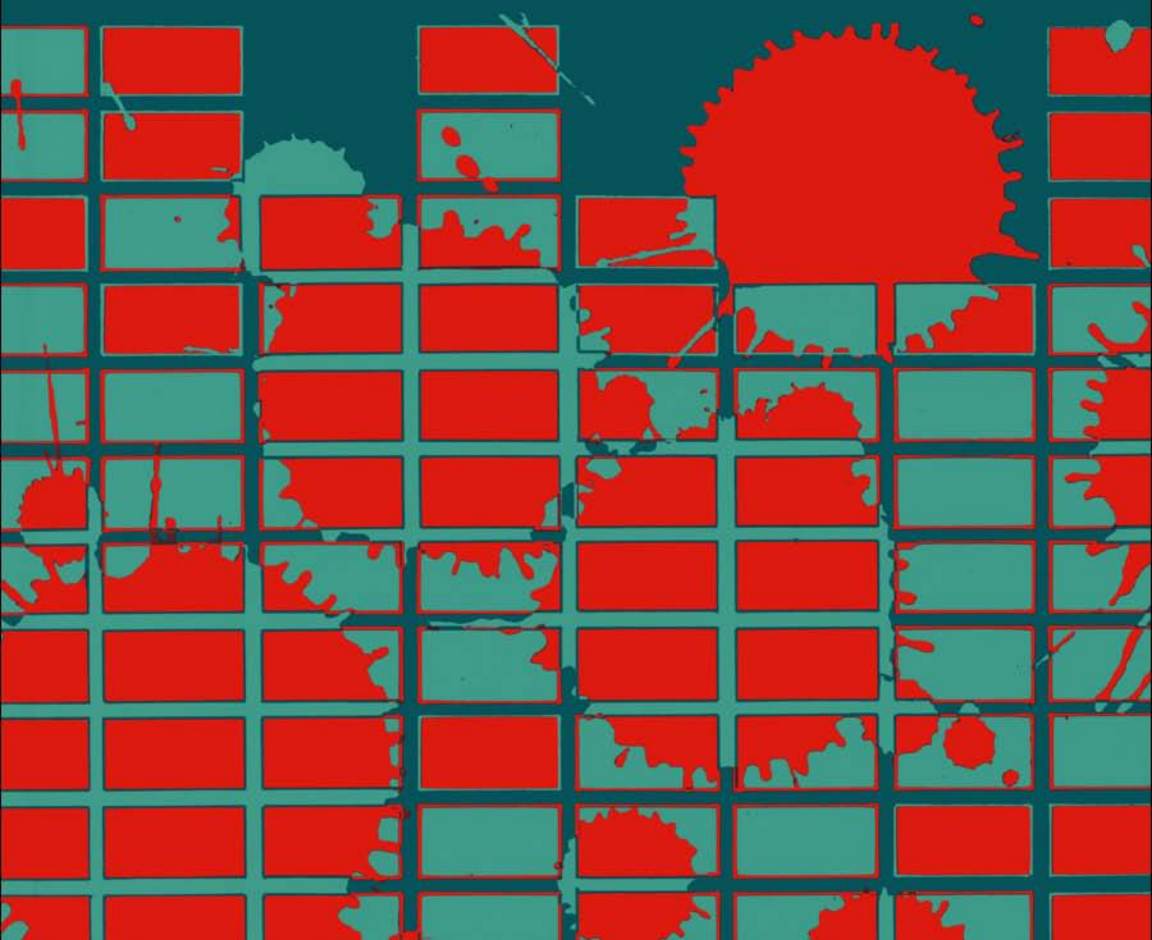
ELLIS HORWOOD SERIES IN FORENSIC SCIENCE

THE USE OF STATISTICS IN FORENSIC SCIENCE

C. G. G. Aitken and D. A. Stoney



THE USE OF STATISTICS IN FORENSIC SCIENCE

ELLIS HORWOOD SERIES IN FORENSIC SCIENCE

Series Editor: JAMES ROBERTSON, Head of Forensic Services Division, Australian Federal Police, Canberra, Australian Capital Territory

The Use of Statistics in Forensic Science
C.G.G.Aitken & D.A.Stoney
The Scientific Examination of Documents
D.Ellen
DNA in Forensic Science: Theory, Techniques and Applications
J.Robertson, A.M.Ross & L.A.Burgoyne

THE USE OF STATISTICS IN FORENSIC SCIENCE

C.G.G.AITKEN B.Sc., Ph.D.
Department of Mathematics and Statistics, The University of Edinburgh
D.A.STONEY B.S., M.P.H., Ph.D.
Forensic Science Programme, University of Illinois at Chicago

ELLIS HORWOOD
NEW YORK LONDON TORONTO SYDNEY TOKYO SINGAPORE

First published in 1991 by
ELLIS HORWOOD LIMITED
Market Cross House, Cooper Street,
Chichester, West Sussex, PO19 1EB, England

This edition published in the Taylor & Francis e-Library, 2002.

A division of
Simon & Schuster International Group
A Paramount Communications Company

British Library Cataloguing-in-Publication Data

Aitken, C.G.G.
The use of statistics in forensic science. —
(Ellis Horwood series in foresnic science)
1. Forensic science. I. Title. II. Stoney, D.A.
363.2560151595
ISBN 0-13-933748-2 (Print Edition)

Library of Congress Cataloging-in-Publication Data

Aitken, C.G.G.
The use of statistics in forensic science/C.G.G.Aitken, D.A.Stoney.
p. cm. —(Ellis Horwood series in forensic science)
Includes bibliographical references and index.
ISBN 0-13-933748-2 (Print Edition)
1. Forensic sciences—Statistical methods. 2. Forensic statistics. 3. Evidence (Law)—
Statistical methods. I. Stoney, David Alan. II. Title. III. Series.
HV8073.A55 1991
363.2´5´015195–dc20 91–22613
CIP

ISBN 0-203-01732-3 Master e-book ISBN
ISBN 0-203-19148-X (Glassbook Format)

Table of contents

To Liz and Penny

Interpretation: a personal odyssey

I.W.Evett
Central Research and Support Establishment, Home Office Forensic Science Service,
Aldermaston, Reading, Berkshire, RG7 4PN, UK

WHY STATISTICS?

In 1966 I joined the Home Office Forensic Science Service as a member of the Documents Section, then based at Cardiff. Most of the work was handwriting examination; then, as today, it was based on painstaking observation of minute detail, recording as much as possible by sketching and note writing. The opinion of the expert stemmed solely from this highly personal and, to my young mind, rather antiquated activity. I had graduated as a physicist, from a world of sophisticated circuitry, instruments, particle accelerators, and precise measurement, so it seemed to me, in those early days, that handwriting examination was long over-due for revolution, a revolution to be brought about through implementing methods of measurement. Fired with enthusiam I embarked on my own exploratory investigations, and, after many hours of patient work with a ruler, I reached that stage which so many experimenters reach sooner or later: 'Now that I've got all these numbers, what do I do with them?' The answer, of course, lay in the science of statistics, but, although my undergraduate course had been heavily mathematical, my ignorance of statistics was absolute. Until then I had successfully avoided involvement with a subject that appeared, with the wisdom of youth, to be fine for opinion polls, market research, and the like, but was not necessary for the kind of science I was engaged in. Now, however, I began to think differently, and, the prospect of a new challenging problem sharpening my appetite, I undertook a serious study of statistics. It was fortunate for me that my Director encouraged my efforts, and it was thanks to him that I was granted a year's special leave to follow a full time postgraduate course at University College, Cardiff.

It was also during my early days as a forensic scientist that I discovered the first of a small number of papers which have determined the broad course of my thinking. Entitled 'The ontogeny of criminalistics' it had been written by Paul L.Kirk (1963), then Professor of Criminalistics† at the University of California, Berkeley. Kirk posed searching questions about the nature of criminalistics—is it a science? Is it a profession? If it is a science then where are the principles which should form the foundation for any pursuit which is claimed to be a science? This is a question which still remains at the forefront of my thoughts; now, I think we are closer to the answer, as I will later explain. Kirk and his pupil, Charles Kingston, wrote other significant papers (1963, 1964, for example) which still make good reading today. They called for constructive collaboration between forensic scientist and statistician, and their work fuelled still further my interest in the latter's discipline.

TWO-STAGE TESTING

After qualifying as a statistician, in the early 1970s, I soon realized that the complexities of the handwriting data were not going to yield to simple methods, and it seemed to me that my time would be better spent tackling problems which involved simpler univariate data. So I turned my attention to the problems of evaluating refractive index measurements on glass fragments. Initially, and throughout the 70s, the approaches that I have followed were rooted firmly in standard statistical hypothesis testing and a two-stage approach to interpretation, proceeding along lines such as the following.

Let y denote a measurement carried out on some material of unknown origin, call it the *trace* material, and let x be a measurement on a material sample of known origin, call it the *control* (see section 4.4.2 for some comments on terminology). The issue is 'did the trace come from the same source as the control?' The first stage of analysis is to set up a null hypothesis that they do have the same source. Under that hypothesis, the statistical distribution of (y-x) can be established, and so the probability of this difference exceeding a chosen threshold can be established. If, for example, we pick a threshold such that the probability of |y-x| *e*xceeding it is 0.05, then that threshold forms the basis of a test: the test is passed if |y-x| is less then the threshold and failed if it exceeds the threshold. If we adopt such testing as policy, then we settle with the knowledge that, in 5% of the cases in which the trace and control had actually come from the same source, the test will show otherwise. In practice, the words 'pass' and 'fail' are not used. It is much more common to use words such as match/non-match or similar/different.

If the test results in a non-match then that is the end of the comparison—a 'negative' case—but if there is a match then it does not necessarily mean that the two samples had the same source. The evidence in favour of that eventually depends on how rare, or how common, material having the observed properties is in the relevant population of sources of such material (see section 2.3 for a discussion of the meaning of 'relevant population'). So the second stage of interpretation requires reference to a sample of some kind from the population to estimate the probability of a coincidental match. The smaller the coincidence probability, the greater the evidential value of the match.

† 'Criminalistics' is a term which is used infrequently in the UK. Within the confines of this discussion it can be used interchangeably with 'forensic science'.

This 'coincidence' approach reflects widespread practice in courts of law, but its logic is incomplete, leading to a classic prosecution/defence dichotomy. I can best illustrate it by reference to a simple example of a bloodstain left at the site of a crime which matches the blood type of the defendant (call this some hypothetical type X1): assume that we know that 1% of the population have this blood type. The classic prosecution argument proceeds as follows. The probability of the scene evidence if someone else had left the bloodstain is only 0.01; therefore, the probability that someone else left the bloodstain is 0.01; therefore, we can be 99% sure that the bloodstain came from the defendant. The defence argument will be along the following lines. In the UK there must be over half a million people with this blood type, so the probability that the blood stain came from the defendant is only one in half a million.

Neither of these arguments, of course, is correct—they have been called the 'prosecutor's' and 'defence attorney's' fallacies by Thompson & Schumann (1987). However, to the layman, it must be extremely difficult to resolve the true position. For my own part I found the dichotomy confusing, and it took me some years to understand the correct perspective.

I will return to this problem later, but before proceeding it is essential that I should make it absolutely clear that the above prosecution argument is completely wrong: the second statement of the prosecution reasoning does not follow from the first!

AN ALTERNATIVE PERSPECTIVE

Most of the research which I carried out in the 1970s and the early 1980s, at first alone, but later in collaboration with my colleague Jim Lambert, was rooted in this significance testing two-stage approach. It was profitable and enjoyable work which led to several improvements in operational procedures: a computer program arising from that period is still being used for certain aspects of our work. However, other developments were to have a gradual, yet profound, effect on my thinking and approach to forensic science problems.

In early 1974 I found that Professor Dennis Lindley had given a lecture entitled 'Probabilities and the law' in Scotland to a local group of the Royal Statistical Society. In response to a letter he kindly sent me a transcript of his lecture. Although it has since been published (1975), I still have that dog-eared, yet treasured, original typescript.

It is difficult for me to convey the excitement I experienced when I first read the paper. That it was Bayesian in its approach was not new—Mosteller & Wallace (1964) had used Bayesian methods in their analysis of the authorship of The Federalist papers, and Finkelstein & Fairley (1970) had advocated a Bayesian approach to evidence evaluation using illustrations based on trace evidence left at a crime scene—what *was* new was that Professor Lindley had devised an analysis for the case where trace evidence is found on the person of a suspect. This is more difficult than the case where trace evidence is left at the scene, because the issues of transfer and persistence become very important. Colleagues of mine, Tony Pounds & Ken Smalldon (1975a, 19755) had been working at the Central Research Establishment (then CRE, now Central Research and Support Establishment— CRSE) on gaining an understanding of the mechanisms underlying the transfer of materials such as glass or fibres from a crime scene to a suspect and the subsequent persistence of such traces. My own work, restricted as it was to the two-stage treatment of measurement error, had given me no inkling as to how the kind of probability distributions which Ken and Tony were establishing could be brought into a formal treatment. Here was the key: Professor Lindley (although, to my knowledge, he was not aware of the work being done at CRE) had a means for taking account of the considerations of transfer and persistence—and also of the findings of clothing surveys such as those previously carried out by Pearson, May, & Dabbs (1971) —in one equation.

That was the start of an exchange of correspondence with Dennis Lindley which continues today; an exchange which, more than anything else, has shaped my views on evidence interpretation.

But although my own ideas were moving forward, my initial attempts to share my enthusiasm with the forensic science world were failures. A paper which I submitted to the *Journal of the Forensic Science Society,* based on Dennis Lindley's methods, was savaged by the referees and rejected without a single word of encouragement. A paper which I presented at a colloquium in Aldermaston met a response which bordered on the abusive. These were hard knocks, but they taught me the vital importance of presentation. Good ideas are worthless if they are not presented in terms which are suitable for the recipient audience. When, several years later, I did succeed in having a Bayesian paper published by the *Journal of the Forensic Science Society* it was given the P.W.Allen Award for the best paper of the year! Yet it contained but a fraction of the ideas in the original rejected paper.

BAYESIAN ANALYSIS OF MEASUREMENTS

During the correspondence with Dennis Lindley I explained the nature of the work that I was doing on glass refractive index measurements. He was far from impressed by the philosophy of the two-stage approach, and I found it difficult to defend certain logical weaknesses. Why should there be an arbitrary cut-off line for matching? (called the 'fall-off-the-cliff effect by Ken Smalldon). Why should the matching test be carried out while deliberately ignoring the population data? Is the null-hypothesis that the two samples have the same source compatible with the presumption of innocence? What is the functional relationship between coincidence probability and evidential strength? Dennis Lindley's response (1977) was to show how the Bayesian likelihood ratio (see section 1.13 and Chapter 3) provides a conceptually simple and coherent one-stage solution to the problem. The numerator of the ratio is a function of the 'closeness' of the trace and control measurements—the greater the difference between them, the smaller the numerator. The denominator is a function of the frequency with which measurements such as the trace and control measurements occur in the population—the more common they are the larger the denominator, and vice versa. Thus evidential value is measured by a continuous function with no arbitrary cut-off. Values of the ratio greater than one tend to confirm or support the assertion that the trace and control have come from the same source; and values less than one tend to confirm or support the assertion that they have come from different sources. The justification for this lies in Bayes' Theorem as is briefly explained in the next section and in much more detail by David Stoney (section 4.5). Dennis Lindley's paper is a classic which is essential reading for all serious students of transfer evidence interpretation.

I did not take to Dennis Lindley's analysis at first. I presented to him a counter argument: two measurements might be far enough apart to be ruled a non-match by the comparison test, yet—particularly if they were from a relatively sparse area of the population distribution—the likelihood ratio calculation could give a value appreciably greater than one. It is almost incredible to me now that I actually presented such an example to demonstrate that the two-stage approach was correct! I still have the letter from Dennis Lindley which patiently explains that my example carried precisely the point he was trying to make: the totality of the evidence might support the hypothesis of common origin for the two samples even if their measurements differed by more than some arbitrary threshold. It was with a sense of *déja vu* that, over ten years later, I saw the same issue arise in the new field of DNA profiling.

THE BAYESIAN APPROACH—A BRIEF SKETCH

Other writers in this book explain the Bayesian approach to evidence evaluation in some detail. A brief decription of its essentials is useful at this point because it will enable me to explain my own views further.

A criminal trial is all about information: some of it good, some bad, much uncertain, and some downright misleading. Our concern is to see how scientific evidence can best be presented in this complex framework. Assume a trial in which the issue is whether there is sufficient evidence to decide that a defendant had committed a particular crime. In the adversary system of justice we can visualize two competing hypotheses: C, the defendant committed the crime; and $\bar{C}$, the defendant did not commit the crime (which carries the automatic implication that someone else did). There will, in general, be items of non-scientific evidence, such as eye-witness reports which will affect the degree of uncertainty about whether C or $\bar{C}$ obtains, and it is most useful to visualize this as 'odds' —in a similar manner to betting odds. Whatever the odds on C were before the scientific evidence, we need to see how they are changed. This is where Bayes' Theorem comes in, because, in this context, it tells us that the odds on C *after* the scientific evidence are simply the odds *before* the scientific evidence multiplied by the factor which is known as the likelihood ratio.

This is:

$$\frac{\text{probability of the evidence if C is true}}{\text{probability of the evidence if } \bar{\text{C}} \text{ is true}}\,.$$

This is a result of fundamental importance because it identifies the class of questions which the scientist should address. Whereas the court must address questions of the type ‘What is the probability (or odds) of C given the evidence?’; the scientist must ask questions of the type ‘What is the probability of the *evidence* given C?’ and ‘What is the probability of the evidence given $\bar{C}$?’

Let us return to the example of the bloodstain left at the crime scene. The probability of the scene evidence if C is the case is one because the blood type X_1 is exactly what would occur if the defendant had left the stain. If $\bar{C}$ is the case, then someone else left the crime stain. It may be that the defence will pose some alternative explanation, but in the absence of a specific explanation we consider this other person as a random selection from the population. The probability that a random member of the population is of type X_1 is 0.01. The likelihood ratio, therefore, is 100.

This is the balanced view which holds the middle ground between the classic prosecution/defence dichotomy. The evidence has been assessed from both viewpoints, C and $\bar{C}$, and whatever the odds on C based on the other evidence, they are multiplied 100-fold.

Detractors of the Bayesian approach will argue that this logic is too difficult to explain to the layman. I am still waiting for the holders of this negative view to show me a better alternative. Are we going to stick with the classic prosecutor’s fallacy as a way of explaining our results? At the end of the day, the problems of interpretation and communication are far from trivial, so we should not expect to find trivial solutions. I return to the problem of explaining the likelihood ratio later.

The present example is the simplest kind of case. That simplicity is reflected in the numerator of one. In real life cases the numerator is unlikely to be one, and the classic coincidence approach cannot give a balanced and logical assessment. I will illustrate this in the next section, but first it is necessary for me to caution that there are issues which I have avoided here in the need to provide a brief and, I hope, simple explanation. In particular, the concept of a random selection from the population is fraught with difficulty (section 2.4), but these are difficulties which do not arise specifically from the Bayesian formulation. On the contrary, a considered Bayesian view is the means for facilitating their resolution.

THE TWO-TRACE PROBLEM

Although I had been heavily influenced by Dennis Lindley’s paper on evidence transfer and measurement analysis I continued to work, in collaboration with Jim Lambert, on the two-stage approach to glass refractive index interpretation. The computer program that we developed enabled comparisons to be made, using t-tests, and also calculated coincidence probabilities where necessary. However, in this latter context, we ran into a conceptual problem which we were unable to resolve. The problem can best be illustrated by means of an analogous example based on bloodstains.

In the simple example used to illustrate Bayes’ Theorem, the evidence consisted of one bloodstain of type X_1 at the crime scene. Assume now that the crime had involved not one but *two* men, each of them leaving a bloodstain at the scene. The two stains are of type X_1 and X_2, which occur with relative frequencies f_1 and f_2 respectively. The investigator has arrested just one suspect, and his blood is of type X_1. How is the evidence to be assessed in this case? Most would see intuitively that the evidence strength should be less than it would be had there been only one stain present. How much less?

The classic coincidence approach which I, and others, had followed runs as follows. If there were only one stain present we would ask 'What is the probability that a person selected at random would match the crime stain?' Given two different stains we ask 'What is the probability that a person selected at random would match one of the crime stains?' This is f_1+f_2.

At first sight this seems fine. The probability is greater than f_1, which is the probability which would be quoted if there were only one stain. However, a problem, soon arises. What if, in another similar case, the suspect matches stain X_2? The evidential strength is still indicted by f_1+f_2. Now imagine that f_1 = 0.01 and f_2 = 0.10. Can the strength of the evidence be the same whichever of the stains the suspect matches? No, it can't be (if the reader doubts my assertion then he should think it through with f_1 = 0.0001 and f_2 = 0.9999).

Once one understands the Bayesian approach it becomes clear why the coincidence approach cannot tackle cases like this. Basically, it is because the coincidence approach addresses only the denominator of the likelihood ratio. The tacit assumption is that the numerator is one. In the simple case involving one stain at the scene, then this is fine, but in any other situation the coincidence approach cannot work.

As it turns out, the solution to the two-trace problem is quite simple. Recall that for the one-trace problem when the suspect and stain were both X_1 the likelihood ratio is 1/f_1. For the comparable two-trace problem it is 1/2f_1. Note that the frequency of the other non-matching blood type is irrelevant. David Stoney explains the derivation of the result (section 4.5.3.1).

TRANSFER FROM THE CRIME SCENE

The examples that I have given have involved trace evidence left at a crime scene. What about trace evidence found on the person of a suspect which may, or may not, have come from the crime scene? Now the interpretive considerations become more complex. Assume that a person has been stabbed to death. A suspect has been taken and a search of his clothing reveals a bloodstain on his jacket which is of the same blood type, X_1, as that of the victim. The suspect is another blood type. Some of the questions to be addressed are:

What is the probability of a bloodstain of the observed shape and size if the suspect had stabbed the victim?

What is the probability of finding a non-self bloodstain of the observed shape and size on the clothing of a person at random?

What is the probability that a non-self bloodstain on the clothing of a person at random would be type X_1?

The answers to such questions depend on using surveys intelligently and also, where surveys are inadequate, on judgement. Given the answers they can be combined logically as shown by David Stoney (section 4.5.7). Here the Bayesian approach again demonstrates its power, as originally shown by Dennis Lindley in the early 1970s.

INTELLIGENT KNOWLEDGE-BASED SYSTEMS

In principle, one can use Bayesian logic to tackle any interpretive problem. In practice, of course, the analysis can become very complicated. Think, for example, of a glass case where a man is suspected of smashing three windows with a hammer. A search of his clothing reveals many glass fragments which can be put into five groups on the basis of their physical properties. Two of these groups match glass from two of the broken windows, but the other three are quite different. Given the circumstances of the crime, are these the sort of findings we would expect if the suspect has been the man responsible? Are they the sort of findings we would expect if he was a random member of the population?

There are many considerations here relating to evidence transfer and persistence, and it is necessary to have some feel for the distribution of glass fragments on the clothing of members of the population. It is possible to frame a series of questions and build the answers into a formula for the likelihood ratio. However, there is no denying that it is a complicated business which is going to differ in detail from case to case. It is too much to expect the busy caseworker to attempt it as a matter of routine.

The solution lies in computer technology; in particular, the area of computer technology which is known as intelligent knowledge-based systems (IKBS). These represent means of providing computer assistance for making judgements and decisions in complex situations where expertise is an important factor. I first became interested in the technology in the early 1980s, and Jim Lambert and I had a short period of collaboration with Queen Mary College, London. However, at that time it seemed as though the hardware and software available were not up to the sort of challenges we faced. Later, the husband and wife team Vina and Ernie Spiehler, while working at CRSE, soon demonstrated the potential for real progress through the use of programming tools which are known as 'expert system shells'.

The major step forward in the area of transfer evidence was taken in 1989 by John Buckleton, of DSIR, New Zealand and a colleague of mine, Richard Pinchin. John was serving a period of attachment to CRSE, and he and Richard decided to use a shell known as CRYSTAL to create an advisory system for glass evidence interpretation. I had misgivings about the project because I thought their goal to be over-ambitious: it was a genine pleasure to be proved wrong!

The computer system—called Computer Assistance for Glass Evidence (CAGE) — is described in some detail by John Buckleton and Kevan Walsh (section 6.3.3), but here is a brief summary of its main elements.

At its simplest level CAGE is just a powerful calculator which takes a series of probabilities and combines them into a likelihood ratio. That, in itself, is a very important, though hardly revolutionary, process. What is innovative is the means by which the probabilities are adduced.

Transfer and persistence are highly relevant to interpreting glass evidence, but how can they be determined? The idea of quantifying such probabilities is strange to operational forensic scientists, but CAGE provides the means for presenting to the scientist the research results which are relevant to his case and also a 'knowledge base' which stores the opinions of experts from previous cases. In the light of such information the scientist is led through an imaginative interface providing displays which encourage him to provide his own estimates for the case in hand. He is free to ignore the stored information if he so wishes, but at the end of the consultation he is asked if he will agree to his estimates being stored for possible future reference in another similar case.

These personal or 'soft' probabilities are brought together in the calculation with other 'hard' probabilities which have come from a survey of glass on clothing which had been carried out at the Belfast laboratory. The ability to combine soft and hard probabilities is one of the central concepts of Bayesian inference, and such applications were foreseen by Kirk & Kingston (1964), writers who, as I have said, have had a considerable effect on my views on forensic interpretation.

CAGE also embodies the facilities for the users to study the elements of the likelihood ratio and thus find out which may be particularly sensitive for the individual case. Indeed, the user is encouraged to take nothing at face value but to explore the consequences of varying the different components of the calculation. At the end of the activity it is the user who decides on the strength of the evidence: the computer's role is advisory.

DNA STATISTICS

Over the last few years I have become increasingly involved in the exciting new world of DNA profiling. Already this is a subject broad enough to justify a book in its own right. Here, however, I will do no more than make a few general observations.

It has seemed remarkable to me that organizations have been eager to invest heavily in the expertise of moleculer geneticists to develop the technology of DNA profiling, yet when it came to the scientific questions of greatest importance—i.e. the inferences to be drawn from the profiles—they were addressed by amateurs with no more than a rudimentary knowledge of statistics. The notorious Castro case was graphically described by Lander (1989). There were all sorts of problems with this case, but many of them could have been resolved by good statistics—as has been demonstrated by Berry (1991). I have already referred to the *déja vu* I experienced. The methodology, such as it was, in the Castro case was the same as the two-stage approach that I had followed several years earlier when dealing with glass. Some of the difficulties arose because comparisons which fell just outside an arbitrary threshold were ruled as matches by the expert. The root of the problem of course was not that they fall outside the threshold—but that a threshold was adopted in the first place.

DNA technology, as I write, is evolving rapidly, and the operational sensitivity will reach levels which were undreamed of five years ago. This has widespread implications for forensic science. How will the availability of such techniques affect the priorities of the forensic science laboratory in the 21st century? What about the issues of transfer, persistence, and contamination? How should one tackle the analysis of data from mixtures? Statistics must play a central role in resolving questions such as these.

COMMUNICATION

It has become clear to me over the years that communication of findings is closely linked to, and as important as, interpretation. For too many years, forensic scientists have taken too much for granted in relation to the powers of comprehension of non-scientists. Arriving at a flawless evaluation of the evidence is not much use unless we can convey that evaluation effectively to a court where most of the audience will have only the vaguest idea of what forensic science is about.

My views have been shaped in part through my involvement in interpretation courses which are run at the Forensic Science Training Unit of the Forensic Science Service. Each of these culminates in an exercise where the evidence must be presented to a lawyer, who is one of the course tutors. Also, over the last year, I have been a member of a team which has made recommendations for improving the ways in which we write our statements.

Of course, there are numerous problems involved in explaining sophisticated scientific techniques simply. But as far as the interpretation side is concerned I believe that there are three main elements of effective communication.

First, in all except the simplest cases, the interpretation of the scientific results depends on the circumstances—or, at least, the perceived circumstances—so it is necessary to detail these in the statement. If this is done properly then the rules of hearsay need not be flouted, and it signals a clear message: if any of the circumstances are changed then it will be necessary to review the interpretation.

Second, it is necessary to state the alternative explanations for the evidence which have been considered. In my experience it is generally feasible to consider two alternatives at any one time, and this makes interpretation and communication much simpler. One of these will be the prosecution alternative, the other the defence alternative. In the event of no specific defence alternative we generally invoke the concept of a ‘random selection from the population’. Obviously we should make that concept clear and also indicate which population we have considered. In this way, as in our explanation of circumstances, we signal our willingness to consider other alternatives if either the circumstances change or the defence wishes us to address hypothetical scenarios. Given this framework we establish the probability of the evidence under each of the alternative explanations. If there are two, then the likelihood ratio then provides the means of weighing the evidence.

The third stage is the presentation of the weight of evidence derived in this way. However, I should make it clear that I am not enthusiastic about the idea of presenting the likelihood ratio in numerical terms, for two main reasons. First, a statement of the form ‘the likelihood ratio is 473.64’ cannot be expected to convey anything to a layman. Second, workers such as Kahneman, Slovic, & Tversky (1982) have shown that people are quite irrational in their use of numbers in the face of uncertainty. It is my view that the expert should express the evidence strength as far as possible in verbal terms. One path to this, as I have advocated elsewhere (1987), is to assign verbal equivalents to increasing ranges of the likelihood ratio. For example, in current work on DNA we are using the following verbal equivalents:

LIKELIHOOD RATIO IN THE RANGE	EVIDENCE STRENGTH
1–33	Weak
33–100	Fair
100–330	Good
330–1000	Strong
1000+	Very strong

However, these are not meant to be tablets of stone. Language is imprecise, and there is no ‘rule’ which says that 99 is ‘Fair’ and 101 is ‘Good’. The convention is meant solely as a guide for improving the comprehension of people who are unused to thinking numerically. I have found that forensic scientists generally react quite favourably to this approach, which has also generated some interest among lawyers.

PRINCIPLES

At the start of this foreword I reiterated Professor Kirk’s question about what were the principles of our science. I believe that Bayesian inference has provided the means of shaping principles for the interpretation process at least.

The first two principles are:

(i) To assess the strength of scientific evidence it is necessary to consider (at least) two explanations for its occurrence;

(ii) The evidence is evaluated by assessing its probability under each of the competing explanations.

Some may regard the first of these as obvious, but, in my experience, many of the problems which people have with interpreting evidence stem from a single-minded concentration on the ‘coincidence’ aspect. They concentrate on the denominator of the likelihood ratio, ignoring the fact that the numerator is not necessarily one. The second is far from obvious, and it demands considerable personal reflection.

If we confine ourselves to the context of the case where there are only two alternative explanations—one defence and one prosecution—then we can enunciate a third principle.

(iii) The strength of the evidence in relation to one of the explanations is the probability of the evidence given that explanation, divided by the probability of the evidence given the alternative explanation.

This is, of course, merely a verbalization of the likelihood ratio, and I would be the first to accept that none of the above three deserves the status of fundamental principles—

such as the great natural principles of Newton, for example. All three flow from other results which, in turn, are built on the axioms of probability (see Chapter 1). It may be that it would be better (and more modest!) to call them precepts, but I have no strong feelings either way.

STATISTICS AND MORAL CERTAINTY?

