Analytic Table of Contents

Chapter 1. *The Federalist* Papers As a Case Study 1

1.1. Purpose 1
To study how Bayesian inference works in a large-scale data analysis, we chose to try to resolve the problem of the authorship of the disputed *Federalist* papers.

1.2. *The Federalist* papers 2
The Federalist papers were written by Hamilton, Madison, and Jay. Jay's papers are known. Of the 77 papers originally published in newspapers, 12 are in dispute between Hamilton and Madison, and 3 may regarded as joint by them. Historians have varied in their attributions.

1.3. Early work 6
Frederick Williams and Frederick Mosteller found that sentence length and its variability within papers did not discriminate. Tables 1.3–1, 2, 3, 4 show that they found some discriminating power in percentage of nouns, of adjectives, of one- and two-letter words, and of *the*'s. Together these variables could have decided whether Hamilton or Madison wrote all the disputed papers, if that were the problem, but the problem is to make an effective assignment for each paper.

1.4. Recent work—pilot study 10
We call marker words those which one author often uses and the other rarely uses. Douglass Adair found *while* (Hamilton) versus *whilst* (Madison). We found *enough* (Hamilton) and *upon* (Hamilton); see Tables 1.4–1, 2 for incidence and rates. Tables 1.4–3, 4, 5 give an overview of marker words for *Federalist* and non-*Federalist* writings. Alone, they would not settle the dispute compellingly.

1.5. Plots and honesty 14
Some say that the dispute is not a matter of honesty but a matter of memory. Hamilton was hurried in his annotation by an impending duel, but Madison had plenty of time. Editing may be a hazard. We want to use many words as discriminating variables.

1.6. The plan of the book 15

Chapter 2. Words and Their Distributions 16

2.1. Why words? 16
Hamilton and Madison use the same words at different rates, and so their rates offer a vehicle for discrimination. Some words like *by* and *to* vary relatively little in their rates as context changes, others like *war* vary a lot, as the empirical distributions in the four tables show. Generally, less meaningful words offer more stability.

2.2. Variation with time 19
In Table 2.2–2, a separate study illustrated by Madison's rates for 11 function words over a 26-year period examines the stability of rates through time. We desire stability because we need additional text of known authorship to choose words and their rates for discriminating between authors. Among function words, some pronouns and auxiliary verbs seem unstable.

2.3. How frequency of use varies 22
For establishing a mathematical model, we need to find out empirically how rates of use by an author vary from one chunk of writing to another.

2.3A. Independence of words from one block of text to another . . . 23
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 *his*, *one*, *only*, and *than*.

2.3B. Frequency of occurrence. 28
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.

2.4. Correlations between rates for different words 35
Theoretical study shows that the correlation between the rates of occurrence for different words should ordinarily be small but negative. An empirical study whose results appear in Table 2.4–1 shows that these correlations are ordinarily negligible for our work.

2.5. Pools of words 37
Three pools of words produced potential discriminators.

2.5A. The function words 39
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.

2.5B. Initial screening study 39
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.

2.5C. Word index with frequencies 42
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.

2.6. Word counts and their accuracies 43
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.

2.7. Concluding remarks 45
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.

Chapter 3. The Main Study 46

In the main study, we use Bayes' theorem to determine odds of authorship 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 probability 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 49
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.

3.1A. An example applying Bayes' theorem with both initial odds and parameters known 52
A simple probability calculation gives the probability of authorship from evidence on one word. Casting the result, the classical Bayes' theorem, into odds form is helpful and gives:

$$\text{Final odds} = \text{initial odds} \times \text{likelihood ratio.}$$

3.1B. Selecting words and weighting their evidence 54
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

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

3.1C. Initial odds 56
Initial odds of authorship reflect the investigator's assessment of the historical evidence. The final odds is a product of the initial odds and the likelihood ratio, and a large likelihood ratio can overwhelm most variation in initial odds. We concentrate on the likelihood ratio. Our serious use of Bayes' theorem lies elsewhere, in our handling of unknown parameters.

3.1D. Unknown parameters 57
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 continuous 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.

3.2. Handling unknown parameters of data distributions 60
We begin to set out the components of our Bayesian analysis.

3.2A. Choosing prior distributions 61
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 σ be the sum of the rates for the two authors and let τ be the ratio of Hamilton's rate to the sum σ. 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 σ and τ.

3.2B. The interpretation of the prior distributions 63
We work with a parametric family of prior distributions, and call its parameters *underlying constants*. By 1984, *hyperparameters* has become the accepted term for them.

3.2C. Effect of varying the prior 63
We do not determine a single choice of the underlying constants, but study the sensitivity of the assessments of authorship to changes in the prior distributions reflecting changes in the underlying constants.

3.2D. The posterior distribution of (σ, τ) 64
For any choice of the underlying constants, the joint posterior density of (σ, τ) follows directly from Bayes' theorem. The mode of the posterior density can be located by numerical methods and gives the *modal estimates* of parameters used for determining the odds of authorship.

