayesian or non-Bayesian is less revealing than categorizing them as inferential or behavioristic, in either of which Bayes' theorem may often be used. The behavioristic approach in our problem calls for establishing a rule for deciding who wrote any disputed paper and evaluating or bounding the frequencies of incorrect classifications if the rule is followed. In the inferential approach, one tries to provide odds or other measures of confidence for (or against) Madison's authorship of any paper. Under these definitions, our methods of Chapters 3 and 6 are clearly inferential, though we could, if necessary, immediately specify a decision rule, for example, by deciding for Hamilton if the log odds are positive, for Madison if they are negative. The methods of Chapter 5 and 7 fall more nearly in the behavioristic approach, though they become more inferential when in Section 5.5 we try to assess the strength of the evidence by computing tail probabilities and by estimating confidence limits on the likelihood ratio.

Neyman also suggested that nonparametric discrimination methods along lines developed by Fix and Hodges (1951, 1952) would be of considerable interest in our problem. Although we have not followed up this suggestion, we hope others may, and have added the papers to our reference list. Our references are gathered at the end of the book.

We are aware of a considerable body of writing both on problems of discrimination generally and on the analysis of style for purposes of deciding authorship. To review these works would require a monograph comparable in size to the present one, and so we have not made the attempt. Hodges' (1950) review of work in discrimination fills part of this void, though a comparable new review would be welcome.

#### Acknowledgments

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

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

We appreciate the careful calculations and other work of Linda Alger, Virgil Archer, Judithe Bailey, Eleanor Beissel, Barbara Block, Mary Blyman, Alf Brandin, John Burnham, Margaret Byron-Scott, Helen Canning, Philip Chartrand, Adelle Crowne, Roy D'Andrade, Abraham Davidson, Roy Dooley, Sara Dustin, Gerald Fowler, Miriam Gallaher, Yoel Haitovsky, Jane Hallowell, Joanna Handlin, Joan Hastings, Elizabeth Ann Hecht, Mervyn Hecht, Ann Hilken, Theodore Ingalls, Helen V. Jensen, Kathryn Karrasik, Vincent Kruskal, Frederick Loo, Christine Lyman, Nancy McCarthy, William Mosteller, Joseph Naus, Loneta Newburgh, Mary Nye, Eva Pahnke, Mitchell Robbins, Susan Rogers, Eleanor Rosenberg, Astrid Salvesen, Epifanio San Juan, Jr., Mary Shillman, Lucy Steig, Ralph A. Stewart, Jr., Ruth M. Stouffer, Victoria Sullivan, George Telford, Elizabeth Thorndike, Henry Tibery, Martha Van Genderen, Bruce B. Venrick, C. Kristine Wallstrom, Druscilla Wendoloski, Richard Wendoloski, Herbert Winokur, and Charles Zimmer.

For work on the manuscripts leading to this work, we thank Mrs. Jane Adams, Mrs. Irene Bickenbach, Mrs. Rita Chartrand, Joan Gobby, Mrs. Vendolyn Harris, Mrs. Angela Klein, Mrs. Mary McQuillin, Janet Mendell, Marguerite O'Leary, Mrs. Helena Smith, Mrs. Cleo Youtz, and Phyllis Zamatha. Several publishers and individuals have kindly given us permission to quote from their copyrighted works:

The quotations in Chapter 1 from the editor's introduction to *The Federalist*, edited by Jacob E. Cooke, copyright © 1961 by Wesleyan University are reprinted by permission of Wesleyan University Press.

The quotation early in Chapter 3 from Julian Lowell Coolidge's book, *Introduction to mathematical probability*, is reprinted here by permission of the copyright owners, the Clarendon Press, Oxford.

The quotation early in Chapter 3 from Egon Pearson's 1925 article in *Biometrika*, "Bayes' theorem, examined in the light of experimental sampling," is quoted with the permission of the author and the Editor of *Biometrika*.

The quotation in Chapter 3 from Joseph Berkson's 1930 article in *The Annals* of *Mathematical Statistics*, entitled "Bayes' theorem," is given with the permission of the author.

We reported publicly on this research at a session of Special Papers Invited by the Presidents of the American Statistical Association, The Biometric Society (ENAR), and The Institute of Mathematical Statistics at the statistical meetings in Minneapolis, Minnesota, September 9, 1962. The prepared discussants were Douglass Adair, F. J. Anscombe, and Jerzy Neyman, with Leo Goodman in the chair. We presented the material in Mosteller and Wallace (1963).

