

# LINGUISTIC-ACOUSTIC FORENSIC SPEAKER IDENTIFICATION WITH LIKELIHOOD RATIOS FROM A MULTIVARIATE HIERARCHICAL RANDOM EFFECTS MODEL: A “NON-IDIOT’S BAYES” APPROACH

Phil Rose<sup>1,2</sup>, David Lucy<sup>2</sup>, Takashi Osanai<sup>3</sup>

<sup>1</sup>Phonetics Laboratory, Linguistics Program, A. N. U.

<sup>2</sup>School of Mathematics, University of Edinburgh.

<sup>3</sup>Forensic Science Laboratory, Miyagi Prefectural Police H.Q., Japan.

## Abstract

The discriminant performance of a likelihood ratio based on a two-level multivariate model is examined on the speech of 60 male Japanese speakers using non-contemporaneous telephone recordings with channel mismatch. The performance is determined for both F-pattern centre frequencies and LPC cepstral coefficients, extracted from three very different phonetic segments only: a vowel, a voiceless fricative and a nasal. The Multivariate Likelihood Ratio (MVLRL) is shown to perform well in discriminating same- from different-speaker pairs, yielding strength of evidence that can be characterised as moderate for F-pattern and, at the least, very strong for the cepstrum. Comparison is made using the same data with a so-called “independence” or “Idiot’s Bayes” LR approach, which ignores correlation between variables. It is shown that, as commonly found, approaches which assume independence outperform the MVLRL. The consequences of this finding for forensic speaker identification are alluded to.

## 1. Introduction

In the real world, Forensic Speaker Identification (FSI) as a prosecutorial tool typically involves the comparison of one or more samples of an unknown voice with one or more samples of a known voice. In the vast majority of cases, samples are recordings intercepted from the telephone, both landline and cellular. Often the unknown voice is that of the individual alleged to have committed an offence, and the known voice belongs to the suspect. The court wants to determine whether the two samples have come from the same person or not, and thus be able either to identify the suspect as the offender or exonerate them. However, practitioners in many different fields of forensic identification are becoming increasingly aware of the fact that (however much the court may desire otherwise) it is not desirable to quote the probability of the hypothesis given the evidence (Aitken and Taroni 2004, Robertson & Vignaux 1995). Applied to FSI this means that it is not possible to say, for example, that one is 80% sure that the samples have come from the same speaker, given the similarities between them (Rose 2002, 2003). Given that this message was first enunciated in Lewis’s (1984) paper, and is established practice in some major areas of forensic identification, for example using

DNA, its application in FSI is taking some time to propagate.

The reasoning for why the expert has difficulty quoting the probability of a hypothesis, given the evidence, applied to FSI, is as follows. 1) The court, and the Law, is faced with decision making under uncertainty – it wants to know how certain it is that the incriminating speech samples have come from the defendant. 2) Probability can be shown to be the best measure of uncertainty (Lindley 2002). 3) Therefore it is necessary to evaluate how much more likely the evidence - i.e. the differences/similarities between the speech samples - shows the defendant to have produced the incriminating samples than not to have produced them. This is shown by the ratio of conditional probabilities at (1), where H = prosecution hypothesis that the incriminating speech samples come from the defendant; ~H = defence hypothesis that the samples come from someone else; E = evidence (similarities/differences between the offender and defendant speech samples).

$$p(H | E) / p(\sim H | E) \quad (1)$$

If this ratio exceeds some previously determined value – beyond reasonable doubt or the balance of probabilities for example – the defendant is considered to have produced the samples. 4) The solution to (1) is given by Bayes’ theorem, the odds form of which (which is more convenient to express) is given at (2).

$$o(H | E) = o(H) * p(E | H) / p(E | \sim H) \quad (2)$$

In words: posterior odds in favour of hypothesis = prior odds in favour of hypothesis \* Likelihood Ratio). 5) The FSI expert is normally not privy to the prior odds  $o(H)$ . 6) Therefore they *cannot usually* evaluate a posterior probability. 7) However, the expert can estimate the strength of the evidence in favour of the same-speaker hypothesis by estimating the likelihood ratio – the ratio of the probability of the evidence assuming samples have come from same speaker to the probability of the evidence assuming the samples have come from different speakers: that is their proper role, and this paper is about estimating Likelihood Ratios from speech for forensic purposes.

