Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

Handbook of Forensic Statistics

Edited by

David Banks
Karen Kafadar
David H. Kaye
Maria Tackett

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Handbook of Forensic Statistics

Chapman & Hall/CRC
Handbooks of Modern Statistical Methods

Series Editor
Garrett Fitzmaurice, *Department of Biostatistics*, *Harvard School of Public Health*, *Boston, MA, U.S.A.*

The objective of the series is to provide high-quality volumes covering the state-of-the-art in the theory and applications of statistical methodology. The books in the series are thoroughly edited and present comprehensive, coherent, and unified summaries of specific methodological topics from statistics. The chapters are written by the leading researchers in the field and present a good balance of theory and application through a synthesis of the key methodological developments and examples and case studies using real data.

Published Titles

Handbook of Design and Analysis of Experiments
Angela Dean, Max Morris, John Stufken, and Derek Bingham
Handbook of Cluster Analysis
Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci
Handbook of Discrete-Valued Time Series
Richard A. Davis, Scott H. Holan, Robert Lund, and Nalini Ravishanker
Handbook of Big Data
Peter Bühlmann, Petros Drineas, Michael Kane, and Mark van der Laan
Handbook of Spatial Epidemiology
Andrew B. Lawson, Sudipto Banerjee, Robert P. Haining, and María Dolores Ugarte
Handbook of Neuroimaging Data Analysis
Hernando Ombao, Martin Lindquist, Wesley Thompson, and John Aston
Handbook of Statistical Methods and Analyses in Sports
Jim Albert, Mark E. Glickman, Tim B. Swartz, and Ruud H. Koning
Handbook of Methods for Designing, Monitoring, and Analyzing Dose-Finding Trials
John O'Quigley, Alexia Iasonos, and Björn Bornkamp
Handbook of Quantile Regression
Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng
Handbook of Statistical Methods for Case-Control Studies
Ørnulf Borgan, Norman Breslow, Nilanjan Chatterjee, Mitchell H. Gail, Alastair Scott, and Chris J. Wild
Handbook of Environmental and Ecological Statistics
Alan E. Gelfand, Montserrat Fuentes, Jennifer A. Hoeting, and Richard L. Smith
Handbook of Approximate Bayesian Computation
Scott A. Sisson, Yanan Fan, and Mark Beaumont
Handbook of Graphical Models
Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright
Handbook of Mixture Analysis
Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert
Handbook of Infectious Disease Data Analysis
Leonhard Held, Niel Hens, Philip O'Neill, and Jacco Walllinga
Handbook of Forensic Statistics
David Banks, Karen Kafadar, David H. Kaye, and Maria Tackett

For more information about this series, please visit: https://www.crcpress.com/Chapman--HallCRC-Handbooks-of-Modern-Statistical-Methods/book-series/CHHANMODSTA

Handbook of Forensic Statistics

Edited by
David Banks
Karen Kafadar
David H. Kaye
Maria Tackett

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data
Library of Congress Control Number: 2020934402

ISBN: 978-1-138-29540-7 (hbk)
ISBN: 978-0-367-52770-9 (ebk)

Typeset in Palatino
by Nova Techset Private Limited, Bengaluru & Chennai, India

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

In memory of Stephen E. Fienberg (1942–2016), a pioneer in the field of forensic statistics and the use of statistics in the courtroom, and his beloved wife, Joyce Fienberg (1943–2018), who tragically perished in the Tree of Life mass shooting in Pittsburgh in 2018.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Foreword

The subject of this book is forensic statistics—the application of statistical and probabilistic reasoning to the discovery and proof of facts in legal settings. It focuses on those aspects of forensic science that are used primarily in criminal cases.

The book begins with thematic and methodological chapters. The first chapter is essential reading for a broad understanding of current developments. It sketches the recent history of forensic inference and statistics, identifies today's issues and interpretive methods, and points to where they are heading.

Chapters 2–8 (Section II) discuss general concepts, methods, and philosophies of statistical inference, including basic introductions (intended for readers with limited knowledge of statistics) to such topics as hypothesis testing, confidence intervals and credible regions, likelihood ratios, and simulation methods. Additional chapters in this section discuss decision theory, Bayesian networks, and validation of likelihood ratios. Within these chapters, a reader can find the underlying statistical theory used by later chapters on applications of statistics to various domains of forensic science.

Chapters 9 and 10 (Section III) are on statistics in the courtroom and the psychological dimensions of statistical testimony. They complete the background for Chapters 11–21 (Section IV) on the statistical aspects of methods employed in specific areas of forensic science. The latter chapters include the analysis of DNA samples, friction ridge skin patterns, shoe prints, glass fragments, firearms and ammunition, and much more. In this section, the authors lay out the current state of statistical knowledge and practice in many different areas, but forensic science is so broad that this handbook by no means exhausts the field. Nonetheless, Section IV addresses domain-specific applications that are commonly encountered in criminal investigations and trials, and the statistical issues that arise in one subfield frequently transfer to others.

The chapters were commissioned, compiled, reviewed, and edited with a broad readership in mind. We believe they will be of value to students, statisticians, forensic scientists, and lawyers. Previous familiarity with statistical thinking and methods will be a strong asset in benefiting from the book as a whole, but every chapter can be read on its own. Some chapters are relatively mathematical and technical; others have no formulas, or just a few. Some chapters are wide ranging; others focus on particular types of forensic-science evidence. Each chapter describes, to varying extents, past literature on completed studies, while some present new research. We hope the handbook will be useful to broad audiences with diverse interests.

David Banks
Duke University

Karen Kafadar
University of Virginia

David H. Kaye
Pennsylvania State University

Maria Tackett
Duke University

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Preface

During the 2015–2016 academic year, the Statistical and Applied Mathematical Sciences Institute (SAMSI) held a research program called Statistics and Applied Mathematics in Forensic Science. That program had many positive scientific outcomes, and the editors hope that this book will count among them.

The SAMSI research effort was part of a broader movement triggered, to a large degree, by the 2009 report from a committee of the National Academies entitled *Strengthening Forensic Science in the United States: A Path Forward*. Among other issues, that report called attention to gaps in the scientific and statistical research into common forensic science methods for analyzing various forms of trace evidence, to problems with the presentation of findings and opinions in the courtroom, and to the need for education and training to enable criminalists and other practicing forensic scientists to understand probability and the limits of decision making under uncertainty.

These issues continue to be central to the transformation of forensic-science disciplines from skilled crafts into fully scientific endeavors. For example, the limited validation of pattern-matching evidence (a category that includes tool marks on firearms and other items, shoe prints, hair microscopy, and handwriting) remains a matter of concern to a growing number of forensic scientists and courts. In many of these fields, examiners are taught to follow broadly defined steps known as "Analysis, Comparison, Evaluation, and Verification", or ACE-V. The steps constitute a high-level description of the process of identifying the features of a pair of items, assessing the degree of similarity between two items, and coming to a decision (or not) about their origins. In latent fingerprinting, for example, the analysis step produces a determination of whether a latent print has enough clear detail to merit further study. The comparison step is a side-by-side examination of the features of the latent print and an exemplar print from a known finger. The evaluation step yields a final determination as to whether the prints come from the same finger, whether they do not, or whether the similarities and dissimilarities in the patterns are inconclusive. Finally, the verification step requires a second practitioner (typically unblinded to the information in the case and to the first examiner's work and conclusions) to evaluate the prints.

Of course, "the devil is in the details". Merely dividing the process into these phases does not make it a reproducible and valid method—particularly when the interpretive process is highly personal and judgmental. Thus, the FBI had to repudiate its thrice-verified conclusion that a latent fingerprint on a bag containing detonators and explosives found in the aftermath of the 2004 train bombing in Madrid belonged to Brandon Mayfield, an Oregon lawyer, who claimed he had not left the United States for ten years and did not own a passport. The retraction came after Spanish National Police matched the latent print to Ouhnane Daoud, an Algerian national living in Spain.

To be sure, there is experimental evidence that supports the ability of examiners following the general ACE-V steps for friction-ridge matching to reach correct conclusions, but only with accuracies below the level that many members of the public would expect. And, the "black box studies", as they are sometimes called, do not reveal the accuracy and reliability of examiners engaged in real casework, all the way from the beginning (evidence collection) to the end (final assessment) of the process. This limitation is especially significant when one considers the wide range of settings in which friction ridge analysis is performed—crime laboratories, consultants, police "identification units", and

non-accredited facilities. And, as the 2009 National Academies report also noted, training may be done formally or through informal mentoring. It can consist of just a short course, and there is no unified curriculum.

As a result, little is known about error rates in practice. Although at least one laboratory has experimented with introducing a small number of blind test prints into the flow of casework as a quality control measure, and although analysts in accredited laboratories (who know when they are being tested) are periodically given proficiency tests, well-powered double-blind experiments using realistically smudged or partial prints to determine a practitioner's error rate are rarely used.

In addition to the issues of validation and estimation of error rates for forensic-science methods and examiners, how to accurately and comprehensibly communicate the probative value of forensic-science findings has moved to center stage. For more than thirty years, the paradigm of presenting firm (or even overtly probabilistic) conclusions on source hypotheses has been questioned. The 2009 National Academies report noted literature advocating having the expert describe the weight of evidence in favor of one hypothesis relative to another instead of deciding between the two. That approach has been adopted in several countries. Yet, issues of validation pertain in all approaches (e.g., frequentist, likelihood, or Bayesian), and unresolved questions remain regarding what type of presentation on weight of evidence is best understood by judges and jurors.

This book contains discussions of all these topics—and more. Nevertheless, as a short handbook rather than an encyclopedia, it does not purport to review every statistical aspect of every method within the sprawling field of forensic science. As editors, we aimed for a collection that would provide cross-cutting statistical and historical background for all readers as well as reviews of the state of the statistical science in certain fields. Although every chapter was peer-reviewed by other experts and revised in response to their comments, inclusion does not necessarily imply that the editors agree with all the views and positions expressed in each chapter. Rather than impose our own opinions, we allowed the authors to speak in their own voices. We hope that the final result is a convenient resource for students and practitioners of statistics, forensic science, and law—and one that will improve the standard of practice and the communication of uncertainty in the courtroom.

The editors thank the referees who volunteered their expertise and time to this task. Besides the four editors, the referees were Colin Aitken, Alicia Carriquiry, Maria Cuellar, Ernest Fokoué, Christopher Glynn, Alice Liu, Lucas Mentch, Roi Naveiro Flores, Cedric Neumann, Karen Pan, and William Thompson.

David Banks
Duke University

Karen Kafadar
University of Virginia

David H. Kaye
Pennsylvania State University

Maria Tackett
Duke University

Editors

David Banks is a professor in the Department of Statistical Science at Duke University. He is a former coordinating editor of the *Journal of the American Statistical Association*, director of the Statistical and Applied Mathematical Sciences Institute, and a Fellow of the American Statistical Association and the Institute of Mathematical Statistics.

Karen Kafadar is a Commonwealth Professor and the chair of the Department of Statistics at the University of Virginia. She is a former president of the ASA; a Fellow of the International Statistics Institute, the ASA, and the AAAS; and a former member of the Forensic Science Standards Board (FSSB) of the Organization of Scientific Area Committees for Forensic Science (OSAC).

David H. Kaye is Distinguished Professor of Law Emeritus at Pennsylvania State University and Regents' Professor of Law and Life Sciences Emeritus at Arizona State University. He is a former editor of *Jurimetrics Journal*; a member of the FSSB; and the 2020 recipient of the Association of American Law Schools' Wigmore Lifetime Achievement Award for contributions to the understanding of the proof process and the rules of evidence.

Maria Tackett is an assistant professor of the practice in the Department of Statistical Science at Duke University.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contributors

Colin Aitken
University of Edinburgh
Edinburgh, Scotland

John A.D. Aston
University of Cambridge
Cambridge, England

Madeline Ausdemore
South Dakota University
Brookings, South Dakota

David Banks
Duke University
Durham, North Carolina

Charles E.H. Berger
Netherlands Forensic Institute
Leiden University
Leiden, the Netherlands

Alex Biedermann
University of Lausanne
Lausanne, Switzerland

Silvia Bozza
Ca' Foscari University of Venice
Venice, Italy

Alicia Carriquiry
Iowa State University
Ames, Iowa

Junqi Chen
University of Virginia
Charlottesville, Virginia

Maria Cuellar
University of Pennsylvania
Philadelphia, Pennsylvania

A. Philip Dawid
University of Cambridge
Cambridge, England

Gary Edmond
University of New South Wales
Sydney, Australia

Benjamin Eltzner
University of Göttingen
Göttingen, Germany

Ewald Enzinger
Aston University
Birmingham, United Kingdom
and
Eduworks
Corvallis, Oregon

Brandon L. Garrett
Duke University
Durham, North Carolina

Joaquín González-Rodríguez
Higher Polytechnic School
Autonomous University of Madrid
Madrid, Spain

Martin J.R. Hall
Museum of Natural History
London, England

James Hamby
International Forensic Science Laboratory & Training Centre
Indianapolis, Indiana

Rudolf Haraksim
European Commission
Ispra, Italy

Jessie Hendricks
South Dakota University
Brookings, South Dakota

Heike Hofmann
Iowa State University
Ames, Iowa

Stephan Huckemann
University of Göttingen
Göttingen, Germany

Alan Julian Izenman
Temple University
Philadelphia, Pennsylvania

Karen Kafadar
University of Virginia
Charlottesville, Virginia

Naomi Kaplan-Damary
University of California Irvine
Irvine, California

David H. Kaye
Pennsylvania State University
University Park, Pennsylvania
and
Arizona State University
Phoenix, Arizona

Alice J. Liu
University of Virginia
Charlottesville, Virginia

Alicia Lozano-Díez
Higher Polytechnic School
Autonomous University of Madrid
Madrid, Spain

Kristy A. Martire
University of New South Wales
Sydney, Australia

Lucas Mentch
University of Pittsburgh
Pittsburgh, Pennsylvania

Didier Meuwly
Netherlands Forensic Institute
and
University of Twente
Enschede, the Netherlands

Geoffrey Stewart Morrison
Aston University
and
Forensic Evaluation Ltd.
Birmingham, United Kingdom

Julia Mortera
Roma Tre University
Rome, Italy

Cedric Neumann
South Dakota State University
Brookings, South Dakota

Karen Pan
University of Virginia
Charlottesville, Virginia

Davide Pigoli
King's College
London, England

Daniel Ramos
Higher Polytechnic School
Autonomous University of Madrid
Madrid, Spain

M.J. (Marjan) Sjerps
Netherlands Forensic Institute
and
University of Amsterdam
Amsterdam, the Netherlands

Cliff Spiegelman
Texas A&M University
College Station, Texas

Hal S. Stern
University of California Irvine
Irvine, California

Maria Tackett
Duke University
Durham, North Carolina

Xiao Hui Tai
Carnegie Mellon University
Pittsburgh, Pennsylvania

Franco Taroni
University of Lausanne
Lausanne, Switzerland

Susan VanderPlas
Iowa State University
Ames, Iowa

Bruce S. Weir
University of Washington
Seattle, Washington

Sarena Wiesner
Israeli Police Force
Jerusalem, Israel

Joanne Yaffe
University of Utah
Salt Lake City, Utah

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Section I

Perspectives on Forensic Statistics

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

The History of Forensic Inference and Statistics: A Thematic Perspective

Colin Aitken and Franco Taroni

CONTENTS

1.1 Introduction

The historical development of forensic inference and statistics is presented through fifteen important themes. The themes have been chosen as the ones that created, and in some cases are still creating, important debates. The choice of themes is a personal choice of the authors and some readers may not agree. It is hoped this form of presentation will help clarify thinking around current problems and suggest ways in which the subject may develop further.