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

May, 1964

F. M. and D. L. W.

# **Analytic Table of Contents**

Chapter 1. The Federalist Papers As a Case Study	1
<b>1.1. Purpose</b>	1
<b>1.2.</b> The Federalist papers	2
<b>1.3. Early work</b>	6
1.4. Recent work—pilot study . We call marker words those which one author often uses and the other rarely uses. Douglass Adair found <i>while</i> (Hamilton) versus <i>whilst</i> (Madison). We found <i>enough</i> (Hamilton) and <i>upon</i> (Hamilton); see Tables 1.4–1, 2 for incidence and rates. Tables 1.4–3, 4, 5 give an over- view of marker words for <i>Federalist</i> and non- <i>Federalist</i> writings. Alone they would not settle the dispute compellingly.	10
1.5. Plots and honesty	14 F
1.6. The plan of the book	. 15

Chap	ter 2. Words and Their Distributions	16
2.1.	Why words?	16
2.2.	Variation with time	19
2.3.	How frequency of use varies	22
	<b>2.3A. Independence of words from one block of text to another</b> A special study of extensive empirical data tests the independence of the occurrences of the same word (for 51 words) in four successive blocks of approximately 200 words of Hamilton text. Table 2.3–1 compares the observed counts with the binomial distributions for the 39 sets of four blocks for each word. Some words give evidence of lack of independence, especially <i>his, one, only,</i> and <i>than</i> .	23
	<b>2.3B. Frequency of occurrence.</b> For 51 words we show in Table 2.3–3 the frequency distribution of occurrences in about 250 blocks of 200. The Poisson distribution does not fit all the empirical distributions of the number of occurrences of high-frequency words, but the negative binomial distribution comes close to doing so. For 10 of these words Poisson and negative binomials are fitted and displayed in Table 2.3–4 for Hamilton and for Madison. The negative binomial distribution allows for higher tails than does the Poisson.	28
2.4.	<b>Correlations between rates for different words</b>	35
2.5.	Pools of words	37
	<b>2.5A. The function words</b> . From a list of 363 function words prepared by Miller, Newman, and Friedman, we selected the 70 highest-frequency and a random 20 low-frequency words without regard to their ability to discriminate authorship. They appear in Tables 2.5–2 and 2.5–3.	39

	<b>2.5B. Initial screening study</b> We used some of the papers of known authorship to cut 3000 candidate words to the 28 listed in Table 2.5–4, based on ability to discriminate.	39
	<b>2.5C. Word index with frequencies</b> From 6700 different words, 103 non-contextual words were chosen from 240 that looked promising as discriminators on papers of known authorship. Of these words, the 48 in Table 2.5–6 were new.	42
2.6.	Word counts and their accuracies Some word counts were carried out by hand using slips of paper, one word per slip. Others were done by a high-speed computer which constructed a concordance.	43
2.7.	<b>Concluding remarks</b> Although words offer only one set of discriminators, one needs a large enough pool of potential discriminators to offer a good chance of success. We need to avoid selection and regression effects. Ideally we want enough data to get a grip on the distribution theory for the variables to be used.	45
Chap	ter 3. The Main Study	46
	In the main study, we use Bayes' theorem to determine odds of author- ship for each disputed paper by weighting the evidence from words. Bayesian methods enter centrally in estimating the word rates and choosing the words to use as discriminators. We use not one but an empirically based range of prior distributions. We present the results for the disputed papers and examine the sensitivity of the results to various aspects of the analysis. After a brief guide to the chapter, we describe some views of prob- ability as a degree of belief and we discuss the need and the difficulties of such an interpretation.	
3.1.	Introduction to Bayes' theorem and its applications We give an overview, abstracted from technical detail, of the ideas and methods of the main study, and we describe the principal sources of difficulties and how we go about meeting them.	49
	<b>3.1A.</b> An example applying Bayes' theorem with both initial odds and parameters known	52
	Final odds = initial odds $\times$ likelihood ratio.	
	<b>3.1B.</b> Selecting words and weighting their evidence Applying Bayes' theorem to several words, and taking logarithms gives the final log odds as the sum of initial log odds and the log likelihood ratios for the separate words. The difference between the expected log likelihood ratio for the two authors is a measure of importance of the	54

word as a discriminator. We discard words with small importances. No bias arises from selection when rates are known.