Since values of the LR greater than one indicate same-subject data, and values less than one indicate different-subject data, it can be treated as a potential function for discriminating same-subject from different-subject data. The extent to which it can do this reflects both the individual-identifying nature of the evidence and the method of evaluating its strength, although it is not easy to separate these factors (Rose 2004). Such testing is vitally important, given that the well-known *Daubert* (1993) rulings on the admissibility of scientific evidence include as one criterion whether the theory or technique can be and has been tested (Black et al. 1994: 783ff) - and in Federal and State Australian courts the practice notes requiring reliability, replicability and transparency on the part of expert testimony are *de facto* adoptions of *Daubert*. The ability of the LR to discriminate same-subject from different-subject data has to date been successfully tested on several different types of forensically common evidence, including DNA (Evetts et al. 1993: 503); elemental ratios in glass fragments (Aitken & Lucy 2003, Aitken et al. Ms.; and speech: Meuwly and Drygajlo (2001: 150) have investigated Swiss French; Gonzalez-Rodriguez et al. (2001) Spanish, Nakasone and Beck (2001) American English.

However, the approach used in some previous speech experiments (Kinoshita 2001, 2002, Rose et al. 2004, Alderman 2004), which is based on Lindley’s (1977) LR formula, is, despite its credible performance, primitive in not conforming to several important characteristics of

speech (Rose et al. 2003). In particular the LR formula used assumes that the variables in question (formant centre frequencies, cepstral coefficients) are normally distributed and have equal variance; and the approach also assumes that the variables are independent (i.e. are not correlated). None of these is of course necessarily true for speech. Approaches which do not control for correlations often go by the unflattering monikers “Idiot’s”, “Simple” or “Naïve Bayes” (Hand & Yu 2001: 386), although “Independence Bayes” is perhaps more neutral. The aim of this paper is to examine whether improvements in discriminatory power result from the use of a more sophisticated LR formula which is able to take into account potential correlation between variables.

## 2. Procedure

The same data were used as in Rose et al’s (2003) experiment on Japanese (q.v), and comprised telephone recordings from two non-contemporaneous sessions, separated by ca. three to four months, of 60 male Japanese speakers from several different prefectures. Recording was made centrally, on the same equipment, from landline telephone calls, but there was no requirement for any speaker to use the same handset, so essentially the channel was uncontrolled. Three very different phonetic segments, in various words, were identified for each speaker using dynamic programming (Osanai et al. 1995). These were 1) the syllable-coda nasal consonant /N/ (which varies mostly between [ʃ] and [N]), e.g. in the word *san three*; 2) the voiceless alveopalatal fricative [ç] e.g. in the word *moshimoshi hello*, and 3) the long back mid-rounded vowel [ɔ], e.g. in the word *ginkoo bank*. The three sounds are referred to below as “N”, “sh” and “oo” respectively. There were twenty tokens each of “sh” and “oo”, and fourteen of “N”, in each recording session. F-pattern centre frequencies (surprisingly, the Japanese phone transmission appeared to allow extraction of F1 thru F5!) and 12<sup>th</sup> order LP cepstral coefficients were automatically extracted from each token of each phonetic segment.

## 3. Multivariate Likelihood Ratio

The multivariate likelihood approach used in this paper was developed as a solution to the non-trivial problem of estimating the strength of evidence when predictor variables may be correlated. It treats the variables for which a LR has to be estimated – for example a set of measurements of frequencies for different vowel formants - as multivariate data (Aitken & Lucy 2003). The

LR =

$$\frac{|2 \quad \left[ (n_1 + n_2) U^{-1} + C^{-1} \right]^{-1} \left| \frac{1}{2} \exp \left\{ -\frac{1}{2} (H_2 + H_3) \right\} \right|}{|2 \quad C \left| \frac{1}{2} \right| \quad \left( n_1 U^{-1} + C^{-1} \right)^{-1} \left| \frac{1}{2} \right| \quad \left( n_2 U^{-1} + C^{-1} \right)^{-1} \left| \frac{1}{2} \exp \left\{ -\frac{1}{2} (H_1 + H_2) \right\} \right|}
 \quad (3)$$