It is only the role of statistics in the evaluation of evidence in criminal cases that is discussed. No reference is made to civil law, such as examples of jury selection and employment discrimination. Preference is given instead to the development since Dennis Lindley's seminal paper in *Biometrika* in 1977 (Lindley, 1977).

Fifteen themes are identified as important in the development of the ideas for probabilistic and statistical reasoning in forensic science. The first theme (Section 1.2) is the recognition in the early 20th century of the need for the interpretation of scientific findings in the administration of criminal justice. The next two themes (Sections 1.3 and 1.4) concern ideas for the integration of scientific information with other relevant information from a particular criminal case and the increasing support of judicial disciplines for the scientific presentation of evidence. These ideas led to recognition of the importance of the separation of evidence from propositions and the correct conditioning of one on the other depending on the role of the person making the judgement (Section 1.5). Various attempts to quantify the value of the evidence before the general acceptance of the likelihood ratio as the best way to do this are described in Section 1.6, followed by the description of the discrediting of the idea of a match (Section 1.7). The advent of DNA profiling in the mid-1980s led to consideration of many new factors in the evaluation of evidence which are outlined in Section 1.8. One factor in particular, that of the possibility of extremely small probabilities for a DNA profile and correspondingly large values of the likelihood ratio merits a section on its own (Section 1.9). The concept of propositions was developed further in the late 1990s with the introduction of differing levels of propositions (Section 1.10). The general use of the likelihood ratio and the difficulty jurists had with its interpretation led to attempts to summarise its numerical value verbally (Section 1.11). Though the role of the likelihood ratio was generally accepted, there could be several different values in a particular case arising from the use of different assumptions and statistical models. Recognition of these differences led to consideration of methods for the assessment of the performance of different models (Section 1.12). The role for the forensic scientist in the investigation of a crime (the *investigative role*) before they are asked to evaluate evidence in a trial (the *evaluative role*) was recognised in a separate development in the 1990s, a role that is described in Section 1.13. Statistical research in the late 20th century led to probabilistic graphical networks for complicated problems of inference. These networks had an intuitively satisfying application in forensic science, in particular for the management of many different pieces of evidence, and this application is described in Section 1.14. Ultimately a decision has to be reached by the jurist (jury or judge) concerning the outcome of a criminal trial. Scientists also have decisions to make, earlier in the process, for example concerning sample size or choice of analysis. The role of decision theory for the scientific process is described in Section 1.15. Finally, the early years of the 21st century have seen questioning of the presentation of a single value for evidential value with the likelihood ratio. The alternative suggestion is that the single value of the likelihood ratio should be replaced by an interval for, or a lower bound on, its value. A comment on this debate is given in Section 1.16.

1.2 Forensic Science and the Evaluation of Evidence

In the early 1960s, the forensic science community started to take a more explicit position with respect to the problems of interpretation and evaluation of scientific data. In a now widely known quote, Kirk and Kingston (1964) from the University of California, Berkeley, note:

> When we claim that criminalistics is a science, we must be embarrassed, for no science is without some mathematical background, however meagre. This lack must be a matter of primary concern to the educator [. . .]. Most, if not all, of the amateurish efforts of all of us to justify our own evidence interpretations have been deficient in mathematical exactness and philosophical understanding. (at pp. 435–436)

Today, interpretation and data evaluation are still a neglected area, mainly in fields that involve so-called *physical* evidence. This neglect continues to exist despite an important paper by Stoney (1984) that gave the relevant questions a scientist should ask in their analyses. This neglect is now acknowledged, for instance, by reports such as that of the National Research Council of the US (National Research Council, 2009) and of the President's Council of Advisors on Science and Technology (PCAST, 2016). In its report the NRC Council notes that '[t]here is a critical need in most fields of forensic science to raise standards for reporting and testifying about the results of investigations' (at p. 185). In many contexts this perception is reinforced by the fact that scientists' assessments of evidential value consist of a largely subjective component with a connotation of arbitrariness. As mentioned by Kirk and Kingston (1964), it was indeed rare at their time of writing that a scientist's opinion was based on a quantitative study. With some notable exceptions, for example transferred material such as DNA, fibres or glass fragments, this is still the situation today even though quantification with the use of probabilities was suggested by earlier forensic scientists, for example Bertillon (1893, 1898) and Locard (1920, 1940) in areas such as questioned documents and anthropology. See, for the sake of illustration, a quote from Bertillon (1898) on the need for a quantification:

> This writing, characterized by the set of unique features we have enumerated, can only be encountered in one individual among a hundred, among a thousand, among ten thousand or among a million individuals. (at p. 20)

Perhaps the first probabilistic approach for evidence evaluation was the approach used in the Howland Will case between 1865 and 1868 (Meier and Zabell, 1980). A general probabilistic summary of the evidence was given by scientists in the early 20th century (Darboux et al., 1908) in the Dreyfus case. This approach was supported later by Kingston (1965a,b) and echoed by Saks and Koehler (2005) who concluded that 'Although obstacles exist both inside and outside forensic science, the time is ripe for the traditional forensic sciences to replace antiquated assumptions of uniqueness and perfection with a more defensible empirical and probabilistic foundation' (at p. 895). However, this approach is still viewed as controversial in some quarters.

The final word in this section comes from an article some 25 years before those of Kingston. Locard (1940) proposed some inspired guidelines for the interpretation of scientific evidence. These guidelines remain pertinent to scientists and lawyers even today.

> The physical certainty provided by scientific evidence rests upon evidential values of different orders. These are measurable and can be expressed numerically. Hence the expert knows and argues that he knows the truth, but only within the limits of the risks of error inherent to the technique. This numbering of adverse probabilities should be explicitly indicated by the expert. The expert is not the judge: he should not be influenced by facts of a moral sort. His duty is to ignore the trial. It is the judge's duty to evaluate whether or not a single negative evidence, against a sextillion of probabilities, can prevent him from acting. And finally, it is the duty of the judge to decide if the evidence is in that case, proof of guilt. (at pp. 286–287)

1.3 The Need for an Interpretative Model

Data are fundamental for the reduction of uncertainty about propositions of interest for a court. Uncertainty should be expressed by the concept of probability. Uncertainty is inevitable because the role of a court is to reconstruct what has happened in the past (i.e., the commission of the crime) based on incomplete knowledge. Such a reconstruction inevitably results in an uncertain representation of the true state of affairs. Information may be gained by enquiry, analysis and experimentation. As a consequence of this, a method is required to adjust existing beliefs in the light of newly acquired evidence. Inferences, if they are to be taken seriously, must be approached within a probabilistic framework. The revision of beliefs should be made according to Bayesian procedures. This is not controversial; it is a logical consequence of the basic rules of probability. Reference will be made often to Bayes's theorem; it is a very important result that helps one understand how beliefs should be adjusted in the light of new evidence. Although the theorem has a history of about 250 years, the associated approach to inference has gained more widespread use only since the end of the 20th century. This is the case even though, historically, practical applications of patterns of reasoning corresponding to a Bayesian approach can be found, for example, as early as the beginning of the 20th century. For example, at the Dreyfus's military trial held in 1908, Henri Poincaré invoked Bayes's theorem as the only way in which the court ought to revise its opinion about the issue of forgery (Darboux et al., 1908). He described its applications as follows:

> An effect may be the product of either cause A or cause B. The effect has already been observed; one wants to know the probability that it is the result of cause A; this is the *a posteriori* probability. But, I'm not able to calculate this if an accepted convention does not permit me to calculate in advance the *a priori* probability for the cause producing the effect; I want to speak of the probability of this eventuality, for one who has never before observed the result.

Given the difficulty - for a scientist - in dealing with the *a priori* probability, Poincaré and his colleagues supported the use of the likelihood ratio, the expression that is the connection between prior and posterior probabilities:

> 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. (at p. 504)

This quotation is a statement of the odds form of Bayes's theorem, namely

$$\frac{Pr(H_p \mid E, I)}{Pr(H_d \mid E, I)} = \frac{Pr(E \mid H_p, I)}{Pr(E \mid H_d, I)} \times \frac{Pr(H_p \mid I)}{Pr(H_d \mid I)}. \tag{1.1}$$

Statements about the probability of the evidence E if the suspect is not the forger H_d, with background information I, are part of the likelihood ratio $Pr(E \mid H_p, I)/Pr(E \mid H_d, I)$ (the X of the statement of Poincaré's and his colleagues). The proposition that the suspect is the forger is denoted by H_p, the proposition that the suspect is not the forger by H_d. '[T]he ratio of the forgery's probability to the inverse probability' is a statement of the odds in favour of H_p. '[F]ollowing the observation of this coincidence, this ratio becomes X times greater than before the observation.' The probability the suspect is the forger given the evidence is the numerator of the posterior odds $Pr(H_p \mid E, I)/Pr(H_d \mid E, I)$. As H_p and H_d are complementary (a situation that does not always hold in the evaluation of evidence), it is possible to determine the probability $Pr(H_p \mid E, I)$ from the posterior odds. The likelihood ratio X and the posterior odds are related by the prior odds $Pr(H_p \mid I)/Pr(H_d \mid I)$.

Whereas Poincaré and his colleagues refused to evaluate prior probabilities, an anonymous commentator of the Dreyfus trial (Darboux et al., 1907) opined that reasoning remained valid and sound if probabilities could be put on past acts. The commentator developed the inferential reasoning adopting a shoe print example:

> Burglary is committed in a house surrounded by a park. A suspect is apprehended because of his appearance and his criminal history record. These elements alone are not sufficient to allow conviction. However, shoe prints are recovered at the scene and they correspond to the soles of the suspect's shoes. This is sufficient proof. The juror's opinion is established and the conviction is delivered. But nothing proves with certainty that two different shoes could not produce an identical shoe print, and that the shoe prints from the scene could not come from shoes worn by someone other than the suspect. The juror's logic is sound only if the probability of other explanations is extremely small. The juror's reasoning is as follows: the perpetrators of the burglary are either the suspect or a person or persons unknown. The *a priori* probabilities of these possibilities are fixed only by moral criteria. This possibility cannot be expressed accurately by a single number, but it can be said to fall within certain limits or a given range. However, after the verification of the shoe prints, everything changes. If the prints are those of the accused, the probability of observing such evidence is very high, the functional equivalent of certainty. Conversely, the probability of finding these prints made by someone else cannot be precisely determined, but is is extremely low. In sum, the *a priori* probabilities have been modified, permitting the 'a beyond reasonable doubt' conviction required by the law. (at pp. 19–20)

Despite the problem encountered by Bertillon during the Dreyfus case (e.g., an early example of what is now known as 'the prosecutor's fallacy', Section 1.6), Bertillon can be nominated as the first Bayesian forensic scientist. In fact, after expressing the need of a quantification of the observed features (Bertillon, 1898), he completed his reasoning by affirming that the only way to accept an expert's categorical conclusions was to consider not only the statistical evidence provided by the examination of the document, but also other information pertaining to the inquiry. He described how the number of people who could be the author of the questioned document size is reduced by the inquiry (i.e., the testimonies and circumstances of the case). This description introduces the general idea of a relevant population, a concept expanded and discussed by Lempert (1991), Robertson and Vignaux (1993a), Champod et al. (2004) and Kaye (2004). An important contribution to the

role of scientific evidence is that of Fienberg et al. (1996). He and his co-authors note that (a) what is treated as a relevant population may only be a conveniently available population and (b) the event that evidence associated with the crime came from the defendant is not necessarily the same as the event that the defendant committed the crime.

Therefore, the evidentiary value of the scientific observations, even if not totally confirmatory of guilt, could supply sufficient information to allow a conviction when the case is considered as a whole. Other examples of similar reasoning were published by Balthazard (1911) and Souder (1934) in the fields of fingermarks and typewriting machines, respectively. A complete historical summary of the relationship of forensic scientists to Bayesian ideas is presented in Taroni et al. (1998).

The Bayesian interpretative model has attracted many supporters in the area of interpretation and evaluation of evidence in forensic science. The model contains all the ingredients necessary for the required inferences.

> [...] the only argument we can adduce is to ask the reader to pursue it and see where it leads - the proof of the pudding is in the eating. (Lindley, 1985, p. 101)

1.4 Support of Judicial Disciplines for a Scientific Presentation of the Value of Evidence

A crucial factor in the progressive acceptance of a probabilistic presentation of the value of evidence has been the support of the judiciary. This support was not gathered by statisticians or forensic scientists but by jurists.