By this stage, some readers may be wondering how my statistics training affected my views on handwriting examination. I have, at various stages of my career, returned to the statistical analysis of handwriting, and I have been much interested in the potential of pattern recognition and image analysis techniques. However, I still find the sheer complexity of the problem daunting. We can now do so much with computers, yet there are some tasks which the trained human eye/brain does incomparably better than the computer. I have seen that graphically demonstrated in a related area in which I have done some work—that of fingerprint examination. Substantial financial investment has led to powerful computers for rapidly retrieving fingerprints similar to an unknown mark from a large collection, but the task of spotting an identification suitable for court purposes can be done only by an expert. Most of the computer ‘hits’ will be rejected by the trained human eye after only an almost cursory glance. Indeed, when I have seen fingerprint officers at work I have been struck with a sense of wonder at the power of the eye/brain to resolve the relevant pattern from an almost overwhelming mess of confounding noise. So, whereas the computer is fine for rapid searching and retrieval, the final, evidentially crucial, stage can be done only by a human being. Such is the complexity of the mental processes, I cannot believe a computer emulation of them to be an economically feasibly prospect for the foreseeable future. I’m confident that the same applies to handwriting comparison. For the medium-term future, our path for improvement follows the bringing together of the strong points of the computer with the strong points of the human eye/brain combination through the medium of IKBS technology.

It is on the middle ground between hard statistics and expert opinion that we come to a conundrum which has exercised my mind more than a little recently. As I have said, I have spent a lot of my time recently working on the new science of DNA statistics. DNA technology, if properly applied, now gives us the potential to quote likelihood ratios in the millions, tens of millions, and even thousands of millions.

The conundrum is this. It is accepted practice for the fingerprint expert and the handwriting expert, given sufficient levels of detail in a particular comparison, to express an opinion which amounts to a statement of moral certainty: such as, for example, ‘…these two samples were written by the same person…’; ‘…these two points were made by the same finger…’. The DNA scientist can do something which the fingerprint expert and the handwriting expert cannot do—he can give a numerical measure of the evidence. When is he in a position to express an opinion of moral certainty?

When I have talked about verbal conventions I have not yet heard anybody disagree when I have suggested that a likelihood ratio of 1000 is very strong evidence. But what about a likelihood ratio of 10000? or 100000? Words seem inadequate for these magnitudes, and the only solution—to which I come reluctantly—may be to settle for presenting the number without verbal comment.

As for the issue of moral certainty, I don't think there can be a numerical equivalent. This may be an expression of defeat, but I think the issue will continue to be debated over the years to come.

REVIEW

My views on the role of statistics in forensic science have changed a lot over twenty-five years. I started out as a member of the 'reductionist' school which believes that everything can ultimately be described and specified through numbers. I now believe that reductionism is essentially a fruitless pursuit. We must continue to exploit the power of the brain/eye combination to carry out recognition tasks, but we must also strive to discipline it. For an expert to say 'I think this is true because I have been doing this job for *x* years' is, in my view, unscientific. On the other hand, for an expert to say 'I think this is true and my judgement has been tested in controlled experiments' is fundamentally scientific. The computer systems that we can now foresee will form the focus of this paradigm. Given a particular interpretive problem, the computer systems must enable the expert to glean what he can from whatever data collections are relevant. On the absence of hard data the computer should assist him to explore his own perceptions critically, analytically, and with the benefit of a distillation of what his peers might have thought in broadly similar circumstances. The computer can also provide the mechanism for ensuring that all of the aspects of the interpretation bind together in a coherent whole.

But it doesn't finish there, because an expanding knowledge base does not necessarily get better as it gets bigger—indeed, if not properly informed it could become an 'ignorance base'! It could be that all of the experts share the same misconceptions about the true state of things. This is where calibration comes in. At intervals we should carry out experiments to test the quality of the contents of the knowledge base under suitably controlled conditions: the results would serve to refine and improve the quality of the pool of expertise.

This is my view of the future—synergistic combination of expert judgement and statistical data. Bayesian inference provides the logic, IKBS the methodology. We are working on such systems at CRSE at present.

In a sense I am doing no more than amplifying the concept expressed by Kirk & Kingston in 1964. I make no apology for that: they provided an outline, I have filled in some of the details from a new perspective.

REFERENCES

Berry, D.A. (1991) Inferences using DNA profiling in forensic identification and paternity cases, *Statistical Science* **6**, 175–205.

Evett, I.W. (1987) Bayesian inference and forensic science: problems and perspectives, *The Statistician* **36**, 99–105.

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

Kahneman, D., Slovic, P., & Tversky, A. (1982) *Judgement under uncertainty: heuristics and biases,* Cambridge University Press.

Kingston, C.R., & Kirk, P.L. (1963) The use of statistics in criminalistics, *Journal of Criminal Law, Criminology and Police Science* **55** 514–521.

Kirk, P.L. (1963) The ontogeny of criminalistics, *Journal of Criminal Law, Criminology and Police Science* **54**, 235–238.

Kirk, P.L., & Kingston, C.R. (1964) Evidence evaluation and problems in general criminalistics, *Journal of Forensic Sciences* **9** 434–444.

Lander, E.S. (1989) DNA fingerprinting on trial, *Nature* **339** 501–505.

Lindley, D.V. (1977) A problem in forensic science, *Biometrika* **64** 207–13.

Lindley, D.V. (1975) Probabilities and the law. In: *Utility, probability and human decision making,* Eds: Wendt, D. & Wlek, C.J., Reidel, Dordrecht, The Netherlands.

Mosteller, F. & Wallace, D.L. (1964) *Applied Bayesian and classical inference—the case of the Federalist papers,* Springer Verlag, New York.

Pearson, E.F., May, R.W., & Dabbs, M.D.G. (1971) Glass and paint fragments found in men's outer clothing—report of a survey. *Journal of Forensic Sciences* **16** 283–300.

Pounds, C.A., & Smalldon, K.W. (1975a) The transfer of fibres between clothing materials during simulated contacts and their persistence during wear, part 1 — fibre transference, *Journal of the Forensic Science Society* **15** 17–18.

Pounds, C.A., & Smalldon, K.W. (1975b) The transfer of fibres between clothing material during simulated contacts and their persistence during wear, part 2 — fibre persistence, *Journal of the Forensic Science Society* **15** 29–38.

Thompson, W.C., & Schumann, E.L. (1987) Interpretation of statistical evidence in criminal trials, *Law and Human Behaviour* **11**(3), 167–187.

Introduction

C.G.G.Aitken ,
Department of Mathematics and Statistics, The University of Edinburgh,
Edinburgh, UK
and
D.A.Stoney
Department of Pharmacodynamics, University of Illinois at Chicago, USA

The purpose of this book is to describe ways of assessing evidence and ways of communicating this assessment to a court of law. It is not our purpose to describe statistical techniques to forensic scientists, except in a limited way.

The differences between these three aspects may be exemplified by considering the analysis of alcohol consumption in a drunk driving case. Methods of back calculation to determine the blood alcohol level of a driver at the time of an accident would be considered statistical techniques. The assessment of this blood alcohol level would include, for example, determination of the probability that the blood alcohol level of the driver was above a legal limit at the time of the accident. If this probability was above some predetermined value (0.90, say) then the driver might be considered guilty. The communication of the assessment would entail, if asked, a clear statement of the assumptions made in the statistical techniques and the reasons for the particular choice of method for the assessment of the evidence.

Certain aspects of Chapter 5 require an understanding of basic statistical terminology such as confidence intervals, but most of the book may be read by those with only a limited knowledge of statistics. Mathematical symbolism is unavoidable if clarity is to be achieved without an undue quantity of words; however, the symbolism is kept to a minimum and is introduced before its first usage.

The book is concerned primarily with descriptions of methods of assessing evidence and with methods of communicating that assessment to a court of law. However, in specific sections, for example those on paternity and alcohol, there may be a need to consider the actual tests employed. This apart, one of the main aims of the book is to provide a clear exposition on probability.

Chapter 1 is an introduction to probability from the Bayesian viewpoint by a leading exponent of this approach and one who has advocated for some considerable time a greater use of probabilities in legal procedures, not just in the assessment of forensic science evidence. Probability is dealt with in a very clear way, laying a very strong foundation for the later chapters.

Chapter 2 emphasizes the importance of the concept of population. Statistics relies heavily on well-defined samples from well-defined populations. These ideas are extremely important for forensic scientists. There are limitations to the ideal population. These are spelled out and put in perspective.

Average probabilities are first mentioned here. They occur elsewhere in this book, in particular in discussions relating to discriminating power. They have a place in the assessment of the worth of a general area of evidence, such as head hairs, where frequency data are unavailable. However, it is very misleading to quote average probabilities in the courts as a measure of the value of the evidence. Other topics discussed include the effects of selection on the representativeness of samples and the assessment of rare events or events which may happen but have not been observed in the experience of the investigating scientist.

Chapter 3 describes what is possibly the best way of evaluating evidence. The likelihood ratio is easy to interpret for a layman. The only probabilities that are relevant to the assessment of evidence are the probability of the evidence if the suspect is guilty and the probability of the evidence if the suspect is innocent. For sound logical reasons the best way of comparing these probabilities is to calculate their ratio.

This chapter, by Professor Good, provides the philosophical foundation for the assessment of evidence. It is a beautiful piece of writing for the philosopher and the mathematician. For others, some words of introduction are necessary, and these can be found in the editors' introductory comments to Chapter 3.

Chapter 4, on transfer evidence, provides a detailed discussion of the issues currently facing forensic scientists endeavouring to analyse transfer or contact evidence. Summary tables provide clear lists of the various issues which have to be considered here.

Chapter 5 is technical. In some cases it is possible to consider the two probabilities for the evidence, one assuming guilt and one assuming innocence. However, in many cases it is not possible to derive these probabilities, and it has been necessary to restrict the examples and discussion to the consideration of significance probabilities and confidence intervals, the so-called frequentist approach. Many people will be more familiar with this approach than with the approach based on the likelihood ratio. However, there are considerable problems of interpretation associated with significance probabilities, as explained by Professor Lindley in the first chapter. As research progresses, methodology based on the likelihood ratio will gradually replace current frequentist methodology. In certain situations, such as bloodgrouping, implementation of the likelihood ratio approach is relatively straightforward; the existence of background data enables this approach to be taken. Other situations, for example in the interpretation of hair evidence, are more complicated. However, even there, it will be seen that careful consideration of the component probabilities of the likelihood ratio leads to a greater understanding of the unre solved problems, and, hence, to a better understanding of how they may be resolved.

Chapters 6 and 7 provide an informal introduction to current developments in expert systems and quality control, both important new areas in forensic science. The main purpose of the chapter on expert systems is to emphasize their potential rather than their current usage which is rather limited. The role of quality control in forensic science is obviously important. One needs only consider current controversy surrounding DNA profiling to appreciate this. Chapter 7 provides a background to this role.

Statistical tables are not provided, as these are readily available elsewhere; see, for example, Lindley, D.V. & Scott W.F. (1988), *New Cambridge elementary statistical tables,* Cambridge University Press.

The editors have striven to ensure uniformity of notation, and to a lesser extent, terminology, throughout the book. For example, *P(x)* is used to denote the probability of an event x; *P(x|y)* is used to denote the probability of the event *x,* conditional on the occurrence of another event *y* or 'given that' another event *y* has occurred. The upper case letter G is used to denote *guilt;* its negation, *not guilty,* is denoted $\bar{G}$, read as G-bar. The upper case letter I is used to denote *information* and is not to be confused with innocence. The upper case letter E is used to denote *evidence,* usually one particular set of evidence, such as fibres or paint fragments. Thus, the expression *P*(G|E,I) is short-hand for 'the probability that a person is guilty, conditional on the particular set of evidence under consideration and on any background information currently available'. Further discussion of this is found in the body of the text.

Readers will note that a certain result is referred to as Bayes' Theorem in parts of this book and as Bayes' Rule in Chapter 1. It is a matter of personal taste whether one wishes to call this result a theorem or a rule; the result is the same under either name, the editors have decided not to interfere with the personal tastes of the authors.

ACKNOWLEDGEMENTS

This book could not have come into being without the contributions from its authors. The editors are very grateful for the efforts of all those who have contributed, G. Gettinby, I.W.Evett, J.Robertson, D.V.Lindley, I.J.Good, W.F.Rowe, G.A. F.Seber, R.N.Curnow, D.A.Berry, J.C.Wright, M.A.Green, J.Buckleton, K. Walsh, T.A.Kubic, and J.Buscaglia, especially for taking time from busy lives to write in response to the pesterings of editors, perhaps from thousands of miles away and perhaps whom they have never met. The editors are grateful also for the implicit support the authors have given to the venture. No book before has attempted to explain the mysteries of statistical interpretation to forensic scientists; time will tell if we have been successful.

Particular thanks are owed to three people. First, thanks are due to Jim Robertson, the series editor, for suggesting the subject area and inviting us to be the editors and who also contributed to the book himself. Second, thanks are due to Ian Evett, from whom many of the original ideas in the application of statistics to forensic science have come, for contributing the Foreword, and, thus, forming the context of the remainder of the book. Third, thanks are due to George Gettinby for compiling Chapter 5, the chapter which considers several areas of application of statistics

within forensic science, and by so doing removing some of the editorial burden from ourselves. These three also provided valuable comments on the book, including the contributions from the editors themselves, as well as much needed moral support.

Finally, thanks are due to the publishers, Ellis Horwood, for their patience while the manuscript suffered many delays and for their understanding of the problems caused by intercontinental communications between the United Kingdom, USA, and New Zealand.

1

Probability

D.V.Lindley
2 Periton Lane, Minehead, Somerset, UK.

1. THE MEASUREMENT OF PROBABILITY

The main point at issue in a trial in a court of law is whether or not the defendant is truly guilty of the crime with which he has been charged. Another issue is the sentence to be delivered in the event of the judgement of guilt; this is not our concern here. It is also true that the charge may be altered during the trial; for example, murder reduced to manslaughter. This is, from the viewpoint discussed here, a minor complication. Fundamentally, there is some charge and it is a question of whether it can be made to stick, that is, whether it can be proven to a standard of proof in the eyes of the jury. We shall always use the term 'guilty' to refer to the event that the defendant is truly guilty of the crime with which he has been charged. (The language may change slightly in a civil case.) We do *not* refer to the *judgement* of guilt. True guilt is an event whose truth or falsity is rarely determined. Nevertheless, it is a useful concept. The hope is that the judgement of guilt will be made only when the defendant is truly guilty.

The opinion of the jury about this central issue of guilt fluctuates during the course of the trial. As the prosecution presents its case, the jury may lean towards guilt, but when the defence appears, guilt seems less reasonable. The feature that we discuss here is the uncertainty surrounding the defendant's guilt. Throughout the trial, the jury is uncertain about this key issue. The hope is that at the end of the proceedings the uncertainty will have been reduced to a very small amount, and the jury will feel convinced of the defendant's guilt or innocence. There are many cases where the uncertainty remains: according to English criminal law, the judgement is then 'not guilty'. But whatever happens, this uncertainty fluctuates during the trial. The cause of this fluctuation is the presentation of scientific, or other, evidence to the court, raising or lowering the uncertainty according to the nature of that evidence. The question of varying uncertainty is at the heart of any trial in a court of law.

The obvious way for a scientist to study any concept is to try to measure it; for measurement, numeracy is central to the scientific method. A great advantage of numbers is that they combine easily according to well-known rules of addition and multiplication. Consequently, if a concept occurs in several contexts, measurement of it enables the contexts to be combined. And this is exactly what is wanted in a trial. There are many causes of uncertainty—was he at the scene?, is he of the right blood type? —and overall judgement is possible only by combining all the different causes, by putting all the pieces of evidence together. As scientists, then, we should like to measure the uncertainty that is omnipresent in the trial. If this could be done, the various uncertainties could be combined by the rules of arithmetic to reach a final uncertainty concerning guilt embracing all the evidence. So the question is: how can uncertainty be measured?

There have been several answers to this question. We are going to discuss one answer, which we believe to be the correct one; consider its consequences and only then refer to other answers and explain why we think they are wrong. How can uncertainty be measured? By probability.

2. A STANDARD FOR UNCERTAINTY

The essence of any measurement process is comparison with a standard. Essentially, the measurement of the width of the desk at which I write is by comparison with a standard metre held in Paris, with copies in other cities. (Nowadays comparison with the wavelength of sodium light may be used: but the point of a standard remains.) Every measurement, whether it be length, temperature, pressure, or whatever, requires a standard. We, therefore, require a standard for uncertainty. Notice that the physical act of measurement is not usually accomplished by reference to the standard: it is not necessary to take the desk to Paris or London. The physical measurement is done indirectly. No one has ever measured the distance between Paris and London. The value is determined by the direct measurement of other distances and angles and their combination according to the rules of geometry. Consequently when, in the next paragraph, we propose a standard for uncertainty, do not imagine that it is suggested that a direct comparison is necessary between the forensic evidence and the standard, any more than that it is necessary to take the desk to Paris. The standard is merely for comparison: actual measurement will be accomplished by indirect methods, many of which will be discussed in later chapters. The immediate task is to measure uncertainty by constructing a standard.

Many standards have been proposed; the following is perhaps the simplest. Imagine an urn, a container, opaque so that the contents cannot be seen, but with a narrow neck, sufficiently wide for a hand to be inserted and some of the contents withdrawn. Suppose the urn contains a number of balls, all as identical as modern mass-production processes can make them, except that some are painted black and the remainder white. Denote by b the proportion of black balls and by w, that of white, so that necessarily $b + w = 1$. Now suppose a hand is inserted into the urn and a single ball drawn *at random.* Let us explain what is meant by 'at random'.

Imagine the balls were numbered consecutively 1, 2, 3 and so on up to N, the total number in the urn: each ball having a different number, and every number up to N occurring. Suppose that you were to be made the following offer: you withdraw a ball from the urn; if the ball is numbered 17 you get £10, otherwise nothing. (We suppose there are at least 17 balls in the urn.) Suppose the number 17 were replaced by any other number up to *N,* say 31, and then suppose you did not mind which number was made in the offer, 17 or 31 or any other. In other words, you do not mind upon which number the £10 is contingent. Then we say the ball is withdrawn 'at random'. Alternatively expressed, for you each numbered ball has the same chance of being drawn. You will appreciate that such a device of an urn and a withdrawal at random can be constructed, just as can a standard metre. Indeed, essentially this device is used in Britain to draw numbers for premium bond prizes: it is called ERNIE (Electronic Random Number Indicator Equipment). There are difficulties, like people thinking 13 is unlucky, but there are difficulties with the standard metre, and these can be overcome or reduced to insignificant amounts. So we now have an urn, with black and white balls (forget the numbering) and the concept of a single ball drawn at random.

The randomly drawn ball can either be black, an event B, or white, an event W. One and only one of these two events must occur. Consider the uncertainty of the event B of a black ball. It depends on b, the proportion of black balls in the urn. If b is small; B is unlikely; if b is large, B is highly likely; if b is near $\frac{1}{2}$, B is about as likely as W. We shall refer to b as the *probability* of getting a black ball on a single, random drawing from the urn. We now have a measured standard for uncertainty: the uncertainty of the ball being black is b, the probability of a black ball, equated to the proportion of black balls. Notice that probability is simply a proportion and can vary between 0, no black balls, and 1, no white balls, and all values between these extremes are possible (for very large numbers of balls in the urn).

Let us return to the uncertainty of the defendant's true guilt at some point of the trial. Denote the event of true guilt by G. Compare the two events, G and B; the events of guilt and of a black ball being drawn at random. In other words, compare the event of interest with the standard. Now just as the desk could be compared with the standard metre and the conclusion made that the desk was 1.6 metres wide, so, we suggest, a comparison could be made between guilt and the urn with a value b for the proportion of black balls found such that the events G and B are equally uncertain. Just as 1.6 standard metres match the desk, so a standard urn with $b = 0.16$ could match the guilt. Remember, there is no suggestion that either the desk or guilt methods are practical possibilities; they are merely conceptual procedures. Nevertheless both could be used and there are some uncertain events whose probability can be measured in this way. If you were about to toss an ordinary coin, you might well take $b = \frac{1}{2}$, judging that the uncertainty of heads matched that of a black ball from an urn with equal numbers of black and white balls. A forensic scientist about to determine the blood type of a bloodstain may well equate the uncertainty of rhesus negative with the proportion of rhesus negative people in the population: black being equated with rhesus negative, and the balls with the individuals in a population (see section 19). So the method can sometimes be used. It cannot easily be used with the event of main interest, the defendant's guilt. Indirect comparisons will be required there, as with the distance between Paris and London, and our next task is to look at such indirect methods. These will depend on the combination of uncertainties, or probabilities, and so the task must be to study just how probabilities combine. Before doing this, let us recapitulate the argument so far.

Basic to the conduct of a trial is uncertainty surrounding the defendant's guilt. This uncertainty fluctuates during the course of the trial as evidence, scientific or otherwise, is presented. As scientists, we wish to measure this uncertainty. All measurement involves comparison with a standard. A possible standard for uncertainty is the withdrawal of a ball at random from an urn with a proportion *b* of black balls. The uncertainty that the ball is black is *b,* and is called the probability of the event B of a black ball. Probability is the standard measurement of uncertainty, and it is proposed to compare all uncertainties with the standard: that is, using probability.

3. THE CONVEXITY RULE

The next task is to discover the rules of probability. In the measurement of distance, the rules are those of Euclidean geometry, and we determine lengths and angles, combining these according to geometrical ideas to produce the final measurement required. So with probability, we easily measure tosses of coins or blood types, and use the rules of probability calculus (rather than Euclidean geometry) to evaluate the probability of interest, that of guilt.

Some notation is needed. For any event I, we write $P(\mathrm{I})$ for the probability of I. The standard says that the event B of a black ball has probability $P(\mathrm{B}) = b$. We wish to determine $P(\mathrm{G})$, the probability of the event G that the defendant is truly guilty. One rule has already been encountered: it is often called the convexity rule.

CONVEXITY. Probability can assume any value between 0 and 1, both inclusive, and only those values. For an event I that you know to be impossible, the probability is zero.

The first sentence follows because the proportion *b* of such balls to which probability is compared must lie between 0 and 1. The second follows similarly because the only way for the event of a black ball to be impossible is for there to be no black balls in the urn, $b = 0$. It is sometimes necessary to strengthen the rule to say a probability can *only* be 0 when the event is known to be impossible. The reason for this will appear in section 14. In our notation, the rule can be restated.

CONVEXITY. For any event I, $0 \leq P(\mathrm{I}) \leq 1$. If I is impossible, $P(\mathrm{I}) = 0$.

The convexity rule is especially simple because it concerns only one event. The two other rules that follow involve several events and their combinations. It has already been explained how important it is to combine uncertainties; for example, these may be of different pieces of evidence. First, it is necessary to combine the events whose uncertainty, or probability, is under discussion. Events can combine in two ways. Let I and J be any two events. One combination is 'I and J', the event that occurs if, and only if, I and J *both* occur. It is called *the conjunction* of I and J. If I is the event that a child's mother is dead and J the event that the child's father is dead, then 'I and J' is the event that the child is an orphan. A second combination is 'I or J', the event that occurs if either I or J (or both) occurs. It is called the *disjunction* of I and J. Let I be the event that the first link in a chain of two links breaks: let J be similarly defined for the second link. The 'I or J' is the event that the chain breaks, since this will occur if either link breaks.

4. THE ADDITION RULE

The second rule of probability concerns two events, I and J, and their disjunction 'I or J', in the special case where their conjunction 'I and J' is impossible. Two events whose conjunction is impossible are called *mutually exclusive*: the occurrence of one excludes the possibility of the other. If I is the event of blood type A, and J that of blood type O, then I and J are mutually exclusive, the event 'I and J' is impossible. To discover the rule, suppose the balls in the standard urn are either red, yellow, or white. Balls that are red or yellow are termed coloured. Let r, y, and w be the proportions of balls of the three colours. Necessarily, $r + y + w = 1$. More importantly, the proportion of coloured balls is $r + y$, the sum of the proportions of red and of yellow. Let I and J be two mutually exclusive events. Then by the standard comparison we can equate the occurrence of I with the drawing of a red ball and, with a suitable choice of r, $P(\mathrm{I}) = r$. Similarly, J may be equated with a yellow ball and $P(\mathrm{J}) = y$. Since I and J are mutually exclusive, the event 'I and J' is impossible and has probability 0 by the convexity rule, agreeing with the fact that no ball is both red and yellow. The equating of I with 'red' and J with 'yellow' means that 'I and J' is equated with 'coloured', an event of probability $r + y$, the sum of the probabilities of I and J. We therefore have the additive rule of probability.

ADDITION. If I and J are mutually exclusive events (their conjunction is impossible), the probability of their disjunction, I or J, is equal to the sum of probabilities of I and J. In symbols,

$$P(\mathrm{I\ or\ J}) = P(\mathrm{I}) + P(\mathrm{J}) \quad \text{if} \quad P(\mathrm{I\ and\ J}) = 0.$$

The addition rule says nothing more than that proportions of mutually exclusive things add. For example, the blood groups A and AB are mutually exclusive, and the proportion of people with antigen-A is the sum of the proportions of type A and of type AB:

$$P(\text{antigen-A}) = P(\mathrm{A}) + P(\mathrm{AB})\ .$$

Notice the importance of the condition that the two events be mutually exclusive. Suppose that some balls were both red and yellow, so that on withdrawal of a random ball both the event 'red' and the event 'yellow' could occur simultaneously. Then the proportion of coloured balls would not be $r + y$ since the balls painted both red and yellow would be counted twice. The interested reader can easily work out that the general addition rule is

$$P(\text{I or J}) = P(\text{I}) + P(\text{J}) - P(\text{I and J})$$

for any events I, J. It is rarely used; the previous form with I and J mutually exclusive is more commonly employed.

5. CONDITIONAL PROBABILITY

The third, and last, rule of probability is subtler than the other two, and its presentation in a paper by Thomas Bayes in 1763 (Bayes, 1764, Barnard, 1958, Pearson & Kendall, 1970) remains a landmark in scientific thinking. Like the addition rule it considers two events, but concentrates on their conjunction, 'I and J' without the restriction of their being exclusive. Before the rule can be stated, it is necessary to repair an omission in the argument and a defect in the notation.

The uncertainty of the defendant's true guilt can conceptually be measured by its probability, P(G). It has been explained that this fluctuates during the trial as new evidence is produced. Yet all we have referred to is *the* probability of guilt. Neither the language nor the notation reflects the fluctuation in P(G). This defect must be remedied. Any uncertainty is assessed in the light of the knowledge possessed at the time of the assessment. Denote this knowledge by K. It is then usual to speak of the probability of an event I, given knowledge K; and to write $P(\text{I}|\text{K})$, the vertical line having the meaning 'given'. Thus at some point in a trial the uncertainty is $P(\text{G}|\text{K})$, where K is the knowledge then available. If additional evidence E is presented, the revised uncertainty is $P(\text{G}|\text{E and K})$, where 'E and K' is the conjunction of E and K.

An alternative, and useful, way of looking at the situation is to recognize that uncertainty, and hence probability, depends on two things: the event I whose uncertainty (probability) is being considered, and the knowledge K that you have when contemplating I. It is strictly nonsense to talk of the probability of I in the absence of K. Reference must be made to the conditions at the time of the assessment. The probability $P(\text{I}|\text{K})$, with two arguments, I and K, is sometimes referred to as *conditional* probability: I being conditional on K. This terminology will not be used here since, in our view, *all* probabilities are conditional, so the adjective is unnecessary. Sometimes K may not be explicitly stated, but it is always there implicitly.

The two rules, convexity and addition, need repair. The reason we could avoid reference to K in stating them is that K is fixed throughout the calculations. The correct, modified form is given in section 7. The third, product or multiplication, rule is different in that the knowledge changes; indeed, it explains how probability is affected by an increase in information. In particular it relates $P(\text{G}|\text{K})$ and $P(\text{G}|\text{E and K})$.

6. THE MULTIPLICATION RULE

To obtain the rule, let us return to the urn with black and white balls with proportions b and w respectively, $b + w = 1$. In addition to the coloration, suppose that each ball is either spotted or plain, with proportions s and p, $s + p = 1$. There are then four types of ball: ‘black, spotted’, ‘black, plain’, ‘white, spotted’, and ‘white, plain’. The proportions are conveniently displayed in a Table:

	Black	White	
Spotted	*c*	*d*	*s*
Plain	*e*	*f*	*P*
	b	*w*	1

Thus the proportion of black, spotted balls is *c*. There are several constraints: necessarily $b + w = s + p = 1$, but also $c + d = s,$ $c + e = b,$ etc. Let B be the event of the ball drawn at random being black: let S similarly refer to it being spotted. The conjunction ‘B and S’ means the random ball was ‘black, spotted’. On identifying probabilities with proportions, we can write $P(\mathrm{B}|\mathrm{K}) = b$, as before, except that now we explicitly introduce the knowledge K.K contains the knowledge of the constitution of the urn and that the drawing was random. Similarly $P(\mathrm{S}|\mathrm{K}) = s$. Also $P(\mathrm{B}\text{ and }\mathrm{S}|\mathrm{K}) = c$.

Now introduce a new idea: suppose that someone else had withdrawn the ball at random and told you truthfully that it was black, event B. You are still uncertain whether it is spotted or plain. What is your probability that it is spotted? In symbols, what is $P(\mathrm{S}|\mathrm{B}\text{ and }\mathrm{K})$? Notice your knowledge has changed; it not only includes K but has the additional fact, B, that the ball is black. Clearly this probability must be, on identifying probability with the standard proportion, the proportion of spotted balls *amongst the black ones*. This is *c/b,* as is easily seen from the Table. Next, consider the trivial result that

$$c = b(c/b).$$

In words, the proportion of balls that are both black and spotted, c, is the proportion that are black, *b,* times the proportion of spotted ones amongst the black, *(c/b).* Turning the proportions into probabilities, we equivalently have

$$P(\mathrm{B}\text{ and }\mathrm{S}|\mathrm{K}) = P(\mathrm{B}|\mathrm{K})P(\mathrm{S}|\mathrm{B}\text{ and }\mathrm{K}).$$

This is the third rule of probability and, because of the multiplication involved, is called the multiplication rule.

7. THE BASIC RULES

The three rules are now stated in the new notation. For events I, J, and knowledge K.

CONVEXITY $\mathrm{O}\leq P(\mathrm{I}|\mathrm{K})\leq 1$ and $P(\mathrm{K}|\mathrm{K}) = 1$;
ADDITION If ‘I and J’ is impossible given K,
$P(\mathrm{I}\text{ or }\mathrm{J}|\mathrm{K}) = P(\mathrm{I}|\mathrm{K}) + P(\mathrm{J}|\mathrm{K})$;
MULTIPLICATION

$$P(\text{I and J}|\text{K}) = P(\text{I}|\text{K})\,P(\text{J}|\text{I and K}).$$

Aside from the explicit reference to K in the statement of the convexity and addition rules, one other change has been made. In the former, reference has been made to the upper end of the scale, 1, rather than 0. If K is known, K is certain, so $P(\text{K}|\text{K}) = 1$.

These three rules describe the basic properties of probability as a measure of uncertainty. All other results in probability can be deduced from the three basic rules. Although the rules are extremely simple, being merely the reflection of elementary properties of proportions of balls in an urn, they lead to results of considerable complexity because they allow probabilities to combine in two different ways; addition and multiplication. There are few quantities for which this is true. Lengths may be added, but their multiplication leads to a new concept, area. This richness of possibilities for combination leads to a beautiful and important calculus of probabilities. Since some probabilities are easy to assess, like tosses of a coin or blood types, these may be used in the rules to calculate other probabilities. Examples of these calculations occur in other chapters of this book. We now consider another question: why measure uncertainty in this way, through proportions leading to probability? Until there is a satisfactory answer to this question our approach has an element of arbitrariness to it.