$U$  = Within-group covariance matrix;  $C$  = between-group covariance matrix

$n_1, n_2$  = See attached

$H_1$  = See attached

$H_2$  = See attached

$H_3$  = ditto

$H_4 = (\mu - \mu^*)^T [(D_1 + C)^{-1} + (D_2 + C)^{-1}] (\mu - \mu^*)$ ;

$H_5 = (\mathbf{I}_1 + \mathbf{I}_2)^T (D_1 + D_2 + 2C)^{-1} (\mathbf{I}_1 - \mathbf{I}_2)$

$\mu^* = \{(D_1 + C)^{-1} + (D_2 + C)^{-1}\}^{-1} [(D_1 + C)^{-1} \mathbf{I}_1 + (D_2 + C)^{-1} \mathbf{I}_2]$

approach accommodates two levels of variance: between- and within-subjects, and is thus somewhat unrealistic for speech, where at least a third level of variance – between non-contemporaneous sessions – must usually be assumed. The variables' distributions can be modelled either with normal curves or kernel densities. The former option was used, since the variables did not, upon examination, appear to depart from normality to an extent that warranted kernel density modelling. The Japanese formant and cepstral data were examined for correlations, and were found to differ considerably. **Check - David can you check these please The formants were basically uncorrelated between speakers, and for within sounds, were uncorrelated both within and between. By contrast the cepstral coefficients were correlated between speaker, sometimes highly.**

As far as the formants were concerned, there was generally little correlation between the variables, either within a sound or between sounds, the exceptions being F2 and F3 in “sh” (as might be expected); F2 and F5 in “oo”, F2 and F3 in “n” ; and F1 in “n” and “oo”. The cepstral coefficients, on the other hand, showed pervasive and often high correlations, especially within-sound **[David is this right? - correct, but between sounds and at the between speaker level. Bare in mind that for between sounds we can only talk about between speaker correlations – within sounds we talk about within and between, or between speaker correlations]**

The formula for the multivariate-normal LR is derived in Aitken & Lucy (2003), where it simplifies to the expression at (3).

#### 4.

#### Experiments

The 60 speakers and two non-

contemporaneous recording sessions gave 60 same-speaker pairings and 1770 different-speaker pairings, and LRs were calculated for each of these pairings. Separate LRs were calculated using the MVLR approach for the F-pattern and for the cepstrum, and both for the individual sounds and for the three sounds combined.

#### 5. Results

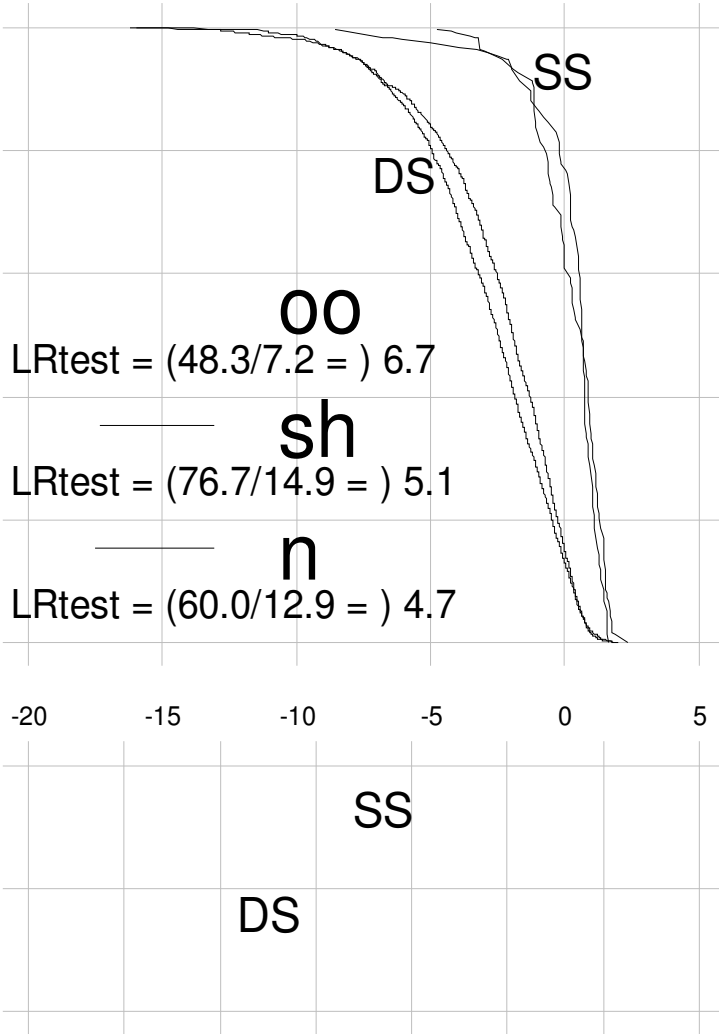
Figure 1 presents the results of the experiment using so-called reliability functions/Tippett plots - a method of representation proposed for forensic DNA analysis (Evet & Bucklelton 1996) and forensic automatic speaker recognition (Drygajlo et al. 2003). Results for the comparison with formants are given on top and for the cepstrum on the bottom (note that very different horizontal scales are used for these two features). Plots for the three individual segments are given on the left, and for all three segments combined on the right. Thus in the top right panel of figure 1 it can be seen that, with formants, there is about a 90% chance of getting a  $\log_{10}(\text{MVLR})$  value greater than -5 assuming that the samples come from the same speaker, compared to a 20% chance of  $\log_{10}(\text{MVLR}) > -5$  assuming that the samples come from different speakers. Thus one would for example be ca. (90%/20% =) 4.5 times more likely to observe a  $\log_{10}(\text{MVLR})$  value of at least -5 assuming that the samples had come from the same rather than different speakers. This would count as “limited” evidence for the prosecution in terms of the values proposed for use in the British Forensic Science Service (Rose 2002: 61).