The support dates back to 1897. Mr Justice Holmes, then of the Supreme Judicial Court of Massachusetts and latterly of the Supreme Court of the United States (Holmes, 1897) wrote

> For the rational study of the law the black-letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics. (at p. 469)

He was echoed almost a century later by Twining who affirmed that

> The lawyer of today needs to be a master of elementary statistics. (Twining, 1994, p. 209)

Other supporters amongst jurists for a probabilistic presentation of the value of the evidence include Sir Richard Eggleston. Through a series of examples he underlined the fundamental role played by probabilities in legal settings. He wrote in the introduction to his book:

> It is plain from this example that probabilities must play a very large part in the decision of cases in the courts. Even in criminal cases, where the jury must be satisfied beyond reasonable doubt, probability theory is often of the highest importance. The acceptance of fingerprint evidence, for example, depends essentially on the 'multiplication rule', to be discussed hereafter. Yet the legal profession as a whole has been notably suspicious of the learning of mathematicians and actuaries, and ignorant of the work of philosophers in this field. (Eggleston, 1983, p. 2)

A justification for the use of probabilities can also be seen through the definition of 'relevant evidence' as defined by The U.S. Federal Rule of Evidence 401. The Rule states:

> Relevant evidence means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.

The use of a probabilistic line of reasoning is supported by Lempert who wrote:

> One of the main areas of interest of the so-called *new evidence* scholarship is the application of probability theory to arguments about facts in legal cases. As a preliminary to making a decision, courts have to 'find facts' which require them to reason under uncertainty. In some cases it may be the reasoning process itself which is examined in an appeal. The result may be a statement by the court about how facts ought to be thought about. Alternatively the way facts are thought about in a particular case may be seized upon as a precedent for future cases. Should there be rules about how facts are to be thought about? And, if so, does probability theory offer a prescription for those rules? (Lempert, 1986, p. 457)

There is a need for a clarification here. Lempert's statement supports the use of probabilistic reasoning by noting it is about the structure of reasoning and not particularly about numbers. This perspective has been reaffirmed by Robertson and Vignaux (1993a). Numbers are not important in themselves: what really matters is that numbers allow us to use powerful rules of reasoning which can be implemented by computer programs. What is important is not whether the numbers are 'precise', whatever the meaning of 'precision' may be in reference to subjective degrees of belief based upon personal knowledge. What is important is that we are able to use sound rules of reasoning to check the logical consequences of our propositions, and consider the consequences with respect to the degree of belief in one proposition of assuming a certain degree of belief in another proposition.

The legal system is concerned with making decisions, and decisions must often be made in a situation of uncertainty, either as to what has happened in the past, or as to what is going to happen in the future. Eggleston emphasised that

> We are not concerned here with uncertainty as to the legal rule, though this is frequently a matter of anxiety, especially to those who have to make the decision whether or not to commence proceedings. Our interest is in uncertainty as to the facts to which the law must be applied.[. . .] [Therefore] students and practitioners, on the one hand, and non-lawyers, on the other, need to understand the judicial approach to probability. (Eggleston, 1983, pp. 3–4, 10)

It seems therefore that jurists support some of the concerns and desiderata expressed by forensic scientists such as Kingston and Kirk (1964) in the 1960s. They wrote:

- It can be fairly stated that there is no form of evidence whose interpretation is so definite that statistical treatment is not needed or desirable.
- The statistical analysis provides the criminalist with a basis for his opinion, and an evaluation of the likelihood that his testimony reflect the truth, rather than his personal belief or bias.

- This is not proposed in the belief that such accepted evaluations will be changed, but more in the hope that firmer lines of reasoning might replace the arbitrary justifications upon which many such evaluations now rest.

The first point is restated by Robertson and Vignaux (1995b, p. 12) in the following terms:

> An ideal piece of evidence would be something that always occurs when what we are trying to prove is true and never occurs otherwise. If we are trying to demonstrate the truth of a hypothesis or assertion we would like to find as evidence something which always occurs when the hypothesis is true and never occurs when the hypothesis is not true. In real life, evidence this good is almost impossible to find.

The necessity of probabilistic reasoning in law has been discussed. The next step is discussion of its use in the presentation of the value for a piece of evidence. The following quote from 1977 is an early suggestion for lawyers of a form of words for the presentation of a numerical value of the likelihood ratio*:

> [...] the defendant's thumb print was found on the gun the killer used. [...] assume that the fact-finder believes that the presence of this evidence is 500 times more likely if the defendant is guilty than if he is not guilty. [...] Now suppose that the prosecution wished to introduce evidence proving that a print matching the defendant's index finger was found on the murder weapon. If this were the only fingerprint evidence in the case, it would lead the fact-finder to increase his estimated odds on the defendant's guilt to the same degree that the proof of the thumb print did. Yet, it is intuitively obvious that another five hundredfold increase is not justified when evidence of the thumb print has already been admitted. (Lempert, 1977, p. 1043)

This idea is novel for a lawyer. However, two forensic scientists had already used such a probabilistic metric for the value of the evidence. A practical example is offered by Kingston and Kirk (1964, p. 514):

> Now consider a problem of evaluating the significance of the coincidence of several properties in two pieces of glass. Suppose that the probability of two fragments from different sources having this coincidence of properties is 0.005, and that the probability of such a coincidence when they are from the same source is 0.999. What do these figures mean? They are simply guides for making a decision about the origin of the fragments [...]

Not all jurists appreciated the statistical approach to the evaluation of evidence. In the early 1970s a counter-argument was well-aired and debated; see Finkelstein and Fairley (1970), Tribe (1971) and Finkelstein and Fairley (1971). Notwithstanding this debate some jurists did appreciate the approach. For example, the likelihood ratio and the role background information plays in the assignment of probabilities was described by Richard Friedman in 1996:

> Suppose, for example, you are sitting in a restaurant when you hear a voice that you do not recognize yell, 'There's been an accident outside!' You know nothing about the declarant and her relationship to what either did or did not happen outside, apart from what you know in general about the world and what you can infer from her voice. But

* A form of words that is not to be confused with a verbal summary of the likelihood ratio, a form of words that is discussed in Section 1.11.

> you know enough to make a preliminary assessment of how likely she would make the statement if it were true, and how likely she would do so if it were not - that is, you have enough information to make an assessment, albeit very tentative, of the likelihood ratio. If you turn to look and find that she is obviously drunk or obviously joking or, on the other hand, shaken and bloodied, you rapidly and radically may reassess the likelihoods in light of this further information. (Friedman, 1996, p. 1817)

In conclusion, it is affirmed that a probabilistic approach and particularly one based on Bayesian modelling should be considered as a valuable tool for reasoning about evidence (Redmayne, 1996); this is a consequence, as suggested by Robertson and Vignaux (1991), of Bayes's theorem as a formalisation of logic and common sense.

Examples of the applications of Bayesian analyses to legal matters include Cullison (1969), Fairley (1973), Finkelstein and Fairley (1970), Finkelstein and Fairley (1971), Fienberg and Kadane (1983), Kaye (1986) and Anderson and Twining (1998).

1.5 Probability of Proposition Given Evidence and of Evidence Given Proposition

One of the most important topics that created debate in the development of forensic statistics in the 1980s is the difference between the probability of a proposition given evidence and the probability of evidence given a proposition. To a statistician, the difference is clear. To a statistical layman the difference does not seem to be so clear. The interpretation of the probability of evidence given a proposition as the probability of the proposition given the evidence has been termed the prosecutor's fallacy (Thompson and Schumann, 1987) or inversion fallacy (Kaye, 1993) or transposed conditional (Evett, 1995). This confusion can be expressed more clearly as the interpretation of a small probability of finding the evidence on a person who is innocent of a crime as a small probability that a person on whom the evidence is found is innocent of the crime.

The error is easily exposed through the use of the odds form of Bayes's theorem. Consider a suspect and evidence E and mutually exclusive propositions:

- H_p: the suspect is guilty;
- H_d: the suspect is innocent.

These propositions are also exhaustive. In general it is not necessary for the evaluation of evidence for the propositions to be exhaustive but it is helpful here to expose the fallacy. For the evaluation of evidence there will always be a framework of circumstances or background information I to bear in mind for the evaluation.

The odds form of Bayes's theorem is given in (1.1). Statements about the probability of the evidence if the suspect is innocent are part of the likelihood ratio $Pr(E \mid H_p, I)/Pr(E \mid H_d, I)$. Statements about the probability of innocence of the suspect given the evidence are part of the posterior odds $Pr(H_p \mid E, I)/Pr(H_d \mid E, I)$. The likelihood ratio and the posterior odds are related by the prior odds $Pr(H_p \mid I)/Pr(H_d \mid I)$.

There is a very good medical analogy. The propositions of guilt and innocence may be replaced with diagnoses of presence or absence of disease. The evidence in the criminal case is replaced with medical symptoms or medical test results. In medicine there are often

data on the incidence of symptoms in the presence of the disease, a so-called *incidence rate* and on the incidence of symptoms in the absence of the disease. However, given these data it is not possible to estimate the probability a patient has the disease without knowledge of the so-called *base rate* of the disease in some background population from which the prior odds can be assigned.

Similarly in forensic science, knowledge of the incidence of a certain characteristic amongst a relevant population as well as in the criminal is not sufficient for a determination of guilt in a possessor of the profile. It is also necessary to have an assignment of the prior probability of guilt, the equivalent of a base rate in the medical analogy.

The propositions of guilt and innocence are what are known as *offence-level* propositions (see Section 1.10). It is a large step, for example, to move from an inference about the DNA profile of a suspect to the guilt of the suspect. It may be that the only inference possible is that the DNA of the suspect was present at the crime scene. Propositions about the source of the DNA are known as *source-level* propositions (see Section 1.10).

There are many variations of the prosecutor's fallacy and these are discussed as other errors of logic in Koehler (1993).

Thompson and Schumann (1987) introduced also a *defence attorney's fallacy* to balance the prosecutor's fallacy. Consider a crime in which a relevant population to which the criminal is deemed to belong is of size 100,000. For example, this could be a rape in which it is thought there are 100,000 males who could have committed the crime. A degraded DNA profile of the criminal obtained from semen found on the victim has an occurrence of 1 in 2,000. A suspect is identified in a manner independent of the profile and found to have a profile which is indistinguishable from that of the one known to have come from the criminal. The prosecutor's fallacy interprets the value of 1 in 2,000 as a probability of 1 in 2,000 that the suspect is innocent. The defence argue that there are 100,000 people who could have been the criminal. The occurrence of the profile is 1 in 2,000, thus there are fifty people in the relevant population who could have this profile. The suspect is one of fifty people so the probability of innocence is forty nine out of fifty. So far, the argument is correct. The argument becomes fallacious when it is extended to argue that the evidence is thus irrelevant. Before consideration of the evidence, the suspect was one of 100,000 men, after consideration of the evidence the suspect is one of fifty men. Such a consequence is very relevant.

1.6 Quantification of the Value of Evidence Using Alternative Numerical Summaries

The use of the likelihood ratio and functions of it, such as the logarithm (Peirce, 1878; Good, 1950), for the evaluation of evidence increased following the seminal paper of Lindley (1977). Before 1977, several attempts had been made to summarise the value numerically.

Consider categorical evidence, such that the evidence manifests itself as one, and only one, of a set K of exhaustive and mutually exclusive categories. The probability a particular evidential item has category k is p_k, $k = 1, \ldots, K : \sum_{k=1}^{K} p_k = 1$. Various suggestions might be offered for the value of evidence of category k found at a crime scene.

- The probability that two people have the same category k is p_k^2;

- The probability that two people have the same category, without specifying the category is

$$p_1^2 + \cdots + p_K^2.$$

- Given that one person, the *control* person, is of category k, the probability another person, chosen at random from a relevant population is also of category k is p_k.

The first two suggestions are of limited importance for the evaluation of evidence. One piece of evidence will have a known source. That evidence is known as *control* evidence. Another piece of evidence which is to be compared with the control evidence will have an unknown source, this evidence is known as *recovered* evidence. It is the third suggestion that is of importance. The numerator of a likelihood ratio is the probability the recovered evidence is of the same category k as the control evidence assuming, for example, that the recovered evidence and the control evidence come from the same source and in an idealised scenario this probability is 1. The denominator of a likelihood ratio is the probability the recovered evidence is of the same category as the control evidence assuming the recovered evidence and the control evidence come different sources; this probability is p_k. Thus p_k^{-1} is the value of the evidence.

A general assessment of the evidential value of a method is *discriminating power* (*DP*). It is related to the second probability with $DP = 1 - (p_1^2 + \cdots + p_K^2)$. *DP* is the probability that two people chosen at random will belong to different categories. For example, if everybody is of the same category, say category 1 without loss of generality, then $p_1 = 1$ and $p_2 + \cdots + p_K = 0$ and $DP = 0$; no-one can be discriminated. Conversely, if all categories are equally likely, $p_1 = \cdots = p_K = 1/K$ and $p_1^2 + \cdots + p_K^2 = 1/K$ and it can be shown that *DP* is maximised.

Discriminating power is a measure of the general worth of an evidential type. A high value for *DP* is indicative of evidence of a good discriminatory type. A low value for *DP* is indicative of evidence of a poor discriminatory type. Discriminating power is not a measure of evidential value in a particular case. An example of the use of discriminating power for hair examinations is given in Gaudette and Keeping (1974) with a critical discussion in Aitken and Robertson (1987).

DNA evidence introduced new challenges, notably that of a DNA mixture. An approach related to discriminating power, known as *random man not excluded* was proposed, discussed and criticised. A debate on this topic is given in Buckleton et al. (2016b).

1.7 Change from Two-Stage Approach to Continuous Approach

Procedures for the evaluation of evidence in forensic science were changed dramatically by a paper by Dennis Lindley in 1997 (Lindley, 1977). Previous to the publication of that paper a common procedure was a two-stage approach.

- Similarity: In a comparison of characteristics of evidence found at a crime scene and in the environment of a suspect, are the characteristics similar or dissimilar?

