Evaluation and interpretation of evidence: what can be done, is done and should be done

Colin Aitken

University of Edinburgh

c.g.g.aitken@ed.ac.uk

SAMSI

August / September 2015

Table of Contents

Contents

Abstract

- Concentration on the statistical basis of forensic science and mainly on the evaluation and interpretation of evidence.
- Review of recent statistical developments; what can be done currently in assisting
  - the forensic scientists evaluate and interpret their evidence,
  - the courts in understanding that evidence.

  Refer to *R. v. George* and *R. v. T* and some publications and activities.
- Proposals for research and education in the context of the legal framework within which forensic scientists work.

Contents

History

CS Peirce (1878)

- Probability is the ratio of favourable cases to all cases.
- Chance is the ratio of favourable to unfavourable.
- Belief is the logarithm of chance and is proportional to the weight of chance; to multiply chances is to add beliefs.
- Balancing reasons: take the sum of all the feelings of belief which would be produced separately by all the arguments *pro* the proposition, subtract from that the similar sum for arguments *con*. The remainder is the feeling of belief which one ought to have on the whole.

History

Darboux, Appell and Poincaré (1908):

> *since it is absolutely impossible for us [the experts] to know the a priori probability, we cannot say: this coincidence proves that the ratio of the forgery's probability to the inverse probability is a real value. We can only say: following the observation of this coincidence, this ratio becomes X times greater than before the observation. (p.504.)*

History

Good(1979)
Summary of statistical ideas of Alan Turing in 1940, 1941.

- Introduction of the expression '(Bayes) factor in favour of a hypothesis' without the qualification 'Bayes': the factor by which initial odds of H must be multiplied to obtain the final odds in favour of H provided by evidence E.
- Sequential analysis and log factors; log factor is the 'weight of evidence'; closely related to the amount of information concerning H provided by E.
- The ban and deciban: the unit by which weight of evidence is measured. A deciban is one-tenth of a ban.
- The weighted average of factors.
- Expected weight of evidence, variance of weight of evidence.

History

Lindley (1977)

$$BF = \frac{\tau}{2^{\frac{1}{2}}\sigma} \exp\left\{ -\frac{(X-Y)^2}{4\sigma^2} \right\} \exp\left\{ \frac{(Z-\mu)^2}{2\tau^2} \right\}.$$

Let $\lambda = \mid X - Y \mid /(2^{\frac{1}{2}}\sigma)$ and $\delta = \mid Z - \mu \mid /\tau$; $\tau/\sigma = 100$.

$$\begin{aligned} \lambda = \delta = 0 &\Rightarrow BF = 70.7; \\ \lambda = 0, \delta = 3.0 &\Rightarrow BF = 6370; \\ \lambda = 6.0, \delta = 0 &\Rightarrow BF = 1/(9.29 \times 10^5) \end{aligned}$$

Evidence evaluation - Bayesian hierarchical models

- multilevel, multivariate random effects models;
- within-group normal;
- between-group non-normal;
- two-level: glass objects and fragments of glass within objects;
- three-level: types of glass objects, objects within types; fragments within objects.

Evidence evaluation - Comparison and classification

Evidence evaluation - Comparison and classification

- Comparison (e.g., Aitken and Lucy, 2004):
  - Control data: source known;
  - Recovered data: source unknown;
  - Background data: database for estimation of parameters in models
- Classification (e.g., Wilson *et al.*, 2015):
  - evidence of unknown origin;
  - example: banknotes: are they associated with crime or not?
  - require two databases, one for each possible origin.

Evidence evaluation - Criticism and debate

Evidence evaluation - Criticism and debate

Criticisms:

- Absence of reliable data for control and recovered evidence;
- Absence of reliable data for background population;
- Use of numbers suggest a level of certainty which does not exist in reality.

Evidence evaluation - Criticism and debate

Criticisms:

- Absence of reliable data for control and recovered evidence;
- Absence of reliable data for background population;
- Use of numbers suggest a level of certainty which does not exist in reality.