8. OTHER MEASURES OF UNCERTAINTY

First, notice that there is a group of people who study uncertainty and do not use probability; namely bookmakers. They are uncertain whether horse H will win the race, and they quote odds, say of 5–1 against. Odds replace probability. But odds and probability are equivalent. To see this we must introduce the event 'not-I', the negation or *complement* of I. It is the event which is true when I is false, and false when I is true. The complement of 'guilt' is 'innocence' (in English law).

ODDS. If an event I has probability $P(\text{I})$, the odds against I is $P(\text{not-I})/P(\text{I})$. (Here reference to K has been omitted for simplicity: all probabilities are given K.)

Since the event 'I or not-I' is certain, it has probability 1 and by the addition law, with 'I and not-I' impossible,

$$P(\text{I}) + P(\text{not-I})=1.$$

Hence the odds against I is [1—P(I)]/P(I). Thus 5–1 against corresponds to a probability of 1/6. A probability of 1/2 yields odds of 1, or evens. Odds against that are less than 1 are usually inverted and quoted as 'odds on'. Thus a probability of 2/3 gives odds against of 1/2 to 1, or odds on of 2 to 1. Bookmakers are therefore effectively agreeing with us in using probabilities, merely transforming them into odds. We shall see later that odds are useful in connection with the multiplication rule: they are inconvenient with the addition rule.

Other transforms of probability are in use. In some technical contexts it is convenient to use logarithms of odds. Meteorologists, at least in the United States, use probabilities as defined here but multiply them by 100. They still use the word probability as when they say ‘the probability of rain tomorrow is 20%’ replacing our 0.2. All these practices are essentially in agreement with the view put forward here. They are transforming probability, just as a surveyor transforms angles to tangents. Our dispute lies with those who use other measures of uncertainty that do not transform to probability.

The major difference between these other measures and probability is that they use different rules of combination from the addition and multiplication rules. Thus a proposal, derived from the logic of fuzzy sets, is to measure uncertainty by possibilities, rather than probabilities. Thus *ps*(I) replaces *P*(I). The addition rule for exclusive events I and J is replaced by

$$ps(\mathrm{I\ or\ J}) = \max\{ps(\mathrm{I}), ps(\mathrm{J})\}.$$

That is, addition of probabilities is replaced by maximization, taking the larger, of two possibilities. Another idea is to use beliefs and refer to the belief in I, *B*(I). Now, it is not true, as it is with probability, that

$$P(\mathrm{I}) + P(\text{not-I}) = 1\ ;$$

we merely have the weaker condition

$$B(\mathrm{I}) + B(\text{not-I}) \leq 1\ .$$

So an alternative phrasing of the question: why use probability to measure uncertainty? is to ask: why use addition and multiplication? Before answering that, let us look at another measure of uncertainty that superficially looks like probability but, in fact, is not. It is necessary to do this because of the measure’s popularity amongst scientists. It is called a significance level. Rather than present a general description, we describe it in a familiar, forensic science context.

9. SIGNIFICANCE LEVELS

Suppose that in forcing entry into a building, a window has been broken. Later, a suspect is apprehended and a fragment of glass is found on his clothing. The uncertainty lies in whether the fragment came from the broken window. Suppose that the investigation uses the refractive indices (r.i.) of the two pieces of glass. Let μ_o be the value for the window and x that for the fragment. The uncertainty is now more specific and concerns whether the fragment could have come from glass of r.i. μ_o, it being recognized that measurements of r.i. of fragments may differ from the mean value of those of the window from which they truly came. This variation is almost always expressed in terms of probability, and here it might reasonably be described by saying that if the fragment came from a window of r.i. μ, x would have a Normal probability distribution with mean μ and standard deviation (s.d.) σ. The original uncertainty can now be described by investigating whether μ has the value μ_o of the broken window. So far there is no divergence from our probability approach.

The significance level argument now proceeds as follows. Suppose $\mu = \mu_o$; that is, the fragment came from a window of r.i. μ_o. Then the deviation of x from μ_o, without regard to sign, should be small. How small depends on the s.d. σ. Tables of the Normal distribution enable one to calculate the probability of a deviation of x from μ_o as large as, or larger than, the one observed. For example, if $\mu_o = 1.516723$, $x = 1.517933$ and $\sigma = 0.0022$; the deviation is 0.00121, or 0.55 times the s.d. Reference to the tables shows that the probability of a deviation at least as large as that observed, 0.00121, is 0.58. Here this probability is quite large and a deviation of that amount quite reasonable. Had the deviation been four times as large, at 0.00484, the probability would have been down to 0.0028 and the deviation is unlikely. The probability here calculated is called the significance level and is used as a measure of the uncertainty of whether $\mu = \mu_{o.}$ If it is high, then μ_o is thought to be likely; if small, then doubt is cast on the value μ_o (and some value of μ nearer to x thought more reasonable).

Here, then, in the significance level, we have a measure of uncertainty that is based on probability, so that it might appear to be in agreement with the urn approach adopted above. But appearances are misleading. The basic uncertainty is whether $\mu = \mu_{o.}$ According to our view, this should be expressed through the probability that $\mu = \mu_o$, given data x and general background knowledge, including Normality and the value σ of the s.d.: $P(\mu = \mu_o|\mathrm{x}, \sigma, \mathrm{K})$. But this is not the probability calculated in a significance level, which is a probability of a deviation given μ_o (and σ, K). Simplifying and abusing the notation and description to emphasize the distinction: we advocate a probability for μ, given *x;* a significance level uses a probability for x, given $\mu = \mu_{o.}$ In other words, the roles of μ and x are reversed in the two probabilities. (We shall return to the distinction between $P(\mathrm{I}|\mathrm{J})$ and $P(\mathrm{J}|\mathrm{I})$ in section 11.)

Consequently, a significance level, although a probability, is not of the same type as our theory would require. That the two are different may also be appreciated by the fact that they combine in different ways. The method here uses the multiplication rule. Statisticians have discussed how significance levels from two different data sets may be combined; the result is not multiplication. The following example provides an extreme illustration. Suppose the r.i. calculations gave a significance level of 0.05. Suppose that it was also possible to measure the concentration (c) of a given element of the glass, that the two measurements were quite independent, and that similar calculations for c also gave a level of 0.05. Thus both levels being small, both cast doubt on the fragment having come from the window. If the data for r.i. are combined with those for c, it can happen that the combined level exceeds 0.05, say 0.07. In other words, although both c and r.i. measurements cast doubt on the fragment having come from the window, in combination they cast less doubt. One would naturally expect that the two pieces of evidence would reinforce one another, as they would with the multiplication rule: with significance levels they do not always do that.

Some time has been spent on significance levels because they are popular in science and have been used extensively in forensic science. The previous discussion shows that although probability based, they are not probabilities of the uncertain quantity (μ), but of something else *(x).* It also shows that levels combine in a way that can be surprising and is not in agreement with the multiplication law. However, neither of these issues addresses the question of which method is correct, or whether some other rules, of possibilities or beliefs, may not be better. Let us look at this matter.

10. SCORING RULES

A simple and effective reason for using probability is that it works. I know of no situation in which probability is inadequate or fails to yield a reasonable answer. Sometimes the calculations are horrendous and cannot at the moment be done: but that is a technical difficulty that adequate research will, one hopes, overcome. Sometimes it is hard to relate the probability model to the actual situation: but again, when it can be done the result is illuminating. So, pragmatism is one reason.

A second reason is that probability has been recognized for three centuries as being the only way to handle uncertainty in games of chance; card games, roulette, etc. For nearly as long it has been successfully used in physics and in actuarial problems. There appears to be no essential difference between the uncertainty that occurs in these subjects with uncertainties in other fields such as politics or law. Indeed, the law does weakly recognize probability as in the phrase 'balance of probabilities' for civil cases, and indirectly reference is made to 'beyond reasonable doubt'. At any rate, no one seems able to differentiate between different types of uncertainty so that one type could be classified as like games of chance and another like law. Occam's razor, the desire to keep things as simple as possible, would also support the notion of a single concept.

This century has seen the emergence of a strong, practical reason for using probability (or a transform thereof, like odds) as the unique measure of uncertainty. There are several forms of the argument; a simple form goes like this. Suppose you wish to measure uncertainty in some way that appeals to you: significance levels, possibilities, or any other way. Suppose now that, on being asked to describe the uncertainty of an event, you provide, using your method, a number t. It is proposed to score you in the following manner. If the event turns out to be true, you will receive a score $(t - 1)^2$: if false t^2. The idea is to get t near 1 if the event is true, and near 0 if false, the score being thought of as a penalty score. The method cannot be used with an event like the defendant's guilt because its truth cannot be unambiguously determined. But it can be, and has been, used for the event 'rain tomorrow' when a meteorologist has measured its uncertainty as t. If $t = 0.8$ (or 80%) the score will be 0.04 if it rains and 0.64 if not. The confidence in rain, expressed by a value near 1, will be heavily penalized by a large score of 0.64 if there is no rain.

Now imagine this is done for several events—three, suitably chosen, may be adequate—and the scores added. Then de Finetti (1974) demonstrated a remarkable result. A person who uses *any* method other than probability could have obtained a smaller score had probability been used; and this is for whatever outcome, true or false, of the events. Here is a little numerical illustration.

Suppose when asked for the uncertainty of I, the value 0.4 is given; and for the uncertainty of 'not-I', the complement of I, the value 0.3 is given. (This is sensible for someone using beliefs where the two values may add to less than 1: for probabilists the sum must be 1.) Let these two evaluations be scored. If I is true the first will score $(0.4\text{-}1)^2 = 0.36$ and the second $(0.3)^2 = 0.09$, a total of 0.45. (In explanation of the second: if I is true, 'not-I' is false so the score is t^2 and here $t = 0.3$.) If I is false, the total is $(0.4)^2 + (0.3 - 1)^2 = 0.65$. Suppose, however, our 'believer' had used 0.55 for I and 0.45 for not-I, adding to 1. With I true, the score is $(0.45)^2 + (0.45)^2 = 0.41$, less than the original 0.45. With I false, the score is $(0.55)^2 + (0.55)^2 = 0.61$, again less than the original 0.65. So when the original assessments of 0.4 and 0.3 were made, the 'believer' knew that a lower score would be bound to arise, whether I was true or false, for the choices 0.55 and 0.45. How absurd then are the belief values. More complicated examples exploit the multiplication law. Consequently the score criterion makes probability a winner over any other system, whatever happens.

The alert reader will immediately raise an objection. Suppose different scores were used: why 1 and 0, strongly suggestive of probability, or the squares $(t - 1)^2$ and t^2? The response is simple. Glossing over a few points of detail, there are two possibilities.

(1) The score function will lead to a transform of probability, like odds. (100 and 0 replacing 1 and 0, but retaining the squares, will give percentages, as used by meteorologists.)

(2) The score function will lead to every event being assigned one of only two numbers. (The numbers might be 0 and 1, so every event would be described as 0 or as 1.)

Now (2) is absurd: it is like describing every uncertain event as 'true', 1, or 'false', 0; a device which fails to capture uncertainty adequately. (Absurd it may be, but it is used in law when the jury's uncertainty has to be described as 'guilty' or 'not guilty': only two values are allowed. In our system, the jury would announce their probability that the defendant is truly guilty. 'Have you reached agreement, members of the jury?' 'We have, your honour, our probability is 0.87.' But this is taking us too far ahead.) There are only two types of scoring system: those that lead to probability, or a transform thereof, and those that yield a dichotomy essentially equivalent to 'true' and 'false'. Dismissing the second, only probability remains.

Let us recapitulate the argument so far. Uncertainty is a common feature of life, and of a law court in particular. Scientific thinking encourages measurement of uncertainty. All measurement is based on comparison with a standard. A readily available standard for uncertainty is probability based on urn models. Probability satisfies three basic rules, convexity, addition, and multiplication. Probability has several advantages over other measures of uncertainty, like significance levels or possibilities. One advantage is pragmatic. The principal advantage lies in the fact that under any reasonable scoring system, probability measurement will achieve a lower score than any other measurement. This is the case for probability.

If probability is accepted as *the* measure then it is necessary to investigate further its properties beyond the basic rules and to gain experience in applying it. These topics occupy the remainder of this book. In the rest of this chapter, we shall discuss the other rules of probability that are important in many applications, describe a simplification that is often sensible, and finally make some miscellaneous comments on probability.

11. BAYES' RULE

First, a slight, but important extension is made to the multiplication rule. The latter says, for any two events, I and J, that

$$P(\mathrm{I\ and\ J}) = P(\mathrm{I})\ P(\mathrm{I}|\mathrm{J})\ .$$

Here explicit reference to the background knowledge K has been omitted from the notation as an aid to clarity. It will stay constant throughout the following presentation, and its omission should not mean that it is forgotten. Now the event 'I and J' is the same as the event 'J and I', so I and J may be interchanged in the rule without affecting the left-hand side. Consequently, the right-hand side must be similarly unaffected by the interchange. Hence

$$P(\mathrm{I})P(\mathrm{J}|\mathrm{I}) = P(\mathrm{J})P(\mathrm{I}|\mathrm{J})\ .$$

(The interested reader may like to think of this result in terms of the proportions of balls in the Table.) If $P(\mathrm{I}) \neq 0$, we can divide by it and obtain

BAYES' RULE

$$P(\mathrm{J}|\mathrm{I}) = P(\mathrm{I}|\mathrm{J})\ P(\mathrm{J})/P(\mathrm{I}).$$

The importance of Bayes' rule lies in its connecting P(J) with P(J|I), showing how the uncertainty about J on the right-hand side is changed by the knowledge of I (in addition to K) to that on the left. This change will be discussed further below: for the moment, let us note that the connection between $P(\mathrm{J})$ and $P(\mathrm{J}|\mathrm{I})$ involves $P(\mathrm{I}|\mathrm{J})$. The probabilities $P(\mathrm{J}|\mathrm{I})$ and $P(\mathrm{I}|\mathrm{J})$ are often confused, so let us make clear the distinction between them. (The point arose when discussing significance levels in section 9.) In $P(\mathrm{J}|\mathrm{I})$, the event J is uncertain whilst I is known or given. In $P(\mathrm{I}|\mathrm{J})$, the uncertainty resides in I, whereas it is J that is known. Here is an example where the two probabilities can be confused.

(1) The death rate amongst men is twice that amongst women.
(2) In the deaths registered last month there were twice as many men as women.

It is sometimes thought that (1) and (2) are alternative expressions of the same fact: in fact, they are quite different. To see this let F be the event of being female, so that not-F is male, denoted M; let D be the event of death. In the language of probability, the statements can be rewritten

(1) $P(\mathrm{D}|\mathrm{M}) = 2P(\mathrm{D}|\mathrm{F})$,
(2) $P(\mathrm{M}|\mathrm{D}) = 2P(\mathrm{F}|\mathrm{D})$,

emphasizing the reversal of the roles of F and D. Statement (2) implies, since M is not-F the probabilities of complementary events add to 1, that P(M|D) = 2/3 and *P*(F|D) = 1/3. No such implication follows from (1). To demonstrate numerically that (1) and (2) are different, consider the modified form of the Table of section 6 with sex replacing colour and death, spotted.

	Male	Female	
Dead	2	1	3
Alive	98	199	297
	100	200	300

(All the proportions have been multiplied by 300 to avoid fractions.) From the first row we see that amongst the dead there are twice as many males as females, in agreement with (2). The vertical margin says the death rate for the month was 1/100, equal to *P*(D), a reasonable figure. The other margin reveals that there are twice as many females as males, *P*(F) = 2/3; again a reasonable figure amongst senior citizens. Looking at the columns in the body of the Table, the death rate for males from the first column is 2/100, *P*(D|M) = 0.02, whereas the second column gives a female death rate of only 1/200, *P*(D|F) = 0.005, a quarter that of the males. Hence *P*(D|M) = 4*P*(D|F) in disagreement with (1). The reader may like to check that (1) and (2) are in agreement only if the two sexes are present in equal numbers, *P*(M) = *P*(F) = 1/2, so that the marginal row of the table has entries 150 and 150 (total 300).

12. ODDS FORM OF BAYES' RULE

Having emphasized the distinction between *P*(I|J) and *P*(J|I), let us return to Bayes' rule as a way of converting *P*(J) into *P*(J|I), using *P*(I|J). To appreciate the rule it is easier to write it in terms of odds rather than probability. Remember, from section 8, the odds *on* J, say, are *P*(J)/*P*(not-J). (Henceforth, odds *on* will be used in preference to odds *against.*) As well as the original form of Bayes' rule

$$P(\mathrm{J}|\mathrm{I}) = P(\mathrm{I}|\mathrm{J})\ P(\mathrm{J})/P(\mathrm{I})$$

there is a form with 'not-J' replacing J:

$$P(\text{not-J}|\mathrm{I}) = P(\mathrm{I}|\text{not-J})\ P(\text{not-J})/P(\mathrm{I})\ .$$

The ratio of the right-hand sides of these two results must be equal to the ratio of the left-hand sides, so

$$\frac{P(\mathrm{J}|\mathrm{I})}{P(\text{not-J}|\mathrm{I})} = \frac{P(\mathrm{I}|\mathrm{J})}{P(\mathrm{I}|\text{not-J})} \cdot \frac{P(\mathrm{J})}{P(\text{not-J})}\ ,$$

P(I) cancelling. Writing O for odds, this result is

BAYES' RULE (ODDS FORM)

$$\mathrm{O}(\mathrm{J}|\mathrm{I}) = \left\{\frac{P(\mathrm{I}|\mathrm{J})}{P(\mathrm{I}|\text{not-J})}\right\} \cdot \mathrm{O}(\mathrm{J}) \; .$$

In this form, the odds on J (given K) are easily changed on receipt of additional information I by multiplying by the quantity in braces. (Readers familiar with logarithms may prefer to work with log-odds, when the change is effected by *addition* of the logarithm of the quantity in braces.)

Important as Bayes' rule is, it is only a simple result for proportions. Consider the original Table in section 6 with black/white and spotted/plain balls where the randomly selected ball was black, this constituting the information, here I. Now

$$\frac{c}{e} = \left\{\frac{c/s}{e/p}\right\} \cdot \frac{s}{p}$$

trivially. Here *s/p* is the original odds on the ball being spotted; and *c/e* is the same odds when the ball is black. One is obtained from the other by multiplying by the quantity in braces. Consider this quantity. It plays an important role and deserves a name: it is called a *likelihood ratio* or *Bayes' factor,* under which name it is discussed further in Chapter 3.

The likelihood ratio is, in general, the ratio of P(I|J) to P(I|not-J); in the urn illustration, the ratio of *c/s* to *e/p*. Both probabilities refer to the same uncertain event I; one when J is true, the other when J is false. Similarly, *c/s* is the proportion of black balls amongst spotted ones, and *e/p* is the proportion of black balls amongst the plain. In order, therefore, to pass from the odds on J (given K) to the new odds, including information I as well, it is necessary to multiply the former by the likelihood ratio, which is the ratio of the probabilities of the information I, both when J is true and when J is false. Notice the inversion here: to calculate odds on J (or equivalently probabilities for J) we need probabilities of I (when J is true and when false). The distinction discussed above between P(I|J) and P(J|I) is vital.

If the information I is twice as likely under J as under 'not-J', the likelihood ratio is 2 and the original odds on J are increased by multiplication by 2. If I is only half as likely under J as under 'not-J', the ratio is 1/2 and the original odds are reduced on division by 2. A vital lesson here is that in considering new information I, you have to think how probable I is when J is true and also when J is false. If these are equal, the ratio is 1 and the information has no effect on the odds. It is a common mistake to think that an event I, unlikely on J, provides evidence against J. This is not so unless additionally I is more likely when J is false, so that the likelihood ratio is small.

A good example occurred whilst this material was being written. An enquiry was being conducted into the possible erection of a nuclear power station near where the author lives. An objector had pointed out the excessive number of leukaemia cases near the existing station. 'Where 4 would have been expected, 10 were observed'. The objector then went on to cite a probability of 0.0081. This is the probability of 10 or more deaths, where the average is 4. (This is a sort of significance level: where $\mu = \mu_o$, here 4, the observed x, here 10, or more, has probability 0.0081.) In our approach, a relevant figure is the probability of the evidence, 10 deaths, given the national average, J of 4. This is 0.0053. But more importantly, the objector's figure takes no account of the probability of I, 10 deaths, given that the rate is *above* the national average, not-J, which is what he was trying to establish. If the local average were 10 ($\mu = 10$), the probability of 10 is 0.1251. This gives a likelihood ratio of 0.0053/0.1251 = 0.0423, dividing the odds on the national average holding near the station by about 24. Actually, this ratio is too high, for there are other possibilities besides $\mu = 10$. The point is that the probability 0.0081 (or the alternative 0.0053) of the observation given the national average is only half the story.

13. FORENSIC EVIDENCE AND BAYES' RULE

Bayes' rule in odds form has important consequences for the interpretation of forensic science evidence. Replace J by G, the event that the defendant is truly guilty. Then the rule enables the court to update the odds on guilt (given background knowledge and previous evidence) to the new odds given additional evidence E. This is done by multiplying the former odds by the likelihood ratio, incorporating the probabilities of the evidence assuming guilt and assuming innocence. So, all the court needs is this ratio. An elaboration, explained in Chapter 3, uses the logarithm of this ratio. The responsibility of the forensic scientist in acting as expert witness is to provide this ratio. In practice, it will be more meaningful to give both probabilities. Notice an important distinction: the court is concerned with the probabilities (or odds) of guilt, and the expert witness is concerned with the probabilities of the evidence. The forensic scientist should not respond to questions involving the uncertainty of G; only of E.

Here is a simple illustration. A crime has been committed in which the victim has bled. A suspect's clothing is found to have a bloodstain of the same blood group as the victim's. Making some simplifying assumptions, the likelihood ratio can be evaluated as follows. If the suspect is guilty, the match between the two blood groups is certain, $P(\mathrm{E}|\mathrm{G}) = 1$. If innocent, the match is coincidental and the probability that a stain will be of the victim's blood group will be p, the proportion of people in the population of that blood type: so $P(\mathrm{E}|\text{not-G}) = p$. The likelihood ratio is $1/p$ and the odds on G are multiplied by $1/p$. Thus, a rare blood type *(p* small) provides stronger evidence than does a more common type *(p* larger).

14. CROMWELL'S RULE

The convexity rule says that if you know an event J is impossible, it has probability 0. We now suggest that it is only impossible events that can realistically be assigned zero probability. To see this consider Bayes' rule on odds form and remember that if $P(\mathrm{J}) = 0$, then $\mathrm{O}(\mathrm{J}) = 0$ also. But if the odds are 0, multiplication by *any* likelihood ratio will produce 0, so that $O(\mathrm{J}|\mathrm{I}) = 0$ for *any* information I. Consequently if you assign probability zero to an event, no evidence will ever alter that probability. This is surely unreasonable. We therefore suggest amending the first rule to read:

CONVEXITY $0 \leq P(\mathrm{I}|\mathrm{K}) \leq 1$ and $P(\mathrm{I}|\mathrm{K}) = 1$ if, and only if, K implies the truth of I.

The additional feature is sometimes called Cromwell's rule because of his advice to the Church of Scotland: 'I beseech you,… think it possible you may be mistaken'. (Cromwell 1650). The law almost violates Cromwell's rule when it says a person is innocent until proven guilty. Innocence cannot mean $P(\mathrm{G}) = 0$ at the beginning of the trial, since that would imply $P(\mathrm{G}|\mathrm{E}) = 0$ for all E, and so innocence throughout the trial. Notice that for any non-zero value, a, $\mathrm{P(G)} = a$ is capable of being raised to $\mathrm{P(G|E)}$ near 1: a can be as small as you like.

15. EXTENSION OF THE CONVERSATION

We now turn to another rule of probability that is much used and interestingly combines the addition and multiplication rules. Events $\mathrm{J}_{1,}$ $\mathrm{J}_2,\ldots\mathrm{J}_n$ form a *partition* if one of them must be true, and only one of them can be true. (An alternative description is that they are *exhaustive* and *mutually exclusive*: they exhaust the possibilities, and the occurrence of one excludes the possibility of any other.) Blood groups provide an illustration: the four types A, B, O, and AB form a partition with $n = 4$. Consider the case $n = 2$ first , then J2 is the complement of J_1. Let I be any other event. The two events 'I and J_1', 'I and J_2' cannot both occur; the event "'I and J_1' and 'I and J_2'" is impossible. Also, the event "'I and J_1' or 'I and J_2'" is simply I. Hence by the addition law

$$P(\mathrm{I}) = P(\mathrm{I} \text{ and } \mathrm{J}_1) + P(\mathrm{I} \text{ and } \mathrm{J}_2).$$

But, by the multiplication law, $\mathrm{P(I \text{ and } J_1) = P(I|J_1)P(J_1)}$ and similarly with J_2. Consequently,

$$P(\mathrm{I}) = P(\mathrm{I}|\mathrm{J}_1)P(\mathrm{J}_1) + P(\mathrm{I}|\mathrm{J}_2)P(\mathrm{J}_2).$$

As usual, an interpretation is possible in terms of proportions. Referring again to the basic Table in section 6,

$$b = (c/s)s + (e/p)p \ (= c + e).$$

In words, the proportion b of black balls (I) is the proportion *(c/s)* of black balls amongst the spotted (J_1) times the proportion *(s)* of spotted, plus the proportion *(e/p)* amongst the plain (J_2) times the proportion *(p)* of plain.

The argument extends to a partition of any size, and the general result is

RULE OF EXTENSION OF THE CONVERSATION
If J_1, J_2, $\ldots\mathrm{J}_\mathrm{n}$ form a partition

$$P(\mathrm{I}) = P(\mathrm{I}|\mathrm{J}_1)P(\mathrm{J}_1) + P(\mathrm{I}|\mathrm{J}_2)P(\mathrm{J}_2) + \ldots + P(\mathrm{I}|\mathrm{J}_\mathrm{n})P(\mathrm{J}_n).$$

The name is not in general use, but is so evocative that its use is attractive. The idea is that instead of talking just about I, the conversation is extended to include the partition. The rule is useful whenever the events J_i, forming the partition, influence I and $P(\mathrm{I}|\mathrm{J}_i)$ is easier to assess than $P(\mathrm{I})$ directly. Remember this is all for a fixed K: in full the rule reads

$$P(\mathrm{I}|\mathrm{K}) = P(\mathrm{I}|\mathrm{J}_1 \text{ and } \mathrm{K})P(\mathrm{J}_1|\mathrm{K}) + \ldots + P(\mathrm{I}|\mathrm{J}_\mathrm{n} \text{ and } \mathrm{K})\mathrm{P}(\mathrm{J}_\mathrm{n}|\mathrm{K}).$$

An illustration is provided by some calculations for blood types. In paternity questions, the blood groups of mother and child are known, and it is useful to extend the conversation to include the unknown father's group. Let F_1 and F_2 be the father's possible groups: here $n = 2$, Rh - and Rh + say. Let the child be Rh -, event C, and the mother Rh -, event M. What is the probability that a Rh - mother will have a Rh - child: $P(\mathrm{C}|\mathrm{M})$? Extending the conversation to include the father

$$P(\mathrm{C}|\mathrm{M}) = P(\mathrm{C}|\mathrm{M} \text{ and } \mathrm{F}_1)P(\mathrm{F}_1|\mathrm{M}) + P(\mathrm{C}|\mathrm{M} \text{ and } \mathrm{F}_2)P(\mathrm{F}_2|\mathrm{M})\ .$$

If both parents are Rh -, events M and F_1, the child is inevitably Rh - and $P(\mathrm{C}|\mathrm{M} \text{ and } \mathrm{F}_1) = 1$. With a Rh + father, event F_2, the corresponding probability $P(\mathrm{C}|\mathrm{M} \text{ and } \mathrm{F}_2) = 1/2$. Notice that both these probabilities are known from genetical theory. If we make the assumption (discussed in section 16) that parents mate at random in respect of rhesus quality, the probability, $P(\mathrm{F}_1|\mathrm{M})$, that the father will be Rh -, given that the mother is Rh -, is just p, the frequency of Rh - in the population. Similarly, $P(\mathrm{F}_2|\mathrm{M})$ is $1\text{-}p$. Inserting these values into the formula,

$$P(\mathrm{C}|\mathrm{M}) = 1(p) + \tfrac{1}{2}(1 + p),$$

for the probability of a Rh - mother having a Rh - child. Notice how the easily obtained probabilities on the right-hand side have been used to evaluate the less obvious on the left.

16. INDEPENDENCE

Probability calculations can become very complicated, but, fortunately, it is sometimes reasonable to make an assumption that simplifies them. The calculations just made with the rhesus factor gave an illustration. It was assumed there that, with respect to that factor, parents mated at random. That is, the mother's group has no influence on the father's group or vice versa. This need not be true; people may subconsciously prefer others of the same type. Or the type may be linked to some feature, like hair colour, which is more plausible as a selective mechanism. In the case of the rhesus factor there would be a survival advantage in parents preferring their own type since Rh + fathers and Rh - mothers can cause difficulties for the child. The mating at random can be alternatively expressed as

$$P(\mathrm{F}_1|\mathrm{M}) = P(\mathrm{F}_1) :$$

in words, the probability of the father being Rh -, F$_1$, is the same whether or not you know the mother is Rh -, M.

In general, we make the definition:

Events I and J are statistically *independent,* given K, if

$$P(\mathrm{I\ and\ J}|\mathrm{K}) = P(\mathrm{I}|\mathrm{K})\ P(\mathrm{J}|\mathrm{K})\ .$$

Since, by the multiplication law,

$$P(\mathrm{I\ and\ J}|\mathrm{K}) = P(\mathrm{I}|\mathrm{K})P(\mathrm{J}|\mathrm{I\ and\ K})\ ,$$

an alternative form of the definition is

$$P(\mathrm{J}|\mathrm{I\ and\ K}) = P(\mathrm{J}|\mathrm{K})$$

agreeing with the rhesus definition (identify M and I). The first form exhibits the symmetry between I and J, so that we can talk of I and J being independent, I independent of J or J independent of I: all are equivalent (given K).

It is most important to recognize that reference *has* to be made to K in the definition. The following urn example shows the importance of the conditions K. Each ball in the urn, in addition to being 'black or white' and 'plain or spotted', is either 'hollow or dense', the last classification abbreviated to 'H or D'. The Table shows the numbers of balls of each of the 8 types.

H	B	W
S	2	1
P	4	2

D	B	W
S	2	6
P	1	3

Clearly, for the hollow balls, colour and spottedness are independent, the probability of being black is 2/3, regardless of S or P. Similarly, for the dense balls, the probability of being black is 1/3 irrespective of the spots. Yet, if the two tables are combined the result is

	B	W
S	4	7
P	5	5

and the probability of being black is 4/11 for the spotted balls but 1/2 for the plain balls. In symbols

$$P(\mathrm{B}|\mathrm{S\ and\ H}) = P(\mathrm{B}|\mathrm{P\ and\ H}),$$

$$P(\mathrm{B}|\mathrm{S}\text{ and }\mathrm{D}) = P(\mathrm{B}|\mathrm{P}\text{ and }\mathrm{D}),$$

yet

$$P(\mathrm{B}|\mathrm{S}) \neq P(\mathrm{B}|\mathrm{D})\ .$$

Reference to the hollow/dense quality is therefore essential.