- Rarity: If the characteristics are dissimilar then the evidence is not considered any further, the evidence associated with the suspect is deemed not to be associated with the crime. If the characteristics are similar then the evidence associated with the suspect is deemed to be associated with the crime. The strength of the evidence under source level propositions (Section 1.10) is measured by the rarity of the characteristic; the more rare the characteristic, the stronger the association.

This description of the two-stage approach begs the questions of what is meant by similarity and what is meant by rarity.

Often, similarity was defined in relation to the result of a significance test. The characteristic takes the form of a continuous measurement such as that of the refractive index of a fragment of glass. The comparison of characteristics was made with a significance test of a null hypothesis of common source for the measurements of the crime scene characteristic and measurements of the characteristic (e.g., refractive index of fragments of glass) found in association with a suspect (e.g., upon their clothing). If the result of the test were such that the null hypothesis was rejected at some pre-specified level (e.g., 5%) then the two pieces of evidence would be deemed dissimilar and the alternative hypothesis of different sources would be accepted, in the sense that a decision would be made not to consider this evidence further. There are two problems with this procedure. The first problem is that the null hypothesis is one of common source. The null hypothesis is conventionally taken as the status quo and it will only be rejected if there is sufficient evidence against it (e.g., at a significance level of 5%). It is normally the prosecution that proposes a hypothesis of common source and it is the prosecution that wishes to show the suspect is guilty. A proposition that the suspect is guilty until there is sufficient evidence to show them innocent is contrary to the presumption of innocence. The second problem is the effect that has been called that of falling off a cliff (Robertson and Vignaux, 1995b; Robertson et al., 2016). Assume a pre-specified level of 5% for rejection of the hypothesis of common source. Evidence for which the comparison gives a result which is significant at the 5.1% level will be deemed to have a common source. Evidence for which the comparison gives a result which is significant at the 4.9% level will be deemed to have different sources. This is unsatisfactory.

Assessment of rarity in such a procedure is difficult. One procedure proposed in the late 1970s is that of a *coincidence probability*; see Evett (1977). This probability was defined as the probability that the characteristics taken from evidence selected at random from some item selected in turn at random from some relevant population of items would be found to be similar, in some sense, to a control item with a particular value of the characteristic. An example is that of the refractive index of glass fragments from a window. The coincidence probability would be the probability that the refractive indexes of a number of fragments selected at random from a window selected in turn at random from some relevant population of windows would be found to be similar, in some sense, to a control window, at a crime scene say, with a set of refractive indexes from a particular sample of fragments from the control window.

This approach can be compared with that of discriminating power. For discriminating power, the recovered and control fragments are both taken to be random samples from some underlying population. The probability is that of two random samples having similar characteristics. For a coincidence probability the data from the control window are taken to be fixed. The concern is with the assignment of the probability that one sample, the recovered sample, is found to be similar to the sample from the control window. Any variability in the values of the data from the control window is ignored.

The problems associated with the two-stage approach were overcome by Lindley (1977). A procedure was developed which accounted for the similarity and the rarity, with associated variation, in one statistic based on the likelihood ratio. The likelihood ratio developed provided a continuous measure of the value of evidence. Consider evidence E which is a set of continuous measurements, E_c and E_r, on control and recovered material, i.e., material for which the source is known and material for which the source is unknown, respectively, with $E = \{E_c, E_r\}$. The propositions for comparison are

H_p: E_c and E_r have the same source;
H_d: E_c and E_r have different sources.

The likelihood ratio is then

$$\frac{f(E \mid H_p)}{f(E \mid H_d)},$$

where the probability Pr of (1.1) is replaced by probability density functions in recognition of the continuous nature of the evidence. Further details are given in Lindley (1977).

1.8 Presentation of Evidence: New Challenges to Solve

In the early 1990s, interest in the probabilistic evaluation of DNA profiling results grew considerably. Topics such as the effect of database searches to select persons of interest, the role of sub-populations in the assignment of conditional probabilities and the consideration of error for false inclusions, were responsible for an increase in papers focusing on forensic inference.

1.8.1 The Island Problem and Results of a Database Selection

One important debate was one that became known as the *island problem* to which various solutions have been proposed (Eggleston, 1983; Yellin, 1979; Lindley, 1987). The problem relates to the determination of the probability of guilt. The problem has often been approached by consideration of a finite population such as may be found on an island, of population size $(N + 1)$ say. A crime is committed and evidence of a characteristic (e.g., a blood stain of DNA profile Γ, with occurrence γ amongst some larger population) is found at the scene of the crime. A person is found who possesses this characteristic and the probability of their guilt is of interest.

Determination of the probability is related to the manner in which this person has been selected (become a person of interest - POI). There are different ways a POI comes into consideration in a criminal investigation. One of them is through selection from a database. The compilation of DNA databases could enable police forces to collect samples taken during investigations of unsolved criminal cases, as well as samples from convicted felons, in order that such stored information could be used to select a person of interest in a way similar to the collection and storage of fingerprint records.

A debate appeared around the question 'should the fact that the person of interest was selected through a database search affect the value of the evidence?' Confusion surrounding the evaluation of the outcome of such a search can arise because the probability of a

match increases as the database gets larger. Robertson and Vignaux (1995a) explain this confusion by stating that '[i]t is commonly claimed that the evidential value of a match, when a POI is selected through a search in a database, is affected by the number of comparisons one has made. Certainly, the larger the database the more likely we are to find a match' (at p. 122). This leads to the erroneous conclusion that the larger the database the weaker the evidence. This erroneous approach was proposed in a report of the National Research Council of the United States (National Research Council, 1996) which published the recommendation:

> When the suspect is found by a search of DNA databases, the random match probability should be multiplied by N, the number of persons in the database. (Recommendation 5.1 at p. 40)

It has been shown that the application of such a recommendation produces illogical results with a drastic dilution of the strength of the evidence. Balding and Donnelly (1996) and Evett and Weir (1998) showed that the likelihood ratio is higher following a search than in a case where the size of the potential criminal population is known and no sequential searches have been performed. Each person who does not match the DNA profile of the recovered trace is excluded. The exclusion of these individuals from the pool of the potential culprits increases the probability of involvement of the individual who matches. The argument was developed further in Dawid and Mortera (1996) and potential solutions were given in Dawid (2001), Balding (2002) and Kaye (2009).

1.8.2 Profile Probability vs Conditional Profile Probability

Imagine the likelihood ratio calculation in a common situation involving a single biological stain where there is the DNA profile E_r (r for 'recovered' as the origin of the sample is not known) of the crime sample and the profile E_c (c for 'control' as the origin of the sample is known) of a POI. Let I represent the background information and let the propositions be

H_p: the POI is the source of the stain,

H_d: another person, unrelated to the POI, is its source; *i.e.*, the POI is not the source of the stain.

Both profiles are of genotype A, say. The likelihood ratio can then be expressed as

$$\frac{Pr(E_r = A \mid E_c = A, H_p, I)}{Pr(E_r = A \mid E_c = A, H_d, I).}$$

Assume that the DNA typing system is sufficiently reliable that two samples from the same person will be found to match when the POI is the donor of the stain (proposition H_p), and that there are no false negatives. If it is known that the POI is of type A and if H_p is assumed true then it follows that the recovered sample is of type A and $Pr(E_c = A \mid E_p = A, H_p, I) = 1$.

It is widely assumed that the DNA profiles from two different people (the POI and the donor of the stain when proposition H_d is true) are independent. Then $Pr(E_c = A \mid E_p = A, H_d, I) = Pr(E_c = A \mid I)$. In such a case only the so-called *profile probability*, γ_A, with which an unknown person would have the profile A is needed. This is a widely accepted over-simplification. In reality, the evidential value of a match between the profile of the

recovered sample and that of the POI needs to take into account the fact that there is a person (the POI) who has already been seen to have that profile (type A). So, the probability of interest is $Pr(E_r = A \mid E_c = A, H_d, I)$ and this can be different from $Pr(E_r = A \mid I)$. In fact, observing one genotype in the population increases the chance of observing another of the same type. Hence, within a population, DNA profiles with matching allele types are more common than suggested by the independence assumption, even when two individuals are not directly related. The conditional probability (also called *conditional profile probability* or *conditional match probability*) incorporates the effect of population structure and other dependencies between individuals. The more common source of dependency is a result of a membership in the same population and having similar evolutionary histories. Populations are finite in size thus two people taken at random from a population have a non-zero chance of having relatively recent common ancestors. Disregarding this correlation of alleles in the calculation of the value of the evidence results in an exaggeration of the strength of the evidence against the compared person. This aspect was presented by Balding and Nichols (1994, 1995) and mainly supported by Weir (1996), Curran et al. (2003), Buckleton and Triggs (2005) and Curran and Buckleton (2007).

1.8.3 Evaluation by Taking Errors into Account

The evaluation of scientific evidence has to consider the role of error. Error has been mentioned by many scholars in scientific and legal literature (i.e., Koehler, 1996, 1997 and reiterated Koehler, 2018), including in the PCAST report to the US President on 'Forensic science in criminal courts; ensuring scientific validity of feature-comparison methods' (President's Council of Advisors on Science and Technology PCAST, 2016), that states:

> Without appropriate estimate of accuracy, an examiner's statement that two samples are similar - or even indistinguishable - is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing - not training, personal experience nor professional practices - can substitute for adequate empirical demonstration of accuracy. (at p. 46)

Therefore, when evaluating the strength of DNA evidence for supporting that two samples have a common source, one must consider two factors. One factor is the conditional profile probability. A coincidental match occurs when two different people share the same DNA profile. The second factor is the probability of a false positive. A false positive occurs when a laboratory erroneously reports a DNA correspondence between two samples that actually have different profiles. A false positive may occur due to error in the collection or handling of samples, misinterpretation of test results, or incorrect reporting of test results (Thompson, 1995). Either a coincidental match or a false positive could cause a forensic scientist to report a DNA match between samples from different individuals. Thus, the conditional profile probability and the false positive probability should both be considered in order to make a fair evaluation of the evidence (Koehler et al., 1995). Proficiency testing performances do not necessarily equate to the false positive probability in a particular case. This aspect represents a first practical difficulty. A second practical difficulty is the presentation of a logical framework which takes account of both probabilities. Various suggestions have been made (Robertson and Vignaux, 1995b; Balding and Donnelly, 1995; Balding, 2000). A likelihood ratio framework for considering the role that error may play in determining the value of forensic DNA evidence in a particular case is presented in Thompson et al. (2003) and Buckleton et al. (2005). Even a small false positive probability can, in some circumstances, be highly influential so serious consideration has to be

given to its estimation. Recognition is needed that accurate assignments for false positive probabilities can be crucial for the assessment of the value of DNA evidence.

1.9 A Minimum Value for the Profile Probability

At the end of the 1990s, the increasing use of DNA evidence put forward new questions. One of those that is still of importance was 'What figure should be presented in court if the aim is to provide the judge and jury with the best numerical measure of the actual probability that the defendant's DNA profile match is accidental?' This is purely a quantitative question.

It was (and still is) common that experts quoted astronomical figures such as 1 in ten billion or trillion and insisted that such figures were both accurate and reliable. It can be agreed that anyone who makes such a statement in Court is either numerically naïve or is deliberately trying to deceive a (possibly numerically unsophisticated) judge and jury because such figures are well beyond the accuracy of human science and technology for several reasons. Scientific and legal literature critically discussed this point giving the practical range of values for a profile probability when a variety of specified alternatives (possible sources of the stain other than the POI), corresponding to individuals who exhibit different degrees of relatedness to the POI, presented full matching profiles. Simulations have been used to generate such values (Foreman et al., 1997; Foreman and Evett, 2001). At that time, using the most common DNA profiling system (10-locus Short Tandem Repeat, or STR, profiling system), the authors recommended general conditional match probability values for use when reporting full profile matches. They supported the use of a value of 1 in a billion (10^{-9}) when an alternative individual (unrelated and coming from the same sub-population of the real donor of the recovered stain) is considered. A general discussion of this topic was clearly presented in Evett et al. (2000a). The same minimum value was also given by Hopwood et al. (2012) using 15-plex STR profiling systems. The use of values more extreme than 10^{-9} for a profile probability or conditional match probability is widely and constantly criticized (see, for example, Kaye, 1993, Lambert et al., 1995, Saks and Koehler, 2008 and Curran, 2010). The main reason for such a criticism is expressed by Hopwood et al. (2012) who noticed that

> Such values [values lesser than 10^{-9}] invoke independence assumptions to a scale of robustness that we cannot demonstrate empirically, given the size of available databases. [...] In addition to the empirical evidence for the reliability of DNA evidence interpretation, we recognise also that such numbers are difficult to conceptualise and require unreasonable real life comparisons. (p. 188)

A related practical problem with which forensic scientists are faced is the estimate of genotype proportions. A method was outlined by Balding (1995) taking advantage of Bayesian statistical methods that was able to deal with situations where the genotype may not appear in the sample at a given locus. A development by taking account of sampling error using what is called the size-bias correction was proposed by Curran et al. (2002). Work is still in progress at the time of writing to deal with other forms of DNA evidence such as mt-DNA or Y-Chromosome (Buckleton et al., 2016a).

1.10 Propositions and Pre-Assessment

In the late 1990s increasing recognition of different aspects, or levels, of forensic examinations and reports of their outcomes, led to series of important papers that introduced a framework known as 'Case Assessment and Interpretation'.