Points for debate:

- What is the role, if any, of a verbal scale?
- Should a likelihood ratio be quoted as a single value or as an interval? See Taroni et al plus discussion (LP, to appear)

Evidence evaluation - Bayes nets

Bayes nets: see Lauritzen and Spiegelhalter (1988)
Decisions are required about

- what to include as nodes;
- what nodes to be linked;
- what values to assign to conditional probabilities.

Evidence evaluation - Bayes nets

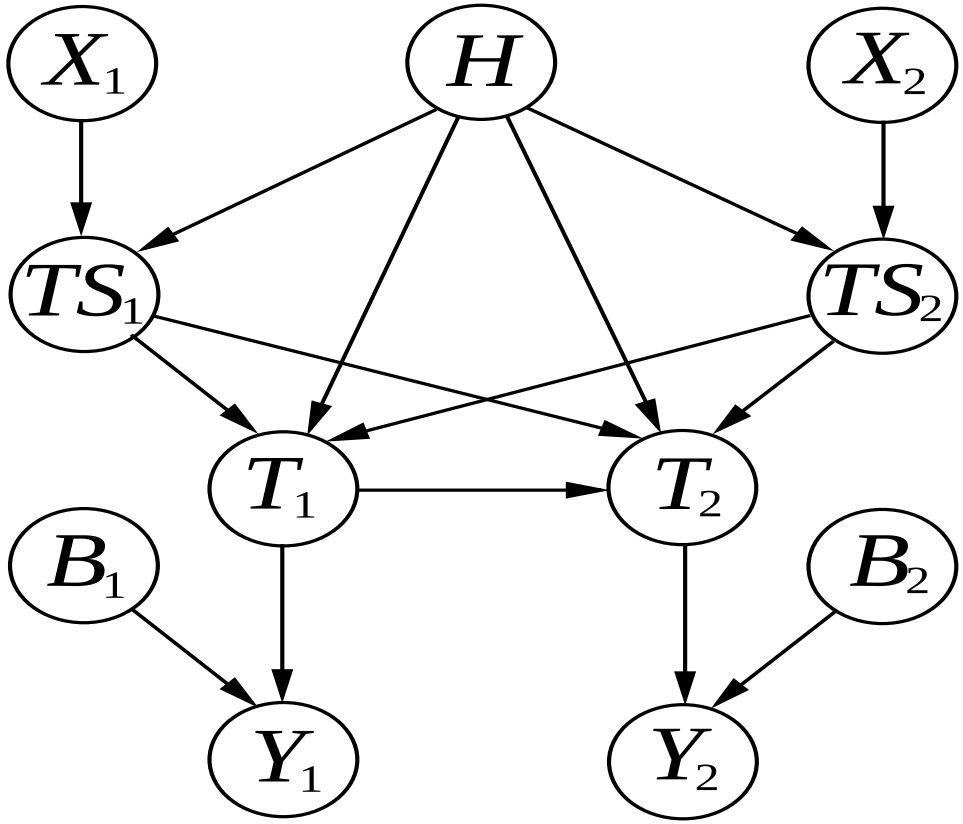


Figure: 1. Cross-transfer network.

Evidence evaluation - score-based and frequency-based models

Given data $\mathbf{x}$ and $\mathbf{y}$, a distance $s(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$, density functions $f(\mathbf{x}, \mathbf{y})$ and $f(s(\mathbf{x}, \mathbf{y}))$, and propositions H_p and H_d, the two methods are

- Frequency based model:

$$\frac{f(\mathbf{x}, \mathbf{y} \mid H_p)}{f(\mathbf{x}, \mathbf{y} \mid H_d)}.$$

- Score-based model:

$$\frac{f(s(\mathbf{x}, \mathbf{y}) \mid H_p)}{f(s(\mathbf{x}, \mathbf{y}) \mid H_d)}.$$

Evidence evaluation - score-based and frequency-based models 2

Comparison of score-based and frequency-based models:

- Need definition of distance.
- Frequency-based models consider similarity and rarity.
- Score-based models consider similarity.