The reason for the importance of independence is that the multiplication law simplifies from the general form

$$P(\mathrm{I}\text{ and }\mathrm{J}|\mathrm{K}) = P(\mathrm{I}|\mathrm{K})\ P(\mathrm{J}|\mathrm{I}\text{ and }\mathrm{K})$$

to

$$P(\mathrm{I}\text{ and }\mathrm{J}|\mathrm{K}) = P(\mathrm{I}|\mathrm{K})\ P(\mathrm{J}|\mathrm{K})\ ,$$

so that in considering J in Bayes' rule, I may be ignored. As an illustration, consider two separate pieces of evidence E_1 and E_2 presented to the court. By Bayes' rule applied to the total evidence 'E_1 and E_2',

$$\mathrm{O}(\mathrm{G}\,|\,\mathrm{E}_1\text{ and }\mathrm{E}_2) = \left\{\frac{P(\mathrm{E}_1\text{ and }\mathrm{E}_2|\mathrm{G})}{P(\mathrm{E}_1\text{ and }\mathrm{E}_2|\text{not-G})}\right\}\ \mathrm{O}\ (\mathrm{G})\ .$$

If E_1 and E_2 are independent, given G and also given not-G, the likelihood ratio in braces may be written as the product

$$\left\{\frac{P(\mathrm{E}_1|\mathrm{G})}{P(\mathrm{E}_1|\text{not-G})}\right\}\left\{\frac{P(\mathrm{E}_2|\mathrm{G})}{P(\mathrm{E}_2|\text{not-G})}\right\}$$

of the two individual likelihood ratios for E_1, and for E_2, separately. Thus the two pieces of evidence can truly be separated. Notice that the condition for this is they be independent *both* on the supposition of guilt *and* on the supposition of innocence.

For three or more pieces of evidence, the situation is even more complicated. Applying the multiplication law to 'E1 and E_2' and 'E_3'

$$\begin{aligned} P(\mathrm{E}_1\text{ and }\mathrm{E}_2\text{ and }\mathrm{E}_3|\mathrm{G}) &= \mathrm{P}(\mathrm{E}_1\text{ and }\mathrm{E}_2|\mathrm{G})\ \mathrm{P}(\mathrm{E}_3|\mathrm{E1}\text{ and }\mathrm{E}_2\text{ and }\mathrm{G}) \\ &= P(\mathrm{E}_1|\mathrm{G})\ \mathrm{P}(\mathrm{E}_2|\mathrm{G})\ P(\mathrm{E}_3|\mathrm{E}_1\text{ and }\mathrm{E}_2\text{ and }\mathrm{G}) \end{aligned}$$

if the previous condition of independence of E_1 and E_2 holds. For the factorization to be complete it is necessary that the last factor has the property

$$P(\mathrm{E}_3|\mathrm{E}_1\text{ and }\mathrm{E}_2\text{ and }\mathrm{G}) = P(\mathrm{E}_3|\mathrm{G})\ .$$

It is not enough that E_3 be independent separately of E_1 and of E_2 given G. Independence is an extremely subtle concept and should be handled with great care. But when it does hold, the calculations are much simplified because of the separation achieved.

17. RANDOM QUANTITIES

Discussion has so far been confined to uncertain *events:* those that may be either true or false. A useful extension is possible. Suppose forcible entry has been made to a building by breaking a pane of window glass. We may consider the event I that the breaker of the window will have fragments of glass from the window on his clothing, and its probability $P(\mathrm{I})$. This is sometimes referred to as the probability of transfer. We may be interested in more than just transfer and enquire how many fragments may be transferred. This is easily handled. Let Ir be the event that exactly r fragments are transferred. (So that I_0 is 'not-I' in the simpler case.) Let $p_r = P(\mathrm{I}_r)$ be the probability of the event I_r; the probability of r fragments being transferred. The events I_r, $r = 0, 1, \ldots$, form a partition and hence $p_0 + p_1 + p_2 + \ldots = 1$. (The sums here are infinite, but in practice an upper limit to r may be introduced.) This leads to a general definition:

A quantity r that can assume several values, 0, 1, 2,... with probabilities p_r *is* called a *random quantity* and p_r is its (probability) *distribution.* The more common term is 'random variable', but this is inappropriate because r does not vary. We may not know its value, but the number of fragments transferred is fixed, not variable, 'uncertain quantity' is logically the better term. Necessarily, each $\mathrm{p_r}$ is non-negative, and their sum is 1. It is possible to extend the definition of a random quantity to a continuous quantity, like the total weight of glass transferred. Since this involves the differential calculus, the extension will not be presented here.

An important feature of a random quantity, r, or its distribution, p_r, is the *expectation* *E(r),* defined as

$$E(r) = Op_0 + 1p_1, + 2p_2 + 3p_3 + \ldots$$

obtained by multiplying the value of the quantity by the probability of that value and summing over all values. See Note 9 of Chapter 3 for the general formula. For example, suppose

$$p_o = 0.1, p_1 = 0.3, p_2 = 0.3, p_3 = 0.2, p_4 = 0.1$$

so that the probability of transfer is 0.9 ($= 1 - \mathrm{p}_0$) and there is no chance of more than 4 pieces being transferred. The expected number of fragments transferred is

$$0 \times 0.1 + 1 \times 0.3 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.1 = 1.9.$$

One says that 1.9 fragments are expected to be transferred. Notice that the expected number (1.9) is not a possible number (0, 1, 2, 3, or 4). Its use is justified by the frequency with which the defining sum occurs in probability calculations. For example, it is typically reasonable to assume that the measurement *x* of refractive index has the property that *E(x)* = μ, the true r.i. The expectation of r, *E(r),* is often called the *mean* of the associated probability distribution p_r.

It is useful to consider the deviation of a random quantity *r* from its expectation, *r-E(r), w*rite this as d. *T*hen the expectation of d^2 is called the *variance* of *r*, V*(r).* We used the idea in section 9 when the measured r.i. *x* was described as having *standard deviation* s, the positive square root of the variance.

18. PROBABILITY AND CHANCE

The preceding discussion explains the concept of probability and develops the calculus of probabilities sufficiently far to provide a basis for more detailed treatment in the rest of the book. Before concluding this chapter, it is essential to make a further, important remark about the meaning of probability.

Probability has here constantly been viewed as a measure of uncertainty of an event I. An alternative description is a measure of belief that I is true. This description is particularly apposite in the case of the defendant's guilt G. At the end of the trial, the jury will, one hopes, have *P*(G|K), where K includes all the evidence presented and all background knowledge, either near 1 or near 0, corresponding to a strong belief that the defendant is guilty or innocent, respectively.

There is, however, another entirely different interpretation of probability: not in terms of belief, but using the concept of frequency. A coin when tossed may be said to have a probability of 1/2 of falling heads because in a long sequence of similar tosses, it will fall heads one half of the time. We shall reserve the term 'chance' for this frequency concept. Thus the chance of heads is the frequency. It is the frequency concept that dominates current thinking in statistics and probability. We have implicitly met it when we referred to a r.i. *x* having a probability distribution with mean μ and standard deviation s. This means that if the experiment of measuring the r.i. of a single piece of glass was repeated a large number of times, the frequency of the result would agree with the probabilities. In particular, the mean result would be the expectation *E(x)* = μ. It is because of adherence to the frequency view that statisticians have used the concept of a significance level instead of the probability of the hypothesis. The former is a chance: the chance, in repetitions of the experiment, of a derivation as great, or greater, than that observed. The latter is a measure in the hypothesis. To treat the chance as a belief is, as explained in section 9, erroneous.

There is nothing wrong with the frequency interpretation, or chance. It has not been used in this treatment because it is often useless. What is the chance that the defendant is guilty? Are we to imagine a sequence of trials in which the judgements, 'guilty' or 'not guilty', are made and the frequency of the former found? It will not work because it confuses the judgement of guilt, but, more importantly, because it is impossible to conceive of a suitable sequence. Do we repeat the same trial with a different jury; or with the same jury but different lawyers; or do we take all Scottish trials; or only Scottish trials for the same offence? The whole idea of chance is preposterous in this context.

There are occasions when probability and chance are numerically the same. A case was encountered earlier. A forensic scientist's belief that a stain will be of a particular blood type may be the frequency of that type in the population. (The population acting as a sequence of people.) There is nothing wrong in this. What is happening is that the (frequency) data on blood types is being used as a basis for the belief. In our notation, if I is the event of having that blood type, $P(\mathrm{I}|\mathrm{K}) = p$, where K is the data on the population and p is the frequency from that data.

There are, as we have seen with the defendant's guilt, occasions where probability exists but chance does not. There are other situations where both exist but are different: here is an example. Suppose you are a white, English male aged 50. Then you could consult actuarial tables and determine the chance that you will die within 10 years. This is a chance based on the frequency with which white Englishmen, aged 50, have been observed to die before their 60th birthday; plus a few actuarial calculations to allow for trends in the death rate. But this chance may not be your belief that you will die. You may be of social class II, with parents still alive and a family record of longevity. Or you may have just been diagnosed as having lung cancer. In the first case, your belief may exceed, and in the second be less than, the chance. You might ask for actuarial evidence about people in social class II from long-lived families, but it is probably not available. In any case, you might include other factors such as being happily married. Everyone is unique. All that you can do is use the chance (here in the form of actuarial information) as data, or evidence, to help assess your belief. In the longevity example, in the absence of ill-health, it provides a reasonable upper bound.

19. SUBJECTIVITY

There is a difficulty with the belief interpretation of probability that frequentists cite in support of their view—an objection that is sound and valid. It is that belief is a property of an individual. It is subjective. Whereas chance is a property of a sequence and all who observe the sequence will agree its value. It is objective. Objectivity is supposed to be the hallmark of science and subjectivity is thought to be undesirable. It is assuredly true that your belief may differ from mine. The 12 members of the jury may have 12 different beliefs.

The genuine difficulty is mitigated by the following observation. Suppose, on knowledge K, two people have different beliefs in the truth of an event G: their probabilities are not the same. Suppose now additional evidence E relevant to G is produced. Then it can be shown rather generally that E will tend to bring the two probabilities closer together, and that, for sufficiently large amounts of evidence, they will agree for all practical purposes. Briefly, additional evidence makes for agreement in beliefs.

This is exactly what happens in the court room; the jurors are brought to agreement by what they hear. Of course the evidence may be insufficient, and the disagreements persist. Then they would argue together, which is, in effect, changing the emphasis that individual jurors have put on parts of evidence. ('Remember, his alibi was weak.') There is nothing to force agreement, but experience shows that agreement is usually reached.

Scientists adopt a similar procedure. If a hypothesis is thought worthy of study and yet individual scientists disagree as to whether it is true or not, then experiments are performed and repeated (analogous to the evidence in a court of law) until there is general agreement.

There is a lot to be said about divergence of beliefs, especially in a democracy, and much that we do not know. But the subjective or personal nature of probability as belief is surely a correct reflection of reality, and any defects it has seem small when set against the near impossibility of providing, in many cases, an adequate sequence of similar things from which to calculate a frequency, or chance.

20. SUMMARY

The chapter concludes with a summary of the main ideas. A key feature of life is uncertainty. Measurement of uncertainty is accomplished by using a standard called probability. Probability obeys three, basic, rules: convexity, addition, and multiplication. It is preferred to other measures of uncertainty for pragmatic reasons and also because, when scored, it does better than any other measure, like significance levels. Two further rules are deduced from these two: Bayes' rule and the extension of the conversation. The former is the main rule to be used in treating the uncertainty in forensic evidence and involves a likelihood ratio, inverting the roles of evidence and guilt. Calculations are simplified by using the subtle concept of independence. We concluded with a discussion of probability as a subjective notion, in contrast with the physical property of a chance. The use of probability ideas in forensic science is subtle and leads to important points and novel calculations. The material in this chapter can only provide a basis for more thorough discussion in the rest of the book.

REFERENCES

Cromwell, O. (1650) Letter to the General Assembly of the Church of Scotland, 3 August 1650. *Oxford Dictionary of Quotations* (3rd edition), Oxford, 1979.

Barnard, G.A. (1958) Thomas Bayes—a biographical note (together with a reprinting of Bayes, 1764). *Biometrika* **45**, 293–315. Reprinted in Pearson and Kendall (1970), 131–153.

Bayes, T. (1764) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London for 1763,* **53**, 370– 418. Reprinted with Barnard (1958), in Pearson and kendal (1970), 131–153.

De Finetti, B. (1974) *Theory of Probability,* Vol. 1. English translation of *Teoria delle Probabilità.* Wiley, London.

Pearson, E.S. & Kendall, M.G. (eds.) (1970) *Studies in the History of Statistics and Probability.* London, Charles Griffin.

2

Populations and samples

C.G.G.Aitken,
Department of Mathematics and Statistics, The University of Edinburgh,
Edinburgh, UK

1. INTRODUCTION

The definitions of a population and of a sample and the relationship between these two concepts are not always well understood. Forensic science poses particular problems since a forensic scientist, working on a case, unlike his academic colleague, is not interested in studying a physical, biological, or chemical phenomenon in order to gain a greater understanding of the workings of the Universe. The forensic scientist is very interested in problems of comparison. He wishes to determine the relative likelihoods that two groups of items could, or could not, have come from the same source and so construct the likelihood ratio as discussed in section 13 of Chapter 1. He is assisted in his attempts at determination by his past experience and by that of his colleagues as published in the literature or as delivered in personal communications. This wealth of experience may be used to construct abstract models or, increasingly so, mathematical models which may be used to estimate the appropriate likelihoods. The experience may be thought of as a sub-population of information, or data, from a much larger population from which the two groups of items may be thought of as samples. The information contained in the sub-population is used to estimate the likelihood values. Various problems arise with the construction of the background to the problem and with the estimation process. It is the purpose of this chapter to discuss the problems relating to the construction of the background. Chapter 4 on transfer evidence discusses problems with the estimation process. Applications of statistics to particular areas of forensic science are discussed in Chapter 5. After the introduction of necessary terminology, the relevance of the population to the construction of the background to the problem and from which a model might be derived is considered.

Often, a suspect is selected after a specific search for a person with specific characteristics. The investigator has to consider the effect on the value of the evidence which relates a suspect to a crime, of the selection process. This is the socalled selection effect and there are important consequences of it.

Not all of a forensic scientist's work is casework. Some is research. Some of this may be experimental, and then it is important to know the size of sample it is necessary to take in order to have a reasonable chance of detecting an effect when such an effect exists. Some discussion of sample size determination is given here.

If background data exist, the task of the forensic scientist in determining the rarity of any items with which he is presented is greatly eased. However, there are situations, as in hair analysis, where such background data do not exist. Some idea of the value of the evidence in these situations may be obtained by considering all possible pair-wise comparisons of items known to be from different sources. The ratio of the number of comparisons in which the items are indistinguishable to the total number of comparisons provides a measure of the value of the evidential process in general, though not of the value of the evidence in a particular case. The ratio is known as an average probability. The use of values of average probabilities in court is not recommended, but they do provide a general measure of the worth of the type of evidence under consideration. They do not provide a measure of the worth of the evidence in a particular case, and therein lies the danger of an indiscriminate use of them.

Occasionally, the forensic scientist may be asked to give an opinion on the relative frequency of occurrence of a particular characteristic when he has never observed such a characteristic. It is possible to give a meaningful response in such a situation, and this provides the final section of this chapter.

There is a certain overlap of interest between the material in this chapter and that in Chapter 4 on transfer evidence. For example, a discussion of discriminating power is given in Chapter 4, but discriminating power may be considered as an extension of the idea of 'the probability of a match' (Smalldon & Moffatt 1973) which is itself just another expression of an average probability.

2. TERMINOLOGY

The concept of a population is important in forensic science. Many references are made to this concept in published papers, and it is impossible to discuss probability without some reference to an underlying population. The terminology used in statistics has subtle meanings, which are not always the same as those used in everyday talk. Before a general discussion of populations and samples is given, some discussion of the terminology used will be instructive.

(i) Population

There are two types of population to be considered, the reference population, sometimes known as the background or target population, and the sampled population. The reference population is that population about which it is desired to make inferences. The sampled population is that population from which a sample is taken. The reference and sampled populations are not necessarily the same population. For example, a data bank of fibres made from fibres occurring in casework is a sample from the population of fibres occurring in casework. It is not a sample from the population of fibres in the country as a whole. Note that the word 'population' need not refer to human beings; it could also refer to animals or objects, such as fibres.

(ii) Sampling unit or experimental unit

This is the item for study, selected from the sampled population; for example it might be a paint fragment, a human head hair, a fibre, or a fragment of glass. The word 'item' is used as a general word to cover all these possibilities and many more.

(iii) Sample

A sample is a subset of a population. Items from a population may be chosen for inclusion in a sample in many ways. The method of choice can have a considerable effect on the inferences which can be made from the sample about the population.

(a) Subjective

An investigator selects those, and only those, items in the population in which he is interested.

(b) Random

Each item in the population has an equal chance of selection. The most well-known example of random sampling is that of ERNIE, the Premium Bond computer (see Chapter 1), when it selects the winning numbers. For those of us who do not have access to anything as sophisticated, tables of random numbers are published, for example, Lindley & Scott (1984), Table 27. These tables are sets of numbers in which 'each digit is an independent sample from a population in which the digits 0 to 9 are equally likely' (Lindley & Scott 1988, p. 78). Random numbers with any required number of digits are obtained by combining individual random digits. Many pocket calculators have a random number function key. The numbers produced by repeated depression of the key are known as pseudo-random numbers, since starting a sequence with a particular number will produce the same sequence each time that number is used. This is not as silly as it may at first seem; often, an experimenter will want to repeat an experiment with the same random selection as used previously in order to ensure direct comparability. This can be done only by using such a sequence of pseudo-random numbers. Similar sequences of pseudo-random numbers are produced by many statistical software packages. Obviously the sequence will begin to cycle after a certain period of generation of the random numbers. However, if the number of digits in each random number is sufficiently large and the generator is well-chosen, the cycle length will be sufficiently large that problems of this kind will not be encountered. Further details of the generation of random numbers is given in Hammersley & Handscomb (1965).

(c) Stratified

The population is divided into sub-populations, or strata, and sampling is conducted within the strata. For example, in a survey of paint fragments from motor cars the population may be stratified by the different layers and by the different car or paint manufacturers.

(d) Systematic

The sample may be chosen at regular points from within a population. For example, in a handwriting study, every tenth word may be examined. Care has to be taken with such a sampling process. Unknown to the investigator there may be a systematic trend in the data which the sample matches exactly, thus giving a false impression of the underlying population.

(e) Cluster

The sampling unit is a group or cluster of smaller units called elements or subunits. For example, in a survey of farm animals it may be cheaper to survey all the animals in each of a selected number of farms in a region than to select individual animals at random from within the whole region. In the latter case, a surveyor may spend much time travelling between farms inspecting a few animals at each farm. It would be more efficient to visit a few farms and to inspect all the animals at these few farms.

In forensic science, there are two possible purposes for sampling. One is to estimate the value of a parameter of interest, such as the mean difference in the results between two repetitions of a breath test for alcohol. The other is to estimate the underlying distribution of a particular measurement or measurements of interest in order that the degree of rarity of a set of these measurements in a particular case may be estimated. For example, an investigator may be interested in the distribution of the refractive index of glass of a particular type.

(iv) Sampling frame

This is a list of items in the sampled population. In some cases, such as the electoral register for use in an opinion poll, this is already in existence, or it is easy to construct a frame from such a list. In other cases, such as a population of human head hairs, it is impossible to construct.

(v) Parameter

A characteristic of the population about which information is desired. For example, the frequency of the AB blood group in leukaemia patients in Brooklyn, New York, is a parameter. It may be estimated by the frequency of the AB blood group in a sample of leukaemia patients in Brooklyn, New York. The sample frequency is known as an estimate of the population frequency. There may be reasons why the sample frequency may be expected to differ from the population frequency, in which case the sample frequency is said to be a biased estimate of the population frequency. For example, the samples may all be taken from one ethnic sub-group within the population of leukaemia patients in Brooklyn (Gaensslen *et al.* 1987a, p. 1020).

(vi) Experiments and observational studies

An experiment is an investigation in which the scientist is able to control the values of some of the parameters of interest. He then studies changes in the values of the other parameters of interest as he alters the values of those variables which he controls. An observational study is one in which the scientist is not able to control the parameters of interest but may only observe them. For example, an agronomist may wish, among other things, to study the effect of rainfall on crop yield. He cannot affect rainfall, except rather generally perhaps by cloud-seeding, but he can observe its effect. An example, from a forensic science context, of an experiment is given by Rawson *et al.*

(1986) who carried out a series of investigations to determine the reliability of a system for the evaluation of bite marks in human skin. The purpose of the technique they described was to enable them to produce a consistent bite pattern reproducible in future experiments. An example of an observational study is given by Sparks *et al.* (1986) who were interested in determining the postmortem interval. Samples from the putamen were harvested from coroners' cases for which the time of death had been determined from eye-witness accounts. For obvious reasons, the observations made on these samples are not repeatable. Future investigations would require different cases.

(vii) Accuracy and precision

There is no definitive meaning for these terms. However, it is useful to think of accuracy as a measure of the closeness of an observation to the 'true' value of a parameter of interest. Measurements may be thought of as precise if their standard deviation is small, whether or not the measurments are close to the 'true' value (Kendall & Buckland 1982, pp. 3, 152).

(viii) Randomization

This is an important concept, and the validity of many statistical tests is dependent on the sample being considered a random sample from the reference population. A justification of randomization was given by Cox (1958, section 5.4). This included two salutary examples of what can go wrong if randomization is not followed. In one, a systematic arrangement of pairs of mice were treated (T) or not treated (U) with a series of stimulating injections, and then all mice were challenged with a solution supposedly containing a standard number of larvae and noting any response. Within each pair the treated mouse was challenged first, followed by the untreated mouse. The order of administration of the challenges was thus TU, TU, TU,.... Unfortunately, the number of larvae increased steadily over the course of the experiment and the untreated mouse in each pair consistently received a greater dose of larvae than the treated mouse.

The other example illustrated what might happen with a subjective arrangement. In 1930, in the schools of Lanarkshire, Scotland, 5000 children received 3\4 pint of raw milk per day, 5000 received 3/4 pint of pasteurized milk, and 10 000 were selected as controls to receive no milk. The children were weighed and measured for height at the beginning and end of the experiment which lasted for four months. However, children were not allocated at random to the three groups. Teachers were unconsciously influenced by the greater needs of the poorer children, and this led to the inclusion of too many ill-nourished children among those receiving milk and too few among the controls. The final observations on the control group exceeded those on the treated group by an amount equivalent to three-months growth in weight and four-months growth in height.

A much more economical and precise way of comparing two treatments, say the types of milk, in this context would have been to work with pairs of identical twins. Each pair would be as alike as possible in terms of genetic similarity and environmental background. Any differences discovered in height or weight at the end of the trial could reasonably be explained by differences in the effects of the two types of milk.

If one feels one is not able to make the assumptions about randomness necessary to use many statistical tests there is the option to use randomization tests (Edgington 1987). Considerable computing power is necessary for most applications of such tests, and a fuller description of these tests is beyond the scope of this book.

Randomization may occur in two ways. One is in the random allocation of treatments to experimental units as was supposed to happen with the mice and the Lanarkshire schoolchildren. The other is in the selection of a sample from a population which is the main application in forensic science.

3. RELEVANT POPULATIONS

Coleman & Walls (1974) discussed the commission of a crime by a ‘member of the relevant population’ and carried on to say that ‘(t)he relevant population are those persons who could have been involved; sometimes it can be established that the crime must have been committed by a particular class of persons on the basis of age, sex, occupation, or other sub-grouping, and it is then not necessary to consider the remainder of, *say, the United Kingdom’* (emphasis the present author’s). It is not clear why the United Kingdom should be considered the original population of interest. At the very least, a distinction between Great Britain and Northern Ireland may be thought possible.

Smith & Charrow (1975), in a paper discussing upper and lower bounds for the probability of guilt, introduced the idea of a ‘suspect population’ which was defined as ‘the smallest population known to possess the culprit as a member’. It could be argued that this is the population of the World, but this would not be very helpful. An example is given by Smith & Charrow of a town with only three inhabitants. Two characteristic traits are present in this town with relative frequencies defined so that the traits may be distributed amongst the three inhabitants in one of only four different ways, and these four ways are considered as four possible suspect populations. The logic of this approach is questioned by Finney (1977) who stated that this argument would hold only if the concept of a ‘superpopulation’ was introduced from which the relative frequencies of the two traits were obtained. This line of argument introduces another weakness noticed by Finney, who stated that Smith & Charrow wanted ‘to have a random selection model for the actual population from some superpopulation, yet to insist on exact agreement of actual trait frequencies with socalled census information’ (from the superpopulation).

Kirk & Kingston (1964) stated that ‘(m)ost evaluations of significance must rest on the study of a population, which ideally would be the total population involved’, but did not state how such a population might be determined. Since it will be almost impossible to examine the total population, even if we knew what it was, a sample should be studied. This study requires considerable planning, both in the choice of the sample and in the design of the method of data collection. Kingston (1965b) referred to a sample from an ‘appropriate population’ without further elaboration. Kingston (1965a) considered estimating the expected number of people in a population which would have a characteristic in question, and concluded that it was best to base calculations ‘on the maximum possible population’.

Stoney & Thornton (1986, 1988) and Kingston (1988) discussed the role of a population in the context of fingerprint evidence. Stoney & Thornton (1986) criticized a model developed by Kingston (1964) for the evaluation of fingerprints. The main thrust of the criticism is as follows. Kingston (1964) was assumed to have chosen a suspect on the basis of fingerprint evidence alone, and thus the underlying population from which the suspect has to be considered as being selected was that of the whole World. Stoney & Thornton (1986) argued, however, that it is rarely the case that a suspect would be chosen purely on the basis of fingerprint evidence. Normally, there would be a small group of suspects which would have been isolated from the World population on the basis of other evidence. The fingerprint evidence would then be considered relative to this small group only. In reply to this criticism, Kingston (1988) disagreed with this restriction to a small group of suspects.

> ‘A fingerprint expert should not base an opinion about the fingerprint match on an evaluation of the other evidence. It is reasonable, therefore, for the fingerprint expert (before actually comparing the prints) to consider that anyone in the World could have made the trace print, and that the prior probability of any specific individual having left the print is $1/P$, where P is the population of the World.’

In rebuttal of this criticism, Stoney & Thornton (1988) explained that the small group of suspects which they postulated for consideration was not to be used to estimate the prior probability that a specific individual may have left the print. Rather it was to be used to provide a measure of the number of fingerprint comparisons required to obtain an estimate of the probability of a false association.

The difference in approach may be described as follows. The prior probability under consideration is the probability that the individual, J say, was at the crime scene, prior to consideration of any evidence. The complementary probability of this event is the probability that J was not at the crime scene, prior to consideration of any evidence. Let us denote the events that J was or was not at the crime scene by the symbols C and $\bar{\text{C}}$ respectively. Thus $P(\text{C})$ denotes the prior probability that J was at the crime scene. This is the prior probability to which Kingston (1988) referred. The probability of a false association is the probability that the print at the crime scene and a print from J’s finger are found to ‘match’, in some sense, when J had not been at the crime scene. Let us denote the event of a ‘match’ between the print at the crime scene and a print from J’s finger by the symbol M. The probability of false association may then be denoted in symbols as $P(\text{M}|\bar{\text{C}})$. This is the probability to which Stoney & Thornton (1988) referred. The two probabilities, $P(\text{C})$ and $P(\text{M}|\bar{\text{C}})$, discussed here concern two very different concepts. In the first situation, the presence of J at the crime scene is the unknown element. In the second situation, the absence of J from the crime scene is assumed known; it is the ‘match’ between the print at the crime scene and a print from J’s finger which is the unknown element.

An example in which the authors admitted that their sample was not representative of a meaningful population was given by Eldridge *et al.* (1984) who described the variation of letter-forms generated in samples of handwriting taken from 61 right-handed people under controlled conditions. In the discussion of their paper they stated that ‘the sample of subjects was too small (61 people), and it cannot be said to be representative of the British population as a whole’. One obvious difference, which the authors mentioned, was that half the subjects had attended a college or a university. The variability in the handwriting of such a group will not be as great as that in the population with which document examiners are usually concerned. Further research of the kind outlined by Eldridge *et al.* (1984) is required to establish the characteristics of differences in handwriting in different groups of people.

Lenth (1986) discussed the relevance of the target population when the evidential value of the likelihood ratio is considered, with reference also to Kingston (1965a). The model used assumed that ‘the alleged source of the evidence is a random selection from those persons having the required characteristics’. An interesting definition of a relevant population is given by Gaensslen *et al.* (1987a, b, c) in a series of papers which discusss the distributions of genetic markers in United States populations. Populations are defined in their studies ‘by the size of the effective interbreeding gene pool rather than by geographical boundaries’. The definition of the reference population requires considerable care in this context, and a discussion by Berry & Geisser (1986, p. 373) with reference to cases of disputed paternity is of particular importance. The genetic frequencies in the race of the alleged father are not the relevant ones to use when calculating the paternity index, so-called by Salmon & Salmon (1980), a measure related to the relative likelihood that a particular person is the true father. The index is the ratio of the probability of the evidence, assuming that the true father is the alleged father, to the probability of the evidence, assuming the true father is not the alleged father. The evidence is the blood grouping data from the child, the mother, and the alleged father. The likelihoods of the bloodgrouping data which have been observed depend on the genetic frequencies, and those, in turn, depend on the definition of the population, or populations, to which the child, mother, and alleged father belong. If the true father is not the alleged father, then he is considered as a ‘random man’ from some population. To assert that this population is, of necessity, the same racial group as the alleged father is patently not true. The true father could come from some other racial group, though which groups are possible are constrained by the bloodgrouping result of the child. Thus one possible population is that defined by the child. Another is that defined by an average over some local population, subject again to the constraints imposed by the results from the child. Further refinements are possible if the defence claims that the true father is from a different racial group to that of the alleged father and also specifies the group, or if the suggested alternative to the alleged father is related to the mother or to the alleged father. In the first instance, the calculations can be made and compared; in the second the concept of a man selected ‘at random’ is affected (Berry & Geisser 1986). For further comments about paternity testing, with particular reference to the exciting developments in DNA fingerprinting, see Chapter 5.

The definition of a population ‘relevant’ to a particular case is very difficult. At present, there is no obvious general set of guidelines available. However, awareness of the existence of the problem is of considerable assistance in determining a solution. At present, if forensic scientists are aware of these problems, this awareness will, of itself, assist in their interpretation of their results.

4. SELECTION EFFECTS

Consider the Collins *(People v Collins* 1968) case, one which has achieved considerable notoriety. Briefly, the crime was as follows. In June 1964 an old lady was pushed to the ground in an alley-way in Los Angeles by someone whom she neither saw nor heard. She managed to look up and saw a young woman running from the scene. Immediately after the incident the lady discovered her purse was missing. A man watering the lawn at the front of his house at the end of the alley was attracted by the commotion. He saw the woman, whom he later described as a Caucasian with her hair in a dark blonde ponytail, run out of the alley and enter a yellow automobile parked across the street from him. The car was driven by a male Negro wearing a moustache and a beard.

A couple answering this description were eventually arrested and brought to trial. The prosecutor called as a witness an instructor of mathematics at a state college in an attempt to bolster the identifications. The witness testified to the product rule for multiplying together the probabilities of independent events, and used this rule to calculate the probability that a couple selected at random from a *population* would exhibit all of a certain selection of characteristics as 1 in 12 million. The values used for the probabilities of the individual characteristics were chosen by the prosecutor without any justification, in particular without reference to any studies which may have been conducted to atttempt to estimate these probabilities. The prosecutor invited the jurors to apply their own values, though there is no record of the jury so doing. The individual probabilities, according to the prosecutor, should be as follows:

Characteristic	*Probability*
Yellow automobile	1/10
Man with moustache	1/4
Girl with ponytail	1/10
Girl with blonde hair	1/3
Black man with beard	1/10
Interracial couple in car	1/1000

There are many reasons why there should be misgivings about these figures. However, the main purpose of the current discussion is to consider the role of the background population in the interpretation of these probability figures. For example, what does it mean when it is said that the probability of a yellow automobile is 1/10 or when it is said that the probability of a couple answering the description is 1 in 12 million? The frequentist approach to inference, based on the concept of long-run relative frequencies, would imply that in random sampling of some population of couples, a couple answering the above description would be selected, on average, once in every 12 million selections. It is not clear, however, how such a population, from which these selections are to be made, should be defined. Should it be the population of couples within the neighbourhood, county, city, state, or country in which the crime was committed? The size of such a population is critical in determining a probability of the type calculated above. Perhaps the jury could be given a range of choices of population with associated probabilities, but such a procedure would just add a further layer of complexity to a process which is already very complicated, and there is nothing to suggest that this will make the jurors' task any easier.