3.2E. Negative binomial 65
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 67
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 69
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.

3.5 Log odds by words and word groups 77

3.5A. Word groups 77
Table 3.5–1 breaks the log odds into contributions by the five word groups for the disputed, joint, and some papers of known authorship. The general consistency of evidence is examined.

3.5B. Single words 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.

3.7. Adjustments to the log odds 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 ± 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

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? 88
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 92

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.

4.1. The negative binomial distribution 93
We review and develop properties of the negative binomial family of distributions.

4.1A. Standard properties 93
For the negative binomial, we set out the frequency function, the cumulant generating function, the first four cumulants, the representation as a gamma mixture of Poisson distributions, and the limiting relationship to the Poisson family.

4.1B. Distributions of word frequency 96
The mixture representation motivates the negative binomial as a distribution of word frequency.

4.1C. Parametrization 96
Several parametrizations of the negative binomial are compared by criteria of interpretability in several modelings, asymptotic orthogonality, and stability of value across applications to different words.

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

4.1D. Estimation 97
Parameter estimation by maximum likelihood has no closed forms (except for the mean when all paper lengths are the same). The method of moments gives initial estimates for use directly or as starting values for iteration. Explicit method-of-moments estimates and approximate standard errors are given.

4.2. Analysis of the papers of known authorship 99
We treat the choice of prior distributions, the determination of the posterior distribution, and the computational problem in finding posterior modes.

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.

4.2B. Object of the analysis 100
The odds factor for authorship of any unknown paper is a ratio of posterior expectations, taken over the distribution of parameters posterior to the data on papers of known authorship. A *modal approximation* is natural and leads to the determination of posterior modal estimates as a principal intermediate goal of the analysis.

4.2C. Prior distributions: assumptions 100
For each word, two negative binomial parameters describe Hamilton's usage, and two describe Madison's usage. These four are reparametrized to a form in which a sampling model for a pool of words is in accord with empirical support from studies of method-of-moments estimates. Table 4.2–1 presents the method-of-moments estimates for 22 function words. A parametric family of prior distributions is assumed with five *hyperparameters* that we call *underlying constants*. The 21 sets of underlying constants used in sensitivity studies are listed in Table 4.2–2.

4.2D. The posterior distribution 103
For any choice of underlying constants, the posterior distributions are independent across words. For each word, the posterior is a four-dimensional density known up to its normalizing constant. The posterior mode and the Hessian matrix of second derivatives of the logarithmic density are determined by a Newton–Raphson iterative algorithm.

4.2E. The modal estimates 106
The posterior modal estimates are the main output of the empirical Bayesian analysis of the papers of known authorship and the main input for assessing the evidence of authorship on any unknown paper. The Hessian matrices are important for exploring the adequacy of approximations. The modal estimates for the 30 final words and one choice of prior were set out in Table 4.2–3.

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.

4.3D. Historical information 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.

4.3F. Prior distributions for many nuisance parameters 114
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 hyperpopulation is modeled as a parametric family of low dimension with parameters we call *underlying constants* but for which *hyperparameter* 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.

4.3G. Summary 117

4.4 Odds factors for the negative binomial model 117
We develop properties of the Poisson and negative binomial families of distributions. The discussion of appropriate shapes for the likelihood ratio function may suggest new ways to choose the form of distributions.

4.4A. Odds factors for an unknown paper 117
The odds factor for an unknown paper is the product, over words, of a ratio of expectations of two negative binomial probabilities, the numerator expectation with respect to the posterior distribution of the Hamilton parameters, the denominator with respect to the posterior distribution of the Madison parameters for the word.

4.4B. Integration difficulties in evaluation of λ 119
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 λ 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.

4.4C. Behavior of likelihood ratios. 120
With known parameters and a single word, the odds factor is a simple likelihood ratio depending on the frequency of the word in the unknown paper. Likelihood ratios whose logarithms are monotone or even linear are popular in statistical theory, and arise for Poisson and other exponential family models. For representing intuitive assessment of evidence, shapes that redescend toward zero for very high (and suspect) frequencies are appealing. The behavior for the negative binomial is examined. It is not linear, but is unbounded, and to prevent any word

contributing excessively, truncation rules were set up to prevent any word from contributing more strongly than the extreme observed in the 98 papers of known authorship.

4.4D. Summary 124
Further work is needed to develop asymptotic expansions and well-designed quadrature and Monte Carlo methods to evaluate the integrals that arise in Bayesian analyses. Also needed is a greater range of appropriate shapes for log likelihood ratios.

4.5. Choosing the prior distributions 124
We give methods for choosing sets of underlying constants to bracket the prior distributions and we explore the effects of varying the prior on the log odds. Choices are based in part on empirical analysis but also on heuristic considerations of reasonableness, analogy, and tractability.

4.5A. Estimation of β_1 and β_2: first analysis 125
The first two hyperparameters β_1 and β_2 measure the spread of the prior distribution of the differential word rate. A variance components analysis of the observed mean word rates in 90 function words stratified according to total frequency of use leads to estimates of β_1 and β_2 for the pool of function words. The stratification can be collapsed to give a better estimate of β_1. We apply the jackknife procedure with eight random subgroups to produce a standard error for the estimated β_1.