	<b>3.1C. Initial odds</b>							
	<b>3.1D. Unknown parameters</b> Even if data distributions were Poisson, we would not know the mean rates. From the known Hamilton and Madison texts, we can estimate the rates, but with important uncertainties: the simple use of Bayes' theorem is not quite right, and the selection effects in choosing the words are not negligible. We treat the rates as random quantities and use the con- tinuous form of Bayes' theorem to determine the posterior distribution to represent their uncertainty. Figure 3.1–1 shows the logical structure of the two different uses of Bayes' theorem. The factor from initial odds to final odds is no longer a simple likelihood ratio, but a ratio of two averaged probabilities, averaged over the posterior distributions of the word rates. The factor can often be approximated by a likelihood ratio for an appropriately estimated set of rates.	57						
3.2.	Handling unknown parameters of data distributions	60						
	<b>3.2A.</b> Choosing prior distributions We expect both authors to have nearly the same rates for most words, we shift to parameters measuring the combined rate and a differential rate. For any word, let $\sigma$ be the sum of the rates for the two authors and let $\tau$ be the ratio of Hamilton's rate to the sum $\sigma$ . Empirical evidence on 90 unselected words illustrated in the Figure 3.2–1 plot of estimated parameters guides the choice of families of prior distributions for $\sigma$ and $\tau$ .	61						
	<b>3.2B.</b> The interpretation of the prior distributions	63						
	<b>3.2C. Effect of varying the prior</b>	63						
	<b>3.2D.</b> The posterior distribution of $(\sigma, \tau)$ For any choice of the underlying constants, the joint posterior density of $(\sigma, \tau)$ follows directly from Bayes' theorem. The mode of the posterior density can be located by numerical methods and gives the <i>modal</i> estimates of parameters used for determining the odds of authorship.	64						

**3.2E.** Negative binomial . . . . . . . . . The negative binomial data distribution underlies our best analysis of authorship. The parametrizations and the assumed families of prior distributions are set forth. The priors are parametrized by five underlying constants. Posterior modal estimates were obtained for all words under each of 21 sets of underlying constants. For one typical set, Table 3.2-3 presents the modal estimates of the negative binomial parameters for the final 30 words used to assess the disputed papers.

3.2F. Final choices of underlying constants . . . . . . 67 Analyses (to be described in Section 4.5) of a pool of 90 unselected words provide plausible ranges for the underlying constants. Table 3.2-2 shows six choices in that range. We interpret the effect of the five underlying constants and describe an approximate data-equivalence for the prior distributions they specify.

#### 3.3. Selection of words

. . . . . . . . . . . The prior distributions are the route for allowing and protecting against selection effects in choice of words. We use an unselected pool of 90 words for estimating the underlying constants of the priors, and we assume the priors apply to the populations of words from which we developed our pool of 165 words. We then selectively reduce that pool to the final 30 words. We describe a stratification of words into word groups and our deletion of two groups because of contextuality.

#### 3.4. Log odds

We compute the logarithm of the odds factor that changes initial odds to final odds and call it simply log odds. The computations use the posterior modal estimates as if they were exact and are made under the various choices of underlying constants and using both negative binomial or Poisson models.

. .

.

**3.4A.** Checking the method . . . . . 69 Table 3.4-1 shows the total log odds over the 30 final words when each of the 98 papers of known authorship is treated as if unknown. It shows the results for six choices of prior for the negative binomial, four for the Poisson. For almost all papers with known author, the log odds strongly favor the actual author. Choice of prior makes about 10 per cent difference in the log odds. Choice of data distribution has far larger effects. Paper length matters, and paper-to-paper variation is huge. 3.4B. The disputed papers

. . . . . . . . 75 For each disputed paper, Table 3.4-2 shows the log odds factors, totaled for the 30 final words, for ten choices of priors, six for the negative binomial and four for a Poisson model. The evidence strongly favors Madison, with paper 55 weakest with an odds factor of 240 to 1. 25 Log odds by words and word groups

