FREDERICK MOSTELLER AND DAVID L. WALLACE, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley Publishing Company, Inc., Reading, 1964. xv + 287 pp. $12.50.

Review by G. S. WATSON[1]

The Johns Hopkins University

Many people may approach this book with the feeling: "If I can wade through all this, I'll know at last whether this Bayesian movement makes sense in practice." In fact the book is so "practical" that it is hard to read, if like this reviewer, one has zero interest in who wrote the twelve disputed Federalist papers in particular and authorship problems in general.

This is as it should be. Practical statistics requires a great deal of reliable and intelligent labor and the problem and its setting ought to determine the solution more than any academic theories of "inference." Different questions, different fields of application allow different methods and more (or less) convincing answers. So there is surely no one method of statistical inference. The subject might be better classified under problems than philosophies of inference. Certainly it is silly to divide statisticians into Bayesians and non-Bayesians! Fisher encouraged statisticians to feel that their role is a noble one. It seems healthier to believe statistics is essential over an ever widening field but, by its very nature, a humble craft. If this is accepted there is less point in seeking the ultimate in any given view. It is worth remarking, parenthetically, that Fisher himself felt that Neyman-Pearson theory was probably satisfactory for "trade"(!), i.e., he realized that the different "systems of inference" were in fact simply the appropriate methods for different situations. Anyway, Mosteller and Wallace are not doctrinaires. They provide Bayesian and non-Bayesian analyses, robust and less robust.

Given then we are dealing with the special classification or discrimination problem, "who actually wrote *each* of the disputed papers," what statistical methods leap to mind? Let x be some vector of word frequencies or whatever data we are given, with density $f_1(x)$ for Hamilton, $f_2(x)$ for Madison, and likelihood ratio $f_1(x)/f_2(x)$. Put in prior probabilities π_1 and π_2 to get $\pi_1 f_1/\pi_2 f_2$, check in Rao's book for details and the job is done. While that is how much statistical consulting *is* done, Mosteller and Wallace are simultaneously the author-seekers, and distinguished statisticians concerned to show us, in all its dreadful details, how to do the best possible investigation their resources permit.

The choice of X, i.e., *the choice of sample space* is the first vital task. Statisticians normally leave that to customers but there are many instances to suggest that this is dangerous. For example, the following two instances of this have

[1] The preparation of this Review was partially supported by Contract Nonr 4010(09) Office of Naval Research awarded to the Department of Statistics, The Johns Hopkins University.

recently been of interest to me. The sample survey literature has taken an interesting theoretical turn because certain obviously true results are false if the sample space consists (as it usually does) of sampling unit names and measurements instead of just measurements. Again "fitting straight lines when both variables are subject to error" turns out not to be an insoluble conundrum when the experimental units are recognized to lead to a triplet (x, y, z) instead of the classical (x, y), where z is an "instrumental variable." It is my conviction that in laboratory experiments, instrumental variables can always be found, though they may not always be recorded ([4], [9], Section 11.3).

In the present case a little thought reveals all sorts of difficulties of definition as well as technical ones. (The authors give a birds-eye view of the book in their earlier paper [6], so few details need to be given here. The book itself is *full* of summaries.) Thus a major part of the study is concerned with the choice of x, the "style indicator." They don't, as papers in this journal so often do, start with $(\Omega, \mathcal{B}, P)$. *For the non-statistician the choice is almost the whole problem!* They do not use older indicators, sentence lengths and words special to either author. Instead they favor the use of the cumulative strength of many filler or function words (unrelated to context). Many side studies of stability, independence over time and context, etc., were made.

In Section 8.2, they give a brief guide to future author-seekers that might well be read before starting the book. It deals mainly with the choice of words. To try to paraphrase this section for quite *different discrimination* problems is worth the effort for it is basic statistical reasoning in a special setting.

The problem then arose of how to choose well a subset of filler words from a large set of such words, known to have different frequencies (per 1000 words of text) in the writings of Hamilton and Madison excluding the disputed papers. They had to watch out that they didn't use up all their material in this quest. This seems (to me) to be theoretically, the most interesting general problem raised, but not solved, by the study. One can think of many similar problems, e.g., the choice of a "good" variety of wheat. Mosteller and Wallace found that, naturally, words chosen for their good performance in exterior materials tended not to be so effective in the Federalist papers. This is called, from Galton's day, a regression effect. They refer (not as explicitly as this reader would like) to "selection and regression effects." Industrial experiments also raise similar "screening problems." What is meant by "selection effect," if distinct from the "regression effect" just mentioned, is not clear to me, though it may well be to the authors. One can imagine that technical and computational difficulties (note the vast army of helpers and all the Grants and Contracts listed in the Preface) did force a cutting down of the list of discriminators. Some of the literature in multivariable analysis is interested in the "diluting" effect of including "null" variates but I can't see how that would happen here.

Mosteller and Wallace say one of their major contributions arises at this point. To cope with the regression effect, due to selecting a subset, they introduce a prior distribution for the parameters of the Poisson or the better fitting negative binomial distribution. (The choice of these sampling distributions is discussed in the next paragraph, that of the priors later). It has been used before in at least two studies of screening, Beale and Mallows (unpublished) [2], Watson [8], and undoubtedly elsewhere. Duncan's studies of multiple comparisons [3], which were originally motivated by the "best varieties" problems, certainly, in their new Bayesian form, use a simultaneous prior distribution for the unknown means.