4.5B. Estimation of β_1 and β_2: second analysis 128
If posterior modal estimates of the word rate parameters were used as if exact to estimate hyperparameters, those measuring variation would be too small because of the shrinking effect of the Bayesian estimation. For an analogous binomial problem, the extent of underestimation is determined and used as an informal guide to the actual situation.

4.5C. Estimation of β_3 130
The hyperparameter β_3, measuring the spread of differential non-Poissonness is assessed informally from the frequency distribution of method-of-moments estimates and from the posterior modal estimates of differential non-Poissonness. These tend respectively to show too much and too little variation and bracket β_3. A weakness from an assumed symmetry is considered.

4.5D. Estimation of β_4 and β_5 131
These two hyperparameters that measure the mean and variance of the composite non-Poissonness are assessed by informal analyses.

4.5E. Effect of varying the set of underlying constants 132
The sensitivity of the final log odds factors to the choice of underlying constants or hyperparameters is examined by selective comparisons among the 21 sets chosen to bracket the priors. An appropriate response measure is a proportional change to the log odds, and a 12 per cent change up or down from the primary choice is judged an adequate allowance. The main effect of changing each hyperparameter is examined as are some interactions.

4.5F. *Upon:* a case study 135
The effect of choice of prior on the estimated rates and non-Poissonness parameters for the highly discriminating word *upon* illustrates some possible strange effects of tail behavior of a prior interacting with a four-dimensional likelihood surface. Use of priors conjugate to the likelihood hold few surprises, even when the priors and likelihood are quite inconsistent, because prior and likelihood effectively represent equivalent and exchangeable data. Our gamma- and beta-like priors have tight tails, and in extreme situations can dominate the broader tails of the negative binomial likelihoods, and strongly shift the parameters from the observed rates. This behavior stands in contrast to analyses with flat priors and tight-tailed data distributions.

4.5G. Summary 138
Sensitivity to choice of priors is modest relative to other source of variation. The study of Section 4.5F suggests a point likely important throughout Bayesian inference: the effect of small tails of the prior is very different from the effect of small tails of the data distribution.

4.6. Magnitudes of adjustments required by the modal approximation to the odds factor 138
We study, by example, the effect of using the posterior mode as if it were exact. To make the assessment we develop some general asymptotic theory of posterior densities.

4.6A. Ways of studying the approximation 138
The odds factor is a ratio of two expectations, usually with respect to a concentrated posterior distribution. An expectation can be approximated by the integrand evaluated at the mean or, to a higher order, by the next delta-method adjustment using covariances. We have only the posterior modes and the Hessian matrix of second derivatives at the mode, and want to use that information to assess the modal approximation.

4.6B. Normal theory for adjusting the negative binomial modal approximation 140
We further transform the four parameters for each word to a form in which a normal posterior is a plausible approximation. We use the mode and inverse of the Hessian matrix in the new parametrization as if they were the exact mean and variance matrix. We apply the first two terms of the delta method approximations to the required expectations, and study the changes in log odds for five words, including the three strongest discriminators and two of the strongest rare words. The modal approximation gives log odds that are too large (in magnitude), and a 15 per cent reduction in total log odds is a rough bound for the effect.

4.6C. Approximations to expectations 146
The delta method is based on means and covariances. The *Laplace integral expansion* for a posterior density gives the equivalent approximation in terms of the posterior mode, and second and third derivatives at the mode. Using modes relative to specially chosen density elements

can reduce the role of the third derivatives. Normal, beta, and gamma distributions motivate choices of density elements. Multivariate extensions are set forth.

4.6D. Notes on asymptotic methods 152
The asymptotic basis of the approximations developed and used in the preceding sections is explored for a general posterior density and related to the Laplace integral expansion. The sampling distribution of the posterior density makes plausible the appropriate asymptotic form, but in any application the shape and behavior of the actual density should be examined.

4.7. Correlations 155
We study the magnitudes of effects of erroneous assumptions: the effects of correlations between rates for different words.

4.7A. Independence and odds 155
Odds calculated assuming independence of word usage are likely to be too high. We seek to assess how much the log odds based on independence would differ from a log odds based on a model incorporating dependence.

4.7B. Adjustment for a pair of words 155
If dependence is modeled by bivariate normality, the needed additive adjustment follows from standard discrimination theory. The adjustment has a normal distribution with mean and standard deviation that allow qualitative and quantitative assessment of effects.

4.7C. Example. The words *upon* and *on* 157
The pairwise adjustment model is applied to the two strongest discriminators, *upon* and *on*, whose use is somewhat complementary. Word rates are transformed to a Freeman-Tukey square root scale to stabilize variance and to improve the normality assumption. The observed correlation is mildly negative, and the adjustment reduces the strong discrimination.