Another factor which has to be considered in the assessment of such probability values is what Finkelstein (1978) called the 'selection effect'. This effect refers to the selection procedure in which characteristics for study in the assessment of evidential value are selected from some larger group of characteristics which might have been considered but were not. Consider a case, cited in Finkelstein (1978, p. 84), in which an expert examined a large number of fibres taken from the accused's clothing and from the scene of the crime and was able to make eleven matches. Using the product rule, he concluded that the probability that such a number of matches would have occurred by chance, that is, if the accused were not guilty, was one in a thousand million. However, care has to be taken in such an analysis that the number of comparisons that did not match were also considered. To report only the number of matches found without reporting also the number of comparisons which were made in total to find that number of matches is bad statistical practice. A hypothetical example from coin tossing will illustrate the point. If a fair coin is tossed 10 times it will be thought unusual if 10 heads occur, since the probability of this occurring, with a fair coin, is 1 in 1024. However, if a fair coin is tossed 1000 times 10 times, it would be not at all unusual for a sequence of 10 consecutive heads to occur. (The probability of at least one sequence of 10 consecutive heads in such an experiment is 0.62.)

One example where a background population was defined is that of *State v Sneed* (Finkelstein 1978, p. 99). The background population in that case was defined by listings in telephone directories. There was evidence that the accused had, on occasion, used the name 'Robert Crosset'. There was also evidence that on the day of the murder in question someone of that name had purchased a hand gun which, apparently, was the murder weapon. Telephone directories in the 'area' of the crime were examined, and no Crosset was found in approximately 129 million listings. The definition of 'area' poses similar problems to the definition of 'population'. On the basis of this evidence a guess of the frequency of Crosset in the population was taken as 1 in a million. This is not a very good guess. Using the formula given in section 7, the value for the frequency of Crosset in the population is estimated to be 1 in 232 million. The frequency of Robert in the population was taken to be 1 in 30. Assuming that the two names are chosen independently of each other, the frequency of Robert Crosset was estimated to be 1 in 30 million. In reversing the conviction of the defendant, the Supreme Court of New Mexico objected to the use of 'a positive number…on the basis of the telephone books when the name Robert Crosset was not listed in those books'. The guess made in this case is not justifiable. However, the list of names in the telephone books could have been considered as a large random sample of people for the general area covered by the books and appropriate calculations made to obtain an upper bound or an expected value for the frequency of occurrence of the name 'Robert Crosset'. Such calculations are described in section 7 on zero occurrences.

A good discussion of the importance of consideration of the selection effect is given by Fienberg (1986) in a comment on the paradox of the gatecrasher, first posed by Cohen (1977). The paradox, stated by Cohen, is as follows.

> 'Consider, for example, a case in which it is common ground that 499 people paid for admission to a rodeo, and that 1000 are counted on the seats, of whom A is one. Suppose no tickets were issued and there can be no testimony as to whether A paid for admission or climbed over the fence. So by any plausible criterion of mathematical probability there is a 0.501 probability, on the admitted facts, that he did not pay. The mathematicist theory would apparently imply that in such circumstances the rodeo organizers are entitled to judgement against A for the admission money, since the balance of probability (and also the difference between prior and posterior probabilities) would lie in their favour. But it seems manifestly unjust that A should lose his case when there is an agreed mathematical probability of as high as 0.499 that he in fact paid for admission.
>
> 'Indeed, if the organizers were really entitled to judgement against A, they would presumably be equally entitled to judgement against each person in the same situation as A. So they might conceivably be entitled to recover 1000 admission moneys when it was admitted that 499 had actually been paid. The absurd injustice of this suffices to show that there is something wrong somewhere. But where?'

Fienberg (1986) gave a detailed exposition of a Bayesian analysis of this problem based on considerations of the odds form of Bayes' rule discussed in Chapter 1. It can be shown that the posterior odds in favour, after presentation of the evidence, of guilt are equal to the prior odds of guilt multiplied by the ratio of two probabilities, the probability of the evidence if the suspect is guilty and the probability of the evidence if the suspect is innocent. This ratio is known as Bayes' factor or as a Bayesian version of the likelihood ratio (see Chapter 3). This relationship emphasizes that any decisions about the guilt or otherwise of the suspect should be based not only on the evidence of the rodeo but also on the prior odds, or prior belief, in his guilt before the rodeo evidence is presented. Denote the evidence, which is defined more explicitly later, by E; denote the guilt of the suspect, that is that he attended the rodeo and did not pay, by G, and the innocence of the suspect, that is that he either did not attend the rodeo or, if he did, that he paid to enter, by $\bar{G}$. The important ratio is then

$$P(E|G)/P(E|\bar{G})$$

with the same notation as was introduced in Chapter 1. Fienberg (1986) considered the evidence as being decomposable into two parts

(i) E_1: the background information, that 1000 people attended the rodeo and 499 paid to enter, and
(ii) E_2: A attended the rodeo,

so that $E = (E_1, E_2)$. With this decomposition, it is reasonable to argue that

(a) $$P(E_1|G) = P(E_1|\bar{G}) = P(E_1)$$

and

(b) $P(E_1|G \& E_2) = P(E_1|\bar{G} \& E_2) = P(E_1)$.

Expression (a) is a mathematical expression of the idea that knowledge of the guilt (G) or innocence ($\bar{G}$) of the suspect does not affect our belief about the numbers of people attending and paying to enter the rodeo. Expression (b) states that knowledge that the suspect attended the rodeo also does not affect this belief. The evidence that A attended the rodeo is vital, however, rather obviously, in the assessment of the likelihood of his guilt. The ratio $\{P(E|G)/P(E|\bar{G})\}$ may be written, using the rules described in Chapter 1, as

$$P(E_1|G \& E_2) \times P(E_2|G)/\{P(E_1|\bar{G} \& E_2) \times P(E_2|\bar{G})\}.$$

However, $P(E_1|G \& E_2) = P(E_1|\bar{G} \& E_2)$ from (b), thus this ratio is reduced to the ratio of $P(E_2|G)$ to $P(E_2|\bar{G})$. The numerator, $P(E_2|G)$, equals 1 since if the suspect is guilty then he attended the rodeo by the definition of guilt above. The interesting probability is the denominator. Discussion of the determination of the denominator requires consideration of a reference population, and it is for this reason that this example is presented here.

Suppose that there is a reference population of size N to which the people who attended the rodeo are considered to belong; that is, that problems of definition of the population and determination of N have been solved. N is assumed to be considerably greater than 1000. There are 501 people in this population who attended the rodeo and did not pay for entry. These comprise the sub-population of guilty people. The remaining (N - 501) people are innocent, from the definition of innocence above, and comprise the sub-population of innocent people. Of these (N - 501) people, 499 attended the rodeo and paid. Thus, the probability that A attended the rodeo, conditional on the fact that he is innocent of the charge of attending the rodeo and not paying, is given by

$$P(E_2|\bar{G}) = 499/(N - 501).$$

The likelihood ratio is then

$$\{P(E_2|G)/P(E_2|\bar{G})\} = 1/\{499/(N - 501)\} = (N - 501)/499.$$

The posterior odds in favour of guilt are obtained by multiplying the prior odds in favour of guilt by this likelihood ratio, using the odds form of Bayes' rule from section 12 of Chapter 1. Cohen's argument implies that the posterior odds in favour of guilt are 501/499 since Cohen has said $P(G|E) = 0.501$ and $P(\bar{G}|E) = 0.499$, implying that the ratio or, by definition, posterior odds is 501/499. The value of the likelihood ratio, (N - 501)/499, then implies that the prior odds, $P(G)/P(\bar{G})$, must be 501/(N - 501) since the prior odds multiplied by the likelihood ratio equals the posterior odds. This ratio in turn implies that the prior probability of guilt, $P(G)$, equals 501/N, which is the proportion of the population who are gatecrashers at this rodeo and is an estimate of the probability that a member of the population selected at random from the population is a gatecrasher at this rodeo. If N is large, this probability is extremely small and not in accord with the prior belief, implicit in Cohen's argument, that the odds in favour of guilt are 1. Put another way, the implicit prior belief is that the suspect is equally likely to be guilty as innocent (which does not accord with the belief that a person is innocent until proven guilty either), and this probability $501/N$ does not accord with this belief.

However, if we were to assume that the prior odds were 1, then the posterior odds in favour of guilt are $(N - 501)/499$ which, if N is large, are themselves large. The strong implication to be taken from this is that if the prior odds of guilt in a civil case are taken to be equal to 1, then paradoxical conclusions may emerge.

The choice of a value for *N,* the size of the reference population, obviously has great bearing on the values finally attached to the posterior odds. If one argues, rather unreasonably, that the suspect is one of a population defined by those, and only those, attending the rodeo, then this implies that $N = 1000$. The likelihood ratio is then

$$(N\text{-}501)/499 = 1.$$

The posterior odds equal the prior odds, and the evidence has no effect on the odds in favour of guilt, whatsoever. Fienberg (1986) also said, very reasonably, that '(w)e might also pick N to ensure a prior odds ratio of 1. If $N = 1002$ then the prior odds are 1, the likelihood ratio is 501/499, and the posterior odds are 501/499. This likelihood …is not very sensible to me because I cannot construct an argument for the choice of $N = 1002$'. Fienberg concludes with the following statement.

> The evidence in the gatecrasher problem is probabilistic only if we make explicit some random selection mechanism whereby defendants are chosen from some population. The probabilistic nature of such selection mechanisms is often not part of the evidence at all, or at least its link to the evidence is not made explicit to the fact-finder. A classic example is in *People v Collins* (1968) where the selection mechanisms used to identify the defendants, an interracial couple, was not made explicit at the trial.'

The idea of 'relevant' evidence is important and worthy of consideration when defining the reference population at a particular point in an investigation. The population which may be called the reference population for a particular piece of evidence, E say, for example, bloodgrouping frequency, is that population which has been determined by the evidence discovered prior to the discovery of E and by all other information I 'relevant' to the investigation. It is not determined in any way by E. The definition of the population may alter after a suspect has been found if other evidence, such as eye-witness evidence, comes to light which is not derived from the suspect alone. Probability values, measures of weights of evidence, and any other statistic which may be calculated to represent numerically some characteristic of E should use the reference population determined for E. It is not legitimate to determine the probability that the suspect is from a particular racial group, R say, if he were picked up precisely because he was a member of that racial group. In that case, I includes E, the evidence that the suspect is from R and thus the probability, $P(\mathrm{E}|\mathrm{I})$, equals 1, independently of the guilt or innocence of the suspect. Further discussion of this is given in Chapter 4.

The reason for the apprehension of a suspect has also to be considered when assessing the weight of the evidence E. The value attached to E if the suspect has been discovered in a general sweep of the local population will be very different from the value which is evaluated for a suspect who was picked up for some reason unconnected with E. These values will, themselves in turn, be different from that procured from a suspect who was picked up because of E itself.

The notion of the reference population is fundamental to much of statistical thought. In the forensic science context when examining evidence, two probabilities are of interest, the probability of the evidence if the suspect is guilty and if he is innocent. One has also to bear in mind, though, that there is much other information, apart from the basic evidence available to the scientist who is trying to determine these probabilities. A discussion of what this may entail is given in Evett (1984) with a particular example relating to a breaking and entering case in which a window has been smashed to gain entry. This 'other information' will help to define the population from which the evidence may be considered a sample and which will be used to derive probability values.

There is another aspect to this discussion. The number of comparisons made before discovering a match, or a similarity, in transfer evidence has to be taken into account when addressing the value of the evidence. A general scanning of some large population, such as in surveys done for genetic fingerprinting investigations, will produce a different evidential value than either a match by chance with the suspect alone examined or a match discovered after several comparisons have been made. This general scanning may well enable a suspect to be identified, but the probabilistic nature of such a scanning procedure may not be part of the evidence and yet is crucial to the evaluation of the probability. The case of *People v Collins* (1968) is a well-known example where the procedure by which the defendants, an interracial couple, were identified was not explained fully at the trial. If other evidence had led the police to the couple and the police had then discovered 'by chance' all the characteristics given by the eye-witness evidence, then the eye-witness evidence would be very strong. Conversely, if the police had searched for a couple as defined by the eye-witness evidence, not considered other evidence, and not stopped till such a couple had been found, then the eye-witness evidence would not be very strong.

The concept of randomness is important in the determination of probabilities. In general, in a forensic science problem one is not concerned with random subsets of particles, fibres, or whatever. One is interested in a subset to which suspicion attaches because it is known to be associated with the victim or crime scene. It is not correct to assume that the frequency of a characteristic in some national or global population may be applicable to some suspects who may be associated with the particular characteristic.

5. SAMPLE SIZES

One form of question which is often asked before an experiment is started is 'How large a sample should I take in order to estimate an effect with a given precision?'

When a statistician is asked such questions he has to reply with some questions of his own. First, how small an effect do you wish to detect or how precise an estimate of a given effect do you wish? Intuitively, small effects require larger sample sizes for their detection than do large effects. Similarly, more precise estimates require larger sample sizes than do less precise estimates. Also, if one wishes to detect an effect of a given magnitude, there is another question which can be asked: what error probabilities can be tolerated? There are two possible errors to be considered, that of detecting an effect when one does not exist, and that of failing to detect an effect when one does exist. The determination of a probability for this second error requires knowledge of the size of the supposed effect. Again, intuitively, if one has a fixed sample size, then the larger the effect which exists the smaller will be the probability of failing to detect it.

Consider, first, the question 'How large a sample should I take in order to estimate an effect with a given precision?' It is assumed that the measurements to be made are Normally distributed with known variance, σ^2, and that the mean of the distribution, which is not known, is the effect to be estimated. Suppose the estimate is required to be correct to within a limit, ± L say, and that the scientist is prepared to tolerate a probability a that the estimate will be further than this limit from the true effect. Let z_ε be the upper 100ε% point of the standard Normal distribution; for example, if ε equals 0.025, z_ε is the upper 2.5% point which is 1.960 (see books of statistical tables for determination of these z values, for example, Lindley & Scott (1988, Table 5, p. 35) or Snedecor & Cochran (1980, Table A4, p. 469)). Then, from the standard theory of confidence intervals, there is a probability (1-a) that the mean $\overline{X}$ of a sample of size n lies within $z_{\alpha/2}\sigma/\sqrt{n}$ of the true mean of the population. Thus, $z_{\alpha/2}\ \sigma/\sqrt{n}$ may be equated to L and the resultant equation solved for n to give a desired sample size of

$$n = z_{\alpha/2}^{2}\ \sigma^2/L^2\ .$$

For example, if $\alpha = 0.05$, $z_{\alpha/2} = 1.960$, $\sigma = 2$, and $L = 1$, then $n = 15.4$ which is rounded up to the nearest integer, as a conservative measure, namely 16. The desired sample size is therefore 16. If $\alpha = 0.05$, $\sigma = 3$, and $L = 0.05$, $n = 13829.8$ and the desired sample size is 13830.

Further information about this discussion is contained in Snedecor & Cochran (1980, p. 53).

The other question which may be asked, 'How large a sample should I take in order to detect an effect?', has more relevance in comparative experiments where one wishes to compare two sets of measurements, one on each of two different populations, to see if there is a real difference in the effects of treatments applied to the two populations. The term 'population' is a technical word in this context. The only difference should be that one 'population' receives one treatment while the other 'population' receives the other treatment. In all other respects, the two 'populations' should be as similar as possible. This constraint is important as it ensures that any differences in the measurements obtained are, as far as is possible, attributable to differences in the effects of the treatments and not to any other differences, known or otherwise.

As an example, a doctor may wish to compare the effect of a new drug with the effect of a standard drug to see if there is sufficient evidence to warrant replacing the standard drug with the new drug. The two groups of patients for this trial should be as alike in all other respects, such as age, sex, severity of illness, as possible. Another example, motivated by Gullberg (1987), may compare the results obtained from a second test of breath for alcohol levels with a first test to see if the results are consistent. This second example differs from the medical example in that the ‘two’ groups of people are in fact the same group to whom two different treatments are applied, namely the first breath test followed by the second breath test. Such an experiment is known as a paired experiment.

The determination of the sample size in the context of the second example is given below. Further details, including the application to the first example, are given in Snedecor & Cochran (1980, p. 103). Determination of the size of a sample requires three pieces of information from the experimenter.

(a) The magnitude (δ) of the effect it is desired to detect.
(b) The power of the test, namely the probability, (1-ß) say, of obtaining a significant result and deciding that the effect exists if the true magnitude of the effect is δ. The power of a test is the probability of detecting an effect of a given magnitude when that effect exists.
(c) The significance of the test, namely the probability α of obtaining a significant result and deciding that the effect exists when there is no real effect.

A general description of the theory required for the derivation of the formula for the desired sample size, *n,* assuming that the population standard deviation, σ_d, of the differences is known, is given by Snedecor & Cochran (1980, p. 103). Suppose it is desired to detect an effect of magnitude δ with probability (1–ß) when such an effect exists and to decide, incorrectly, with probability α, that it exists when it does not. Then the necessary sample size to fulfil these conditions is given by

$$n = (z_{\beta} + z_{\alpha/2})^2 \, \sigma_d^2/\delta^2 \ .$$

For example, suppose it is desired to estimate the sample size necessary to detect an effect of 1.5 units with probability 0.8 and one is prepared to accept that this effect may be claimed to exist, when in fact it does not, with probability 0.05, and that the standard deviation of the differences is assumed to be 3 units. Thus

$$\delta = 1.5,\ \sigma_d = 3.0,\ 1\text{-ß} = 0.8,\ \alpha = 0.05$$

and so

$$\text{ß} = 0.2 \text{ and } z_{\text{ß}} = 0.842$$

$$a = 0.05 \text{ and } z_{\alpha/2} = 1.960.$$

The desired sample size is then given by

$$n = (0.842 + 1.960)^2\ 3.0^2/1.5^2 = 31.4.$$

This should be rounded up to the nearest integer, as a conservative measure. Thus the desired sample size is 32.

For binomial data, one can use the Normal approximation to the binomial distribution so long as one starts with the premise that the sample size will be large, at least double figures say, and that the proportion to be estimated is thought to be close to one half. From standard theory, a confidence interval for the true proportion, *p*, is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\{\hat{p}(1-\hat{p})/n\}}$$

where $\hat{p}$ is the sample proportion. The half width, *L* say, of this interval is then $z_{\alpha/2}\sqrt{}\{\hat{p}(1-\hat{p}/n\}$. By equating *L* to this expression and rewriting the equation in terms of *n,* the following expression for *n* is obtained

$$n = [z^2_{\alpha/2}\{\hat{p}(1-\hat{p})\}/L^2]\ .$$

In order that this equation may be solved for *n*, the investigator is required to specify the following information:

(i) an upper limit *L* to the amount of error which can be tolerated in the estimate,
(ii) the desired probability, 1-α, that the estimate will lie within this limit of error of the true proportion, *p*, and
(iii) a guess, $\hat{p}$, at the true value of the proportion.

This last requirement may be waived if another approximation is permitted. The expression $[\hat{p}(1-\hat{p})\}$ takes its maximum value when $\hat{p} = \frac{1}{2}$ at which point it takes the value $\frac{1}{4}$. Thus a conservative estimate of the sample size n is given by

$$n = z^2_{\alpha/2}/4L^2\ .$$

As an example, consider α = 0.01 and *L* = 0.05. Then $z_{\alpha/2} = z_{0.005}$ = 2.5758, and $n = 2.5758^2/(4 \times 0.05^2) = 663$. This is almost certainly too large a sample size to be practical. The investigator has then to consider whether to relax the requirement for the error probability or for the limit on the amount of error. Suppose a is increased to 0.1 and then $z_{\alpha/2}$ = 1.6449 and $n = 1.6449^2/(4 \times 0.05^2) = 271$. In addition, suppose *L* is increased to 0.10. Then, $n = 1.6449^2/(4 \times 0.10^2) = 68$ which is a more manageable number, but one which is obtained at the price of a reduction of precision, represented by an increase in *L*, and an increased probability, a, of error in that the probability that the true proportion will lie within *L* of the estimate has been decreased.

All of the above discussion has been concerned with absolute error, namely the absolute difference between the estimate and the true value of the characteristic of interest, for example, $|\hat{p}\text{-}p|$ in binomial sampling. An alternative approach considers the relative error of an estimate. The relative error is determined by considering a statistic known as the coefficient of variation which is defined as the ratio of the standard deviation to the mean. This statistic may be used to answer questions of the form: 'Is the estimate "likely" to be correct to within, say, 10% of the total?' Consider this question in the context of binomial sampling. Let us assume that the proportion of interest is thought to be approximately 0.2. Let the desired sample size be n and let $q = 1\text{-}\ p = 0.8$. Then, the mean is np and the standard deviation is $\sqrt{(npq)}$. The coefficient of variation is the ratio of the standard deviation to the mean which, in this context, is $\sqrt{(q/np)}$, and this is set to 0.1. The following equation is then solved for n.

$$\begin{aligned}
\sqrt{(q/np)} &= 0.\ 1 \text{ where } p = 0.2,\ q = 0.8\\
\sqrt{(0.8/0.2n)} &= 0.1\\
\sqrt{(4/n)} &= 0.1\\
\sqrt{n} = 10\sqrt{4} &= 20\\
n &= 400.
\end{aligned}$$

This is quite a large sample. No account is taken of the probabilities of error which were used when absolute errors were discussed. Note the informal use of the word 'likely' in the initial question. It has been interpreted as meaning that the standard deviation is 10% of the mean and thus an interval estimate of the true proportion of the form (mean ± standard deviation) will be correct to within 10%.

Another approach can be considered when the characteristic of interest is thought to be rare. This approach was first described by Haldane (1945), and a summary is given in Cochran (1977, p. 77). This latter reference, Cochran (1977), contains a wealth of detail about many aspects of sampling techniques and is recommended for further reading on this topic. The method used is known as inverse sampling. Let the proportion of the characteristic of interest in the underlying population be p, and it is assumed that p is small. Sampling is conducted until m items have been found with the characteristic of interest. Let n be the sample size at which the mth rare item has appeared ($m > 1$). For n very large and p small, with $m \geq 10$, a good approximation to the variance of the estimate of p can be shown to be $mp^2(1\text{-}p)/(m\text{-}1)^2$. The coefficient of variation is, then, equal to $\sqrt{\{m(1 - p)\}}/(m - 1)$. Since p is assumed small, $(1 - p)$ is close to, but still less than, 1. Thus, an upper limit on the coefficient of variation is given by $\sqrt{(m)}/(m - 1)$. Note that this is independent of p. Solution, for m, of the equation $\sqrt{(m)}/(m - 1)$ = coefficient of variation, where the coefficient of variation is assumed known, tells you for how long samples must be taken. For example, if the coefficient of variation equals

0.2	$m = 27$
0.1	$m = 102$
0.05	$m = 401.$

This process could lead to very large sample sizes. The figures given are for the number of items with the characteristic, which is known to be rare. The total number of items to be considered will, in general, be much larger than this. As a rule of thumb, the total sample size needed will be m/p where p is the unknown proportion. Thus, if $m = 401$ and $p = 0.1$, the sample size will be 4010.

The above theoretical discussion assumes that the sample is small relative to the size of the underlying population. Consider an infinite population in which the population variance of the characteristic of interest is σ^2. Given a sample of size n, the sample mean has a variance σ^2/n. If the population is finite, of size N say, then the variance of the sample mean is now

$$\frac{(N-n)}{N}\frac{\sigma^2}{n},$$

(Cochran 1977, pp. 24–25). The only change is the introduction of the factor $\{(N - n)/\mathrm{N}\}$ for the variance. The standard deviation of the mean, or standard error as it is sometimes known, is altered, correspondingly, by the factor $\surd\{(N\text{-}n)/\mathrm{N}\}$. These factors are called the finite population correction (fpc) factors. The ratio, n/N, is known as the sampling fraction. So long as this fraction remains low, these factors are close to one and the size of the underlying population has no direct effect on the standard error of the mean. In practical applications, the fpc can be omitted when the sampling fraction is less than 0.1. In such a case the fpc of the standard error equals $\surd 0.9$ or 0.95. The small influence of the sampling fraction has the remarkable implication that the precision of the mean of a sample of size 100, as an estimator of the mean of an underlying population, is almost as good if the population is of size 200 000 as if the population is of size 2000. The following examples, where the fpc for the standard error is given for a sample size of 100 and various population sizes, illustrate the point.

Sample size	Population size	sampling fraction
n	N	$\surd\{(N\text{-}n)/N\}$
100	2 000	0.9747
	20 000	0.9975
	200 000	0.9998

Assume the standard error, without the fpc, is $\sigma/\sqrt{n}$ or $\sigma/10$ when $n = 100$. When the fpc is included the estimate of the standard error is reduced by a factor of 0.9747 when $N = 2000$. When $N = 200\,000$, the estimate is reduced by a factor of 0.9998. The precision obtained by a sample of size 100 from a population of 2000 is thus $0.9747\sigma/10$. To obtain this precision from a population of size 200 000 will require a sample of size 10 000, namely an increase by a factor of 100. By using a sample of only size 100, the precision has worsened from $0.9747\sigma/10$ to $0.9998\sigma/10$. An increase in the sample size from 100 to 10 000 will improve the precision of the estimate of the sample mean by a factor of (0.9998–0.9747)/0.9998 or 2.5%. One has to question the relative merits of an improvement in precision of an estimate by a factor of 2.5% and the increased cost of taking a sample of size 10 000 rather than a sample of size 100.

6. BACKGROUND DATA AND AVERAGE PROBABILITIES

It is often the position that a forensic scientist has found similarities between characteristics of trace evidence found at the scene of a crime and trace evidence found on a suspect. The forensic scientist wishes to assess the value of this similarity. If background data were available from some relevant population (see Chapter 4 for a further discussion of this, and Chapter 5 (Paternity) for an example of a bad choice of background population), then an estimate of the Bayes' factor (see section 4 and Chapter 3) may be made. This factor compares the likelihood of the evidence if the suspect was guilty with the likelihood of the evidence if the suspect was innocent. A brief review of background data collections has been given by Evett (1990), and references therein are given for data sets relating to glass fragments, fibres, and genetic markers. However, it may be that such background data are not available. Examples in which this may be the case are hair comparisons where there is no agreed characterisation of hair (see Chapter 5), paint fragment comparisons (Tippett *et al.* 1968), or sole tread pattern comparisons (Groom & Lawton 1987). To obtain a feel for the strength of the evidence, forensic scientists use the following empirical approach.

Use is made of whatever objects of relevance are to hand, such as head hairs from students or laboratory assistants, paint samples from buildings in South Wales, Mid-Wales, and Monmouthshire, or production records from a shoe factory in New Zealand. Pairwise comparisons are then made of items known to be from two different sources. The number of comparisons made in practice can be quite large. Gaudette & Keeping (1974) made 366 630 comparisons of pairs of hairs known to be from different sources. The number, r say, of such comparisons in which the two items of interest (hairs, paint fragments, shoe imprints, etc.) are not found to be distinguishable is noted and compared with the total number, n say, of comparisons which have been made. The ratio, r/n, may then be used as a measure, in some sense, of the strength of the evidence. It might be said to be an estimate of the probability that two items chosen at random from some population will be found to be similar; it is questionable if this probability is of relevance in a criminal investigation. An extra degree of sophistication may be added as was done by Gaudette & Keeping (1974) by comparing several members, known to be dissimilar, of one source with one member of the other. Gaudette &

Keeping chose six to eleven hairs from one source and compared these hairs with one from another source. They were able to distinguish hairs from different sources in all but nine comparisons. Their value for *r/n* was, therefore, 9/366 630 or 1/40 737. It was then argued that if nine dissimilar hairs were independently(?) chosen to represent the hairs from one person, the chance (their word) that a single hair from a different person would be distinguishable from all nine of the first person's would be $(1 - (1/40\ 737))^9$, or approximately 1-(1/4500). It can then be argued that the probability is about 1/4500 that in at least one of the nine cases the two hairs examined would be indistinguishable.

This line of reasoning depends on an assumption of independence; though, as explained below, this assumption is not an important one. The 'chance' that a single hair from one person would be distinguishable from a single hair, chosen at random, from a different person was calculated by Gaudette & Keeping to be 1 - (1/40,737), (1 - *p*), say. It is then argued that if nine hairs were chosen, at random, from this first person and compared with the single hair from the second person, then the probability that all nine hairs from the first person would be distinguishable from the single hair from the second person is the product of the nine individual probabilities, namely $\{(1\text{-}p) \times (1\text{-}p) \times \ldots \times (1\text{-}p)\}$ or (1-*p*)9. This argument makes use of the independence assumption since the individual probabilities are assumed to be unaffected by the knowledge that previous hair comparisons between hairs from the two heads have been shown to be distinguishable. Also, the individual probabilities are all assumed to be equal, and the effect of this assumption is discussed later.

The independence assumption is not all that important in the particular case discussed by Gaudettte (Gaudette 1982). If the independence assumption is dis-pensed with, the figure of 1 in 4500 remains more or less the same. This is so because of an inequality known as the Bonferroni inequality which states that the probability that at least one of several events occurs is never greater than the sum of the probabilities of the occurrences of the individual events. If the events are denoted by E1, E2…, E_n, then the inequality states that *P* (at least one of $E_1, E_2,\ldots, E_n$ occurs)

$$\leq P(E_1) + P(E_2) + \ldots + P(E_n).$$

In Gaudette's example, there are nine events, so $n = 9$, corresponding to the inability to distinguish hairs in comparisons made of each of nine hairs known to be from one source with the one hair known to be from another. The events E_1, $E_{2\ldots}$, E_9 denote these events symbolically. Thus, the probability of interest, which is the probability of at least one indistinguishable pair in these nine comparisons, is the probability that at least one of $E_1, E_2, \ldots, E_9$ occurs. By the Bonferroni inequality, this probability is never greater than the sum of the individual probabilities. From the discussion above, these individual probabilities are all taken to be equal to 1/40 737. The sum of nine of them then equals 9/40 737 which is approximately 1/4526. This is very close to the figure of 1/4500 quoted from the original experiment and shows that even if independence is not assumed there is very little change in the probability figure quoted as a measure of the value of the evidence.

Tippett *et al.* (1968) conducted a similar experiment in relation to paint fragments. They made 1 937 496 pair-wise comparisons of paint fragments known to be from different sources and found only two comparisons in which the fragments were found to be indistinguishable. Thus, they argued, if ‘two individuals each took a paint sample, at random, from anywhere in the area (of the study), then the chance that the two samples would match is 968 748 to one against.

One has to be very careful about the interpretation of these ressults. Tippett *et al.* (1968) made no further claim for their results. However, Gaudette & Keeping (1974) went on to say ‘(i)t is estimated that if one human scalp hair found at the scene of a crime is indistinguishable from at least one of a group of about nine dissimilar hairs from a given source, the probability that it could have originated from another source is very small, about 1 in 4500’. This argument is placing the conditioning the wrong way round, as has been discussed in Chapter 1. The figure 1/4500 refers to the probability that one hair from the head of person, A say, is not found to be dissimilar to one or more hairs from a sample of nine from the head of a different person, B say. This may be understood better with the help of some symbols. Thus, let

X denote the event that the two sets of hairs, one consisting of one hair, the other consisting of nine hairs, come from different sources,

Y denote the event that the hair from the set consisting of only one hair is not found to be dissimilar to one or more hairs from the other set, consisting of nine hairs.