9.9	Log ouus by words and	1 W	ora	grou	ps .	•	•	•	•	•	•	•	•	•	•	-77
	3.5A. Word groups								•				•			77
	Table 3.5-1 breaks t	he	log	odd	s int.	o ce	ontri	but	ions	by	' th	le f	ive	wo	rd	
	groups for the disput	ted	, joi	nt, e	and s	ome	e pap	oers	of	kno	wn	au	tho	rshi	p.	
	The general consister	ncy	ofe	evid	ence	is e:	xami	inec	1.						•	

65

67

69

77 Tables 3.5-2A, B, C show the contributions to the log odds from single words: 9 high-frequency words, 11 Hamilton markers, 9 Madison markers. The gross difference between behavior of Poisson and negative binomial models for extreme usages of rare words is illustrated. 3.5C. Contributions of marker and high-frequency words . . 81 Table 3.5-3 shows how papers with words at the Madison mean rate, at the Hamilton mean rate, and at the average would be assessed; also how papers with all or none of the Hamilton or of the Madison markers would fare. The comparisons support the fairness of the final 30 words. 3.6. Late Hamilton papers . . . . . 83 We assess the log odds for four of the late Federalist papers, written by Hamilton after the newspaper articles appeared and not used in any of our other analyses. The log odds all favor Hamilton, very strongly for all but the shortest paper. 84 Through special studies, we estimate the magnitude of effects on the log odds of various approximations and imperfect assumptions underlying the main computations and results presented in Section 3.4. Percentage reductions in log odds are a good way to extrapolate from the special studies to the main study. **3.7A.** Correlation 84 The study of correlations among words suggests that log odds based on independence should be reduced by an amount between 6 per cent and 12 per cent. 3.7B. Effects of varying the underlying constants that determine the prior distributions . . . . . . . . . . . . . . . . . . 84 The choice of prior distribution used in most of the presented results is in the middle of the estimated range of the underlying constants. Other choices might raise or lower the log odds, but not likely by more than  $\pm 12$  per cent. 3.7C. Accuracy of the approximate log odds calculation . . . . 85 A study of the approximation for five of the most important words suggests that the modal approximation tends to overstate the log odds and that a 15 per cent reduction is indicated. **3.7D.** Changes in word counts . . . . . . . . . 86 Some word changes between the original newspaper editions and the McLean edition we used for making our word counts require adjustment. Two changes involving upon and whilst reduce the log odds for Madison in two disputed papers. Other errors, including counting errors, are smaller and nearly balanced in direction. 3.7E. Approximate adjusted log odds for the disputed papers . . . 88 Table 3.7–2 shows the log odds for the disputed papers after making the specific adjustments for the major word changes, and with three levels of a composite adjustment for other effects. Even the extreme

88

92

adjustment leaves all but two papers with odds of over 2500 to 1 favoring Madison, and the two weakest at 33:1 (paper #55 with log odds -3.5) and 180:1 (paper #56 with log odds -5.2).

**3.7F.** Can the odds be believed? . . . The odds, even after adjustment, are often over a million to one, and on average about 60,000 to 1. We note that all forms of statistical inference have the equivalents of such strong evidence, but in different forms from the Bayesian odds calculations. We discuss the believability of such odds from the standpoint of statistical models, and then from a broader viewpoint external to the model, allowing for what we call outrageous events. We examine how one can ever justify strong evidence for discrimination, and how independent evidence can be built up. We see how the evidence from upon is reasonable and more defensible for a pro-Madison finding than it would have been in a pro-Hamilton finding. We note some potential failings such as computational and other blunders, fraud and serious errors, which can never be absolutely ruled out. We offer evidence for the implausibility of Madison's having edited Hamilton's papers to look like his own writings in the way we assess his style. A probability calculation shows how a small probability of an outrageous event has little impact on weak evidence from a statistical analysis, but does put a bound on strong evidence.

Chapter 4. Theoretical Basis of the Main Study

This chapter is a sequence of technical sections supporting the methods and results of the main study presented in Chapter 3. We set out the distributional assumptions, our methods of determining final odds of authorship, and the logical basis of the inference. We explain our methods for choosing prior distributions. We develop theory and approximate methods to explore the adequacy of the assumptions and to support the methods and the findings.	•
<b>4.1. The negative binomial distribution</b>	93
<b>4.1A. Standard properties</b>	93 , , ;
<b>4.1B. Distributions of word frequency . . . . . . . . . .</b>	96 1
<b>4.1C. Parametrization</b>	. 96 7 -

We choose the mean and a measure of deviation from the Poisson that is not the usual choice.