4.7D. Study of 15 word pairs 159
The pairwise model is applied to 15 pairs among the 11 words of highest frequency for which the observed correlations exceeds .15, and the mean and standard deviation of the adjustment is shown in Table 4.7–2. The expected adjustments would increase discrimination in nearly half of the pairs but the total of the adjustments would reduce the composite discrimination modestly.

4.7E. Several words 159
The adjustment of Section 4.7B extends directly under multivariate normality. Table 4.7–3 gives the correlation matrix for 11 words. The mean and standard deviation of the needed adjustment are obtained assuming multivariate normality. The observed correlation matrix is used directly, and also after two shrinkage adjustments that allow for the sampling errors in the estimated correlations. The composite results are comparable to the pairwise analysis.

4.7F. Further theory 162
The matrix inversion needed for the Section 4.7E analysis was accomplished by power expansion. Although the computational methods are superseded by readily available facilities for matrix operations, the expansion allows examination of the relation of the full multivariate and the pairwise methods.

4.7G. Summary 163
The adjustments for correlations modestly reduce the discrimination indicated from independence. Open questions on methodology are raised.

4.8. Studies of regression effects 163
To study the adequacy of assumptions, we compare the performance of the log odds for the disputed papers with theoretical expectations.

4.8A. Introduction 163
If all assumptions were correct, and the posterior modal estimates of parameters were exact, we could compute the mean and standard deviation of the log odds that would be obtained for a fresh paper by Hamilton or for one by Madison. Table 4.8–1 displays these expected log odds for the 30 final words for a 2000-word paper. In total, the expected log odds under the negative binomial model is about 14 for a Hamilton paper, about −14 for a Madison paper.

4.8B. The study of word rates 165
For each of the 11 words of highest frequency, we show in Table 4.8–3 the frequency distributions of observed rates in the 48 Hamilton papers, the 50 Madison papers, and the 12 disputed papers. In Table 4.8–2, we show the mean rates in the three groups of papers, as well as their expectations under the assumed Poisson and negative binomial models. Only the disputed papers are independent of the fitted models, and for them, the standard deviation of the rates are also shown, and the lack of regression effects explored.

4.8C. Total log odds for the final 30 words 172
The log odds calculation is applied to each of the known and disputed papers, and an adjustment is made to estimate what the log odds would have been had each paper been 2000 words in length. The mean log odds for the 48 Hamilton papers is compared with the expected log odds for a fresh Hamilton paper as discussed in Section 4.8A. Table 4.8–5 shows the great variability in the observed log odds.

4.8D. Log odds by word group 175
The comparison of observed and expected log odds is made for each of six groups of words, and applied to the known papers used in the estimation and modeling, to the disputed papers, and to four late Hamilton *Federalist* papers not used in any preceding analysis. In three of the five groups there is regression for the known papers to the fresh papers. There is little regression from the expected log odds, for which the prior distributions has made allowance for the selection effects.

4.8E. Theory for the Poisson model 177
Under a Poisson model, log odds are linear in the observed rate and expectations are easily calculated. As preparation for the harder negative binomial model, we note that the expectation can be obtained exactly as a weighted sum of the log odds for any two observed rates. The expected log odds is proportional to paper length and a proportional adjustment to a standard paper length is appropriate.

4.8F. Theory for the negative binomial model 178
Calculating the expected log odds for the negative binomial is not easy, and we use a two-point weighted sum like that for the Poisson. We propose and use an *ad hoc* adjustment to adjust a log odds to what would have been obtained for a standard length paper.

4.8G. Two-point formulas for expectations of negative binomial log odds 179
The basis for the two-point approximations are explored for the negative binomial. The specific uses made would be obviated by modern computational capabilities.

4.9. A logarithmic penalty study 180
We explore a way of assessing the validity of methods for making strong probabilistic predictions when only limited test materials are available. We set up a scoring scheme that penalizes probabilistic predictions (quantitative statements about the probability of a future event) on the basis of outcomes of the events. The approach is not closely tied to authorship and is generally applicable to the evaluation of probabilistic forecasts.

4.9A. Probability predictions 181
A method for making probability predictions is a rule for taking data on a trial, for example, word rates for one unknown paper, and producing a probability distribution over a discrete range of outcomes, for example, authorship. Methods may be Bayesian or not, but we restrict attention to methods that predict separately for each unknown and allow no feedback from results of preceding trials.

4.9B. The *Federalist* application: procedure 181
The study is carried out separately for predictions based on each of three words: *by, on, to*. As test trials, we use the papers of known authorship, specifically, 48 Hamilton and 48 Madison papers. The methods of prediction to be studied allow choice of Poisson or negative binomial models, choice of the underlying constants, choice of initial odds, and choice of an adjustment factor. The latter two are mostly kept at neutral values of 1. The role of reusing parameters estimated from all the data including the test case is discussed.