Then, the figure 1/4500 refers to $P(\mathrm{Y}|\mathrm{X})$. However, what the quoted statement from Gaudette & Keeping refers to is the event $(\mathrm{X}|\mathrm{Y})$, and the probability of this event $P(\mathrm{X}|\mathrm{Y})$ may be very different from $P(\mathrm{Y}|\mathrm{X})$, as was explained in Chapter 1, section 11.

Of course, the research of Gaudette & Keeping is valuable, and it showed that hair comparison can distinguish hairs from different sources. The danger lies in the fact that the wrong interpretation is placed on the quantitative results. Awareness of this danger was illustrated by Groom & Lawton (1987). In their investigation there were 90 possible non-matching pairs of shoes. They compared 120 soles and found only 2 which were indistinguishable with regard to the features deemed to be of interest, a rate of 1 in 60. Therefore, they argued, ‘the chance of obtaining a pair of shoes with both soles exhibiting the same characteristics as the shoeprints found at the crime scene would be approximately 1 in 3600’. Little more is said of a quantifiable nature. Their conclusion was

> ‘As there were only 90 possible non-matching pairs of shoes and at least 3600 possible variations, the likelihood of the sole pattern on any of those 90 pairs of shoes being duplicated was very remote. The probability that the suspect’s shoes caused the bloodied shoeprints was, therefore, extremely high.’

Probabilities constructed in this way are known as average probabilities. A simple numerical example will illustrate the reason for the name ‘average’ and the misleading nature of these probabilities for interpretation. Much of the following discussion is summarised in Aitken (1987) with related comments in Gaudette (1987). Consider the ABO system of bloodgrouping. The four phenotypes and their frequencies for a white California population (Berry & Geisser 1986, p. 357) are as follows.

Phenotype	O	A	B	AB
Phenotypic frequency	0.479	0.360	0.123	0.038

Assume, for the purposes of argument, that a sample of 1000 white Californians, with blood group frequencies in exactly the above proportions, are available for study. Thus there are 479 people with phenotype O, 360 with phenotype A, 123 with phenotype B, and 38 with phenotype AB. All possible pairs of people in this sample are considered and their phenotypes compared. There are 1000 × 999/2 = 499 500 different pairings. This is the '*n*' referred to earlier. The number of pairings in which the phenotypes of the two people are the same may be calculated by considering each of the four different phenotypes separately and summing the four answers. Thus there are:

(479 × 478/2) pairings of people who are both phenotype O,
(360 × 359/2) pairings of people who are both phenotype A,
(123 × 122/2) pairings of people who are both phenotype B,
(38 × 37/2) pairings of people who are both phenotype AB.

There are, thus,

$(479 \times 478/2) + (360 \times 359/2) + (123 \times 122/2) + (38 \times 37/2) = 114481 + 64620 + 7503 + 703 = 187\,307$

pairings of people who have the same phenotype. This is the '*r*' referred to earlier. The ratio, *r/n,* is the measure of the strength of the bloodgroup evidence. In this case,

$$r/n = 187\,307/499\,500 = 0.375.$$

A figure derived in this way is an estimate of a probability which is sometimes called an 'average probability'. It is so-called for the following reason. A blood stain of a particular phenotype, O say, is found at the scene of a crime. A suspect is found, and his phenotype is O also. If the suspect is innocent, the blood stain at the crime scene may be thought of, effectively, as one from a person chosen at random from some large reference population, the problems of definition of which were discussed in section 3. The probability of a match with the phenotype of the suspect is, thus, the relative frequency of the O phenotype in this population, say 0.479 for the white California population above. A similar argument provides probability figures of 0.360, 0.123, and 0.038 for matches between blood stain phenotypes and suspect phenotypes of type A, B, and AB, respectively. What is known as the average probability is the average of these four values. It is not sufficient to calculate the simple arithmetic average {(0.479 + 0.360 + 0.123 + 0.038)/4}, however, since these four situations (matches of scene stain phenotype and suspect phenotype for each of O, A, B, and AB) are not equally likely. In fact, their relative likelihood is just their relative frequencies in the general population, namely 0.479, 0.360, 0.123, and 0.038. The average probability uses the relative likelihoods as weighting factors, and is

$$\begin{aligned}&(0.479 \times 0.479) + (0.360 \times 0.360) \times (0.123 \times 0.123) + (0.038 \times 0.038)\\&= 0.479^2 + 0.360^2 + 0.123^2 + 0.038^2\\&= 0.376.\end{aligned}$$

This is very similar to the value 0.375 calculated for *r/n* above. The factor 1/2 in the numerator and in the denominator there cancelled out and instead of 478/1000, 359/1000, 122/1000, and 37/1000 there, the figures 0.479, 0.360, 0.123, and 0.038 have been used. See also Smalldon & Moffatt (1973) and the discussion on discriminating power in Chapter 4.

The value 0.376 is the exact value for this so-called average probability and can be determined only if the relative frequencies are known exactly. This is the case for phenotypic frequencies. However, it is not often so. If relative frequencies are not known, then resort has to be made to empirical evidence based on pair-wise comparisons between individuals known to be different. The more pair-wise comparisons which can be made the better. The reason for the discrepancy between the exact average probability, 0.376, and the empirical value, 0.375, based on a hypothetical sample of 1000 was the relatively small value, 1000, for the sample size. As the sample size increases, so the empirical value approaches the true value. For example, if the sample size is taken to be 1 million, the empirical value is calculated to be 0.3756 which should be compared with a true value, to four significant figures, of 0.3756 and an empirical value, based on a sample of size 1000, of 0.3750, again to four significant figures. Average probabilities are mentioned again briefly in Chapter 5 (Evidence from blood).

An artificial numerical example, in the context of hair comparisons, taken from Aitken & Robertson (1987), illustrates the misunderstanding that might arise. Gaudette & Keeping (1974) took an initial sample of 80 to 100 hairs from an individual and then chose six to eleven mutually dissimilar hairs for their comparison purposes. This selection procedure may well destroy the randomness of the original sample and alter the estimate of the probability of the observed similarity in two hair samples, given they did not come from the same source. This will be so because the sampling procedure takes no account of the relative frequencies of each hair type within the head of any particular individual.

Consider two people, A and B. Each has hair on their head of nine different types. The types on the head of A are labelled a_1, a_2,…, a_9 and the types on the head of B are labelled b_1, b_2,…b_9. Types a_1 and $b_1$represent 25% of the hairs on the heads of A and B, respectively. Types a_2,…, a_9 and b_2,…, b_9 are each represented equally in the remaining 75%, that is, with relative frequency (1/8) (3/4) = (3/32). Suppose that one hair type of A is indistinguishable from one hair type of B, using the techniques currently available for distinguishing hair types. Various possible estimates may be obtained for the probability that a hair selected from A is found to be indistinguishable from a hair selected from B. For ease of exposition, denote by M the event that ‘a hair from A is found to be distinguishable from a hair from B’.

Using Gaudette’s sampling procedure, each of the nine mutually dissimilar hairs from A and from B will be equally represented for the comparison purposes. There will be a sample of nine hairs from each of A and B, both consisting of one and only one hair from each of the nine types to be found on the heads of A and B. There will be $9^2 = 81$ different comparisons of hairs from the two heads. Among these 81 different comparisons, there will be one and only one in which the two hairs are found to be indistinguishable. The estimate of the probability of M in this case is then 1/81 or 0.0123.

Consider, now, sampling procedures in which the sampling is representative of the proportion of each hair type within the individual's head. For example, 32 hairs could be selected from each of A and B. The sample from A would consist of 8 hairs of type a_1 and 3 hairs from each of the eight types a_2 to a_9, making $8 + (3 \times 8)$ or 32 hairs in total. The sample from B would consist of 8 hairs of type b_1 and 3 hairs from each of the eight types b_2 to b_9, again making 32 hairs in total. The total number of comparisons is $32^2 = 1024$. These 1024 comparisons may be divided into four groups.

(i) There are 8 × 8, equals 64, comparisons of a hair of type a_1 with a hair of type b_1.
(ii) There are 8 × 24, equals 192, comparisons of a hair of type a_1 with a hair of one of the types b_2 to b_9.
(iii) There are 24 × 8, equals 192, comparisons of a hair of type b_1 with a hair of one of the types a_2 to a_9.
(iv) There are 24 × 24, equals 576, comparisons of a hair of one of the types a_2 to a_9 with a hair of one of the types b_2 to b_9.

Groups (ii) and (iii) may themselves be subdivided into eight sub-groups each, as follows. There are 8 × 3, equals 24, comparisons of a hair of type a_1 with a hair of type b_2. Similarly there are 24 comparisons of a hair of type a_1 with each of one of the types a_2 to a_9. Group (iv) may be subdivided into 64 sub-groups as follows. There are 3 × 3, equals 9, comparisons of a hair of one of the types a_2 to a_9 with a hair of one of the types b_2 to b_9. (Remember that the selection procedure ensured that there were three hairs of each of the types a_2 to a_9 and b_2 to b_9.)

There are three possible outcomes.

(i) Types a_1 and b_1 are indistinguishable. There are 64 comparisons out of 1024 in which a hair of type a1 is compared with a hair of type b_1. Thus, the estimate of the probability of M from this outcome is 64/1024 = 0.0625.
(ii) Type a_1 is indistiguishable from one of b2 to $b_{9.}$ There are 24 comparisons out of 1024 in which a hair of type a_1 is compared with a hair of one of the types b_2 to b_9. Thus, the estimate of the probability of M from this outcome is 24/1024 = 0.0234.
(iii) One of the types a_2 to a_9 is indistinguishable from one of the types b_2 to b_9. There are nine comparisons out of 1024 in which a hair of one of the types a_2 to a_9 is compared with a hair of one of the types b_2 to b_9. Thus, the estimate of the probability of M from this outcome is 9/1024 = 0.00879.

Notice that these three probabilities are all very different from the figure 0.0123 based on Gaudette's sampling procedure.

The average probability of Gaudette is obtained when each of the three outcomes described above is weighted by the number of times it would occur when using Gaudette's sampling procedure. In Gaudette's procedure it is assumed that each hair type is equally represented. Thus there would be one comparison of an A hair with a B hair where both were of a common type, a_1 and b_1 respectively. There would be sixteen pairs where one hair was common and the other rare, namely a hair of type a_1

compared with each of the eight representatives of b_2 to b_9 and each of the eight representative of a_2 to a_9 compared with b_1. The remaining 64 pairs out of the 81 in total would involve comparisons of a rare hair type (a_2 to a_9) with a rare hair type (b_2 ro b_9). Thus, the probability of M would be evaluated as

$$[(0.0625 \times 1) + (0.0234 \times 16) + (0.00879 \times 64)]/81 = 0.0123.$$

This is an average probability, and the method of its construction explains the reason for the terminology. However, the average probability does not bear much resemblance to any of the other probabilities calculated when sampling was assumed to be proportional to the relative frequency of the hair types.

The probabilities quoted by Gaudette & Keeping (1974), Tippet *et al.* (1968), and Groom & Lawton (1987) are all examples of average probabilities. The average probability is used as a measure of the probability of the evidence of association, conditional on there being no association, in fact. Remember that the calculation for the phenotype example included an assumption that the suspect was innocent and, hence, that the blood stain at the crime scene was, effectively, selected at random from some general population. However, this is not the situation in practice. To return to the phenotype example, the probability of a match, given that the suspect is of phenotype AB, is the probability that the criminal is of phenotype AB, conditional on the suspect being of phenotype AB. If it is assumed that the suspect is innocent, then this probability is just the probability that a person, the person who left the bloodstain, chosen effectively at random from some general population, is of phenotype AB. This probability is 0.038; it bears very little resemblance to the average probability 0.376.

I.W.Evett has drawn an analogy between the use of average probabilities in forensic science for the evaluation of evidence and the use of a broken clock for timekeeping. The broken clock will tell the right time twice a day. It is not known when it will be right, but at least it will be right sometimes. Similarly, an average probability will be right sometimes but it is not known when. Gaudette (1987) explained how even a broken clock may be of some use when there are few other sources of guidance available. ‘Assume a man crossing a desert notes that his watch has stopped. A while later he encounters another man who, having no watch, asks him the time. The first man could give one of three possible answers:

(a) “I don’t know. My watch is broken.”
(b) “My watch says 2 o’clock.”
(c) “When my watch stopped it was 2 o’clock. I can’t give you the exact time now, but based on the position of the Sun in relation to what it was when my watch stopped, I would estimate that it is a few hours after 2.”’

Average probabilities, based on the empirical evidence obtained from the pairwise comparisons of items of interest, such as hairs, paint fragments, and so on, do provide some idea of the value of particular types of evidence. Thus, the use of the equivalent answer to (a), namely that the forensic scientist does not have any idea about the value of the evidence, is not reasonable. The use of the equivalent answer to (b) would be very misleading. Compare the average probability of 0.376 for the blood group example with a particular value of 0.038 obtained when the phenotype was

AB. Answer (c) is attractive and is perhaps possible in the case of a wanderer in the desert. However, there is a danger that the words 'a few hours after', or their analogy in a particular legal case, may be forgotten. Another danger is that the hair examiner is not able to quantify the discrepancy between the case in question and the average probability. He may not be aware of the calculations and assumptions that have been made when the figure for the average probability was obtained. Thus, he may not be able to explain the value of the hair evidence in the particular case in question relative to the 'average' case. There is a possibility that average probabilities, such as 1/ 4500, obtained by Gaudette & Keeping, may be used indiscriminately without caveats. Such use will be very misleading.

Where frequency data are not available, an approach to the evaluation of evidence may be made by using pair-wise comparisons and calculating average probabilities. However, it has to be realized that this is only an approach. Some idea of the worth of the evidence for identification purposes, in general, may be obtained. The use of average probabilities in particular cases is a different matter. They should be used with great care, if at all.

7. ZERO OCCURRENCES

It is possible that an expert witness may be asked to comment on the relative frequency of a characteristic when, in his own experience, he has not observed such a characteristic. For example, in comparing fragments of glass of a particular refractive index, found on a suspect with the refractive index of glass found in a window at a crime scene, the value of the refractive index may be one which the forensic scientist has not seen before but about which he may be asked to provide a value for its relative rarity. Alternatively, he may be asked for his opinion about the relative frequency of a particular fingerprint characteristic in some population when he has not seen any previous occurrences of such a characteristic before.

Suppose the scientist has examined n fingerprint characteristics; normally this would be a large number. Suppose that the unknown relative frequency of the characteristic, characteristic X say, in question is p; normally this would be expected to be small or else the scientist would have seen it before. This relative frequency may be interpreted as the probability of observing the characteristic X when the scientist observes one characteristic chosen at random. The probability of not observing X is then the complementary probability, namely $(1 - p)$, q say. The probability of not observing any occurrences of X in the observation of n characteristics is then q^n. Another example in which this approach may be useful is that of *State v Sneed* discussed in section 4. In that case, an expert scanned telephone books and found no occurrences of the name Crosset in approximately 129 million listings. The problem is now to find a reasonable value for p, the frequency of Crossets in the population defined by that population of which the telephone listings might be thought a representative sample. One has observed 129 million listings; this value is taken as n. The probability of observing that number, n, of listings without finding a Crosset, if the true frequency of Crossets is p, is q^n, denoted symbolically by $P(n\backslash q)$. What is of interest, however, is the probability of q, and hence of p, given that 129 million listings have been observed without any occurrence of a Crosset, denoted symbolically by $P(q\backslash n)$. To do this, a version of Bayes' rule as described in Chapter 1 is required to relate $P(q|n)$ and $P(n|q)$. This version requires use to be made of information about q known before the search of the telephone listings was made. This information is represented by what is known as a prior distribution for q which represents our ideas about the likely values of q with probability figures, represented symbolically by $P(q)$. In this case, and in almost all other cases, larger values of q will tend to be more likely than smaller values of q, since this technique will be of most use when are events are being considered. However, one can act in a conservative manner by assuming that all values of q are equally likely, indicating this assumption by letting $P(q)$ equal a constant throughout the range of possible values of q, namely 0 to 1. This is conservative for the following reason. Using this constant form for $P(q)$, smaller values of q are considered more probable than if a more realistic form for $P(q)$ had been chosen. An accused is favoured by small values of q and hence large values of p. This is so since a large value of p implies that the relative frequency of the characteristic in question (for example, that one's name is Crosset or that one has a characteristic X within one's fingerprint) within the general population is large, and hence that the probability of finding such a characteristic on an innocent person is also large . This use of a constant value for $P(q)$ over the range of q should therefore not be controversial since its use favours the accused. $P(n|q)$ and $P(q)$ can be combined to give an expression for $P(q|n)$. This expression is then used to provide a useful single value for q. One sensible value to use is that value, q_0 say, such that the probability the true value of q is less than q_0 is some small value, ε say. This implies that there is only a probability ε that the true value of p, the probability of interest, is greater than 1 -q_0. This value of p is, in this probability sense, an upper bound on the estimate of the true value of the relative frequency of the characteristic of interest. The value of ε is chosen in advance and represents the error that the scientist is prepared to tolerate in his estimate. If the courts disagree with the choice of e then it is a simple matter to substitute another value for ε in the formula and determine a new value for $q_{0.}$ Given e and n, it can be shown (Finkelstein 1978, p. 98.) that $q_0$is given by

$$q_0=\varepsilon^{1/(n+1)},$$

and, hence, that $p = 1 - q_0$. Table 1 shows how the value of p varies for different values of n and ε.

For the case of *State v Sneed* described above, if ε = 0.05 and n = 129 million, an estimate of the proportion of Crossets in the population is 1 in 232 million. With ε = 0.01, the proportion is calculated to be 1 in 356 million.

8. SUMMARY

The relative importance of an event may be measured by considering what is known as its likelihood. To measure the likelihood of an event it is necessary to refer to a

Table 1 —The estimates for the upper bound (p) for the relative frequency of the characteristic of interest, given the number, n, of observations in which the characteristic has not been observed and the error probability, ε, that the true value of the relative frequency may be greater than p.

n	ε	p
100	0.05	0.03
	0.01	0.04
200	0.05	0.01
	0.01	0.02
500	0.05	0.006
	0.01	0.009
1000	0.05	0.003
	0.01	0.005
10 000	0.05	0.0003
	0.01	0.0005

population, known as the relevant population, of which this event is a member and from which likelihoods may be determined. However, the definition of such a population is not always easy. Usually, it is not practical, even if the population can be defined, to consider the relevant population in its entirety. In such circumstances a sample is taken from the population. There is a well-established methodology of sampling, and this chapter introduces some of the more frequently used terminology. In some cases, once the population is defined the construction and use of a sampling frame may also be difficult.

The process of selection of events to report causes effects of which it is necessary to beware. It is very easy to make a large number of comparisons and, when it comes to report your conclusions, to report only the number of unsatisfactory comparisons which were also made but now forgotten. Also, the procedure by which a defendant was identified could have considerable relevance; the Collins case is a prime example of a case in which this procedure was not explained fully at the trial.

Probability statements concerning the frequency of events will often be based on the results of an experiment designed to investigate the event in question. Two common and important questions that an experimenter will ask before starting are ‘how large a sample is required in order to estimate an effect with a given precision?’ and ‘how large a sample should I take in order to detect an effect?’ A discussion of possible answers to these questions is given.

The ability to use statistical techniques is greatly enhanced if background data exist. Such data enable the relative importance of different characteristics or measurements to be measured. However, it is often the case that background data do not exist or that their collection is not practical. In such cases some idea of the worth of a particular evidential type may be obtained by determining an average probability. An average probability is very much an average, though, and as such needs careful interpretation.

The final section discusses the problems which may arise if a forensic scientist is asked to comment on the relative frequency of a characteristic when he has never observed such a characteristic. It is not necessary, or desirable, to say that the relative frequency is zero if it is perfectly possible for the characteristic to exist; instead, an upper bound may be calculated for the relative frequency, and certain examples are given.

REFERENCES

Aitken, C.G.G. (1987) The use of statistics in forensic science. *Journal of the Forensic Science Society* **27** 113–115.

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

Berry, D.A. & Geisser, S. (1986) Inference in cases of disputed paternity. In: *Statistics and the law,* (DeGroot, M.H., Fienberg, S.E., & Kadane, J.B. eds) John Wiley and Sons Inc., New York.

Cochran, W.G. (1977) *Sampling techniques* (3rd edition). John Wiley and Sons Inc., New York.

Cohen, L.J. (1977) *The probable and the provable.* Clarendon Press, Oxford, p. 75.

Coleman, R.F. & Walls, H.J. (1974) The evaluation of scientific evidence, *Criminal Law Review* 276–287.

Cox, D.R. (1958) *Planning of experiments.* John Wiley and Sons, Inc. New York.

Edgington, E.S. (1987) *Randomization tests* (2nd edition), Marcel Dekker Inc., New York.

Eldridge, M.A., Nimmo-Smith, I., & Wing, A.M. (1984) The variability of selected features in cursive handwriting: categorical measures. *Journal of the Forensic Science Society* **24** 179–219.

Evett, I.W. (1984) A quantitative theory for interpreting transfer evidence in criminal cases. *Applied Statistics* **33** 25–32.

Evett, I.W. (1990) The theory of interpreting scientific transfer evidence. In: *Forensic Science Progress* **4**, Maehly A. & Williams, R.L. (eds) Springer Verlag, 141–180.

Fienberg, S.E. (1986) Gatecrashers, blue buses and the Bayesian representation of legal evidence. *Boston University Law Review* **66** 693–699.

Finkelstein, M.O. (1978) *Quantitative methods in law.* The Free Press, pp. 83–84.

Finney, D.J. (1977) Probabilities based on circumstantial evidence, *Journal of the American Statistical Association* **72** 316–318.

Gaensslen, R.E., Bell, S.C., & Lee, H.C. (1987a) Distributions of genetic markers in United States populations: I. Blood group and secretor systems. *Journal of Forensic Sciences* **32** 1016–1058.

Gaensslen, R.E., Bell, S.C., & Lee, H.C. (1987b) Distributions of genetic markers in United States populations: II. Isoenzyme systems. *Journal of Forensic Sciences* **32** 1348–1381.

Gaensslen, R.E., Bell, S. C, & Lee, H.C. (1987c) Distributions of genetic markers in United States populations: III. Serum group systems and hemoglobin variants. *Journal of Forensic Sciences* **32** 1754–1774.

Gaudette, B.D. (1982) A supplementary discussion of probabilities and human hair comparisons. *Journal of Forensic Sciences* **27** 279–289.

Gaudette, B.D. (1987) The use of statistics in forensic science. *Journal of the Forensic Science Society* 27 117–118.

Gaudette, B.D. & Keeping, E.S. (1974) An attempt at determining probabilities in human scalp hair comparison. *Journal of Forensic Sciences* **19** 599–606.

Groom, P.S. and Lawton, M.E. (1987) Are they a pair? *Journal of the Forensic Science Society* **27** 189–192.

Gullberg, R.G. (1987) Duplicate breath testing: statistical vs forensic significance of differences. *Journal of the Forensic Science Society* **27** 315–319.

Haldane, J.B.S. (1945) On a method of estimating frequencies. *Biometrika* **32** 222–225.

Hammersley, J.M. & Handscomb, D.C. (1965) *Monte Carlo methods.* Methuen, London.

Kendall, M.G. & Buckland, W.R. (1982) *A dictionary of statistical terms* (4th ed). Longman Group Ltd, London.

Kingston, C.R. (1964) Probabilistic analysis of partial fingerprint patterns. D.Crim dissertation, University of California, Berkeley, USA.

Kingston, C.R. (1965a) Applications of probability theory in criminalistics *Journal of the American Statistical Association* **60** 70–80.

Kingston, C.R. (1965b) Applications of probability theory in criminalistics—**II**, *Journal of the American Statistical Association* **60** 1028–1034.

Kingston, C.R. (1988) Discussion of 'A critical analysis of quantitative fingerprint individuality models'. *Journal of Forensic Sciences* **33** 9–11.

Kirk, P.L. & Kingston, C.R. (1964) Evidence evaluation and problems in general criminalistics. Presented at the Sixteenth Annual Meeting of the American Academy of Forensic Sciences, Chicago, Illinois, in a symposium: *The Principles of Evidence Evaluation.*

Lenth, R.V. (1986) On identification by probability. *Journal of the Forensic Science Society* **26** 197–213.

Lindley, D.V. & Scott, W.F. (1988) *New Cambridge elementary statistical tables.* Cambridge University Press, Cambridge, UK.

People v Collins (1968) *California Reporter* **66** 497.

Rawson, R.D., Vale, G.L., Sperber, N.D., Herschaft, E.E., and Yfantis, A. (1986) Reliability of the scoring system of the American Board of Forensic Odontology for human bite marks. *Journal of Forensic Sciences* **31** 1235–1260.

Salmon, D. & Salmon, C. (1980) Blood groups and genetic markers polymorphisms and probability of paternity. *Transfusion* **20** 684–694.

Smalldon, K.W. & Moffatt, A.C. (1973) The calculation of discriminating power for a series of correlated attributes. *Journal of the Forensic Science Society* **13** 291–295.

Smith, R.L. & Charrow, R.P. (1975) Upper and lower bounds for the probability of guilt based on circumstantial evidence. *Journal of the American Statistical Association* **70** 555–560.

Snedecor, G.W. & Cochran, W.G. (1980) *Statistical methods* (7th ed) The Iowa State University Press, Ames, Iowa, USA.

Sparks, D.L., Slevin, J.T. & Hunsaker, J.C. (1986) 3-Methoxytyramine in the putamen as a gauge of the postmortem interval. *Journal of Forensic Sciences* **31** 962–971.

Stoney, D.A. & Thornton, J.I. (1986) A critical analysis of quantitative fingerprint individuality models, *Journal of Forensic Sciences* **31** 1187–1216.

Stoney, D.A. & Thornton, J.I. (1988) Response to letter from C.R. Kingston, *Journal of Forensic Sciences* **33** 11–13.

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

Editors' introduction to Chapter 3

The forensic scientist is concerned with the value of his evidence. The meaning of probability is discussed in Chapter 1. There are sound mathematical reasons why the function of the two probabilities, the probability of the evidence given guilt and the probability of the evidence given innocence, that best measures the value of evidence depends only on the ratio of these probabilities. These reasons, for those who wish to study them, are given in the Appendix of Chapter 3. The function of these probabilities which is most sensible to use is the logarithm of the ratio, since this ensures that various independent pieces of evidence contribute in an additive way. (Note that this is quite different from the logarithm of the odds referred to in section 1.8). It is this expression, the logarithm of the ratio of the two probabilities, that is given the name 'weight of evidence' by Professor Good. It should be emphasized that weights of evidence are not probabilities. Thus they are not restricted to lie between 0 and 1; they may be negative and they may have magnitudes considerably in excess of 1. Confusion has arisen in the past because probability itself, that is the probability of guilt given the evidence (together with background information), has been used as a measure of weight of evidence. The word 'weight' suggests that two independent pieces of evidence can be added when both pieces of evidence are taken into account. However, this would be impossible if, for example, each piece of evidence, was conclusive by itself: the sum would be 2 and hence not a probability.

Various parts of Chapters 1 and 3 cover similar ground though in slightly different ways. It is useful to emphasize these areas of similarity so that readers' understanding may be strengthened. Both chapters discuss the relationship between probability and odds. Chapter 1 discusses odds in the context of 'another measure of uncertainty' and derives the Odds Form of Bayes' Theorem in the context of drawing balls from an urn. In Chapter 1 it is presented as a simple result for proportions. A brief discussion of the forensic context then follows. Chapter 3 provides the theoretical derivation of the Odds Form of Bayes' Theorem in the forensic science context.

A distinction is drawn in Chapter 3 between likelihood and Bayesian likelihood. The meaning of likelihood as given by R.A.Fisher, one of the founding fathers of modern statistical thinking, required the existence of a probability model, and this meaning is associated with the frequency interpretation of probability as discussed in Chapter 1. The term 'Bayesian likelihood' refers to the context of subjective probability (of Chapter 1) or 'epistemic' probability.

Both Chapters 1 and 3 refer to 'expectation'. Chapter 1, section 17, gives a particular example, Note 3 of Chapter 3 gives more detail where the term 'expected value' is also used.

Professor Good also mentions utilities and quasi-utilities. A utility is a numerical representation of ones tastes and preferences (DeGroot 1970). A quasi-utility is a utility in which the numerical representation is not precise and is used when the true utilities are difficult to estimate. In the context of this book, weight of evidence is used as a quasi-utility since we are trying to decide whether a hypothesis (guilt) or its negation (not guilt) is true on the basis of evidence. In many applications in forensic science the possible distribution of the characteristics of interest if a suspect is not guilty is quite vague, and the concept of a quasi-utility is more useful than that of a utility.

Later, Professor Good mentions a result of Turing's that the expected weight of evidence can never be negative. (Note that this is quite reasonable in the context of a ratio of two probabilities; the neutral value for the ratio is 1, values greater than one strengthen the hypothesis of guilt; values less than one lessen the hypothesis of guilt.) Professor Good also discusses a result which he says is 'horrifying' in relation to radar. The result to which he refers follows from a realization that weight of evidence is the logarithm to base 10 of the Bayes' factor. Thus a median Bayes' factor of 40 corresponds to a weight of evidence of $\log_{10}(40) = 16$ decibans, with a standard deviation which is three times the square root of the expectation, namely $3\sqrt{16} = 12$. The conventional 95% confidence interval (mean ± two standard deviations) is then $16 \pm 24 = (-8,40)$ db $= (-0.8,4)$ bans. The corresponding Bayes' factors are $10^{-0.8}$ and 10^4 or (1/6, 10000). Thus, if the median Bayes' factor is 40, it is entirely possible by chance alone that the observed Bayes' factor could be as low as 1/6 or as high as 10000. The conclusions drawn from these two extremes would be very different.

Finally, note on p. 97, Professor Good's version of Popper's statement, namely: 'I regard the doctrine that the weight of evidence cannot be a probability as one of the most interesting findings of the philosophy of knowledge'. Certain confusions in courts of law may well have been avoided if it had been recognized that the value of evidence presented in court could not be measured by probabilities of guilt or innocence.

REFERENCE

DeGroot, M.H. (1970) *Optimal statistical decisions.* McGraw-Hill Book Company, New York, p. 86.

3

Weight of evidence and the Bayesian likelihood ratio

I.J.Good
Department of Statistics, Virginia Polytechnic Institute and State University,
Blacksburg, Virginia, 24061, USA.

Before discussing weight of evidence it is worth while to point out that evidence and information have distinctive meanings in ordinary English, though many writers on technical subjects have confused them. You might ask a game-player for information about the game of Go or evidence about whether it is a more difficult game than chess, but you wouldn't ask simply for evidence about Go. Evidence pertains to whether some hypothesis, theory, or statement is true, but information has a wider meaning. In this chapter we shall be mainly concerned with the concept of weight of evidence in its normal usage in good English, and with a corresponding simple technical concept which captures the normal usage in a simple formula. I shall not use the expression to mean the weight of the paper on which some evidence is printed.