1.10.1 The Choice of Propositions

For an inferential process to be balanced, or in the words of some authors, impartial (Jackson, 2000), attention cannot be restricted to only one side of an argument. Evett (1996) noted, for instance, that

> a scientist cannot speculate about the truth of a proposition without considering at least one alternative proposition. Indeed, an interpretation is without meaning unless the scientist clearly states the alternatives he has considered. (at p. 122)

The requirement for the consideration of alternative propositions is a general one that applies equally in many instances of daily life, but in legal contexts its role is fundamental. This criterion requires a scientist to consider at least two competing propositions. The exact phrasing of propositions is important, an importance that underlies a concept known in the context as the level of propositions or *hierarchy of propositions* (Cook et al., 1998a). The reasons for this are twofold. Firstly, a proposition's content crucially affects the degree to which that proposition is helpful for the courts. For example, the pair of propositions 'the suspect (some other person) is the source of the crime stain' (known in this context as a *source-level* proposition) addresses a potential link between an item of evidence and an individual (that is, a suspect) on a rather general level. Generally, *activity-level* (e.g., 'the suspect (some other person) attacked the victim') or *offence-level* (e.g., 'the suspect (some other person) is the offender') propositions tend to meet a court's need more closely. Secondly, the level of proposition defines the extent of circumstantial information that is needed to address a proposition meaningfully. For example, when reasoning from a source- to an activity-level proposition, phenomena such as transfer of material during the alleged or alternative actions should be taken into account. Under an offence-level pair of propositions, consideration needs to be given to the relevance of a crime stain, in other words, whether or not it has been left by the offender (Stoney, 1991a, 1994). The concept of relevance is not necessarily needed when attention is confined to a source-level or an activity-level proposition.

The scientist should insist on a well-defined framework of propositions. This is in sharp contrast to the opinion that data should be allowed to 'speak for themselves', a suggestion that evidential value represents some sort of attribute that is intrinsic to data and independent of circumstances. Such an opinion is viewed cautiously in legal reasoning, where the following position has been reached:

> In court, as elsewhere, the data cannot 'speak for itself'. It has to be interpreted in the light of the competing hypotheses put forward and against a background of knowledge and experience about the world. (Robertson and Vignaux, 1993a, p. 470)

As may be seen, the concept of levels for propositions is important because it is closely tied to the notion of evidential value. The latter is a subjective function of the former, in the sense that the value assigned to evidence by a particular individual depends on (a) the propositions amongst which the individual seeks to discriminate and (b) contextual

information that is available to the individual. Evidential value is to be seen neither as an abstract property of the external world nor as one that can be elicited in a uniquely defined way (Evett et al., 2000b).

1.10.2 The Pre-Assessment

An evaluation process starts when the scientist first meets the case. It is at this stage that the scientist thinks about the questions that are to be addressed and the outcomes that may be expected. The scientist should attempt to frame propositions of interest and think about potential outcomes and the value of evidence that is expected (Evett et al., 2000c). There is a wide tendency to consider evaluation of evidence as one of the last steps of a casework examination, notably at the time of preparing the formal report. This is so even if an earlier interest in the process would enable the scientist to make better decisions about the allocation of resources. A first approach to decision-making in an operational forensic science problem has been proposed by Cook et al. (1998b). It is based on a model embodying the principle of likelihood ratio as a measure of the value of evidence. In that spirit, Cook et al. (1998b) proposed a model for enhancing the cost-effectiveness of a casework activity from initial contact with the customer. The aim is to enable the customer to make better decisions. In routine work, an assignment of the expected likelihood ratio is often requested by forensic laboratories, before, for example, the performance of any analytical tests. Such an assignment will help the scientist to support a better decision for the customer. Imagine, for the sake of illustration, a situation in paternity testing where the alleged father is unavailable but a cousin of the alleged father could be considered and tested. In such a case, the two propositions of interest may be of the form of H_p: the tested person is a cousin of the true father, and H_d: the tested person is unrelated to the child. Two questions are of interest: (1) can we obtain a value supporting the hypothesis H_p or H_d in this scenario?, and (2) how can the laboratory or the customer take a rational decision on the necessity to perform tests after an assignment of possible values of likelihood ratio? The first question refers to the pre-assessment process, the second to decision-making (Section 1.15). Answers to the first question are proposed in Cook et al. (1998b, 1999) and Evett et al. (2000c). A review of case assessment is given in Jackson et al. (2015).

1.11 Translation of a Numerical Value into a Verbal Equivalent

A verbal scale for a numerical ratio of probabilities in the context of hypothesis testing was discussed by Jeffreys (1983) (first edition in 1939). The ratio, denoted as K by Jeffreys, is that of the probability of the null hypothesis given evidence and background information to the probability of the alternative hypothesis given evidence and background information. Jeffreys takes the prior probabilities of the two hypotheses to be equal so that the posterior odds equals the likelihood ratio. The verbal summary is then phrased in the form of support provided by the evidence against the null hypothesis. Jeffreys comments that K need not be known with much accuracy. If $K > 1$ the null hypothesis is supported. Jeffreys further comments that interest is with the values of $K < 1$ when the null hypothesis may be rejected. A logarithmic scale with so-called grades and the associated verbal descriptors proposed by Jeffreys is given in Table 1.1.

TABLE 1.1

Verbal Scale of Support K for a Null Hypothesis Proposed by Jeffreys (1983) Where K is the Ratio of the Probability of the Null Hypothesis Given Evidence and Background Information to the Probability of the Alternative Hypothesis Given Evidence and Background Information

Grade	Value of K	Verbal Descriptor
0	$K > 1$	Null hypothesis *NH* supported
1	$1 > K > 10^{-1/2}$	Evidence against *NH* but not worth more than a bare mention
2	$10^{-1/2} > K > 10^{-1}$	Evidence against *NH* substantial
3	$10^{-1} > K > 10^{-3/2}$	Evidence against *NH* strong
4	$10^{-3/2} > K > 10^{-2}$	Evidence against *NH* very strong
5	$10^{-2} > K$	Evidence against *NH* decisive

TABLE 1.2

Verbal Scale of Support for a Likelihood Ratio Proposed by Nordgaard et al. (2012) and Nordgaard and Rasmusson (2012)

Scale	Interval of Likelihood Ratio LR	Degree of Support
+4	$10^6 \leq LR$	extremely strong support
+3	$6000 \leq LR < 10^6$	strong support
+2	$100 \leq LR < 6000$	support
+1	$6 \leq LR < 100$	support to some extent
0	$1/6 \leq LR < 6$	support neither . . . nor
−1	$1/100 \leq LR < 1/6$	support to some extent
−2	$1/6000 < LR \leq 1/100$	support
−3	$1/10^6 < LR \leq 1/6000$	strong support
−4	$LR \leq 1/10^6$	extremely strong support

Another approach is to use an ordinal scale for the likelihood ratio. In forensic science such a procedure was introduced by Evett (1991) for ease of communication. A similar scale was proposed in Nordgaard et al. (2012) and Nordgaard and Rasmusson (2012) with a response for the likelihood ratio close to 1 and then a 4-point scale for the likelihood ratio > 1 with its reciprocal for the likelihood ratio < 0 and is given in Table 1.2. The choice of a verbal scale is initially subjective but not arbitrary. However, if the scale is to have credibility and be acceptable to the courts then a particular choice has to be agreed amongst the scientists and ideally published in a peer-reviewed journal.

A 6-point verbal scale for values of the likelihood ratio greater than 1, reciprocated for values of the likelihood ratio less than 1 is given as an illustration in ENFSI (2015) with six adjectives for support of weak, moderate, moderately strong, strong, very strong and extremely strong and corresponding numerical ranges for the logarithm of the likelihood ratio of $\{0-1, 1-2, 2-3, 3-4, 4-6, >6\}$.

Note that while it is permissible to interpret a numerical value verbally, it is not meaningful to interpret a verbal scale numerically. Also, there are still several aspects of the use of verbal scales to be considered. First, there is the nature of the assistance that a verbal scale might offer to the fact-finder (judge or jury). The second is whether the numerical value of a likelihood ratio is a sufficient summary of the value of the evidence. Thirdly, the

limitations of the use of verbal scales need to be recognised. A discussion of the disadvantages of verbal scales is to be found in Marquis et al. (2016). At the time of writing the topic is still under discussion (Berger and Stoel, 2018).

1.12 Assessment of Performance

The development of models for the evaluation of evidence led inevitably to the development of measures for the assessment of their performance. An assessment of their performance enabled the scientist to justify their choice of model and hence their value for the evidence. Measures of performance are generic. They are applied to the general performance of the model with measures obtained from datasets where the correct answer is known. It is not possible to assess a performance in a particular case since the correct answer is not known. The quality of a model is defined as the ability of the model to support the correct result.

The choice of a model and the assessment of the performance of a model require the existence of at least one and preferably two datasets in which the correct answers are known. The first dataset is used as a training set to determine the best model for the data and for estimation of parameters, if any, in the chosen model. The second dataset is used as a validation set to assess the performance of the chosen model and parameter values. In the absence of two datasets then the training set can be used as the validation set with due care to allow for any bias in the results that may arise from this double use.

Consider source propositions with two sets of trace evidence, one set nominally from a known source and one set nominally from an unknown source. The prosecution proposition H_p is that the unknown source is the known source. The defence proposition H_d is that the unknown source is a different source to the known source. An example of such trace evidence is that of window glass evidence, with measurements of refractive index and elemental compositions from fragments within the same windows and with many windows that provide the opportunity to compare measurements from fragments from different windows. Comparisons may then be made of measurements taken from the validation dataset on two sets of different fragments from within the same window (same-source comparisons) and on two sets of fragments from between different windows (the fragments then of necessity being different and the comparisons being of different-source). For each comparison one set may be chosen in the model as the one of known source and the other may be chosen in the model as the one of unknown source (even though its source is actually known). As the source (window) of each set is known, the correct proposition in any comparison, H_p, same-source, or H_d, different-source, is known. For each comparison, a likelihood ratio using a model determined from the training set, is calculated. A value of the LR greater than 1 is said to support the prosecution proposition H_p. A value of the LR less than 1 is said to support the defence proposition H_d. However, unlike a court case where it is not known which proposition is correct, the correct answer is known. Performance of the chosen model can then be assessed with a comparison of the results (supports) with the type of comparison (same-source or different-source) which had been made.

Several measures of performance have been developed.

- *Rates of misleading evidence:* these are
  (a) for comparisons, there are two rates, first, the number of same-source comparisons with $LR < 1$ divided by the total number of same-source comparisons and, second, the number of different-source comparisons with $LR > 1$ divided by the total number of different-source comparisons; if a conclusion that the two items of evidence being compared come from the same source is considered a positive result and if a conclusion that the two items of evidence being compared come from different sources is considered a negative result then a same-source comparison with $LR < 1$ may be thought of as a *false negatives* and a different-source comparison with $LR > 1$ may be thought of as a *false positive*;
  (b) for discrimination, the number of members of the validation set that are allocated to the wrong group, divided by the total number of allocations. For discrimination between two groups, A and B say, two rates can be determined, first, the number of members of A allocated to B divided by the total number of members of A and, second, the number of members of B allocated to A divided by the total number of members of B. If there are more than two groups, various possible combinations of rates may be calculated.

  Histograms of the values of $\log(LR)$ from all possible same-source comparisons from the validation dataset and, separately, of the values of $\log(LR)$ from all possible different-source comparisons from the validation dataset can be drawn on the same axes. The quality of a method at a particular value of $\log(LR)$ is the amount of overlap of the histograms at that value. Ideally, all values of $\log(LR)$ for same-source comparisons will be greater than 0, all values of $\log(LR)$ for different-source comparisons will be less than 0 and there will be no overlap. An analogous use of histograms may be made for discrimination.
- *Tippett plots:* these are generalisations of rates of misleading evidence for comparisons (Evett and Buckleton, 1996; Tippett et al., 1968). They are the complement of empirical cumulative distribution functions for same-source and different-source comparisons. The plots come in pairs, one for same-source comparisons and one for different-source comparisons. The $\log(LR)$ is plotted on the x-axis and, for a particular value x_0 of the $\log(LR)$, the y-axis is the relative frequency of the number of comparisons greater than x_0. For same-source comparisons, it is to be hoped that all $\log(LR)$ values are greater than 0. Thus for $x < 0$, it is hoped the corresponding value on the y-axis will be 1 (or 100%). Similarly, for different-source comparisons, it is to be hoped that all $\log(LR)$ values are less than 0. Thus for $x > 0$, it is hoped the corresponding value on the y-axis will be 0 (or 0%).

  The distance from the intersection of the same-source plot with the line $\log(LR) = 0$ and the line $y = 1(100\%)$ is the rate of misleading evidence for same-source comparisons, the proportion of same-source comparisons that have a value of $\log(LR) < 0$ ($LR < 1$). The distance from the intersection of the different-source plot with the line $\log(LR) = 0$ and the line $y = 0(0\%)$ is the rate of misleading evidence for different-source comparisons, the proportion of different-source comparisons that have a value of $\log(LR) > 0$ ($LR > 1$).

- *Empirical cross-entropy:* The measure used here is known as a *score* and the definition is such that low scores are good. In a comparison of the performance of two models, the model with the lower score is deemed to be the better model. A quadratic scoring rule was used by Lindley (1991) to justify the use of probability as the only measure of uncertainty. The scoring rule used for evaluation of evidence is the logarithmic rule (Good, 1952).

  In the context of forensic science, consider the prosecution and defence propositions H_p and H_d, respectively, and in this context, assume $\Pr(H_p) = 1 - \Pr(H_d)$. For evidence evaluation, the logarithmic rule with base 2 is used for reasons associated with information theory where the common unit of information is the bit. Given a particular model, let p be the posterior probability obtained for H_p given evidence E and background information I. Then $(1 - p)$ is the posterior probability for H_d given evidence E and background information I. The logarithmic scoring rule states that

  (a) If H_p is true, score $-\log_2 p = -\log_2 \Pr(H_p \mid E, I)$;

  (b) If H_d is true, score $-\log_2(1 - p) = -\log_2 \Pr(H_d \mid E, I)$.