A quantification of the results in terms of the  $LR_{test}$

value is also given by the probability that the threshold requirement will be met. This probability is assumed to be 0.5 (LR=1). So by using combined evidence, ca. 55% of the samples are greater than the percentage of 1.8% exceeded by about (55/1.8) = 30.6 times (MVL) greater than a same-speaker sample.

The results for resolution for relatively good samples can not be compared with the two samples compared which rates of 13.5% the MVL approach (although it is like this the El the threshold use of the LR

attention is on the strength of evidence that can be achieved by such an approach.) Thus it can be noted that with formants one is about 30 times more likely to get a LR value greater than threshold assuming that the samples have come from the same speaker than not – this counts as “moderate” evidence in support of the prosecution hypothesis. With the cepstrum the strength of evidence is very much stronger – in this case infinitely so in fact, since different-speaker pairs are absolutely discriminated with the MVL approach. It has already been pointed out that this is an advantageous result from the point of view of avoiding incriminating innocent parties (Aitken et al. Ms). (It is worth noting that one does not, in fact, need the “n” segment at all for this, since absolute different-speaker discrimination can be achieved with just “sh” and “oo”, as well as a concomitant improvement in the same-speaker discrimination from 36.7% to 45%.)



The results in figure 1 thus show once again the clear pattern of cepstral superiority already demonstrated in Rose et al. (2003), with the cepstrum far outperforming the F-pattern as a forensic parameter. The general similarity in overall profile between the cepstral and formant ogives reflects perhaps the fact that the formants are the primary determinant of the overall spectral shape modelled by the cepstrum.

As far as the results for the individual segments are concerned, “oo” is particularly noteworthy. One would be about 290 times more likely to observe a  $\log_{10}(\text{MVL})$  value greater than 0 assuming that the “oo”s had come from the same rather than different speakers. This is also some (290/7 = ) 40 times greater than with “oo” F-pattern. The comparative results for the other segments are not so spectacular, but nevertheless good: (86/5 = ) 17 times greater with “sh” cepstrum than with its formants, and

(52/5 = ) 10 times greater with “N” cepstrum. The superiority of the cepstrum will not come as a surprise to the automatic speech community, but may be of interest since the results are derived from specific phonetic speech segments, and not from a global approach which does not distinguish segments. The results are also of course of interest because they show that segmental acoustics can also have considerable evidentiary power in FSI.

The superiority of the cepstrum is clearly achieved by its performance in discriminating different-speaker data -it generally does badly in discriminating same-speaker pairs, which is one aspect in which the F-pattern is consistently better – by about 10% to 20%.