The usual concept of weight of evidence dates back at least as far as ancient Greek mythology. Themis, the goddess of justice, was associated with a pair of scales (Larousse 1968, p. 136), and clearly these scales were intended to weigh the arguments for and against some thesis, such as an accusation of guilt. Thus the ancient Greeks must have had an informal concept of a quantity of evidence having some kind of additivity. In the English language the expression 'weight of evidence' usually expresses the same concept of a balance of the evidence. This concept can be expressed in terms of probability by means of a simple formula that will be shown, under natural assumptions, to be essentially unique. We shall also discuss briefly the history of the formula. Because probabilities are often difficult to estimate, the formula does not solve all problems of evaluating evidence, but it is a useful step forward of considerable philosophical interest. To emphasize the difficulties of probability estimation, the formula could be called 'semi-quantitative'. In legal applications the difficulties are generally greater than those in medical diagnosis where research on 'expert systems' has made big advances. In the law, the greatest benefits of a semi-quantitative approach might come from the philosophical and therefore practical insights and the emphasis on intuitive ways of expressing judgements not yet familiar to the legal profession.

Let us imagine that a man is accused of some crime and that he is either guilty (hypothesis G) or innocent (hypothesis $\overline{G}$, meaning not G). We here ignore the possibility of partial guilt, and by guilty we mean that he actually committed the crime, not that he was 'found guilty'. That is, we are not using the expression in the familiar newspaper headline sense JONES GUILTY which means that a jury or judge asserted that Jones was guilty as if the law always arrives at a correct verdict. Similarly, by 'not guilty' I mean that he did not commit the crime, not merely that evidence was insufficient for a conviction. Our usage is consistent with the definitions in the *Oxford English Dictionary* but conflicts somewhat with the cliché that a man is 'presumed innocent until he is found guilty' preferably said in a deep voice and fairly slowly. The foreman of a jury is not encouraged to state the probability of guilt although he would sometimes be more honest if he did. He would then be accused of contempt of court (Good 1983a, 1986). At present, the law has a yes-or-no hang-up.

It will be convenient to imagine a trial by a single judge or magistrate without a jury. (Or we could imagine the thinking of a detective.) This will enable us to avoid complicating the issue by discussing multisubjective probabilities, although that is a worthwhile topic. (See, for example, DeGroot 1988.) Let us assume further that the judge is rational in the sense that he behaves as if he accepts the usual axioms of subjective (personal) probability. For a discussion of these axioms see, for example, Chapter 1 of this book or Good (1950, 1982a, b, 1983b, 1987). In the present chapter the main axiom is the product axiom, $P(C\&D) = P(C)P(D|C)$ which is a shorthand for $P(C\&D|F) = P(C|F)\,P(D|C\&F)$, where C, D, and F are propositions asserting that various events occurred.(1)† Note that the vertical stroke, which is a standard notation (not an oblique stroke), always separates what occurs to its left and right, so it is unnecessary, for example, to write ($C\&D$) in place of $C\&D$. If preferred, we can think of C, D, and F as denoting events instead of propositions, where *event* is to be understood in a wide sense.

Let us denote by E the evidence presented in court, and by I all the background knowledge or information available to the judge. We wish to express, in terms of probability, *the weight of evidence in favour of guilt G provided by the evidence E given the background information I* all along. (The probabilities will usually be regarded as subjective or personal.) Denote this weight of evidence by W$(G: E|I)$, where the colon is read *provided by* and the vertical stroke by *given*. Thus W depends upon three propositions.

So far, we have merely introduced some notation, but now we make a critical assumption, namely that $W(G: E|I)$ can depend only on the probability of E given that the man is guilty and on the probability given that he is innocent (really innocent, not merely 'found not guilty'). To state this condition more precisely, and in symbols, we assume that W$(G: E|I)$ depends only on $P(E|G\&I)$ and $P(E|\overline{G}\&I)$. This assumption is, I believe, made throughout this book. It seems to us to be no more than common sense, but there are still some philosophers who suggest interpretations of

† Notes are at the end of the text.

W$(G{:}E|I)$ inconsistent with this assumption. At any rate, based on this piece of common sense, it can be proved, as a theorem, that W$(G{:}\ E|I)$ depends only on the *ratio* $P(E|G\&I)/P(E|\overline{G}\&I)$. Various proofs of this simple important theorem, and allied theorems, have been been given (Good 1968, 1984, 1989a, b, c). We give the simplest proof (Good 1989c) in an appendix to this chapter, a proof that does not require any mention of the probabilities of guilt or innocence, and can therefore sometimes be interpreted as a 'non-Bayesian' proof. But for the moment we shall be thoroughly 'Bayesian' and shall present a historically earlier argument in which the probability of guilt is of the essence of the argument.

It is first convenient to recall the most convenient and natural technical definition of odds. If p denotes any proba)bility, then the corresponding odds are defined as $o = p/(1\text{-}p)$. Equivalently $p = o/(l + o)$. Sometimes odds of say 100 are described as 100 to 1 on or 100:1, and odds of 7/2 as 7 to 2 on, while odds of 2/7 can be described as 7 to 2 against, and so on. (The innumerate sometimes get it wrong, as in the quotation The odds of anything coming from Mars are a million to one'.) A slightly different meaning is used when odds of 7 to 2 are given in a bet. To describe the bet as *fair* would mean that the *physical* odds really *are* 7/2 in a sense that we shall not analyse further.[2] Examples of the relationship between probability and odds are shown in the following table which the reader should verify before proceeding.

Probability:	0	1/100	1/5	1/3	1/2	2/3	4/5	99/100	1
Odds:	0	1/99	1/4	1/2	1	2	4	99	∞

Odds of 1 are known as *evens*. If G has odds of 7/2, for example, then $\overline{G}$ has odds of 2/7, a simple reciprocal relationship.

Some people, especially most gamblers, find it easier to think in terms of odds than in terms of probabilities, some prefer probabilities, while some find it helpful to their judgement to switch from one to the other.

Many of us believe that to 'find' a man guilty should mean that the posterior odds of his guilt exceeds some threshold, though the threshold depends on the nature of the punishment. The threshold should be especially high if the punishment is death. These opinions, though obvious to a 'Bayesian', are not universally accepted. One well-known 'non-Bayesian' statistician, with whom I discussed the matter, said he would just decide whether to vote for or against a conviction. I suppose he had in mind the idea that if he judged many cases he would hope that his 'errors of the first and second kinds' (convicting an innocent man or acquitting a guilty one) would not be too large. But I think it would be criminal to condemn an accused person without believing that one's verdict was very probably correct.[3]

Let us denote the odds of C given D by O$(C|D)$, so that

$$\mathrm{O}(C|D) = P(C|D)/P(\overline{C}|D)$$

because $P(\overline{C}|D) = 1\text{-}\ P(C|D)$. The *odds form of Bayes's theorem* is

$$\frac{O(G|E)}{O(G)} = \frac{P(E|G)}{P(E|\bar{G})} \tag{1}$$

where we have taken the background information *I* for granted to make the formula appear more friendly. The real meaning of (1) is

$$\frac{O(G|E\&I)}{O(G|I)} = \frac{P(E|G\&I)}{P(E|\bar{G}\&I)}\,. \tag{1'}$$

The proof of (1), that is, of (1?), is very simple, but let us express it in words. To do this, we need to remind the reader that Fisher (1922, p. 310) used the word *likelihood* in a technical sense. He called *P(C|D)*, the likelihood of *D* given *C*. (Note the interchange of *C* and *D*.) I have here simplified somewhat; for one thing, as an anti-Bayesian, he used this definition in practice only when *P(C|D)* had a definite value determined by a probability model.[4] To avoid changing Fisher's meaning of *likelihood,* as he used the term in practice, and as it is normally used by statisticians, I shall call *P(C|D)* the *Bayesian likelihood* of *D* given *C* when Fisher's constraint does not apply. This occurs when *P(C|D)* is an 'epistemic' probability, that is, either subjective (personal) or 'logical'. One could use the expression 'epistemic likelihood', but 'Bayesian likelihood' seems more homely. When it is not specified whether Fisher's constraint applies it is appropriate to put *Bayesian* in parentheses.

The left side of (1) is called the *Bayes factor* in favour of *G* (provided by evidence *E*) because it is equal to the factor by which the prior or initial odds of *G* are multipled to get the posterior of final odds. We can now state (1) verbally as:

The Bayes' factor in favour of guilt provided by evidence E is equal to the ratio of the (Bayesian) likelihood of guilt to that of innocence.[5] We leave it as an exercise for the reader to express (1´) verbally (or even orally because talking to oneself might be justified when one has something important to say).

To prove (1) we use the product axiom four times, thus:

$$P(E)P(G|E) = P(E\&G) = P(G)P(E|G) \tag{2}$$

$$P(E)P(\bar{G}|E) = P(E\&\bar{G}) = P(\bar{G})P(E|\bar{G}) \tag{3}$$

Divide (3) into (2) (omitting the middle expressions) and we obtain (1). It is equally easy to derive the result from Bayes' theorem. The more explicit form, (1´), of the theorem can of course be proved by conditioning on *I* throughout the argument.

Note that a Bayes' factor, or a weight of evidence, when computed by formula (1) or its logarithm, does not depend on the initial probability or odds of *G*. This is a useful property because there can be 'large' differences of opinion concerning the value of log O*(G)* (even when the odds are judged by a single subject). This is one reason why the evidence presented in court might not be enough for a jury to reach unanimity. (Other reasons are that half the jurymen are liable to be below average intelligence, and some are not good listeners.)

The odds form of Bayes' theorem is especially convenient when we wish to discriminate between two hypotheses, not necessarily the guilt or innocence of an accused person. It can, for example, be applied to problems of differential diagnosis in medicine. For one medical example among many see Bernstein *et al.* (1989) and, for the statistical details, Good *et al.* (1989).

As a simple exercise on Bayes' factors, consider the familiar puzzle of the three prisoners *A, B,* and *C.* For a recent discussion of this old chesnut see Zabell (1988) and Bar-Hillel (1989). The prisoners are told that just one of the three will be executed and are then placed in three separate cells. The information is assumed to be symmetrical so it is reasonable to assign a probability of 1/3 to each of the possibilities, at any rate this is to be assumed. The prisoner *A* says to the warden 'Tell me the name of one of the others who will *not* be executed. If neither of them will be, choose at random which of the other names, *B* and *C*, to tell me, by spinning a coin. This information will not affect my subjective probability of being executed'. The warden replies 'But if I do that, it will reduce the number of 'candidates' to 2, the relative probabilities of which will be, for you, the same as they were before, so your estimate of the probability that you will be executed will be 1/2 instead of 1/3'. Which was right, the prisoner or the warden?

The prisoner was right. For suppose the warden does what *A* requests and announces to *A* that *B* will not be executed. This is *A*'s new evidence. Its probability is $\frac{1}{2}$ assuming the hypothesis that *A* will be executed, and is also $\frac{1}{2}$ assuming that *A* will not be executed. Hence, from (1), the Bayes factor is 1, so *A*'s subjective probability of personal disaster from a 'cheap and chippy chopper on a big black block' is unaffected. This conclusion depends on the assumption of equal probabilities $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$.

It is obvious to a strict Bayesian, from (1) or (1?), that the weight of evidence in favour of *G* provided by *E* should be defined as some monotonic function of the Bayes' factor, because (from the Bayesian point of view) the point of collecting the evidence is to convert the prior odds into the posterior odds. Whether to *find* a man guilty depends on more than just the odds, especially on the 'utility' losses in making errors of the first and second kinds. For various reasons, one of which is given in the appendix, even a non-Bayesian should agree that the likelihood ratio ($\mathrm{P}(E|G\&I)/P(E|\bar{G}\&I)$ exhausts all the useful evidence when this *is* a likelihood ratio (in the non-Bayesian sense). But (1) or (1´) gives the likelihood ratio an intuitive meaning that could be appreciated by the woman in the street. The reader is cautioned that the expression *likelihood ratio* is often used by non-Bayesians to mean a ratio of *maximum* likelihoods. This is a further reason for distinguishing between *likelihood ratio* and *Bayes' factor.*

Knowing that the weight of evidence must be some function of the Bayes' factor, we must define $\mathrm{W}(G{:}E|I)$ as the logarithm of the Bayes factors if we wish to obtain an additive property that is required by Themis, and by ordinary English.[(6)] This additive property is

$$\mathrm{W}(G{:}E\&F|I) = \mathrm{W}(G{:}\ E|l) + \mathrm{W}(G{:}\ F|E\&I) \tag{4}$$

which is obtained at once by taking logarithms on both sides of the equation expressing the multiplicative property of Bayes' factors.

$$\mathrm{Fr}(G{:}E\&F|I) = \mathrm{Fr}(G{:}E|I)Fr(G{:}F|E\&I) \tag{5}$$

where Fr stands for *Bayes' factor.* (To avoid additional parentheses, the colon is taken to be a strong separator, but less strong than the vertical stroke.) We leave the trivial proof of (5) to the reader. It expresses, in a multiplicative manner, the fact that, if evidence is taken in two bites, *E* and *F*, then the posterior odds of *G* after the first bite becomes the prior odds just before the second bite is taken into account (and it could be called the *intermediate* odds). Equation (4) expresses the same fact, additively, for log-odds. When it is read from left to right it states that the weight of evidence concerning *G*, provided by the conjunction *E&F,* is equal to that provided by *E* plus that provided by *F* when *E* is *given* (or assumed).

When *E* and *F* are statistically independent given G & I and also given $\bar{\mathrm{G}}$ & I, and if we take I for granted and omit it from the notation, then (4) and (5) simplify to

$$\mathrm{W}(G{:}\ E\&F) = \mathrm{W}(G{:}\ E) + \mathrm{W}(G{:}\ F) \tag{4'}$$

and

$$\mathrm{Fr}(G{:}\ E\&F) = \mathrm{Fr}(G{:}\ E)\mathrm{Fr}(G{:}\ F). \tag{5'}$$

For example, if all we know of a coin is that it has been properly tossed ten times and gave heads each time, then we get a Bayes' factor of $2^{10} = 1024$ in favour of the coin's being double-headed. Each toss, yielding a head, provides the same weight of evidence, log 2, but a single tail gives log 0 = - ∞ and disposes of that hypothesis.

Turing, in 1940 or 1941, suggested that, when the base of the logarithms in 10, the unit should be called a *ban,* and that one tenth of it should be called a *deciban* (in other words the base of the logarithms is $10^{0.1}$), abbreviated as db. The analogy with the acoustic decibel (defined in the same manner for the ratio of two powers) is fairly strong because a deciban is about the smallest unit of weight of evidence that can be apprehended intuitively.[7] The word *ban* originated as the first syllable of *Banbury,* a town in Oxfordshire where sheets were printed for use in the cryptanalytic process, 'Banburismus', for which Turing invented the deciban. Turing's name for weight of evidence was the somewhat clumsy 'decibannage' or simply 'score'. Also he didn't propose a symbolic notation such as W*(G: E|I).* I believe it would be possible for judges, detectives, and doctors to develop their judgement so as to be able to say that a weight of evidence is such-and-such a number of decibans, and that it would be useful to do so. Some might prefer to put the emphasis on the Bayes' factors, and some might prefer to use both of these closely allied forms of judgement. Bayesians are supposed to make judgements of probabilities, probability inequalities, utilities, etc., but there is no reason why they should not also make other kinds of judgements such as those of weights of evidence. The aim is still to build up a body of beliefs that is reasonably consistent with the axioms.

Of course Bayes' factors and weights of evidence are directly applicable for comparing hypotheses (such as guilt and innocence) two at a time, but even when there are more than two hypotheses it might still be convenient to consider them in pairs.[8]. Otherwise it is necessary to use Bayes' therorem in its more general form.

Because of the additive property one can think of weight of evidence as a quasi-utility. It is not a strict utility because, for example, by the time the odds reach a thousand million one might as well announce a verdict. In this respect, weight of evidence resembles money which is also only a quasi-utility.

By thinking of weight of evidence as a quasi-utility, it is natural to consider its expected value.[9] We shall give a brief history of that subject below, but let us now consider an example concerned with the use of polygraph as a 'lie detector'. The topic was considered by Toulmin (1984) who pointed out that the expected weight of evidence provided by a lie detector is the negligible quantity of 1½ decibans.

Toulmin based his calculation on statistics obtained by Kleinmuntz & Szucko (1984). They had a sample of fifty confessed thieves and fifty innocent people and the 100 polygraphs were evaluated independently by six professional interpreters. Each interpretation was a yes-no judgement of whether the person was guilty. (The instrument might well be of more value if these judgements were replaced by probability estimates, but this is not yet the standard usage.) The sample is not large but, as an approximation, we take it as its 'face value', after lumping together the verdicts) of the six interpreters. We obtain

'Valid positive', $P(E|G) = 0.76$
'False negative', $P(\bar{E}|G) = 1\text{-}P(E|G) = 0.24$
'False positive', $P(E|\bar{G}) = 0.37$
'Valid negative', $P(\bar{E}|\bar{G}) = 1\text{-}P(E|\bar{G}) = 0.63$,

where E means a verdict of lying or guilt. Thus

$$\mathrm{Fr}(G\colon E) = 76/37 = 2.1, \text{ and } \mathrm{Fr}(G\colon \bar{E}) = 24/63 = 1/2.6\ .$$

In brief, the judgements of the interpreters, though presented as 'failed or passed the test', are really worth Bayes' factors for or against lying of only about 2 or 3. (In this experiment the initial odds were evens.) The weights of evidence are $W(G\colon E) = \log(76/377) = 3.1$db (or 31 'centibans'), $W(G\colon\bar{E}) = \log(24/63) = -4.2$db; and of course $W(\bar{G}\colon E) = -3.1$ db, and $W(\bar{G}\colon E) = 4.2$db. Finally, the expected weight of evidence[10] in favour of guilt when the accused is guilty, is

$$P(E|G)W(G\colon E) + P(\bar{E}|G)W(G\colon \bar{E})$$

$$= 0.76 \times 3.1 - 0.24 \times 4.2 = 1.4\text{db}$$

while that in favour of innocence when the accused is innocent is

$$P(E|\bar{G})W(\bar{G}\colon E) + P(\bar{E}|\bar{G})W(\bar{G}\colon \bar{E}) = 1.5 \text{ db}.$$

(The approximate equality of these two amounts is not a coincidence and can be explained mathematically.) It is hardly worth while to carry out a test when the expected weights of evidence are so small.

The use of a lie detector is just one example of extracting a chunk of evidence to be discussed approximately numerically. Another example is given by Lindley (1977) who considers the situation where material, such as broken glass, is found at the scene of a crime and similar material is found on a suspect. Lindley assumes that the distribution of the refractive indexes of the glass from the two sources, the scene of the crime and elsewhere, are known and that the likelihood ratio can therefore, be computed. The likelihood ratio can then be given an intuitive meaning, a meaning that could be made intelligible to the woman in the street, by using the odds form of Bayes' theorem. In such situations a statistician might be used as an 'expert witness' in the court proceedings. It is unfortunate that there is some opposition to the Bayesian philosophy because this weakens the statistical evidence in the eyes of judge and jury. It is to be hoped that this difficulty will eventually be overcome by elementary education.

Other potential applications depend on blood groups in paternity cases (for example, see Good 1950, p. 74), partial fingerprints, and 'lie detectors' although in this last example the 'chunk of evidence' is small, as we have seen.

The technique of extracting a chunk of evidence is an especially convenient method when this chunk is effectively statistically independent of the rest of the evidence, in which case the weight of evidence from the chunk can simply be added to or subtracted from the log-odds, but the technique is surely useful even if the condition of independence is not satisfied. For example, it is convenient to begin an investigation by making a *prima facie* case. Then again, considerations of motivation are especially relevant when judging an initial or immediate probability of guilt.

In more general circumstances, the concept of *expected weight of evidence* has many applications in statistics, and has several different names, one of which is *cross-entropy.* In effect, it is a generalization of the concept of entropy (or more precisely of negentropy) in the sense in which that term is used in Shannon's theory of communication (Shannon & Weaver 1949). To discuss these applications would take us too far afield.

The reason for wanting to break evidence up into chunks is to try to alleviate the difficulty of coping with complexity. Also, if E is defined as *all* the evidence presented in court it will usually be found that $P(E|G)$ and $P(E|\overline{G})$ are both exceedingly small; they might well be as small as 10^{-100} for example. To apprehend such small probabilities is more difficult than obtaining an intuitive feeling for very small and very large numbers in physics. In physics and astronomy it is often done by breaking the numbers into meaningful factors to be multiplied together. But in our context it will usually be a better strategy to make direct judgements of ratios or their logarithms, in other words judgements of Bayes' factors and weights of evidence. This can be done for chunks of the evidence or for the evidence as a whole, or both as a check on one's consistency of judgements. In other perceptual activities, ratios are often easier to judge than absolute values[(11)], as when we recognize an object by its shape, or when we appreciate music implicitly by the ratios of the frequencies of the notes rather than by apprehending absolute pitch. One would expect that the difficulty of handling of these ratios by the nervous system would be alleviated if the neurons' responses to stimuli were logarithmic so that the 'computations' could be done by subtraction.[(12)] This would help to explain why there is some truth in the Weber—Fechner logarithmic law of perception, although its absolute validity has been questioned by Stevens (1951 p. 36).

One aid to the judge is the use of judgements of approximate irrelevance so as not to clutter the judge's mind and obscure the salient points. This presumably is why some kinds of information, such as the outcomes of 'lie detector' tests, are often thrown out of court.

The legal system seems to require that magistrates and jurymen are capable of judging the posterior probability $P(G|E)$, at least implicitly. Perhaps, in the ordinary affairs of life (and in court cases), we are better at judging posterior probabilities than prior ones. The posterior probabilities are more important for making decisions, that is, as a basis for immediate action, so we get more practice at judging them.[13] Yet we undoubtedly also allow implicitly for prior probabilities; for example, it is easier to recognize a shop assistant in his shop than if we happen to meet him in the street or in a foreign country (Good 1950, p. 68). We are all informal Bayesians to some extent, even those who think they are non-Bayesians.

I believe our ability to make such judgements could be improved by training. This topic could be the basis for a sizable research project.

We all make implicit judgements of probabilities many times a day. Most of these judgements cannot yet be reduced to explicit calculation; if they could be, we wouldn't call them *judgements* (Good 1959, p. 16). They are presumably made by neural networks whose workings are poorly understood. If we understood them, the problems of Artificial Intelligence would be solved. (They might also be 'solved', in a practical sense, by using artificial neural networks that we do *not* understand: Good 1965.) The expression of Bayes' factors and weights of evidence in terms of probability was a step in this direction.

Probabilistic and causal networks were used in my paper on causality (Good 1961/62) and in recent years have been much used in the work on expert systems. For a recent book on the topic, with many references see Pearl (1988). See also Lauritzen & Spiegelhalter (1988).

The expected weight of evidence in favour of a true theory, from any experiment, is always non-negative, a result that I think was first noticed by Turing (1941), though the corresponding *mathematical* fact, which is simple, was known to Gibbs (1902, Chap. XI) and even earlier references are mentioned by Good (1983/85, p. 256). (For the corresponding property of utilities, see Savage 1954, p. 125, Good 1967, 1974.) Turing also discovered a somewhat deeper and surprising theorem which was independently noticed in relation to radar (Peterson, Birdsall, & Fox, 1954). The theorem, in our terminology, is that if weight of evidence has a Normal (bell-shaped or Gaussian) distribution (when G is true) then its variance is double its expectation if the units are natural bans, that is, when natural logarithms are used. (A generalization, dealing with approximate normality, is given by Good (1961) where other references are mentioned.) When Turing's theorem is re-expressed in terms of decibans we find that the standard deviation is about 3 × the square root of the expectation. (The factor 3 arises as an approximation to $\sqrt{(20 \log_{10} e)} = 2.947$.) This shows that it is not unusual for an experiment to be misleading, at least when the weight of evidence has a Normal distribution. For example, if the *median* Bayes' factor in favour of the truth is say 40, then the weight of evidence has a mean of 16 db and a standard deviation of about 12 db. On a given occasion it might well be as high as 40db or as low as - 8db, corresponding to Bayes factors of 10000 or 1/6, a remarkably wide spread. This fact, horrifying in relation to radar, is of philosophical interest in the law because it suggests how easily a miscarriage of justice can occur without any mistakes in reasoning.

If the ‘lie detector’ were interpreted probabilistically, it would be interesting to find out whether the standard deviation of the weight of evidence was approximately $3\sqrt{1.5} = 3.7$ db. If so, then it would sometimes provide a useful amount of evidence.

The idea of thinking of a weight of evidence as a quasi-utility at once suggests that the maximization of its expectation might be a guiding principle in experimental design,[14] and in principle one could imagine a detective using this principle when deciding how to proceed with an investigation. It is less difficult for a physician to use this principle for diagnosis. (See, for example, Good & Card 1971, and Card & Good 1970/1988, where many references are given.)

For a more technical review of the topic of weight of evidence, up to 1983, see Good (1983/85).

We now briefly discuss a relationship between weights of evidence and probabilistic causality, another topic of great interest in the philosophy of the law. Hart & Honoré (1959) discuss causation in the law, with many engrossing examples. They make many references to probability, but they do not attempt to define a measure of probabilistic causality. In Good (1961/62) I used the desideratum-explicatum approach, as applied to causal networks, to analyse the concept of the degree to which one event *F tends* to cause a later one *E*. In this analysis the probabilities are supposed to be physical, but the judgement of them is not. The explicatum attained was $\mathrm{W}(\bar{F}: \bar{E}|U)$, the weight of evidence against F provided by the non-occurrence of *E*, given the state *U* of the universe just before *F* occurred. To avoid the long mathematical argument, a simpler one can be given (Good 1988, p. 394): it seems reasonable to assume that the required tendency canbe expressed by one of the four weights of evidence $\mathrm{W}(E: F|U)$, $\mathrm{W}(F: E|U)$, $\mathrm{W}(\bar{E}:\bar{F}|U)$, and $\mathrm{W}(\bar{F}: \bar{E}|U)$. The first three of these can be ruled out by simple arguments and we are left with the one that was produced by a different argument in Good (1961/62).

It is essential to distinguish between the *tendency of F to cause E*, and the degree to which *F did* cause E, both concepts being highly relevant in legal cases. The distinction, which is sometimes overlooked, is exemplified by that between attempted murder, and actual murder. Another dramatic example of the distinction is given by Good (1961/62) involving a little story about Holmes, Watson, and Moriarty, but I shall not repeat it here. Yet another distinction must be made between a person’s *tendency* to cause some event, and his *estimate* of that tendency. Only the latter, which involves subjective probabilities, is relevant to a measure of his *ethical* as distinct from his *legal* responsibility.

It was shown by Salmon (1985/88) that my original explicatum of the degree to which *F actually caused E* was faulty. I attempted to rectify it in Good (1985/88b), but my analysis covers only a limited class of causal networks.[15] More work is needed on this difficult topic. But I believe the explication of *tendency* to cause is well justified, and it could be used as one more kind of judgement available to the Bayesian.

Although Hart & Honoré (1959) discuss probabilistic matters in twenty places in their book they say on page 432 ‘With the recognition of this distinction [between partial and complete knowledge] must disappear the whole claim to elucidate causal judgements with the help of the notion of increased probability, or indeed of probability at all’. I would have agreed with this remark if, instead of ‘causal judgements’, they had said ‘strict causality’, for this would convert the statement into a truism. But most physicists no longer believe in strict causality (determinism). Moreover, even if the World is deterministic, it behaves as if it were indeterministic relative to a given depth of analysis (Good 1985/88a, p. 24). This remark, obvious enough to common sense, is supported in a new way by the recent work in chaotics (which is *not* obvious).

One can usefully provide a formula for ‘explicativity’ in terms of weight of evidence (Good 1977). Roughly speaking, explicativity is a measure of explanatory power. It has applications in statistics and might be useful in legal philosophy.

Bayesian thinking, and specifically the concept of weight of evidence, thus sheds much light on the philosophy and hence on the practice of science and of the law. Perhaps one day some of these concepts wil be taught in elementary education, and decibans, or at least Bayes’ factors, might become part of the jargon of judges, jurymen, detectives, and doctors. Themis, Shen Nung, and I would be pleased.

FURTHER HISTORICAL COMMENTS

The following historical comments are based primarily on Good (1975, 1987/89), and some of my discussion is here quoted verbatim from the latter paper, with the permission of the author and editor.

As already mentioned, the ancient Greeks must have had an informal concept of additive weights of evidence. The earliest quotation involving ‘weight of evidence’ in the *Oxford English Dictionary* is a remark by T.H.Huxley in 1878 (‘The weight of evidence appears strongly in favour of the claims of Cavendish.’), but the *concept* was used in English literature long before that. For example, Alexander Pope, in a religious poem entitled *Messiah,* published in 1711, says

> All crimes shall cease, and ancient fraud shall fail,
> Returning Justice lift aloft her scale,
> Peace o’er the world her olive wand extend
> And white-robed Innocence from heaven descend.

A concept so well entrenched in human thought is worth expressing as a formula, and used in a ‘calculus’, if possible. David Hume, in a well known attack on a belief in miracles, in 1748, quoted in Good (1987/89), also shows clearly that Hume was conscious of positive and negative weights of evidence.

Laplace (1820, pp. 446–461), when considering the probability of testimony, almost explicitly anticipated the concept of a Bayes factor (compare Poisson 1837, p. 319, De Morgan 1837 and 1847, Section 181). Peirce (1878) came very close to the best definition of weight of evidence, only ‘very close’ because a careful analysis of Peirce’s obscurely written paper shows that he was assuming that the prior probability of the hypothesis [*G* here] was 1/2 (Good 1983a). Peirce’s isolated remark about weight of evidence was somewhat of a throw-away line because he was against Bayesianism which he called the ‘conceptualist’ interpretation of probability. The concept of a Bayes’ factor is explicit in Wrinch & Jeffreys (1921, p. 387) (except that they did not use the expression ‘Bayes’ factor’ and ‘odds’), and almost explicit in Fisher (1922, p. 326) where he denigrated a form of the Bayesian approach. Jeffreys knew Fisher well and might have discussed the formula with him.

The optimal property of the Bayes’ factor, when discriminating between two simple statistical hypotheses, was proved, in terms of the concept of the power of a statistical test, by Neyman & Pearson (1933, Section III). They made Bayesian remarks but did not cite Wrinch & Jeffreys (1921). In other circumstances, that is, when either *G* or $\bar{G}$ is not a simple statistical hypothesis, the Bayes factor cannot be computed by a non-Bayesian while wearing a non-Bayesian hat.

Jeffreys (1936) called weight of evidence ‘support’, which is a good alternative term, but it doesn’t occur in any edition of his book on probability (1939/1961). This was because he aimed at ‘objective Bayesianism’, at least in the first edition of his book, and so usually took the initial probability of a hypothesis as 1/2 just as Peirce had done sixty years earlier. In this case the Bayes’ factor is equal to the posterior odds.

Turing, in an important cryptanalytic application in 1941, introduced the concept of the deciban, combined with sequential analysis. In the same year I mentioned Turing’s idea (not the application) to G.A.Barnard, when I happened to meet him in London. He told me he was doing something similar at the Ministry of Supply, for quality control. (*Quality* is here of course a noun, not an adjective.) Barnard does not remember the meeting.

As Turing’s main statistical assistant, I recognized the importance of these concepts and, somewhat obsessively, wrote about them in numerous publications, beginning with Good (1950). For references to 32 of these publications see Good (1983b, p. 159). Kemeny & Oppenheim (1952), in a paper I had overlooked, introduced the expression *factual support for a hypothesis.* This is a special case of weight of evidence. They laid down thirteen desiderata, one being the desire for a simple formula, and from these arrived at the explicatum (in our notation).

$$\frac{P(E|G) - P(E|\bar{G})}{P(E|G) + P(E|\bar{G})}\,.$$

This is equal to tanh {W*(G: E)/*2}. (I wrote sinh by mistake for tanh in Good, 1983b, p. 160 and in 1975, p. 253.) They didn’t obtain an additive explicatum because they wanted their explicatum to range from - 1 to 1, whereas W ranges from - ∞ to ∞. It is entertaining that the arctanh transformation is the same as that in the Special Theory of Relativity for transforming velocities into ‘rapidities’, a term attributed to A.A.Robb by Eddington (1930, p. 22). No material velocity can exceed that of light, whereas the rapidities are unrestricted and are additive, so the analogy with the relationship between factual support and weight of evidence isn't bad (Good 1989c).