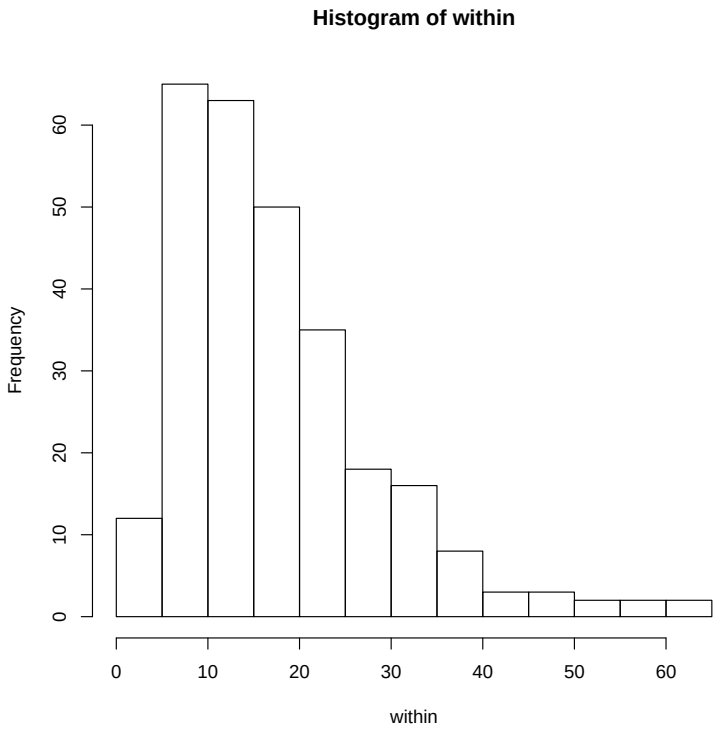
Score-based models used in handwriting (Davis *et al.*, 2012; Hepler *et al.* 2012), speech comparison (Enzinger and Morrison, 2015), soil analysis (Quaak and Kuiper, 2011), and fingerprinting analysis (Neumann *et al.*, 2012).

Evidence evaluation - score-based and frequency-based models 3

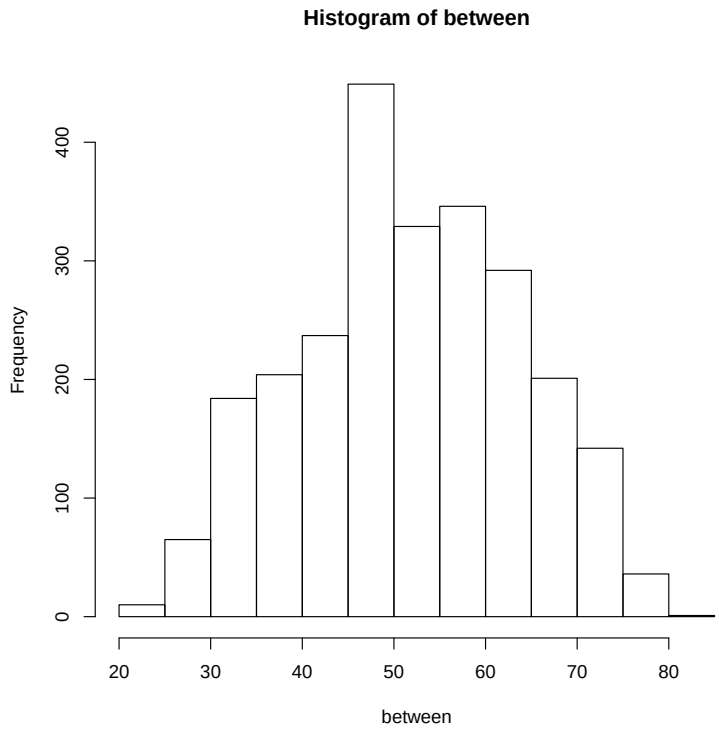
Score-based model: example of striation marks from screwdrivers (Petraco *et al.*, 2012). These are hierarchical binary data which are the striated marks made by nine screw drivers. Each of the line-typed marks is divided into 140 intervals, and if there is a mark in one interval, the measurement of this interval is 1, otherwise the measurement is 0. Hence each observation is a 140-dimensional vector of elements being 1 or 0. There are 75 obser- vations coming from the 9 screw drivers with different numbers of observations (8, 6, 9, 8, 9, 9, 8, 9, 9) from each screw driver.