4.9C. The *Federalist* application: the penalty function 183
For any test paper and method, penalize the probability forecast by an amount equal to the negative logarithm of the probability predicted for the actual author. Better prediction methods should get smaller total penalties. The *penalty score* is like a log likelihood function over the space of "methods". As one calibration of the penality score, we table

how a prediction that was right with constant probability would fare. A correct method should score as well as its data allow, and also should predict how incorrect methods would score. We compare the observed score for one method against the expected score assuming the correctness of a second method, both in units of penality score and in standard deviation units.

4.9D. The *Federalist* application: numerical results 185
For predictions based on rates of high-frequency words, Tables 4.9–2 and 4.9–4 show the observed scores for several Poisson and negative binomial methods and also the expected scores if the respective models were correct. For most words, negative binomial methods get lower (better) scores than do Poisson methods. The negative binomial observed scores are close to what would be expected if their method were correct. The scores for Poisson methods are much worse than what would be expected were their methods correct, showing again that Poisson log odds are severely inflated. More detailed study of *by* develops further support for the correctness of the negative binomial predictions and some support for our choices of the underlying constants. The effect of initial odds is examined briefly.

We comment on the minimization of the penalty score as a criterion for estimation of parameters, and some of its disturbing properties. That method has come into wide use as a conditional maximum likelihood fitting of logistic models (linear for Poisson, but not for negative binomial). Computational capabilities have made the method possible, but the inability of the method usefully to handle very strong discrimination remains a severe limitation.

4.9E. The *Federalist* application: adjusted log odds 189
Would a further multiplicative adjustment factor on the predicted log odds improve performance beyond the Bayesian modeling or whatever was built into the predictive method? For each of three words studied, the factor that minimizes the observed penalty when the predictions are applied to the known papers is very close to 1 for a negative binomial method, and close to .5 for a Poisson method. The discounted Poisson odds are very close to the negative binomial odds, but that does not hold for rare words like *whilst*.

4.9F. The choice of penalty function 190
A proper scoring rule encourages predictions with the correct probabilities when these are known. We show that for predictions of three or more possible outcomes, the logarithmic penalty is the only proper scoring penalty that depends only on the probability predicted for the outcome that obtains. For two outcomes, the logarithmic choice is not unique, but is related to Shannon information, and to likelihood interpretations. Our approach to scoring rules was influenced by work of L. J. Savage that has been published as "Elicitation of Personal Probabilities and Expectations" (*J. Amer. Statist. Assoc.* **66** (1971), 783–801).

4.9G. An approximate likelihood interpretation 192
We develop an approximate representation of the total penalty score as a conditional likelihood for assessing models and methods, conditional on the observed data on the papers.

4.10. Techniques in the final choice of words 195
This section provides details of a special difficulty, and its possible general value lies in illustrating how to investigate the effects of a split into two populations of what was thought to be a single population.

4.10A. Systematic variation in Madison's writing 195
A major part of the known Madison papers comes from a long essay *Neutral Trade* that differs in time and form from his other writings and from the disputed papers. To avoid distortion of the discrimination on the disputed papers, we want to eliminate words for which Madison's patterns of usage change importantly between the two sources of writings. Table 4.10–1 shows for 27 semi-final words the mean log odds for Hamilton versus Madison, and the expected log odds for discriminating between "composite" Madison and "early" or "*Federalist*" Madison.

4.10B. Theory 198
To discriminate between Early and Late Madison writings is a problem equivalent to the main discrimination problem. To get the numbers needed without recourse to major computation, we used a variety of approximations to simplify the problem: linearization of the negative binomial log likelihood ratio, simplified parameter estimates, and Bayes' adjustments. The results are used only to identify major offending words and great precision is not needed.

Chapter 5. Weight-Rate Analysis 200

5.1. The study, its strengths and weaknesses 200
Using a screening set of papers, we choose words and weights to use in a linear discriminant function for distinguishing authors. We use a calibrating set to allow for selection and regression effects. A stronger study would use the covariance structure of the rates for different words in choosing the weights; we merely allow for it through the calibrating set. The zero-rate words also weaken the study because we have not allowed for length of paper as we have done in the main study and in a robust one reported later.

5.2. Materials and techniques 201
Using the pool of words described in Chapter 2, we develop a linear discriminant function $\tilde{y} = \Sigma W_i \tilde{x}_i$, where W_i is the weight assigned to the ith word and $\tilde{x}_i$ is the rate for that word. The W_i are chosen so that $\tilde{y}$ tends to be high if Hamilton is the author, low if Madison is. Ideally the weights are proportional to the difference between the authors' rates and inversely proportional to the sum of the variances. By a

simplified and robust calculation, an index of importance of a word was created. We use it to cut the number of words used to 20.

5.3. Results for the screening and calibrating sets 203
The 20 words, their weights, and estimated importances are displayed in Table 5.3–1, *upon* being outstanding by a factor of 4. Table 5.3–2 shows the results of applying the weights to the screening set of papers. Hamilton's 23 average .87 and all exceed .40, while Madison's 25 average −.41 and all are below −.19. For the calibrating set Hamilton averages .92 and Madison −.38. The smallest Hamilton score is .31, and the largest Madison is .15 (zero plays no special role here).