## 6. Comparison with independence assumption Bayes approach

The second aim of this paper was to see the effect of using a more appropriate LR in FSI – one which was able to take into account potential correlations between predictor variables. In order to determine this, the data were discriminated using an Idiot’s – or Independence – Bayes approach, similar to that used in Rose et al (2003). This method used the LR formula derived in Lindley (1977), and calculated the overall LR for a comparison as the product of the individual LRs on the assumption that they were independent. Thus there were (3 sounds \* 5 formants = ) 15 independent variables assumed in the formant comparison, and (3 sounds \* 13 CCs = ) 39 independent variables in the cepstrum comparison. The results are shown in figure 2, with the formants on the left and the cepstral coefficients on the right. In each panel is reproduced from figure 1 the Tippett plots for the MVLR for the combined 3 segment data, compared with the Tippett plots for the Idiot’s Bayes analysis.

It is readily apparent from figure 2 that the effect of using the Idiot’s Bayes approach is a dramatic leftwards shift in both same-speaker and different-speaker ogives. As can be seen, this actually improves the discrimination, from the point of view of both the EER and the  $LR_{test}$  values. The EER for the formants is now 8% compared to 13.5% for the MVLR approach, and for the cepstrum it is 2% compared to the previous 7.5%. The improvement in the formant  $LR_{test}$  values is from 30.6 to 318. Neither  $LR_{test}$  result is defined for the cepstrum and so they cannot be compared.

Hand & Yu (2001:386 et pass.) point out that the Idiot’s Bayes approach very frequently outperforms more sophisticated analysis, and the results presented here seem to constitute another example. We are therefore

confronted by a rather serious forensic problem. Do we base our estimate of the strength of evidence on what we know to deductively correct – that is, on the MVLR – or do we base it on our practical ability to discriminate – that is, on the Idiot’s Bayes’ LR?

## 7. Summary

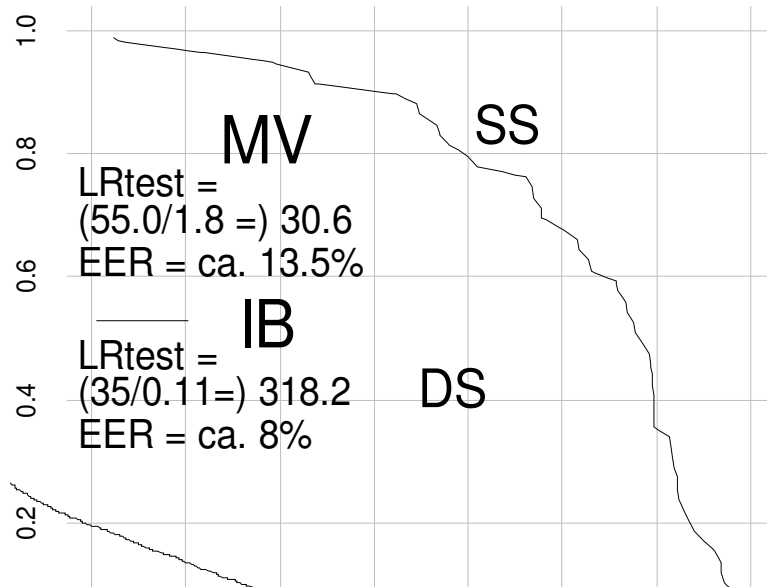
The two-level MVLR approach was tested on forensically realistic Japanese speech data and found to perform reasonably well in discriminating same- from different-speaker pairs, although the better performance in resolution of different-speaker pairs was again noted. The cepstrum considerably outperformed the formants, in particular showing an absolute discrimination of different-speaker pairs which lowers the possibility of false positives. An Idiot’s Bayes discrimination which did not take any correlation between the variables into account was shown to outperform the MVLR, which raises problems in estimating the strength of forensic speaker identification evidence.

### 1. Acknowledgements

The statistical research for this paper was carried out whilst the first author was a British Academy Visiting Professor at Edinburgh University’s *Joseph Bell Centre for Forensic Statistics and Legal Reasoning*. This was made possible by an award of £2700 from the British Academy.