Score-based LR

Within group distances



Between group distances



Case assessment and interpretation

Levels of proposition

- Sub-source;
- Source;
- Activity;
- Offence;

Case assessment and interpretation

Levels of proposition

- Sub-source; Suspect is the source of the DNA at the crime scene.
- Source;
- Activity;
- Offence;

Case assessment and interpretation

Levels of proposition

- Sub-source; Suspect is the source of the DNA at the crime scene.
- Source; Window is the source of the fragments of glass on the clothing of the suspect.
- Activity;
- Offence;

Case assessment and interpretation

Levels of proposition

- Sub-source; Suspect is the source of the DNA at the crime scene.
- Source; Window is the source of the fragments of glass on the clothing of the suspect.
- Activity; Suspect gained fragments of glass on his clothing when breaking a window at the crime scene: require to take account of probabilities of transfer and background of fragments of glass.
- Offence;

Case assessment and interpretation

Levels of proposition

- Sub-source; Suspect is the source of the DNA at the crime scene.
- Source; Window is the source of the fragments of glass on the clothing of the suspect.
- Activity; Suspect gained fragments of glass on his clothing when breaking a window at the crime scene: require to take account of probabilities of transfer and background of fragments of glass.
- Offence; Suspect gained fragments of glass on his clothing whilst committing an offence including the breaking of a window at the crime scene: require to take account of relevance.

Pre-assessment and interpretation: Oesterhelweg et al. (2008)

Cadaver dog signals in the presence or absence of cadaver scent.

Dog's response	Scent present	Scent absent
Positive signal	224	4
No signal	11	115
Total	235	119

Conditional probabilities and associated LRs;
H: scent present, $\bar{H}$: scent absent.

Dog's response E	$Pr(E \mid H)$	$Pr(E \mid \bar{H})$	LR
Positive signal	0.953	0.034	28
No signal	0.047	0.966	1/21

Statistics - sampling

Statistics - sampling

Given a consignment of white tablets, how many should be examined to satisfy court of the illicit nature of the consignment?

Statistics - sampling

Given a consignment of white tablets, how many should be examined to satisfy court of the illicit nature of the consignment?
Determine criteria as to what will satisfy the court. It could be that a the proportion of illicit tablets in the consignment is greater than a prescribed amount. If so, what is the amount?

Statistics - sampling

Given a consignment of white tablets, how many should be examined to satisfy court of the illicit nature of the consignment?
Determine criteria as to what will satisfy the court. It could be that a the proportion of illicit tablets in the consignment is greater than a prescribed amount. If so, what is the amount?
Given the prescribed amount, how certain does the court need to be that the proportion is greater than that amount?

Statistics - sampling

Given a consignment of white tablets, how many should be examined to satisfy court of the illicit nature of the consignment?
Determine criteria as to what will satisfy the court. It could be that a the proportion of illicit tablets in the consignment is greater than a prescribed amount. If so, what is the amount?
Given the prescribed amount, how certain does the court need to be that the proportion is greater than that amount?
Develop a procedure using Bayesian inference, with a prior probability, with certain assumptions,e.g. homogeneity of tablets, to satisfy the criterion: we wish to be $100(1-\alpha)\%$ confident that greater than $100\theta\%$ of the consignment is illicit.