5.4. Regression effects 208
As a rough measure of separation, we use the number of standard deviations between the Hamilton and Madison means. For the whole set of 20 words, the separation regresses from 6.9 standard deviations in the screening set to 4.5 in the calibrating set. In Section 5.3, we see almost no change from screening to calibration set in the average separations; the loss comes from increased standard deviations. In a general way, as the groups of words become more contextual the regression effect is larger. Group 1, the word *upon*, actually gains strength from screening to calibration set.

5.5. Results for the disputed papers 210
After displaying the numerical outcome of the weight-rate discriminant function for the disputed papers in Table 5.5–1, we carry out two types of analyses, one based on significance tests and one based on likelihood ratios. In Table 5.5–2 we show two t-statistics and corresponding P-values for each paper, first for testing that the paper is a Hamilton paper, and second for testing that the paper is a Madison paper. We compute

$$t_j = \frac{y - \bar{y}_j}{s_j\sqrt{1+(1/n_j)}},$$

where j = Hamilton or Madison, y is the value for the disputed paper from Table 5.5–1, s_j is the standard deviation for author j for the calibrating set, and $n_j = 25$, the number of papers in each calibrating set. Except for paper 55, the P-values for the Hamilton hypotheses are all very small (less than .004); the P-values for the Madison hypotheses are large, the smallest being .087. Paper 55 is further from Madison than from Hamilton but both P-values are significant.

Table 5.5–3 gives log likelihood ratios for the joint and disputed papers, assuming normal distributions and using the means and variances in the calibrating set. To allow for the uncertainty in estimating the means and variances, conservative 90 per cent confidence limits are shown for the log likelihood ratio, and a Bayesian log odds is calculated using the t-distribution. Except for paper 55, which goes slightly in Hamilton's favor, the odds favor Madison for the disputed papers.

Chapter 6. A Robust Hand-Calculated Bayesian Analysis 215

6.1. Why a robust study? 215
Because the main study leans on parametric assumptions and heavy calculations, we want a study to check ourselves that depends less on distributional assumptions and that has calculations that a human being can check. This robust approach, based on Bayes' theorem, naturally sacrifices information. It dichotomizes the observed frequency distributions of occurrences of words. For choosing and weighting words, it uses both a screening set and a validating set of papers.

6.2. Papers and words 216
Using a screening set of 46 papers of length about 2000 words, we selected the words shown in Table 6.2–2 for the robust Bayes study.

6.3. Log odds for high-frequency words 217
For each of the 64 high-frequency words, we divide the rates of the 46 papers in the 2000-word set into two equal parts, highs and lows. For each word, we form a 2×2 table for the high and for the low rates. To estimate the odds (Hamilton to Madison) to be assigned to a word, we first add 1.25 to the count in each of the four cells of the word's 2×2 table. We explain the theoretical framework for this adjustment which is based on a beta prior distribution. We use the adjusted counts to estimate the odds for the high and for the low rate for that word.

6.4. Low-frequency words 220
For low-frequency words, we use the probability of zero occurrence and must adjust the Hamilton-Madison odds according to the length of paper.

6.5. The procedure for low-frequency words 220
Following the theory of Section 6.6, this section explains the arithmetic leading to log odds for each word appropriate to the length of the paper. Ultimately we sum the log odds.

6.6. Bayesian discussion for low-frequency words 222
Theoretical development required for the procedure given in Section 6.5.

6.7. Log odds for 2000-word set and validating set 225
For each of the five groups of words in Table 6.2–2 and in total, Table 6.7–1 shows the log odds for each paper in the 2000-word set used to choose the words and create the odds. All Hamilton papers have positive log odds (averaging 14.0) and all Madison papers have negative log odds (averaging −14.2). Table 6.7–2 gives a more relevant assessment: the same information for the validating set of papers not used to develop the odds. The corresponding averages are 10.2 for 13 Hamilton papers and −8.2 for 18 Madison papers. One Hamilton paper has log odds of 0 or equivalently even odds of 1:1.

6.8. Disputed papers 228
Table 6.8–1 gives the detailed data parallel to the previous tables for the unknown papers. Only paper 55 is not ascribed to Madison. The strength of attribution is, of course, much weaker than in the main study.

Chapter 7. Three-Category Analysis 229

7.1. The general plan 229
By categorizing rates into three categories—low, middle, and high—and estimating log odds for each category, we can get a score for each unknown paper. This study defends against outlying results and failures of assumptions though it does a crude job of handling zero frequencies.

7.2. Details of method 229
For a given word, the rates in 48 papers (23 Hamilton and 25 Madison) were ranked with the lowest 18 papers giving the cutoff for "low" and the highest 18 papers the cutoff for "high". Table 7.2–2 gives the cut-points so determined and the log odds for 63 words. To get a score for a paper, sum the log odds. Some special rules killed some words and pooled categories in others.

7.3. Groups of words 234
After applying the rules of Section 7.2, we had 63 words left, grouped as before by perceived degrees of contextuality.