### 2. References

- Aitken, C.G.G. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*. [2<sup>nd</sup> Ed.], Wiley: Chichester.
- Aitken, C.G.G. and Lucy, D. (2002) “Evaluation of trace evidence in the form of multivariate data”, *Applied Statistics* 53/4: 109-122.
- Aitken, C.G.G. Lucy, D. Zadora, G and Curran, J.M. (Ms) “Evaluation of trace evidence for three-level multivariate data with the use of graphical models. **In submission to: computational statistics and data analysis**
- Alderman, T. (2004). Refining the Likelihood Ratio Approach to Forensic Speaker Identification - Effects of Non-Normality in the Background Distribution as Modelled with the Bernard Data for Australian English. Unpublished First Class Honours Thesis, Australian National University.
- Black, Bert Ayala, Francisco J. & Saffran-Brinks, Carol (1994) “Science and the Law in the Wake of *Daubert*: A New Search for Scientific Knowledge”. *Texas Law Review* 72/4: 715-802.



Daubert (1993) *Daubert vs Merrell Dow Pharmaceuticals, Inc.* 113 S Ct 2786.

Drygajlo, Andrzej, Meuwly, Didier & Alexander, Anil (2003) "Statistical methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition". *Proc. 8<sup>th</sup> European Conf. on Speech Communication and Technology (EUROSPEECH)*.

Evelt, I.W. Scrange, J. and Pinchin, R. (1993) "An Illustration of the Advantages of Efficient Statistical Methods for RFLP Analysis in Forensic Science", *American Journal of Human Genetics* 52: 498-505.

Evelt, I.W. & Buckelton, J. (1996) "Statistical Analysis of STR Data". *Advances in Forensic Haemogenetics*, vol 6: 79-86. Springer Verlag: Heidelberg.

Gonzalez-Rodriguez, J. Ortega-Garcia, J. and Lucena-Molina, J.J. (2001) "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", *Proc. of the 2001 Speaker Odyssey Speaker Recognition Workshop*.

Hand, D.J. & Yu, Keming (2001) "Idiot's Bayes – Not So Stupid After All?" *International Statistical Review*, 69/3:385 – 398.

Kinoshita, Y. (2001) Testing Realistic Forensic Speaker Identification in Japanese: A Likelihood Ratio Based Approach Using Formants. Unpublished Ph.D. Thesis, the Australian National University.

Lewis, S. R. (1984) "Philosophy of Speaker Identification", *Proc. Institute of Acoustics* 6/1, 69-77.

Lindley, D.V. (1977) "A problem in forensic science". *Biometrika* 64/2:207-13.

Lindley, D.V. (1991) "Probability" in Aitken & Stoney (eds.) *The Use of Statistics in Forensic Science*. Ellis Horward, Chistester 27-50.

Meuwly, D. and Drygajlo, A. (2001) "Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM)", *Proc. of the 2001 Speaker Odyssey Speaker Recognition Workshop*.

Nakasone and Beck (2001) "Forensic Automatic Speaker Recognition", *Proc. of the 2001 Speaker Odyssey Speaker Recognition Workshop*.

Osanai, T. Tanimoto, M. Kido, H. and T. Suzuki, T. (1995) "Text-Dependent Speaker Verification using Isolated Word Utterances based on Dynamic Programming." [In Japanese]. *NRIPS Report*, vol. 48/1: 15-19.

Robertson, B. and Vignaux, G.A. (1995) *Interpreting Evidence*. Wiley: Chichester.

Rose, P. (2002) *Forensic Speaker Identification*. Taylor and Francis: London.

Rose P. (2003) *The Technical Comparison of Forensic Voice Samples*. Issue 99, *Expert Evidence*, (series eds. Freckelton, I. & Selby, H.). Thomson Lawbook Company, Sydney, 2003.

Rose, P. (2004) "Technical Forensic Speaker Identification from a Bayesian Linguist's Perspective". Keynote paper, *Forensic Speaker Recognition Workshop, Speaker Odyssey '04*.

Rose, P. Osanai, T. & Kinoshita, Y. (2003) "Strength of Forensic Speaker Identification Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold", *Forensic Linguistics* 10/2: 179-202.

Royal, R (2000) "On the Probability of Observing Misleading Statistical Evidence", J. Am. Statistical Assn. 95/451: 760-768.