The choice of the distributions f_1, f_2 is clearly vital—that is what distinguishes the two authors. In the statistical approach, Hamilton $\equiv f_1$, Madison $\equiv f_2$. Now in some problems (e.g., the tea-tasting lady) this distribution choice hardly arises if one asks a limited question and has conducted the experiment in a suitably randomized way. In other cases (e.g., the present one) one can examine sample distributions and so, in the spirit of applied mathematics, select a distribution for f_1 and f_2 which is reasonable in the light of this data and perhaps other experience. In other cases the choice cannot be guided by empirical considerations and is usually a matter of taste and/or tradition sometimes bolstered by theory. The new subject of mathematical linguistics is not mentioned, no doubt for the good reason that it can not yet offer any guidance. So there is no theory to suggest the sample space or distributions on it. Clearly the final inference differs in conviction in these cases—and no piece of mathematical virtuosity will make it otherwise. Common sense demands it be so. Hence a major part of the study is concerned with this choice. The word frequencies are found to be reasonably independent and roughly Poisson. The negative binomial distribution provides a better fit but leads to more difficult manipulations. Some of their data could be used for good class exercises on distributions.

With this ground work they attack the authorship of each disputed paper separately. A Bayesian approach here has been quite usual in the past, e.g., in Rao [7]. The rule is: attribute to Hamilton if $\pi_1 f_1 > \pi_2 f_2$, toss-up if $\pi_1 f_1 = \pi_2 f_2$ and attribute to Madison if $\pi_1 f_1 < \pi_2 f_2$. In principle they show that for each paper f_2/f_1 is so large that it exceeds any conceivable π_1/π_2. They use $\log (f_1/f_2)$ or log odds (christened *lods* by someone) because the f's are products. Because the f's contain unknown parameters that are given a prior distribution, it is much harder than this to execute.

Another, and very justifiable, claim of the authors is that they have had to develop many new approximations to complete their study. It is possible that these may remain buried in the text, so that I would like to draw attention to Sections 4.6 A, B, C, D.

One of the blessings of Bayesian problems is that one does not have to integrate over the sample space, usually of many dimensions. However in Bayesian problems too one must integrate, though only over a few dimensions. Unless the prior distribution is suitably related to the frequency function, this may be difficult. Attempts to avoid the difficulty may often lead to the use of inappropriate priors. In the parametric study the likelihood ratio for a particular unknown paper, if the parameters for each word were known, is the ratio of products of negative binomial (or Poisson) probabilities—as many terms as there are filler words used and the parameters involve the total number of words in the unknown paper. The Bayesian study must average the numerator and denominator over the posterior distribution of the parameters, given the known papers. In this case numerical integration must be used for accurate results no matter what the prior is. Saddle point or "modal" methods can yield approximations and Section 4.6 is a discussion of their accuracy. *Mosteller and Wallace believe strongly that one should base ones "choices of priors on data, even if feebly."* Inspection of the high peaks of the likelihoods of each word at the modal or ml values of the parameters, suggested that the prior distribution of the parameters could well be taken as flat. But introspection says that some measure of the distance between the parameters for each author should be highly peaked about zero, otherwise the discrimination problem would be trivial. To see what the data says, they take the estimated parameters for each of a group of filler words in the known writings of Madison and Hamilton, and calculate the average parameters and a distance between them. This study verified their hunch and suggested a plausible range of priors. In other words an "empirical-empirical Bayes Method" was used in this part of the work.

Another section of the theoretical Chapter 4, Section 4.9, is also of general interest for the appraisal of probability forecasting methods. If an experiment has k outcomes, occurring with probabilities $\pi_1, \cdots, \pi_k$, and a prediction method, M, suggests probabilities $q_1, \cdots, q_k$, then from the well-known inequality

$$\sum_{i=1}^{k} \pi_i \log q_i \leqq \sum_{i=1}^{k} \pi_i \log \pi_i ,$$

it is reasonable to use $-\sum \pi_i \log q_i$ as the "logarithmic penalty" associated with M. This is applied to evaluate their methods for choosing words.

Not satisfied with all this they have attempted to check that their analysis is robust against every conceivable flaw in their assumptions, e.g., instead of using the actual rates, they classed them as high, medium and low.

It is easy to overlook the originality of this book—perhaps also the courage of its authors. There is no other book in which one or more statisticians have put all their material, and prior knowledge of it, clearly before their readers and then made an analysis in public and with such soul searching honesty. It is a notable addition to statistical literature almost all of which has an odour of unreality. Most authors can't wait to get away from the problem and on to mathematics. This book should forever remain a model, not only for authorship research, but for real statistics.

REFERENCES

[1] BEALE, E. M. L. and MALLOWS, C. L. (1958). Unpublished report of the Statistical Research Group, Princeton Univ.

[2] CARLSON, F. D., SOBEL, E. and WATSON, G. S. (1965). Linear relationships between variables affected by errors. To appear in *Biometrics*.

[3] Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7** 171–222.

[4] Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. Roy. Statist. Soc. Ser. B* **17** 269–278.

[5] Godambe, V. P. (1965). A review of the contributions towards a unified theory of sampling from finite populations. *Rev. Internat. Statistist. Inst.* **33** 242–258.

[6] Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *J. Amer. Statist. Assoc.* **58** 275–309.

[7] Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research.* Wiley, New York.

[8] Watson, G. S. (1961). A study of group screening. *Technometrics* **3** 371–388.

[9] Williams, E. J. (1959). *Regression Analysis.* Wiley, New York.