7.4. Results for the screening and calibrating sets 235
The scoring system was applied to the screening set of papers. As shown in Table 7.4–1, all Hamilton papers scored positive averaging 20.54, all Madison negative averaging −31.24. To see the regression effect, the same scheme was applied to a calibrating set as shown in Table 7.4–2 with average log odds for Hamilton of 8.54 and for Madison of −19.30.

7.5. Regression effects 239

7.5A. Word group 239
For each word group we show in Table 7.5–1 the regression effect from screening to calibrating set. Generally speaking, the more the group is perceived as contextual, the greater its regression effect. The word *upon* improved from screening to calibrating set.

7.5B. The regression effect by single words 241
Table 7.5–2 gives the numerical results.

7.6. Results for the joint and disputed papers 241
As in the analysis of Chapter 6, all disputed papers but paper 55 lean strongly toward Madison, and that paper falls on the fence.

Chapter 8. Other Studies 243

8.1. How word rates vary from one text to another 243
For 165 words we give rates in Table 8.1–1 from six sources: Hamilton, Madison, Jay, Miller-Newman-Friedman, Joyce's *Ulysses*, and the *Bible*.

8.2. Making simplified studies of authorship 249
To begin an authorship study we advise: Edit for quotations and special usage; make counts for separate pieces, using a list of words of moderate length; obtain the rates; assess variation and discard words; get statistical help if the problem is delicate; use natural groupings; use a high-speed computer; see Chapter 10 for some new variables.

8.3. The Caesar letters 251
As a little example, we explore the possibility that Hamilton, as opposed to someone else, wrote the Caesar letters. Table 8.3–1 shows the rates for 23 high-frequency words for the Caesar letters, and for Hamilton, for Madison, and for Jay. For 13 of the words, the Caesar rate differs from the Hamilton rate by two or more standard deviations under Poisson theory. If we apply the log odds computation of Chapter 3 for Hamilton versus Madison to the Caesar letters, we get −4.2, instead of positive log odds in the teens or twenties as we would expect if Hamilton were the author. The results are strongly against Hamilton, though not in favor of Madison, but of some unknown author.

8.4. Further analysis of Paper No. 20 252
Among the three papers we classified as having joint Hamilton-Madison authorship, paper No. 20 is most nearly on the fence. We hunted for Hamilton's contribution. Some Hamilton markers could be traced not to him but to the writing of Sir William Temple, from which Madison drew extensively for this paper. We abandoned the analysis.

8.5. How words are used 253
Joanna F. Handlin made an elaborate study of the various dictionary meanings of 22 marker words and *probably*. In 15 appearances of *upon*, Madison had 3 usages that Hamilton never used in 216 appearances. Table 8.5–1 gives detailed data for the occurrences of 13 meanings of *of* in several papers for each author—a study carried out by Miriam Gallaher.

8.6. Scattered investigations 256
We hunted for useful pairs of words like *toward-towards* with little success. Use of comparatives and superlatives showed great variation. Words with emotional tone gave no discrimination. How Hamilton and Madison handled enumerations led nowhere. A study of conditional clauses failed because of unreliability in classification. Relating strength of discrimination to proportion of original material, although suggestive, was not useful. Length of papers offered some discrimination, but we feared it because of contextuality and because of newspaper constraints.

8.7. Distributions of word-length 259
The earliest discrimination analyses by Mendenhall used word length as a discriminator. Robert M. Kleyle and Marie Yeager display the distribution of word length for eight Hamilton and seven Madison papers in Table 8.7–1 and in three figures. The chi-squared statistic for goodness of fit in Table 8. 7–2 shows so much variation that we cannot use it for discrimination. The Hamilton papers fit the Madison averages as well as do the Madison papers.

Chapter 9. Summary of Results and Conclusions 263

9.1. Results on the authorship of the disputed *Federalist* papers . . . 263
Except for paper No. 55, the odds are strong for Madison in the main study. For No. 55 they are about 90 to 1 for Madison.
The choice of data distribution mattered a great deal, the Poisson log odds were about twice those of the negative binomial, but several studies discredit the Poisson results while supporting the negative binomial. Variations in prior distributions mattered less than other sources of variation.

9.2. Authorship problems 265
Function words offer a fertile source of discriminators. Contextuality must be investigated. See also Chapter 10 for further variables.

9.3. Discrimination problems 265
A large pool of variables systematically explored may pay off when obvious important variables are not available. Contextual effects have counterparts in other situations. Selection effects must be allowed for.

9.4. Remarks on Bayesian studies 266
We recommend sensitivity studies made by varying the priors. We like priors that have an empirical orientation. Data distributions matter. We need simple routine Bayesian methods.

9.5. Summing up 267
We tracked the problems of Bayesian analysis to their lair and solved the problem of the disputed *Federalist* papers.

Chapter 10. The State of Statistical Authorship Studies in 1984 . . . 268

10.1. Scope 268
We treat the time period since 1969, emphasizing prose disputes almost exclusively. This chapter discusses both technological advances and empirical studies.