Statistics - sampling

Given a consignment of white tablets, how many should be examined to satisfy court of the illicit nature of the consignment?
Determine criteria as to what will satisfy the court. It could be that a the proportion of illicit tablets in the consignment is greater than a prescribed amount. If so, what is the amount?
Given the prescribed amount, how certain does the court need to be that the proportion is greater than that amount?
Develop a procedure using Bayesian inference, with a prior probability, with certain assumptions,e.g. homogeneity of tablets, to satisfy the criterion: we wish to be $100(1-\alpha)\%$ confident that greater than $100\theta\%$ of the consignment is illicit.
For example, if $\theta = 0.5$ and $\alpha = 0.05$, the answer is 4: if we wish to be 95% confident that at least 50% of the consignment is illicit, sample 4 tablets and if all are illicit, then the criterion is satisfied.
(Compare the probability of achieving four heads in a row from a toss of a fair coin. This equals $1/16 = 0.0625$.)

Contents

Presentation of evidence in court: case assessment and interpretation

R. v. George [2007] EWCA Crim 2722:

Presentation of evidence in court: case assessment and interpretation

R. v. George [2007] EWCA Crim 2722:

Firearm discharge residue (FDR) at crime scene. A year later a single particle (11.5 microns) of FDR found in the pocket of a coat belonging to Barry George. The particle matched the constituent elements of FDR found at the crime scene. This evidence was an important piece of evidence at the trial of Barry George: the particle was consistent with having come from the gun used in the killing.

Presentation of evidence in court: case assessment and interpretation

R. v. George [2007] EWCA Crim 2722:
Firearm discharge residue (FDR) at crime scene. A year later a single particle (11.5 microns) of FDR found in the pocket of a coat belonging to Barry George. The particle matched the constituent elements of FDR found at the crime scene. This evidence was an important piece of evidence at the trial of Barry George: the particle was consistent with having come from the gun used in the killing.
Report from FSS to CCRC: two propositions:

- Mr George is the man who shot Ms Dando;
- Mr George had nothing to do with the incident.

In our opinion, the probability of finding a single particle of FDR in Mr G's coat pocket would have been the same, regardless of which of the above propositions was true. The FDR evidence is thus inconclusive. In our opinion it provides no assistance to anyone asked to judge which proposition is true

Presentation of evidence in court: presentation of likelihood ratio

R. v. T [2010] EWCA Crim 2439:

Debate about the relative merits of an evaluation based on a likelihood ratio calculation, with an associated verbal scale and of an evaluation based on a subjective opinion of source.

Presentation of evidence in court: presentation of likelihood ratio

R. v. T [2010] EWCA Crim 2439:
Debate about the relative merits of an evaluation based on a likelihood ratio calculation, with an associated verbal scale and of an evaluation based on a subjective opinion of source.

An opinion that a shoe 'could have made the mark' is not in our view the same as saying that 'there was moderate [scientific] support for the prosecution case'. The use of the term 'could have made' is a more precise statement of the evidence; it enables a jury better to understand the true nature of the evidence than the more opaque phrase 'moderate [scientific] support'. (para. 73.)

Reports - other than NAS report

Law Commission of England and Wales: Expert evidence in criminal proceedings (2011)

Reports - other than NAS report

Law Commission of England and Wales: Expert evidence in criminal proceedings (2011)

- It is also important that appropriate training on how to determine evidentiary reliability, particularly in relation to evidence of a scientific nature, should be undertaken by all judges and lawyers involved in criminal proceedings. (para. 1.43)

Reports - other than NAS report

Law Commission of England and Wales: Expert evidence in criminal proceedings (2011)

- It is also important that appropriate training on how to determine evidentiary reliability, particularly in relation to evidence of a scientific nature, should be undertaken by all judges and lawyers involved in criminal proceedings. (para. 1.43)
- Ideally, law students would in due course receive instruction on scientific methodology and statistics as part of their undergraduate courses, and the CPD requirements for practising solicitors and barristers who undertake work in criminal law would be amended to require attendance at approved lectures covering the same areas (in the context of criminal proceedings). (Footnote to para. 1.43)