  If p is high and H_p is true then the score is low. If p is high and H_d is true then the score is high. For example, consider $p = 0.9$ and H_p true; the score is $-\log_2(0.9) = +0.15$. If H_d true; the score is $-\log_2(1 - 0.9) = +3.32$*

  The measure of performance for evidence evaluation is then a weighted average value of the logarithmic scoring rule, and is known as the *empirical cross-entropy* (*ECE*) empirical cross-entropy:

$$
\begin{aligned}
ECE = &-\frac{\Pr(\theta_p \mid I)}{N_p} \sum_{i \in \text{true } \theta_p} \log_2 \Pr(\theta_p \mid E_i, I) \\
&- \frac{\Pr(\theta_d \mid I)}{N_d} \sum_{j \in \text{true } \theta_d} \log_2 \Pr(\theta_d \mid E_j, I) \\
= &\frac{\Pr(\theta_p \mid I)}{N_p} \sum_{i \in \text{true } \theta_p} \log_2 \left(1 + \frac{1}{LR_i \times O(\theta_p)}\right) \\
&+ \frac{\Pr(\theta_d \mid I)}{N_d} \sum_{j \in \text{true } \theta_d} \log_2 \left(1 + LR_j \times O(\theta_p)\right),
\end{aligned}
$$

  where E_i and E_j denote the evidence and LR_i and LR_j denote the corresponding LR values in the training set (or validation set if one exists) with N_p members when θ_p is true and N_d members when θ_d is true and $O(\theta_p)$ denotes the prior odds $Pr(H_p \mid I) = Pr(H_d \mid I)$. (Meuwly et al., 2017) and (Ramos and Gonzalez-Rodriguez, 2013).

  This measure indicates better performance when the likelihood ratio leads to the correct decision. The numerical value will be lower as the performance increases. The value of the ECE can be plotted showing its value for a certain range of priors.

* Note for calculation purposes, $\log_2 x = \log_{10}(x)/\log_{10}(2)$.

1.13 Role for Likelihood Ratio as a Measure for Investigation as Well as for Evaluation

The description of the role of the likelihood ratio for the evaluation of evidence in the form of continuous measurements by Lindley (1977) led to an upsurge of interest in the area amongst statisticians and forensic scientists. Another role for the likelihood ratio in forensic science was introduced in the late 1990s by Cook et al. (1998b). This role was to provide assistance to the police in the investigation of a crime. When the scientist was presenting evidence in court, their role was that of evidence evaluation and the mode of operation was said to be *evaluative*. When the scientist was assisting police in an investigation, their role was said to be *investigative* (Jackson et al., 2015).

The investigative role of the scientist is part of the procedure known as *case assessment and interpretation* (CAI), much of which is outlined in Section 1.10. The CAI procedure has clarified considerably the information that is required by investigators to aid the provision of the characteristics of balance, logic, transparency and robustness (Association of Forensic Science Providers, 2009).

The evaluative role of the scientist provides a measure of support, either numerically or verbally (see Section 1.11), which is in the form of a likelihood ratio. The numerator and denominator of the likelihood ratio are the probabilities of the evidence given the prosecution and defence propositions, respectively. The scientist offers no opinion on the probabilities of these propositions; opinion on these is the role of the fact-finder be they judge or jury.

In contrast, the investigative role of the scientist is to aid in the investigation of a crime. The aid is provided with the use of abductive reasoning to generate explanations for what is being discovered during the course of the investigation. Abduction as defined by Jackson et al. (2015) is the intellectual and imaginative process of generating possible explanations to account for an expert's actual or anticipated scientific observations. Explanations generated by abduction can then be tested against observed data. Note that the word 'explanation' is used during an investigation to describe particular hypotheses for what may have happened to produce what is discovered during the investigation. The word 'proposition' is used during the evaluative stage of the process, for example for the interpretation in court. The propositions are the hypotheses put forward by the prosecution and defence.

During the investigative phase it is permissible for a scientist to provide what may be called posterior probabilities for explanations. A likelihood ratio for evaluation is best used with two and only two mutually exclusive propositions. With more than two propositions it is necessary to include prior probabilities for the propositions to obtain an evaluation. For investigations, the scientist can use prior probabilities for explanations, even if there are only two. Only relative values for the posterior probabilities will be obtained. It is unlikely a scientist can be certain that the choice of explanations is not only mutually exclusive but also exhaustive. Overall, the purpose of the investigative use of the likelihood ratio is to reduce uncertainty about events material to the investigation and thus to help direct the investigation.

Sometimes it may be possible to assess the potential evidential value or investigative impact of a test with the use of experimental data in which the test has been used with known results. If the value or impact is high then the investigators may deem it worthwhile to conduct the test. An example is described in Jackson et al. (2015) based on a German

case reported by Oesterhelweg et al. (2008). The case involved the use of a cadaver dog. Data were available from experiments with cadaver dogs from which it was possible to provide estimates of the probabilities of a positive (+) signal from the dog if a cadaver scent were or were not present (H_p or H_d) and the probabilities of a negative (−) signal from the dog if a cadaver scent were or were not present (H_p or H_d), using the appropriate proportions in the resulting table. The corresponding likelihood ratios, $Pr(+ \mid H_p)/Pr(+ \mid H_d)$ and $Pr(- \mid H_p)/Pr(- \mid H_d)$ can be calculated, for the value of a positive or negative result. The investigator can then decide whether to use the test or not, given the relative values of the likelihood ratios.

The investigative role of the likelihood ratio has also been used for the examination of questioned documents, notably when a person of interest is not available. The evidence is described by trace characteristics only. The forensic scientist can assign conditional probabilities under the assumptions of pairs of explanations. For example, these might be that the writer was male or was female, or that the writer was left-handed or was right-handed. The likelihood ratios may then be calculated for each of these pairs of explanations and the results used to inform the investigation (Taroni et al., 2012).

1.14 Probabilistic Graphical Models

Starting from 1989, a series of papers showed that intricate frameworks of circumstances - situations involving many variables - require a logical assistance and should be approached in a formal way, pointing out the utility of graphical methods that deal with an analysis of rational thinking.

1.14.1 Bayesian Networks

Methods of formal reasoning to assist forensic scientists to understand better the various dependencies which may exist among different aspects of evidence have been developed. Probabilistic graphical models, in particular Bayesian networks (or, Bayes nets for short), emerged from such research as an important approach, capable of providing a valuable aid for representing relationships among target characteristics and propositions (hypotheses) in situations of uncertainty.

They can assist the user not only in describing challenging practical problems, and communicating information about their structural properties, but also in actually computing the effect of knowing the truth or otherwise of one proposition, or one item of evidence, on the probability of other propositions - avoiding possibly difficult algebraic case-by-case calculations. In addition, Bayesian networks have the potential of clarifying the rationale behind particular probabilistic solutions.

Based on the elements of graph and probability theory, Bayesian networks can roughly be defined as a pictorial representation of the dependencies and influences (represented by arcs) among variables (represented by nodes) deemed to be relevant for a particular probabilistic inference problem.

1.14.2 Bayesian Networks to Manage 'Masses' of Evidence

The advances in formalization and computational support for rational thinking are highly valuable because they contribute to the coherent use of forensic information in the legal

process. However, at their current level of development, probabilistic approaches still focus essentially on single items of evidence. Difficulties with the combination of evidence have been discussed under the concept of conjunction; see Cohen (1977, 1988) and Dawid (1987). The inability to evaluate more than a few items of evidence is currently felt as a major limitation, as already noticed some time ago by Schum (1994):

> What is clear is that no probabilist had ever given attention to the task of weighing entire masses of evidence given at trial. As Wigmore noted (1937 at p. 9) 'The logicians have furnished us in plenty with canons of reasoning for specific single inferences; but for a total mass of contentious evidence in judicial trials, they have offered no system'. (at p. 61)

Lindley (2004) reiterated this point when he wrote that a

> [. . .] problem that arises in courtroom, affecting both lawyers, witnesses and jurors, is that several pieces of evidence have to be put together before a reasoned judgement can be reached. [. . .] probability is designed to effect such combinations but the accumulation of simple rules can produce complicated procedures. Methods of handling sets of evidence have been developed: for example, Bayes nets. (at p. xxiv)

'Bayesian networks provide a solution.'

1.14.3 Bayesian Networks in Judicial Contexts

The study of representational schemes for assisting reasoning about evidence in legal settings has a remarkably long history. In this context the charting method developed by Wigmore (1913) is a frequently referenced (though essentially deterministic) predecessor of modern network approaches to inference and decision analyses that can be traced back to the beginning of the 20th century. It is only about three decades ago that researchers have begun to show an interest in graphical approaches with the valid incorporation of probability theory. Examples include decision trees and a modified, more compact version of these, called 'route diagrams' (Friedman, 1986a,b). Since the early 1990s, however, it is Bayesian networks that have advanced to a preferred formalism among researchers and practitioners engaged in the joint study of probability and evidence in judicial contexts, notably because of the efficient representational capacity of computer systems to handle multiple pieces of information. Thus, researchers in law – compared to those in other domains of applications - were among the pioneers who realized the practical potential of Bayesian networks. In judicial contexts, two different ways in which Bayesian networks are used as a modelling technique can be distinguished. Legal scholars focus on Bayesian networks as a means for structuring cases as a whole whereas forensic scientists concentrate primarily on the evaluation of selected issues that pertain to scientific evidence (Robertson and Vignaux, 1993b). Many studies with an emphasis on legal applications thus rely on Bayesian networks as a method for the retrospective analysis of complex and historically important *causes celèbres*, such as the Collins case (Edwards, 1991), the Sacco and Vanzetti case (Kadane and Schum, 1996), the Omar Raddad case (Levitt and Blackmond Laskey, 2001) or the O.J. Simpson case (Thagart, 2003).

1.14.4 Bayesian Networks in Forensic Science: Particular Case Modeling

The first study in print on Bayesian networks applied to forensic case settings was published in the late 1980s (Aitken and Gammerman, 1989). It focused on a hypothetical murder scenario where the authors showed how a network approach might be applied to

cases involving several, possibly complicated, interrelated issues. They provided a detailed discussion on how (i) relevant propositions can be extracted from a scenario, (ii) relationships between propositions are represented qualitatively in terms of a directed graph, and (iii) subjective beliefs are incorporated as probabilities and used for inference. The authors noticed that

> [a] probabilistic reasoning system has been developed and implemented [. . .] The ideas behind the current development of the system have an obvious application in the assessment of evidence which is illustrated with an artificial example in the criminal legal field [. . .].

This paper (and a second one focused on specific case analysis (offender profiling) Aitken et al., 1996) have opened up new horizons in law and forensic science.

1.14.5 Bayesian Networks in Forensic Science: Generic Patterns of Inference

Instead of focusing on a particular scenario, as outlined in the previous section, it is also possible to pursue a modelling approach that aims at a standard analysis of recurrent patterns of inference concerning scientific evidence. This perspective concentrates on some more generic and fundamental issues that forensic scientists should account for if they seek to evaluate their evidence in the light of propositions that are of judicial interest. The modelling concentrates on aspects of case settings that determine the general pattern of inference (e.g., the relevance of evidence (Garbolino and Taroni, 2002)), irrespective of details about a particular situation.

Many inferential problems in the analysis of DNA profiling were solved by Dawid et al. (1999, 2002). Within the branch of DNA evidence, an extensive body of knowledge (accepted biological theory) is available and on which one can rely during network construction. For example, consideration of Mendelian laws of inheritance allows one to obtain clear indications on how nodes in a network ought to be combined. In this way, basic submodels have been proposed and repeatedly used for logically structuring larger networks. An extension to deal with an important category that covers studies focusing on small quantities of DNA has been discussed by Evett et al. (2002). Finally, a hierarchical approach, notably where analyses lead to large network topologies (e.g., when information pertaining to different genetic markers needs to be combined), has been proposed in Dawid (2003) and Mortera et al. (2003).

Forensic applications of Bayesian networks range from offender profiling to single and complex configurations of different kinds of trace evidence. A detailed collection of models is available in (Taroni et al., 2014). Bayesian networks allow their users to engage in probabilistic analyses of much higher complexity than what would be possible through traditional approaches that mostly rely on rather rigid, purely arithmetic developments. The graphical nature of Bayesian networks facilitates the formal discussion and clarification of probabilistic arguments.

1.15 Not Only Inference: The Way to Make a Decision

Courts typically seek to reduce the uncertainty of their knowledge about a defendant's true connection with a criminal act. Often, part of this search is based on the evidence

offered by forensic scientists. According to this view, inference provides contributory information to judicial decision-making (e.g., the decision as to whether a defendant should be found guilty of the offence of which they have been charged). Assessment of this contributory information reflects the intention to promote accurate decision-making. This aspect to judicial decision-making was recognised in the legal literature in a seminal paper published by Kaplan (1968). Later, Kaye (1979), Fienberg and Schervish (1986) and Kaye (1988) described the decision-making process, a process that plays a key role in everyday routine life. This process consists of the rational choice, given personal objectives, between two or more possible outcomes when the consequences of the choice are uncertain. Decision analysis helps individuals better to understand the problem they are faced with and to make clearer and more consistent decisions. The approach has also been applied in forensic science to deal with situations where decisions are required (Taroni et al., 2005; Biedermann et al., 2008). These situations include the identification process, earlier discussed by Stoney (1991b), later by Champod (2000), then further explored by Cole (2009) and supported by Champod et al. (2016).

1.15.1 The Objectives and Ingredients of Decision Theory

Given a set of beliefs about states of the world which cannot be known, the general objective is to identify an available course of action that is logically consistent with a person's personal preferences for consequences. This is an expression of a view according to which one decides on the basis of essentially two ingredients: one's beliefs about past, present or future happenings and, secondly, one's valuation of the consequences. The former ingredient will be expressed by probability and the latter ingredient will be captured by invoking an additional concept, known as *utility*. Both concepts can operate within a general theory of decision that involves the practical rule which says that one should select that decision which has the highest expected utility (or, alternatively, which minimizes expected loss). When the class of such operations is based on beliefs that have been informed with Bayesian updating (statistical inference), then this process is called Bayesian decision analysis.