10.2. Computers, concordances, texts, and monographs 268
The computer and its software leading to easy compilation of concordances have been the major technological advance. Scholars have produced several monographs but few statistical texts in stylistics.

10.3. General empirical work 269
Morton studies sentence length further, and like Ellegård, uses proportional pairs of words (the fraction that the occurrences of word *U* make up the total occurrences of word *U* and word *V*). Morton introduces collocation variables to expand the number of potential discriminators. (A collocation consists of a keyword like *in* and has associated words that precede or succeed it). The ratio of the number of times the associate word occurs with the keyword to the number of times the keyword appears is the measure of collocation. Position of a word in a sentence (especially first or last) offers additional discriminators.

10.4. Poetry versus prose 270
To examine a possible systematic difference between poetry and prose, Williams looks at the Shakespeare-Bacon controversy. He takes samples of Shakespeare (who wrote only poetry), Bacon (who wrote only prose), and as a control samples of both poetry and prose from Sir Philip Sidney. Williams uses words of length 3 and 4 as discriminators. Table 10.4–1 shows the comparisons. He concludes that poetry and prose produce differing distributions of word lengths, and that the difference between Shakespeare and Bacon could be regarded as a poetry-to-prose effect rather than an authorship effect.

10.5. Authorship studies similar to the Junius or *Federalist* studies . . . 271

10.5A. *And Quiet Flows the Don* 272
We review the dispute about the authorship of the Russian novel *And Quiet Flows the Don*. An anonymous critic, D*, in a book with preface by Solzhenitsyn, regards Mikhail Sholokhov, the reputed author, as having plagiarized much of the work of the anti-Bolshevik author Fyodor Kryukov, who died before publishing his work on the Don Cossacks. Roy A. Medvedev reviews the issues, concluding that Sholokhov probably had access to some Don Cossack writings.

The comparisons Kjetsaa gives for *The Quiet Don*, for Sholokhov's other writings, and for Kryukov's support Sholokhov more than Kryukov.

10.5B. *Kesari* 273
In discriminating between two possible authors of certain editorials published in the Indian newspaper *Kesari*, Gore, Gokhale, and Joshi use the variables word length, sentence length, and the rate of use of commas as discriminators. They reject the hypothesis that word length follows the log normal distribution. They find sentence length to be approximately log normal, but unfortunately unstable for material from the same author, and so not helpful. Their new variable, rate of use of commas, offers some discrimination.

10.5C. *Die Nachtwachen* 274
The author of this pseudonymous romantic German novel has been hotly sought since its publication in 1804. Wickmann uses transition frequencies from one part of speech to another as discriminators and

concludes that among several candidates only Hoffmann is a reasonable possibility.

10.5D. Economic history 274
O'Brien and Darnell tackle six authorship puzzles from the field of economics. They use the collocation method and the first words of sentences to decide authorship in a book-length sequence of studies.

10.6. Homogeneity problems 275
In the simplest homogeneity problem, we have two pieces of text and we ask whether they were produced by the same author.

10.6A. Aristotle and *Ethics* 276
Kenny analyzes two versions of a book on ethics reputed to be by Aristotle, using their common material as a standard to decide which version was more similar to the common material. He uses many special studies in his book and concludes that the version which at one time was regarded by scholars as not the more mature version is closer statistically to the common material.

10.6B. *The Bible* 276
The studies of *The Bible*, both Old and New Testament, have become so extensive that they cannot readily be discussed here. We indicate studies that try to settle whether each of *Isaiah*, *Zechariah*, and *Genesis* was written entirely by a single author. The latter two studies have led to controversies, and we cite some papers that deal instructively with the issues. Students of authorship studies will find them helpful.

10.7. Anonymous translation 277
Michael and Jill Farringdon deal with the most unusual authorship problem we found in our literature search. Did Henry Fielding, the English novelist, translate the military history of Charles XII from French into English? By using as discriminators the rates of words that critics had used to parody Fielding and by looking at pooled rates of a variety of other authors, they conclude that Fielding did the translation. They use especially word pairs like *whilst* versus *while*. We think this problem deserves further study because of its challenge.

10.8. Forensic disputes 278

10.8A. Morton 278
Morton writes about his troubling experiences in giving authorship testimony in court.

10.8B. Bailey 278
Bailey gives three requirements before a legal authorship dispute can be decided. His attempts to introduce authorship stylometrics into the Patricia Hearst case were denied by the judge.

10.8C. Howland will 279
Although this dispute concerns the authorship of the signature of a will, it brings out many of the issues that arise in other authorship problems.

The main actors are famous: mathematician Benjamin Peirce, his scientist son Charles Sanders Peirce, and the woman who latter became a multimillionaire when a million was real money, Hetty Green. Meier and Zabell's treatment is most instructive.

10.9. Concluding remarks 280
Although some new variables for use as discriminators have been introduced, the level of statistical analysis in authorship studies has not generally advanced. We suggest that some additional empirical studies might help in future English authorship disputes. We especially need more data on variability within and between authors.

Appendix 283

References 285

Index 291