Reports - other than NAS report

Law Commission of England and Wales: Expert evidence in criminal proceedings (2011)

Recommendation on expert reports

We recommend that Part 33 of the Criminal Procedure Rules be amended to include the following:

(2) a rule requiring an expert's report to include -

(c) a rule that where an expert witness is called by a party to give a reasoned opinion on the likelihood of an item of evidence under a proposition advanced by that party, the expert's report must also include, where feasible, a reasoned opinion on the likelihood of the item of evidence under one or more alternative propositions (including any proposition advanced by the opposing party);

(Para. 7.21)

Reports - other than NAS report

RSS / Nuffield: Communicating and interpreting statistical evidence in the administration of criminal justice (2010 - 2014):

- Fundamentals of probability and statistical evidence in criminal proceedings.
- Assessing the probative value of DNA evidence.
- The Logic of Forensic Proof - inferential reasoning in criminal evidence and forensic science.
- Case assessment and interpretation of expert evidence.

See http://www.rss.org.uk/statsandlaw
Co-authors: Paul Roberts, Graham Jackson, Roberto Puch-Solis, Sue Pope.

Reports - other than NAS report

ENFSI Guideline for evaluative reporting in forensic science (2014) (with support of the Prevention of Fight against Crime Programme of the European Union European Commission - Directorate - General Justice, Freedom and Security):

- Provision of a standard framework for the formulation of evaluative reports and related requirements for the case file.
- Provision of an assessment of the strength to be attached to the findings in the context of alleged circumstances.
- The conclusion shall be expressed by a value of the likelihood ratio and / or the use of a verbal scale related to the value of the likelihood ratio. The verbal equivalents shall express a degree of support for one of the propositions relative to the alternative. (para. 3.14)

Proselytise

Proselytise

- Hold conferences and workshops: ICFIS and FORSTAT;

Proselytise

- Hold conferences and workshops: ICFIS and FORSTAT;
- Liaise with policy makers in law, and judiciary (e.g., through the RSS Section on Statistics and the Law; ASA ad hoc Advisory Committee on Forensic Science, Committee on Law and Justice Statistics.

Proselytise

- Hold conferences and workshops: ICFIS and FORSTAT;
- Liaise with policy makers in law, and judiciary (e.g., through the RSS Section on Statistics and the Law; ASA ad hoc Advisory Committee on Forensic Science, Committee on Law and Justice Statistics.
- Attend law and forensic science conferences and workshops: (e.g., The Hague Institute for Global Policy: Evidence-based fact-finding).

Contents

What should be done: Education

- Statisticians in the ways of law and forensic science;
- Judiciary in the ways of statistics and probabilistic reasoning;
- Provision of primers;
- Courses on evaluation and interpretation in undergraduate law degree programmes.

Research

Peer-reviewed scientific research:

- Funding stream;
- Recognition of forensic science as a research topic;
- Research institute: research programmes.

Research

Peer-reviewed scientific research:

- Funding stream; NIJ / SRP for various aspects of forensic science; European Union Monopoly project with ENFSI.
- Recognition of forensic science as a research topic;
- Research institute: research programmes.

Research

Peer-reviewed scientific research:

- Funding stream; NIJ / SRP for various aspects of forensic science; European Union Monopoly project with ENFSI.
- Recognition of forensic science as a research topic;
- Research institute: research programmes. NIST Centre of Excellence in Forensic Science.

Research

Peer-reviewed scientific research:

- Funding stream; NIJ / SRP for various aspects of forensic science; European Union Monopoly project with ENFSI.
- Recognition of forensic science as a research topic;
- Research institute: research programmes. NIST Centre of Excellence in Forensic Science.

University departments lack 'the in-depth experience of taking ideas through from concept to a 'rugged and validated process that can withstand the rigours of the adversarial judical process'. (Kevin Sullivan of FSS on the closure of the FSS; former Head of R & D, and latterly the Standards and Validation Manager for the FSS in evidence to the House of Commons Science and Technology Committee investigation into the closure of the FSS.)