A decision-based approach can help (i) to clarify the fundamental differences between the value of evidence as reported by an expert and the final decision that is to be reached by a jurist, and (ii) to provide a means to show a way ahead as of how these two distinct roles (evaluation and decision) can be conceptualized to interface neatly with each other. These are topics that are unfortunately viewed differently rather than in a unified manner as already suggested by De Finetti (1968):

> Probabilities are chiefly tools for inference (induction) and for decisions (under conditions of uncertainty) [. . .] The subjectivistic approach is simple and natural (it seems common-sense): every new information leads to the inference of a new distribution of probabilities from the old one (according to Bayes' theorem) and so we have at any moment a probability distribution which gives the basis (i.e., the weights for the expected utility to be maximized) for every decision including that of deferring the final decision in order to collect previously, in any specific way, useful additional information. Inference and decisions are thus (as they must be) logically independent problems, only related by the output of the first serving as input for the latter. (at p. 48)

1.15.2 Graphical Models

An extension of Bayesian Networks provides the scientists with an aid to support decision-making as illustrated in Taroni et al. (2014) and Gittelson (2013). The addition of an explicit

representation of the decisions under consideration and the value (utility) of the resulting outcomes (the states that may result from a decision) leads to Bayesian decision networks, also called influence diagrams. These networks combine probabilistic reasoning with the utility theory to make decisions using the criteria of maximizing the expected utility. Therefore, an influence diagram consists of three types of nodes: (1) the chance nodes which represent random variables (as in Bayesian networks); (2) the decision nodes with the states of a decision node being the different actions that are the outcomes of the various decisions amongst which the decision-maker must choose; and (3) the utility nodes which represent the decision-maker's utility function. They are characterized by utility tables specified for every outcome.

1.16 The Existence or Otherwise of a True Value of the Evidence

Recent (late 2010s) debate has concerned the inclusion, or otherwise, of a measure of uncertainty for a statement about the value of evidence. The best measure of the value of evidence is the likelihood ratio. The determination of a value for the likelihood ratio requires choices by the analyst of (a) the model to be used and, if a parametric model, (b) the the parameters of the model, and (c) the dataset for training purposes for the model and the parameters. A Bayesian approach to evaluation will also involve the choice of a prior distribution and associated hyperparameters. The argument in favour of the provision of a measure of uncertainty is that all these choices suggest possible variability in the value ultimately calculated. This variability should then be represented in a summary of the value of the evidence. For example, such a summary might be a point estimate and a lower confidence bound (e.g., 95%) on this estimate in order to favour the defence. A series of papers in *Law, Probability and Risk* (Nordgaard, 2016; Sjerps et al., 2016; Taroni et al., 2016a,b) and in *Science and Justice* (Morrison, 2016) debate the issue.

There are inferential difficulties with the incorporation of a measure of uncertainty in the summary of the value of the evidence. Determination of the value of the evidence already incorporates the uncertainty in model choice, parameter choice and choice of training data. The addition of a another layer of uncertainty is akin to asking for a probability on a probability: a person is uncertain about the occurrence of an event and then uncertain about their uncertainty.

> It is nonsense for you to have a belief about your belief if only because to do so leads to an infinite regress of beliefs about beliefs about beliefs . . . (Lindley, 2006, p. 115)

Secondly, the example of the provision of a lower confidence bound requires justification for the strength of the confidence. Thirdly, it is also difficult to incorporate a value based on a lower confidence bound with the values obtained from other evidence. Even provision of an estimate of variability separate from a point estimate means difficulty for a fact-finder wishing to combine these two estimates. Finally, it is not possible to know the true value of the evidence in a particular case. All that can be done is determination of the best estimate of the value. This is done with the likelihood ratio, as justified by Good (1952).

Acknowledgments

Colin Aitken gratefully acknowledges the support of the Leverhulme Trust with an Emeritus Research Fellowship, grant number EM2016-027. Franco Taroni gratefully acknowledges the support of the Swiss National Science Foundation through grant No. IZSEZ0-19114.

References

Aitken, C.G.G., Connolly, T., Gammerman, A., Zhang, G., Bailey, D., Gordon, R., and Oldfield, R. Statistical modelling in specific case analysis. *Science & Justice*, **36**, 245–255, 1996.

Aitken, C.G.G. and Gammerman, A. Probabilistic reasoning in evidential assessment. *Journal of the Forensic Science Society*, **29**, 303–316, 1989.

Aitken, C.G.G. and Robertson, J. A contribution to the discussion of probabilities and human hair comparisons. *Journal of Forensic Sciences*, **32**, 684–689, 1987.

Anderson, T. and Twining, W. *Analysis of Evidence: How to do Things with Facts based on Wigmore's Science of Judicial Proof*. Northwestern University Press, Evanston, IL, 1998.

Association of Forensic Science Providers. Standards for the evaluation of evaluative forensic science expert opinion. *Science & Justice*, **49**, 161–164, 2009.

Balding, D.J. Estimating products in forensic identification using DNA profiles. *Journal of the American Statistical Association*, **90**, 839–844, 1995.

Balding, D.J. Interpreting DNA evidence: can probability theory help? In J.L. Gastwirth, editor, *Statistical Science in the Courtroom*, pages 51–70. Springer-Verlag, New York, 2000.

Balding, D.J. The DNA database search controversy. *Biometrics*, **58**, 241–244, 2002.

Balding, D.J. and Donnelly, P. Inferring identity from DNA profile evidence. *Proceedings of the National Academy of Sciences USA*, **92**, 11741–11745, 1995.

Balding, D.J. and Donnelly, P. Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Sciences*, **41**, 603–607, 1996.

Balding, D.J. and Nichols, R.A. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, **64**, 125–140, 1994.

Balding, D.J. and Nichols, R.A. A method for quantifying differentiation between populations at multiallelic loci and its implications for investigating identity and paternity. In B.S. Weir, editor, *Human Identification: The Use of DNA Markers*, pages 3–12. Kluwer Academic, 1995.

Balthazard, V. De l'identification par les empreintes digitales. *Comptes rendus des séances de l'Académie des sciences 1864*, **25**, 683–684, 1911.

Berger, C.E.H. and Stoel, R.D. Letter to the Editor – Response to 'A study of the perception of verbal expression of the strength of evidence'. *Science & Justice*, **58**, 76–77, 2018.

Bertillon, A. Instructions Signalétiques. *Imprimerie Administrative*, Melun, 1893.

Bertillon, A. La comparaison des écritures et l'identification graphique. In *Revue Scientifique, Dec. 18, 1897–Jan. 1, 1898*. Typographie Chamerot et Renouard, Paris, 1898.

Biedermann, A., Bozza, S., and Taroni, F. Decision-theoretic properties of forensic identification: underlying logic and argumentative implications. *Forensic Science International*, **177**, 120–132, 2008.

Buckleton, J.S., Taylor, D., Bright, J.-A., and Curran, J.M. Sampling effect. In J.S. Buckleton, J.-A. Bright, and D. Taylor, editors, *Forensic DNA Evidence Interpretation*, pages 181–202. CRC Press, Boca Raton, 2016a.

Buckleton, J.S., Taylor, D., Gill, P., Curran, J.M., and Bright, J-A. Complex profiles. In J.S. Buckleton, J.-A. Bright, and D. Taylor, editors, *Forensic DNA Evidence Interpretation*, pages 229–276. CRC Press, Boca Raton, FL, 2016b.

Buckleton, J.S. and Triggs, C.M. Relatedness and DNA: are we taking it seriously enough? *Forensic Science International*, **152**, 115–119, 2005.

Buckleton, J.S., Triggs, C.M., and Walsh, S.J. *Forensic DNA Evidence Interpretation*. CRC Press, Boca Raton, FL, 2005.

Champod, C. Identification/individualization. In J.A. Siegel, P.J. Saukko, and G.C. Knupfer, editors, *Encyclopaedia of Forensic Sciences*, pages 1077–1084. Academic Press, San Diego, USA, 2000.

Champod, C., Evett, I.W., and Jackson, G. Establishing the most appropriate databases for addressing source level propositions. *Science & Justice*, **44**, 153–164, 2004.

Champod, C., Lennard, C., Margot, P., and Stoilovic, M., editors. *Fingerprints and Other Ridge Skin Impressions* (2nd ed.). CRC Press, Boca Raton, FL, 2016.

Cohen, L.J. *The Probable and the Provable*. Clarendon Press, Oxford, UK, 1977.

Cohen, L.J. The difficulty about conjunction in forensic proof. *The Statistician*, **37**, 415–416, 1988.

Cole, S.A. Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification. *Law, Probability and Risk*, **8**, 233–255, 2009.

Cook, R., Evett, I.W., Jackson, G., Jones, P.J., and Lambert, J.A. A hierarchy of propositions: deciding which level to address in casework. *Science & Justice*, **38**, 231–239, 1998a.

Cook, R., Evett, I.W., Jackson, G., Jones, P.J., and Lambert, J.A. A model for case assessment and interpretation. *Science & Justice*, **38**, 151–156, 1998b.

Cook, R., Evett, I.W., Jackson, G., Jones, P.J., and Lambert, J.A. Case pre-assessment and review in a two-way transfer case. *Science & Justice*, **39**, 103–111, 1999.

Cullison, A.D. Probability analysis of judicial fact-finding: a preliminary outline of the subjective approach. *University of Toledo Law Review*, **1**, 538–598, 1969.

Curran, J.M. Are DNA profiles as rare as we think? Or can we trust DNA statistics? *Significance*, **6**, 62–66, 2010.

Curran, J.M. and Buckleton, J.S. The appropriate use of subpopulation corrections for differences in endogamous communities. *Forensic Science International*, **168**, 106–111, 2007.

Curran, J.M., Buckleton, J.S., and Triggs, C.M. What is the magnitude of the subpopulation effect? *Forensic Science International*, **135**, 1–8, 2003.

Curran, J.M., Buckleton, J.S., Triggs, C.M., and Weir, B.S. Assessing uncertainty in DNA evidence caused by sampling effects. *Science & Justice*, **42**, 29–37, 2002.

Darboux, J.G., Appell, P.E., and Poincaré, J.H. *Le rapport de MM. Darboux, Appell et Poincaré*. L'Action Française, Paris, 1907.

Darboux, J.G., Appell, P.E., and Poincaré, J.H. Examen critique des divers systèmes ou études graphologiques auxquels a donné lieu le bordereau. In *L'affaire Dreyfus - 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, 1908.

Dawid, A.P. The difficulty about conjunction. *The Statistician*, **36**, 91–97, 1987.

Dawid, A.P. Comment on Stockmarr's 'Likelihood ratios for evaluating DNA evidence, when the suspect is found through a database search'. *Biometrics*, **57**, 976–978, 2001.

Dawid, A.P. An object-oriented Bayesian network for estimating mutation rates. Technical Report 226, Department of Statistical Science, University College London, 2003.

Dawid, A.P. and Mortera, J. Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society*, Series B, **58**, 425–443, 1996.

Dawid, A.P., Mortera, J., Pascali, V.L., and van Boxel, D. Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, **29**, 577–595, 2002.

Dawid, A.P., van Boxel, D.W., Mortera, J., and Pascali, V.L. Inference about disputed paternity from an incomplete pedigree using a probabilistic expert system. *Bulletin of the International Statistical Institute*, **58**(Book 1), 241–242, 1999.

De Finetti, B. Probability: the subjectivistic approach. In R. Klibansky, editor, *La Philosophie Contemporaine (Tome 2) – Philosophie des Sciences*, pages 45–53. La Nuova Italia Editrice, Firenze, Italy, 1968.

Edwards, W. Influence diagrams, Bayesian imperialism, and the Collins case: an appeal to reason. *Cardozo Law Review*, **13**, 1025–1074, 1991.

Eggleston, R. Evidence, *Proof and Probability* (2nd ed.). Weidenfeld and Nicolson, 1983.

ENFSI. ENFSI guideline for evaluative reporting in forensic science, 2015. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

Evett, I.W. The interpretation of refractive index measurements. *Forensic Science International*, **9**, 209–217, 1977.

Evett, I.W. Interpretation: a personal odyssey. In C.G.G. Aitken and D.A. Stoney, editors, *The Use of Statistics in Forensic Science*, pages 9–22. Ellis Horwood, Chichester, 1991.

Evett, I.W. Avoiding the transposed conditional. *Science & Justice*, **35**, 127–131, 1995.

Evett, I.W. Expert evidence and forensic misconceptions of the nature of exact science. *Science & Justice*, **36**, 118–122, 1996.

Evett, I.W. and Buckleton, J.S. Statistical analysis of STR data. In A. Carracedo, B. Brinkmann, and W. Bär, editors, *Advances in Forensic Haemogenetics* **6**, pages 79–86. Springer Verlag, Berlin, 1996.

Evett, I.W., Foreman, L.A., Jackson, G., and Lambert, J.A. DNA profiling: A discussion of issues relating to the reporting of very small match probabilities. *The Criminal Law Review*, pages 341–355, 2000a.

Evett, I.W., Gill, P., Jackson, G., Whitaker, J., and Champod, C. Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks. *Journal of Forensic Sciences*, **47**, 520–530, 2002.

Evett, I.W., Jackson, G., and Lambert, J.A. More on the hierarchy of propositions: exploring the distinction between explanations and propositions. *Science & Justice*, **40**, 3–10, 2000b.

Evett, I.W., Jackson, G., Lambert, J.A., and McCrossan, S. The impact of the principles of evidence interpretation and the structure and content of statements. *Science & Justice*, **40**, 233–239, 2000c.

Evett, I.W. and Weir, B.S. *Interpreting DNA Evidence*. Sinauer Associates Inc., Sunderland, 1998.

Fairley, W.B. Probabilistic analysis of identification evidence. *Journal of Legal Studies*, II, 493–513, 1973.

Fienberg, S.E. and Kadane, J.B. The presentation of Bayesian statistical analyses in legal proceedings. *The Statistician*, **32**, 88–98, 1983.

Fienberg, S.E., Krislov, S.H., and Straf, M.L. Understanding and evaluating statistical evidence in litigation. *Jurimetrics Journal*, **36**, 1–32, 1996.

Fienberg, S.E. and Schervish, M.J. The relevance of Bayesian inference for the presentation of statistical evidence and for legal decision making. *Boston University Law Review*, **66**, 771–798, 1986.

Finkelstein, M.O. and Fairley, W.B. A Bayesian approach to identification evidence. *Harvard Law Review*, **83**, 489–517, 1970.