**4.2A.** The data: notations and distributional assumptions . . . . . 99 Notation and formal distributional assumptions are set out for all words and all papers of known authorship for negative binomial and Poisson models.

100

4.2F. An alternative choice of modes 106 . . . . . . . Modes of asymmetric densities are not ideal for approximating posterior expectations. Some inexpensive improvements come from using modes of densities relative to a measure element other than Lebesgue measure. For the gamma- and beta-like prior densities used here, these relative modes are equivalent to a change in the underlying constants.

4.2G. Choice of initial estimate . . 108 . . . . Iterative procedures require starting values; method-of-moment estimates are natural candidates but are inadequate for low-frequency words where the shrinking effect of the prior density is strong. An approximate data equivalent of the prior leads to weighted initial estimates of good quality. Combining tight-tailed priors with long-tailed data distributions gives rise to special needs that must be faced in the absence of sufficiency or conjugacy.

4.3. Abstract structure of the main study. 111 . . We describe an abstract structure for our problem; we derive the appropriate formulas for our application of Bayes' theorem and give a formal basis for the method of bracketing the prior distribution. The treatment is abstracted both from the notation of words and their distributions and from numerical evaluations.

4.3A. Notation and assumptions . . . . . . . . . . . . . 111 Four initial assumptions model the probabilistic relations among the observables (the data on the disputed papers and the data on the known-author papers) and the non-observables (the parameters of the data distributions and the authorship of the disputed papers). The authorship is the goal of the analysis of The Federalist. The basic application of Bayes' theorem represents the final odds of authorship as the product of the initial odds of authorship and an odds factor that involves the data on the known papers.

4.3B. Stages of analysis . . . . 112 The factorization in Section 4.3A divides the analysis into three stages: choosing data distributions and estimating their parameters, evaluating the odds factors for the disputed papers, and combining the odds factors with initial odds of authorship. The first two are heavily statistical.

4.3C. Derivation of the odds formula 112 The fundamental factorization result of Section 4.3A is derived from four assumptions.

113 Historical evidence bears on authorship and can be treated as logically prior to the analysis of the linguistic data. A fifth assumption sets out what is needed for the statistical evidence that determines our odds factors to be independent of and acceptable to historians, regardless of how they assess the historical evidence. This subjective element is isolated to the assessment of the initial odds.

**4.3E.** Odds for single papers . 114 Odds factors for authorship of a single paper are interesting and important.

<b>4.3F. Prior distributions for many nuisance parameters</b> Our data consist of word frequencies for more than a hundred words. Modeling each as distributed independently as a negative binomial leads to four parameters per word. Estimating hundreds of parameters with the available data cannot be done safely using a flat prior, or with any non-Bayesian equivalent such as maximum likelihood. Here, we consider the abstract notion of modeling the behavior of the word- frequency parameters as sampled from a hyperpopulation. The hyper- population is modeled as a parametric family of low dimension with parameters we call <i>underlying constants</i> but for which <i>hyperparameter</i> has come into common use by 1984. In lieu of an infeasible full Bayesian analysis, we propose to carry out the main analysis conditional on assumed known values of the hyperparameters. The hyperparameters are estimated in a separate analysis and the sensitivity of the main results to the assumed hyperparameters is explored. The method is empirical, and the Bayesian logic is examined. Some similarities and some distinctions from Robbins' "empirical Bayes procedures" are noted.	114
<b>4.3G. Summary</b>	117
<b>4.4 Odds factors for the negative binomial model</b>	117
<b>4.4A. Odds factors for an unknown paper</b>	117
<b>4.4B. Integration difficulties in evaluation of <math>\lambda</math></b> For any word, the posterior distribution for the four parameters is determined up to a normalizing constant. To get the marginal distributions of the two Hamilton or of the two Madison parameters would require quadrature or other approximation. The calculations of the exact odds factor $\lambda$ for any word and unknown paper then is a ratio of two four-dimensional integrals, a formidable calculation that we bypass by the modal approximation.	119
<b>4.4C. Behavior of likelihood ratios</b>	120

examined. It is not linear, but is unbounded, and to prevent any word