Legal Framework - Lord Thomas CJ; 2015

'In whatever system forensic evidence is given it is necessary to ensure that

- the expert evidence has a reliable scientific base;
- the scientists giving evidence are themselves reliable;
- the ambit of the expert's opinion is properly understood (issue to be addressed and the strength of the evaluative opinion);
- the system for collecting the evidence and safeguarding it during analysis provides clear continuity and
- the expert evidence is explained to the judge or jury in a way that they can properly assess it.'

Legal Framework - Lord Thomas CJ; 2015

Strength of an evaluative opinion

'A scientist is entitled and in most cases must express an evaluative opinion as to the conclusion to be drawn from the primary facts on which he gives evidence.'

'More difficult, however, is the question as to the extent to which such an evaluative opinion can be based on a numerical approach ... It is an issue, however, that needs to be addressed.'

Contents

Bibliography 1

- Aitken,C.G.G. and Lucy,D. (2004) Evaluation of trace evidence in the form of multivariate data. Applied Statistics, 53, 109-122; with corrigendum 665-666.
- Darboux,J.G., Appell,P.E. and Poincaré,J.H. (1908) Examen critique des divers systèmes ou études graphologiques auxquels a donné lieu le bordereau. In *L'affaire Drefus - La révision du procès de Rennes - enquête de la chambre criminelle de la Cour de Cassation.* Ligue française des droits de l'homme et du citoyen, Paris, France, 499-600.
- Davis,L.J., Saunders,C.P., Hepler,A. and Buscaglia,J. (2012) Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic Science International*, 216,146-157.
- Enzinger,E, and Morrison,G.S. (2015) Mismatched distances from speakers to telephone in a forensic-voice-comparison case. *Speech Communication*, 70, 28-41.

Bibliography 2

- Good,I.J. (1979) Studies in the history of probability and statistics. XXXVII: A.M.Turing's statistical work in World War II. *Biometrika*, 66, 393-6.
- Lauritzen,S.L. and Spiegelhalter,D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *JRSSB*, 50, 157 - 224.
- Law Commission of England and Wales (2011) Expert evidence in criminal proceedings. (Law Com No 325).
- Lindley,D.V. (1977) A problem in forensic science. *Biometrika*, 64, 207-213.
- Lord Thomas CJ (2015) The legal framework for more robust forensic science evidence. *Phil. Trans. R. Soc. B*, 370: 2014258. http://dx.doi.org/10.1098/rstb.2014.0258.
- Neumann,C., Evett,I.W. and Skerrett,J. (2012) Quantifying the weight of evidence from a fingerprint comparison: a new paradigm (with discussion). *JRSSA*, 175, 371-415.

Bibliography 3

- Oesterhelweg,L. et al. (2008) Cadaver dogs - a study on detection of contaminated carpet squares. *Forensic Science International*, 174, 35.
- Peirce,C.S. (1878) The probability of induction. *Pop. Sci. Monthly* Reprinted (1956) in *The World of Mathematics*, volume 2, Ed. J.R. Newman, pp. 1341-54. New York: Simon and Schuster.
- Petraco,N.D., Shenkin,P. Speir,J., Diaczuk,P., Pizzola,P.A., Gambino,C. and Petraco,N. (2012) Addressing the National Academy of Sciences' challenge: a method for statistical pattern comparison of striated tool marks. Journal of Forensic Sciences, 57, 900-911.
- Quaak,F.C.A. and Kuiper,I. (2011) Statistical data analysis of bacterial t-RFLP profiles in forensic soil comparisons. *Forensic Science International*, 210, 96-101.
- Wilson,A., Aitken,C.G.G., Sleeman,R. and Carter,J. (2015) The evaluation of evidence for autocorrelated data with an example relating to traces of cocaine on banknotes. Applied Statistics. 64, 275-298.