Finkelstein, M.O. and Fairley, W.B. A comment on 'Trial by mathematics'. *Harvard Law Review*, **84**, 1801–1809, 1971.

Foreman, L.A. and Evett, I.W. Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. *International Journal of Legal Medicine*, **114**, 147–155, 2001.

Foreman, L.A., Smith, A.F.M., and Evett, I.W. A Bayesian approach to validating STR multiplex databases for use in forensic casework. *International Journal of Legal Medicine*, **110**, 244–250, 1997.

Friedman, R.D. A close look at probative value. *Boston University Law Review*, **66**, 733–759, 1986a.

Friedman, R.D. A diagrammatic approach to evidence. *Boston University Law Review*, **66**, 571–622, 1986b.

Friedman, R.D. Assessing evidence. *Michigan Law Review*, **94**, 1810–1838, 1996.

Garbolino, P. and Taroni, F. Evaluation of scientific evidence using Bayesian networks. *Forensic Science International*, **125**, 149–155, 2002.

Gaudette, B.D. and Keeping, E.S. An attempt at determining probabilities in human scalp hair. *Journal of Forensic Sciences*, **19**, 599–605, 1974.

Gittelson, S. *Evolving from inferences to decisions in forensic science*. PhD thesis, The University of Lausanne, School of Criminal Justice, Lausanne, 2013.

Good, I.J. Rational decisions. *Journal of the Royal Statistical Society, Series B*, **14**, 107–114, 1952.

Good, I.J. *Probability and the Weighing of Evidence*. Griffin, London, 1950.
Holmes, O.W. Path of the law. *Harvard Law Review*, **10**, 457–478, 1897.
Hopwood, A.J., Puch-Solis, R., Tucker, V.C., Curran, J.M., Skerrett, J., Pope, S., and Tully, G. Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts. *Science & Justice*, **52**, 185–190, 2012.
Jackson, G. The scientist and the scales of justice. *Science & Justice*, **40**, 81–85, 2000.
Jackson, G., Aitken, C.G.G., and Roberts, P. *Case Assessment and Interpretation of Expert Evidence*. Royal Statistical Society, 2015. www.rss.org.uk/statsandlaw; www.maths.ed.ac.uk/~cgga.
Jeffreys, H. *Theory of Probability* (3rd ed.). Clarendon Press, Oxford, 1983.
Kadane, J.B. and Schum, D.A. *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. John Wiley & Sons, New York, 1996.
Kaplan, J. Decision theory and the factfinding process. *Stanford Law Review*, **20**, 1065–1092, 1968.
Kaye, D.H. Probability theory meets res ipsa loquitur. *Michigan Law Review*, **77**, 1456–1484, 1979.
Kaye, D.H. Quantifying probative value. *Boston University Law Review*, **66**, 761–766, 1986.
Kaye, D.H. What is Bayesianism? In P. Tillers and E.D. Green, editors, *Probability and Inference in the Law of Evidence, The Uses and Limits of Bayesianism* (Boston Studies in the Philosophy of Science), pages 1–19. Springer, Dordrecht, 1988.
Kaye, D.H. DNA evidence: probability, population genetics, and the courts. *Harvard Journal of Law & Technology*, **7**, 101–172, 1993.
Kaye, D.H. Logical relevance: problems with the reference population and DNA mixtures in people v. pizarro. *Law, Probability and Risk*, **3**, 211–220, 2004.
Kaye, D.H. Rounding up the usual suspects: a logical and legal analysis of DNA trawling cases. *North Carolina Law Review*, **87**, 425–503, 2009.
Kingston, C.R. Application of probability theory in criminalistics. *Journal of the American Statistical Association*, **60**, 70–80, 1965a.
Kingston, C.R. Application of probability theory in criminalistics – II. *Journal of the American Statistical Association*, **60**, 1028–1034, 1965b.
Kingston, C.R. and Kirk, P.L. The use of statistics in criminalistics. *The Journal of Criminal Law, Criminology and Police Science*, **55**, 514–521, 1964.
Kirk, P.L. and Kingston, C.R. Evidence evaluation and problems in general criminalistics. *Journal of Forensic Sciences*, **9**, 434–444, 1964.
Koehler, J.J. Error and exaggeration in the presentation of DNA evidence at trial. *Jurimetrics*, **34**, 21–39, 1993.
Koehler, J.J. On conveying the probative value of DNA evidence: frequencies, likelihood ratios, and error rates. *University of Colorado Law Journal*, **67**, 859–886, 1996.
Koehler, J.J. Why DNA likelihood ratios should account for error (even when a National Research Council report says they should not). *Jurimetrics Journal*, **37**, 425–437, 1997.
Koehler, J.J. Forensics or Fauxrensics? ascertaining accuracy in the forensic sciences. *Arizona State Law Journal*, **49**, 1369–1416, 2018.
Koehler, J.J., Chia, A., and Lindsey, S. The random match probability in DNA evidence: irrelevant and prejudicial? *Jurimetrics Journal*, **35**, 201–219, 1995.
Lambert, J.A., Scranage, J.K. and Evett, I.W. Large scale database experiments to assess the significance of matching DNA profiles. *International Journal of Legal Medicine*, **108**, 8–13, 1995.
Lempert, R.O. Modelling relevance. *Michigan Law Review*, **75**, 1021–1057, 1977.
Lempert, R.O. The new evidence scholarship: analyzing the process of proof. *Boston University Law Review*, **66**, 439–477, 1986.
Lempert, R.O. Some caveats concerning DNA as criminal identification evidence: with thanks to the Reverend Bayes. *Cardozo Law Review*, **13**, 303–341, 1991.
Levitt, T.S. and Blackmond Laskey, K. Computational inference for evidential reasoning in support of judicial proof. *Cardozo Law Review*, **22**, 1691–1731, 2001.
Lindley, D.V. A problem in forensic science. *Biometrika*, **64**, 207–213, 1977.
Lindley, D.V. *Making Decisions* (2nd ed.). John Wiley & Sons, Chichester, 1985.
Lindley, D.V. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, **2**, 17–24, 1987.

Lindley, D.V. Probability. In C.G.G. Aitken and D.A. Stoney, editors, *The Use of Statistics in Forensic Science*, pages 27–50. Ellis Horwood, Chichester, 1991.

Lindley, D.V. Foreword. In Aitken, C.G.G. and Taroni, F. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, New York, 2nd edition, 2004.

Lindley, D.V. *Understanding Uncertainty*. John Wiley & Sons, Hoboken, 2006.

Locard, E. *L'enquête criminelle et les méthodes scientifiques*. Flammarion, Paris, 1920.

Locard, E. *L'enquête criminelle. Traité de criminalistique*. Desvigne, Lyon, 1940. Tome septième, Livre VIII.

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W.D., Taroni, F., and Hicks, T. Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, **56**, 364–370, 2016.

Meier, P. and Zabell, S. Benjamin Peirce and the Howland will. *Journal of the American Statistical Association*, **75**, 497–506, 1980.

Meuwly, D., Ramos, D., and Haraksim, R. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, **276**, 142–153, 2017.

Morrison, G.S. Special issue on measuring and reporting the precision of forensic likelihood ratios: introduction to the debate. *Science and Justice*, **56**, 371–373, 2016.

Mortera, J., Dawid, A.P., and Lauritzen, S.L. Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205, 2003.

National Research Council. *NRC II – The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, DC, 1996.

National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. National Academy Press, Washington, DC, 2009.

Nordgaard, A. Comment on 'Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio'. *Law, Probability and Risk*, **15**, 17–22, 2016.

Nordgaard, A., Ansell, R., Drotz, W., and Jaeger, L. Scale of conclusions for the value of evidence. *Law, Probability and Risk*, **11**, 1–24, 2012.

Nordgaard, A. and Rasmusson, B. The likelihood ratio as value of evidence – more than a question of numbers. *Law, Probability and Risk*, **11**, 303–315, 2012.

Oesterhelweg, L., Kröber, S., Rottman, K., Willhöft, J., Bfraun, C., Thies, N., Püschel, K., Silkenath, J., and Gehl, A. Cadaver dogs – a study on detection of contaminated carpet squares. *Forensic Science International*, **174**, 35–39, 2008.

Peirce, C.S. The probability of induction. In J.R. Newman, editor, *The World of Mathematics, 1956*, volume 2, Simon Schuster, New York, 1878.

President's Council of Advisors on Science and Technology (PCAST). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, DC, 2016. https://www.nitrd.gov/pubs/PCAST-NITRD-report-2015.pdf.

Ramos, D. and Gonzalez-Rodriguez, J. Reliable support: measuring calibration of likelihood ratios. *Forensic Science International*, **230**, 156–169, 2013.

Redmayne, M. Science, evidence and logic. *The Modern Law Review*, **59**, 747–760, 1996.

Robertson, B.W. and Vignaux, G.A. Extending the conversation about Bayes. *Cardozo Law Review*, **13**, 629–645, 1991.

Robertson, B.W. and Vignaux, G.A. Probability – the logic of the law. *Oxford Journal of Legal Studies*, **13**, 457–478, 1993a.

Robertson, B.W. and Vignaux, G.A. Taking fact analysis seriously. *Michigan Law Review*, **91**, 1442–1464, 1993b.

Robertson, B.W. and Vignaux, G.A. DNA evidence: wrong answers or wrong questions? In B.S. Weir, editor, *Human Identification: The Use of DNA Markers*, pages 145–152. Kluwer Academic, Dordrecht, 1995a.

Robertson, B.W. and Vignaux, G.A. *Interpreting Evidence. Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, Chichester, 1995b.

Robertson, B.W., Vignaux, G.A., and Berger, C.E.H. *Interpreting Evidence. Evaluating Forensic Science in the Courtroom* (2nd ed.). John Wiley & Sons, Chichester, 2016.

Saks, M.J. and Koehler, J.J. The coming paradigm shift in forensic identification science. *Science*, **309**, 892–895, 2005.

Saks, M.J. and Koehler, J.J. The individualization fallacy in forensic science evidence. *Vanderbilt Law Review*, **61**, 199–219, 2008.

Schum, D.A. *Evidential Foundations of Probabilistic Reasoning*. John Wiley & Sons, Inc., New York, 1994.

Sjerps, M.J., Alberink, I., Bolck, A., Stoel, R.D., Vergeer, P., and van Zanten, J.H. Uncertainty and LR: to integrate or not to integrate: that's the question. *Law, Probability and Risk*, **15**, 23–29, 2016.

Souder, W. The merits of scientific evidence. *Journal of the American Institute of Criminal Law and Criminology*, **25**, 683–684, 1934–1935.

Stoney, D.A. Evaluation of associative evidence: choosing the relevant question. *Journal of the Forensic Science Society*, **24**, 473–482, 1984.

Stoney, D.A. Transfer evidence. In C.G.G. Aitken and D.A. Stoney, editors, *The Use of Statistics in Forensic Science*, pages 107–138. Ellis Horwood, New York, 1991a.

Stoney, D.A. What made us ever think we could individualize using statistics? *Journal of the Forensic Science Society*, **31**, 197–199, 1991b.

Stoney, D.A. Relaxation of the assumption of relevance and an application to one-trace and two-trace problems. *Journal of the Forensic Science Society*, **34**, 17–21, 1994.

Taroni, F., Biedermann, A., Bozza, S., Garbolino, P., and Aitken, C.G.G. *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*. Statistics in Practice. John Wiley & Sons, Chichester, 2nd edition, 2014.

Taroni, F., Bozza, S., and Aitken, C.G.G. Decision analysis in forensic science. *Journal of Forensic Sciences*, **50**, 894–905, 2005.

Taroni, F., Bozza, S., Biedermann, A., and Aitken, C.G.G. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, **15**, 1–16, 2016a.

Taroni, F., Bozza, S., Biedermann, A., and Aitken, C.G.G. Rejoinder. *Law, Probability and Risk*, **15**, 31–34, 2016b.

Taroni, F., Champod, C., and Margot, P. Forerunners of Bayesianism in early forensic science. *Jurimetrics Journal*, **38**, 183–200, 1998.

Taroni, F., Marquis, R., Schmittbuhl, M., Biedermann, A., Thiéry, A., and Bozza, S. The use of the likelihood ratio for evaluative and investigative purposes in comparative handwriting examinations. *Forensic Science International*, **214**, 189–194, 2012.

Thagart, P. Why wasn't O.J. convicted? Emotional coherence and legal inference. *Cognition and Emotion*, **17**, 361–383, 2003.

Thompson, W.C. Subjective interpretation, laboratory error and the value of DNA evidence: three case studies. *Genetica*, **96**, 153–168, 1995.

Thompson, W.C. and Schumann, E.L. Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behaviour*, **11**, 167–187, 1987.

Thompson, W.C., Taroni, F., and Aitken, C.G.G. How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Sciences*, **48**, 47–54, 2003.

Tippett, C.F., Emerson, V.J., Fereday, M.J., Lawton, F., and Lampert, S.M. The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society*, **8**, 61–65, 1968.

Tribe, L. Trial by mathematics: precision and ritual in the legal process. *Harvard Law Review*, **84**, 1329–1393, 1971.

Twining, W. *Rethinking Evidence*. Northwestern University Press, Evanston, Ill. USA, 1994.

Weir, B.S. *Genetic Data Analysis II*. Sinauer Associates Inc., Sunderland, MA, 1996.

Wigmore, J.H. The problem of proof. *Illinois Law Review*, **8**, 77–103, 1913.

Yellin, J. Book review of *Evidence, Proof and Probability*, 1st edition, Eggleston, R. (1978). Weidenfeld and Nicolson. *Journal of Economic Literature*, **583**, 583–584, 1979.

Section II

General Concepts and Methods

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Frequentist Methods for Statistical Inference

David H. Kaye

CONTENTS

2.1 Introduction

Statistics textbooks promise that methods for statistical inference can assist in "making valid generalizations from samples" (Freedman et al., 1998, p. xvi); in "draw[ing] conclusions about a population or process based on sample data" (Moore and McCabe, 1993, p. 427); or in answering the question, "[g]iven the outcomes, what can we say about