Popper (1954, 1959) used the desideratum-explicatum approach for 'corroboration' and obtained at least two formulae that would satisfy the desiderata. He says (1959, p. 394) 'I regard the doctrine that the *degree of corroboration or acceptability cannot be a probability* as one of the most interesting findings of the philosophy of knowledge.' I agree with this sentiment, but with the words 'or acceptability' deleted and with 'corroboration' taken as a synonym for weight of evidence. With these glosses, the finding had been taken for granted by some earlier writers, but not by all. Carnap (1950), for example, had confused himself and many other philosophers by defining 'configurations' as a probability, and the confusion continues among many philosophers even after forty years have elapsed. Good terminology is important in philosophy. I appeal to philosophers to *stop using 'confirmation' to mean a probability.* Philosophers should aim at clarification, not obfuscation.

In Good (1960) I modified Popper's desiderata and obtained all possible explicata that would satisfy them. The most useful of these was weight of evidence.

The approach of Kemeny & Oppenheim is so different from the one used in our Appendix that it would require a separate publication to explain the differences (and similarities). Even their explanans is a little different from ours. The most essential difference in the approaches is that in our Appendix we do not refer to the probabilities of hypotheses, and we assume that $W(G: E)$ is a function of $P(E|G)$ and $P(E|\overline{G})$ alone. (Their explicatum satisfies this condition but it was not one of their desiderata.)

The expression 'weight of evidence' was again independently used in the same sense by Minsky & Selfridge (1961). J.M.Keynes (1921, p. 71ff) used the expression 'evidential weight of an argument' in the uninteresting sense of the total mass of the evidence whether for or against a hypothesis. He added, appropriately enough (p. 76), 'I do not feel sure that the theory of "evidential weight" [in this feeble sense] has much practical significance'. And on his page 71 he says, in the spirit of Themis, Pope, or Hume, The magnitude of the probability of an argument… depends upon a balance between…the favourable and the unfavourable evidence'.

Good (1968), and more lucidly in (1984), using a simpler approach than in 1960, showed that some fairly compelling desiderata for $W(H: E|G)$, the weight of evidence in favour of *H provided by E* given *G* all along, leads necessarily to the explicatum as the logarithm of a Bayes' factor. Simpler arguments were given by Good (1989a, b) and the method of the latter paper is used in our Appendix. Card & Good (1974) and Good & Card (1971) used the concept in relation to medical diagnosis, and Spiegelhalter & Knill-Jones (1984) say 'We argue that the flexible use of "weights of evidence" overcomes many of the previous criticisms of statistical systems while retaining a valid probabilistic output'. Spiegelhalter (1983) mentioned that physicians find the concept appealing. Magistrates should do so *a fortiori* because they so often have just two hypotheses to consider. It seems that jurymen, magistrates, and physicians make implicit judgements of the probabilities of guilt and disease states. These must again involve implicit judgements of Bayes' factors or weights of evidence and must also take initial probabilities into account. The neural circuitry of our brains is much more complex than the formulae in the present paper!

More and more uses have been made for the concept of *expected* weight of evidence, thus treating weight of evidence as a quasi-utility. Expected weight of evidence has a variety of names, one of which is cross-entropy. In the discrete case it looks like $p_1 \log(p_1/q_1) + p_2 \log(p_2/q_2) + \ldots$as mentioned in Note 10. The part depending on the p's has its sign opposite to that of 'entropy', $- p_1 \log p_1 - p_2 \log p_2 \ldots$, so the name 'cross-entropy' is better than 'relative entropy' unless the sign is changed. As a mathematical expression it was apparently first used in statistical mechanics by Willard Gibbs around 1900 and was used for cryptanalysis in World War II by Turing and myself. Turing was probably unaware of its previous usage. It was used for a different purpose by Jeffreys (1946). The concepts of maximum entropy and minimum cross-entropy, in ordinary statistical inference, were introduced and emphasized by E.T.Jaynes and by the cryptanalyst Kullback in the fifties. A more precise title for the book by Kullback (1959) would have been *Statistics and Expected Weight of Evidence.* Jaynes used maximum entropy for constructing priors—an aspect of 'objective Bayesianism'. Good (1968/69) asserted that these procedures, as well as Jeffrey's invariant prior (Jeffreys 1946), could also be regarded as minimax procedures, when weight of evidence is regarded as a quasi-utility, and that this was one way to explain why they have invariant properties (besides suggesting the possibility of generalizations). Again, Good (1963) showed that the principle of maximum entropy, reinterpreted as a method for generating hypotheses, and applied to multidimensional contingency tables, leads to the same null hypotheses that had been suggested in the theory of loglinear models for intuitive reasons. Thus the ancient concept of weight of evidence, captured in a simple formula, has had a somewhat unifying influence on statistical theory.

Acknowledgements
This work was supported in part by a grant from the U.S. National Institutes of Health. My thanks are also due to Leslie Pendleton Miller for typing a manuscript in which the handwriting often needed cryptanalytic ability for decipherment.

APPENDIX. PROOF OF THE MAIN THEOREM (Good, 1989b)

We have assumed that $W(G{:}\ E)$ is a function, say, f, of $x = P(E|G)$ and $y = P(E|\overline{G})$ alone, that is, $W(G{:}\ E) = f(x,y)$. (We can of course condition throughout on the background information I, but I am omitting I from the notation merely to make it easier on the eye.) Now suppose that some event or proposition F is entirely irrelevant to G and E. Write $P(F) = \lambda$. Then $P(E\&F|G) = \lambda x$, $P(E\&F|\overline{G}) = \lambda y$, and hence $W(G{:}\ E\&F) = f(\lambda x, \lambda y)$. But it is also obvious that we must have $W(G{:}\ E\&F) = W(G{:}\ E)$ because F is assumed to be irrelevant (and a judge would object to its being presented as evidence concerning G). (I have previously overlooked that this is essentially the twelfth desideratum in Kemeny & Oppenheim, 1952, apart from their restriction to 'factual support'.) Thus $f(\lambda x, \lambda y) = f(x,y)$. This has to be true for all ? in the closed interval [0,1]. (There is enough freedom in the choice of F to ensure this). It follows that f is a function *of* x/y alone. Thus $W(G{:}\ E)$ must be a function of $P(E|G)/P(E|\overline{G})$. Since additivity is desirable to justify the word *weight,* clearly the best definition of

W is the logarithm of the Bayes’ factor, the base of logarithms being greater than 1 to make this function increasing rather than decreasing. The base merely determines the unit in terms of which W is measured so it is appropriate to say that the explicatum for W is unique.

NOTES

(1) A proposition is sometimes defined as the meaning of a statement. The question arises whether uncertain testimony, together with its probability, can be interpreted as a proposition. A similar question arises if you observe an object in poor light so that you are uncertain what you have seen (Jeffrey 1965). I consider that this experience can correspond to a proposition, such as the meaning of the sentence ‘An event occurred which made your conditional subjective probability of *C* equal to *x*.’ If such propositions are allowed, one can attack the problem of evaluating the weight of evidence provided by uncertain testimony (Good 1981) without adding new axioms to the foundations of probability. Jeffrey does not permit the kind of proposition just mentioned and does add a new axiom.

For the early history of the evaluation of uncertain testimony see Zabell (1988).

(2) For a discussion of kinds of probability see, for example, Good (1959).

(3) For some algebraic relationships between weights of evidence and errors of the first and second kinds see Good (1980).

(4) To support this comment, note that Fisher (1956, p. 68) says that ‘the likelihood of *A* or *B’* [given *E*] is undefined, where *A* and *B* are mutually exclusive. But a strict Bayesian would equate it to

$$\mathrm{P}(E|A \text{ or } B) = \frac{\mathrm{P}(A)\mathrm{P}(E|A) + \mathrm{P}(B)\mathrm{P}(E|B)}{\mathrm{P}(A) + \mathrm{P}(B)}$$

and would have to make a judgement of the value of P*(A)*/P*(B).* Thus the Bayesian likelihood of *‘A* or *B’* is a weighted average of the likelihoods of *A* and of *B*.

The concept of likelihood has a long history, but Fisher developed it in some detail and recognized its importance enough to give it a name so it is usually associated with him.

To avoid a break in continuity in the exposition, I omitted to say that, if *E* is known and more than one hypothesis is under consideration, all the likelihoods can be multiplied by, for example, 19.387, that is, the likelihoods are defined only up to proportionality.

Bayes’ theorem can be expressed in the words ‘posterior probability is proportional to prior probability times (Bayesian) likelihood’. Hence, *if a statistical model is not questioned,* it follows that the likelihoods contain all the information contained in an observation or experimental result, provided that the prior probabilities are unaffected by the experiment. This is an interpretation of the expression *likelihood principle.* The estimation of a parameter, or the choice among hypotheses, by the method of *maximum* likelihood, is *not* what is usually meant by the likelihood principle. The method of maximum likelihood is useful but not sensible in all circumstances because it ignores the prior probabilities.

(5) Note that in a likelihood ratio, the probability factor mentioned in Note 4 cancels.

(6) Tribus (1969) used ‘evidence’ or ev to mean log-odds.

(7) In my book Good (1950) I used the name ‘decibel’ in place of ‘deciban’, in case the use of *deciban* would be regarded, so soon after the war, as a breach of security. But this possible confusion with acoustics has made me revert to the original name. (Tribus, 1969, used ‘decibel’.) Note that the magnitude of star brightness is also based on a logarithmic measure.

It is somewhat convenient that the Bayes’ factors 2, 4, 5, 8, and 10 correspond closely to integer values of the weights of evidence, namely 3, 6, 7, 9, and 10 db, the last of which is of course exact.

(8) When restricting one’s attention to a pair of hypotheses, H_1 and H_2, an alternative notation to O$(H_1|H_1$ or $H_2)$ is O(H_1/H_2), the odds of H_1 *as contrasted with* H_2. Similarly, instead of W[H_1: E I&(H_1 or H_2)], one can write W$(H_1/H_2\text{: }E|I)$, read the weight of evidence in favour of H_1, *as contrasted with H2,* provided by *E,* given *I* all along. A similar notation can of course be applied to Bayes’ factors. This convention distinguishes the vertical and oblique strokes.

(9) ‘Expected value’ or ‘mathematical expectation’ is part of the jargon of statistics. Roughly speaking, it means the long-run average value (of a random number). It is calculated by an expression of the form $p_1x_1 + \mathrm{p}_2\mathrm{x}_2 + \ldots$, where x_1, x_2, $x_{3,}\ldots$are the possible values that the random number can take while $p_{1,}$ $p_{2,\ldots}$are the corresponding probabilities (see section 1.17). In many problems it is impossible for a random number to be equal to its expectation; for example, if a die is cast, the expected number of spots is $3\frac{1}{2}$. Thus ‘expected value’ is not a satisfactory expression but it is too firmly entrenched to be dislodged except perhaps by ‘mean value’.

(10) Symbolically this is $\xi\{\mathrm{W}(G\text{: }E|B)|G\}$. The corresponding mathematical expression (in the ‘discrete case’) is of the form $p_1\, log(p_1/q_1) + p_2 \log(p_2/q_2) + \ldots$, (where p_1,p_2,…$_)$ and (q_1,q_2,...) are two sets of multinomial probabilities.

(11) The dependence of perception on the judging of ratios was part of the reason for the name *Ratio Club,* a small exclusive club in 1949–55 at the National Hospital of Nervous Diseases in London to discuss neurophysiological and cybernetic topics, while eating dinner in armchairs, drinking beer, and ratiocinating. The membership was W.R.Ashby, H.Barlow, J.A.V.Bates (founder), G.D.Dawson, T.Gold, I. J.Good, W.E.Hick, V.I.Little, D.M.Mackay, T.MacLardy, P.Merton, J. Pringle, W.H.A.Rushton, H.Shipton, J.Sholl, E.T.O.Slater, A.M.Turing, A. M.Uttley, W.G.Walter, J.Westcott, P.M.Woodward. One of the rules of the club was that no full professor could be invited to membership! The Latin word *ratio* has many meanings including 'theory' and 'knowledge'. A philosophical periodical called *Ratio* started publication in 1957, not long after the Ratio Club became defunct.

(12) For this speculation see Uttley (1959, p. 144) and Good (1961, p. 128). Uttley refers also to an unpublished paper by G.Russell. In our context we need additions as well as subtractions (but addition can be done by repeated subtraction).

(13) If we are better at estimating posterior probabilities than prior probabilities, the question arises whether there is some way of deducing the prior probabilities given a selection of posterior probabilities. A simple example was suggested by Good (1950), as an example of 'the Device of Imaginary Results', and some less simple ideas were proposed by Good (1986).

(14) This principle of experimental design is usually expressed in terms of information theory (Cronbach 1953, Lindley 1956, Tribus 1969, Good 1955/56). But Shannon's rate of transmission of information, as used in his work, can also be expressed, for a discrete channel, as the expected weight of evidence, per observation of an input-output pair, for determining whether the input and output are statistically dependent (when they are). It might not be necessary to define 'amount of information' in Shannon's sense. Compare Good & Toulmin (1968), who, among other things, express the proof of one of Shannon's coding theorems in terms of expected weight of evidence instead of expected amount of information. I believe this increases the intuitive appeal of the proof because weight of evidence has an immediate interpretation in terms of inference.

Turing's interest in the expectation and variance of a score (weight of evidence) arose for deciding whether a specific experiment was worth while, so the concept of choosing among experiments by maximizing the expected weight of evidence was not of immediate concern to him, and as far as I know he never mentioned it.

Cronbach, in his long technical report, considers Shannon's 'rate of transmission of information' as a criterion for experimental design, but eventually decided that it does not have 'the characteristics desirable for the evaluation of [psychometric] tests'. He prefers a more Bayesian approach based on expected utility.

(15) An arithmetic error in Good (1985/88b) has been pointed out by William L. Harper, but it does not affect the argument. Details can be supplied on request.

REFERENCES

Bar-Hillel, Maya (1989). How to solve probability teasers. *Philosophy of Science* **56**, 348–358.

Bernstein, L.H., Good, I.J., Holtzman, G.I., Deaton, M.L., & Babb, J. (1989). ‘The diagnosis of acute myocardial infarction from two measurements of creatine kinase isoenzyme MB with use of nonparametric probability estimation’, *Clinical Chemistry* **35**, 444–447.

Card, W.I. & Good, I.J. (1970 / 1988). ‘A mathematical theory of the diagnostic process’, Technical Report No. 88–20, Department of Statistics, Virginia Tech, 148 pp.

Card, W.I. & Good, I.J. (1974). ‘A logical analysis of medicine’, in *A companion to medical studies* (ed. R.Passmore and J.S.Robson; Oxford, Blackwell’s), Vol. 3, Chap. 60.

Carnap, R. (1950). *Logical foundations of probability.* University of Chicago Press.

Cronbach, L.J. (1953). ‘A consideration of information theory and utility theory as tools for psychometric problems.’ Technical Report No. 1, Contract N6ori-07146, with the Office of Naval Research. College of Education, University of Illinois, Urbana, Illinois, mimeographed, pp. 65.

DeGroot, M.H. (1988). A Bayesian view of assessing uncertainty and comparing expert opinion. *Journal of Statistical Planning and Inference* **20**, 295–306.

Eddington, A.S. (1930). *The mathematical theory of relativity.* Cambridge University Press.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **A222**, 309–368.

Fisher, R.A. (1956). *Statistical methods and scientific inference.* Edinburgh: Oliver & Boyd.

Gibbs, J.W. (1902). *Elementary principles in statistical mechanics,* New York; Scribner. Reprinted by Dover Publications.

Good, I.J. (1950). *Probability and the weighing of evidence.* London, Charles Griffin; New York, Hafners.

Good, I.J. (1955 / 56). Some terminology and notation in information theory. *Proceedings of the Institution of Electrical Engineers C***103** (1956), 200–204. Also Monograph 155R (1955).

Good, I.J. (1959). Kinds of probability. *Science* **129**, 443–447. (Italian translation by Fulvia de Finetti in *L’Industria,* 1959.) (Reprinted in *Readings in applied statistics,* William S.Peters, ed. [New York, Prentice-Hall, 1969], 28–37; and in Good, 1983b).

Good, I.J. (1960). Weight of evidence, corroboration, explanatory power, information, and the utility of experiments. *Journal of the Royal Statistical Society* ***B*** **22**, 319–331; **30**(1968), 203.

Good, I.J. (1961). Weight of evidence, causality, and false-alarm probabilities, *Information Theory, Fourth London Symposium* (Butterworth, London), 125–136.

Good, I.J. (1961 / 62). A causal calculus. *British Journal for the Philosophy of Science* **11** (1961), 305–318; **12** (1961), 43–51; **13** (1962), 88. Largely reprinted in Good (1983b).

Good, I.J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contigency tables. *Annals of Mathematical Statistics* **34**, 911–934.

Good, I.J. (1965). Speculations concerning the first ultra-intelligent machine. *Advances in Computers* **6**, 31–88.

Good, I.J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science* **17**, 319–321. Reprinted in Good (1983b).

Good, I.J. (1968). Corroboration, explanation, evolving probability, simplicity, and a sharpened razor. *British Journal for the Philosphy of Science* **19**, 123–143.

Good, I.J. (1968 / 69). What is the use of a distribution? in *Multivariate Analysis II* (ed. P.R. Krishnaiah; New York: Academic Press), 183–203.

Good, I.J. (1974). A little learning can be dangerous. *British Journal for the Philosphy of Science* **25**, 340–342. Reprinted in Good (1983b).

Good, I.J. (1975). Explicativity, corroboration, and the relative odds of hypotheses. *Synthese* **30**, 39–73. Reprinted in Good (1983b).

Good, I.J. (1977). Explicativity: a mathematical theory of explanation with statistical applications. *Proceedings of the Royal Society (London) A* 354, 303–330. Reprinted in *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (A.Zellner, ed.; Amsterdam: North Holland) and largely in Good (1983b).

Good, I.J. (1980). Another relationship between weight of evidence and errors of the first and second kinds. C67 in *Journal of Statistical Computation and Simulation* **10**, 315–316.

Good, I.J. (1981). The weight of evidence provided by uncertain testimony of from an uncertain event. C96 in *Journal of Statistical Computation and Simulation* **13**, 56–60.

Good, I.J. (1982a). The axioms of probability. *Encyclopaedia of Staistical Sciences,* Volume 1 (S.Kotz and N.L.Johnson, eds.; New York; Wiley), 169–176.

Good, I.J. (1982b). Degrees of belief. *Encyclopaedia of Statistical Sciences,* Volume 2 (S.Kotz & N.L.Johnson, eds; New York: Wiley), 287–293.

Good, I.J. (1983a). A correction concerning my interpretation of Peirce, and the Bayesian interpretation of Neyman-Pearson ‘hypothesis determination’. C165 in *Journal of Statistical Computation and Simulation* **18**, 71–74.

Good, I.J. (1983b). *Good thinking: The foundations of probability and its applications.* University of Minnesota Press.

Good, I.J. (1983 / 85). Weight of evidence: a brief survey. In *Bayesian Statistics* **2**: *Proceedings of the Second Valencia International meeting September* 6/10, 1983 (J.M.Bernardo, M.H.DeGroot, D.V.Lindley, & A.F.M.Smith, eds.; New York: North Holland), 249–269 (including discussion).

Good, I.J. (1984). The best explicatum for weight of evidence. C197 in *Journal of Statistical Computation and Simulation* **19,** 294–299; **20**, p. 89.

Good, I.J. (1985 / 88a). Causal tendency: a review. In *Causation, chance, and credence* (William Harper & Brian Skyrms, eds.; Dordrecht: Reidel, 1988), 23–50. Proceedings of a conference on probability and causation at Irvine, California, July 15–19, 1985.

Good, I.J. (1985 / 88b). Response to Wesley Salmon's critique, being also an extension of the appendix of Good (1985 / 88a). In the same volume, pp. 73–78.

Good, I.J. (1986). The whole truth. *The Institute of Mathematical Statistics Bulletin* **15,** 366–373. (An editor's invited column.)

Good, I.J. (1987). Subjective probability. *The new Palgrave: A dictionary of economics,* Volume 4 (J. Eatwell, M.Milgate, & Peter Newman, eds.; New York: Stockton press), 537–543.

Good, I.J. (1987 / 89). Speculations concerning the future of statistics. In the conference 'Foundations and Philosophy of Probability and Statistics: an International Symposium in Honor of I.J.Good on the Occasion of his 70th Birthday', May 25–26, 1987. The proceedings are to be published in a special issue of the *Journal of Statistical Planning and Inference.*

Good, I.J. (1988). The interface between statistics and the philosophy of science, *Statistical Science* **3**, 386–412 (with discussion). A shorter version was presented by videotape at the Eighth InternationalCongress on the Logic, Methodology, and Philosophy of Science, Moscow, August, 1987, and will be published in the *Proceedings.*

Good, I.J. (1989a). Yet another argument for the explication of weight of evidence. C312 in *Journal of Statistical Computation and Simulation* **31,** 58–59.

Good, I.J. (1989b). Weight of evidence and a compelling metaprinciple. C319 in *Journal of Statistical Computation and Simulation* **31,** 121–123.

Good, I.J. (1989c). On the combination of pieces of evidence. C311 in *Journal of Statistical Computation and Simulation* **31,** 54–58.

Good, I.J. & Card, W.I. (1971). The diagnostic process with special reference to errors. *Methods of Information in Medicine* **10,** 176–188.

Good, I.J., Holtzman, G.I., Deaton, M.L., & Bernstein, L.H. (1989). Diagnosis of heart attack from two enzyme measurements by means of bivariate probability density estimation: statistical details. C328 in *Journal of Statistical Computation and Simulation* 32, 68–76.

Good, I.J. & Toulmin, G.H. (1968). Coding theorems and weight of evidence, *Journal of the Institute of Mathematical Applications* **4**, 94–105.

Hart, H.L.A. & Honoré, A.M. (1959). *Causation in the law.* Oxford: Clarendon Press.

Jeffrey, R.C. (1965). *The logic of discussion.* New York: McGraw-Hill.

Jeffreys, H. (1936). Further significance tests. *Proceedings of the Cambridge Philosophical Society* **32**, 416–445.

Jeffreys, H. (1939 / 1961). *Theory of probability.* Oxford: Clarendon Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A* **186**, 453–461.

Kemeny, J.G. & Oppenheim, P. Degrees of factual support. *Philosophy of Science* **19,** 307–324.

Keynes, J.M. (1921). *A treatise on probability.* London: Macmillan.

Kleinmuntz, B. & Szucko, J.J. (1984). A field study of the fallibility of polygraphic lie detection. *Nature* **308**, 449–450.

Kullback, S. (1959). *Information theory and statistics.* New York: Wiley.

Laplace, F.S. de (1820). *Thèorie analytical des probabilités,* 3rd edn. Paris: Courcier.

Larousse (1968). *New Larousse Encyclopedia of Mythology.* London: Hamlyn.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society* **B51,** 157–224 (with discussion).

Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics* **27**, 986–1005.

Lindley, D.V. (1977). A problem in forensic science. *Biometrika* **64**, 207–213.

Minsky, M. & Selfridge, O.G. (1961). Learning in random nets. In *Information theory* (Colin Cherry, ed.) London: Butterworths.

de Morgan, A. (1847). Theory of probabilities. *Encyclopaedia of Pure Mathematics, Part II,* pp. 393–400. Also in *Encyclopaedia Metropolitana* (London, 1837).

Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosphical Transactions of the Royal Society* **A231,** 289–337.

Pearl, J. (1988). *Probababilistic reasoning in intelligent systems.* San Mateo, California: Morgan Kaufman.

Peirce, C.S. (1878). The probability of induction. *Popular Science Monthly.* Reprinted in *The World of Mathematics* **2** (James R. Newman, ed.). New York: Simon & Schuster (1956), 1341–1354.

Peterson, W.W., Birdsall, T.G., & Fox, W.C. (1954). The theory of signal detectability. *Transactions of the Institute of Radio Engineers* **PGIT-4,** 171–212.

Poisson, S.D. (1837). *Recherche sur la probabilités des jugements.* Paris: Bachelier.

Popper, K.R. (1954). Degree of confirmation. *British Journal for the Philosophy of Science* **5**, 143–149.

Popper, K.R. (1959). *The logic of scientific discovery.* London: Hutchinson.

Salmon, W.C. (1985 / 88). Intuitions—good and not-so-good. In *Causation, chance, and credance,* Vol. 1, 51–71. Also in published in Italian in *Epistemologia ed Economia,* 189–209.

Savage, L.J. (1954). *The foundations of statistics.* New York: Wiley.

Shannon, C.E. & Weaver, W. (1949). *The mathematical theory of communication.* Urbana, Illinois: University of Illinois Press.

Spiegelhalter, D.J. (1983). Private communication.

Spiegelhalter, D.J. & Knill-Jones, R. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application to gastroen-terology (with discussion). *Journal of the Royal Statistical Society* **A147,** 1–34.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In *Handbook of experimental psychology.* New York: Wiley, 1–49.

Toulmin, G.H. (1984). Lie detector lies. *Nature* **309** (May 17), 203.

Tribus, M. (1969). *Rational descriptions, decisions, and design.* New York: Pergamon Press.

Turing, A.M. (1941). Private communication.

Uttley, A.M. (1959). Conditional probability computing in the nervous system. In *Mechanization of thought processes,* N.P.L. Symposium no. 10 (London: H.M.S.O.), 119–147.

Wrinch, D. & Jeffreys, H. (1921). On certain fundamental principles of scientific discovery. *Philosophical Magazine, series* 6, **42**, 369–390.

Zabell, S.L. (1988). The probabilistic analysis of testimony. *Journal of Statistical Planning and Inference* **20**, 327–354.

4

Transfer evidence

David A.Stoney
Assistant Professor, Department of Pharmacodynamics, University of Illinois at Chicago, Chicago, Illinois, USA

4.1 INTRODUCTION

When two items come in contact with one another there is a potential for transfer of either substance or pattern. Transfers between items are an important means to infer the occurrence of past events, and as such they become evidence when the past events are of legal concern. Transfer evidence results when, from observations on the separate items, we can infer that a contact between the two items did take place or may have taken place.

Here we are concerned directly with the strength of this inference—with the interpretation of transfer evidence. The problem is central to forensic science and yet, because of its complexity and vagary, a comprehensive treatment of the problem has not emerged. The present effort is a step toward this goal, which promises to be an elusive one.

In the sections that follow we will identify concepts that are important for proper considerations of transfer evidence. These concepts retain their importance even if we cannot, or choose not to, quantify the strength of our inference. Specifically, they guide our thought process, helping to avoid errors in formulating the problem. Within the broad conceptual framework we can better appreciate the limitations on our ability to infer contact and the relationship of this inference to the legal issues under consideration. Section 4.2 serves as a general introduction and describes the transfer evidence problem.

Section 4.3 discusses match frequencies, where the occurrence of matching properties between two randomly selected samples is used to evaluate transfer evidence. Weaknesses inherent in match frequencies lead to consideration of correspondence frequencies in Section 4.4. Here we take one sample as given and seek the frequency of randomly encountering a corresponding sample. Application of correspondence frequencies to transfer evidence requires careful attention to the types of databases that are used and the precise formulation of the questions to be answered. Limitations of correspondence frequencies lead us to Bayesian methods as a necessary and logical extension. These methods are presented in Section 4.5, using a series of hypothetical cases to explore the difficulties and intricacies that are encountered.

In section 4.6, fundamental limitations of these quantitative methods are briefly discussed with reference to the reality of evidential development in case investigations. Section 4.7 concludes the chapter with a brief review of the more important concepts.

4.2 BASIC CONCEPTS IN TRANSFER EVIDENCE INTERPRETATION

Transfer evidence supports an inference of contact between specific source and receiving objects. Interpretation of transfer evidence is an evaluation of the strength of this inference, based on our knowledge and experience, taking into account any pertinent case circumstances. In this section we will consider the problem of transfer evidence interpretation, eliciting concepts that will guide our choice and evaluation of statistical methods.

The aspects of the problem before us comprise case specific data, a hypothesis regarding contact, and general background information (see Table 4.1). *Case-specific data* are the two objects at issue, the substance (or pattern) that may have been transferred, the match criteria, and the sample histories of these items. The *hypothesis of contact* states that contact with transfer occurred in a particular time and manner. *Background information* relates to the causes and dynamics of transfer, the persistence of transferred substances or patterns, and the commonness of various objects, materials, and patterns.

Table 4.1 —Basic aspects of the transfer evidence problem

Case specific data:

1. Possible source object
2. Possible receiving object
3. Substance (or pattern) that may have been transferred from 1 to 2
4. The match criteria
5. Known histories of 1, 2, and 3

Hypothesis of contact:

1. Contact did occur
2. Time of contact (relative or absolute)
3. Manner of contact

Background information:

1. Transfer causes and dynamics
2. Persistence of transferred materials (or patterns)
3. Commonness of the various objects and materials (or patterns)

The case-specific data allow appropriate definition of the hypothesis and determine the type of background information that is necessary to evaluate support for the hypothesis.

The proper perspective on case-specific data follows only after recognition of the intricacies of the generalized transfer evidence problem. Almost always the problem can be considerably simplified; so much so that in routine work one can develop a naive and inadequate perspective. In this chapter we begin with fairly simple case circumstances where the interpretation is straightforward and intuitive. Through progressively more complex examples the intricacies of the generalized transfer evidence problem become apparent.

Transfer evidence does not come to be unless there has been a laboratory analysis, and quite naturally one focuses on this portion of the problem. Typically two items are submitted to a laboratory along with an examination request. The request itself often defines what particular type of evidence will be sought. In a bloody homicide case, for example, a request might read, ‘Examine item 1 (suspect’s jacket) for bloodstains, type any blood found, and compare bloodtypes with item 2 (known blood from the victim).’ Alternatively, where more extended contact occurred, there may be a more general instruction to ‘look for any evidence that items 1–4 (suspect’s clothing) came into contact with items 5–8 (victim’s clothing)’. Based on the request, any supplementary information and discussions with investigators, the laboratory analyst conducts tests and searches for materials that may have been transferred from one item to another. If extraneous material found on one item could have come from another item, then transfer evidence exists and the problem of interpretation follows.

From the laboratory analyst’s perspective certain tests were applied and a correspondence in test results followed. A natural focus for interpreting this correspondence is to ask how often it would occur by chance. As stated, however, this question is incomplete. Before we can proceed we must define precisely what we mean by the ‘chance event’ (Stoney 1984a). In the next several sections we will consider the usefulness of a few alternative definitions.

4.3 MATCH FREQUENCIES

In this section we describe Discriminating Power and Double Coincidence Methods. Both of these utilize match frequencies, where the occurrence of like properties between two samples is taken as the chance event of interest in evaluating transfer evidence.

4.3.1 Discriminating power of a technique

The significance of an occurrence of like properties (or match) depends to a great extent on how critically the two items are compared. If we use a fairly simple test criterion, samples from two different sources might match frequently. More rigorous test criteria make such fortuitous agreement less common. Discriminating power (Smalldon & Moffat 1973, Jones 1972) is a simple quantitative measure for the frequency of random sample matching, given a particular analytical protocol.

It will facilitate discussion to consider a hypothetical burglary case involving glass fragments as transfer evidence. Suppose glass samples from a broken window at a burglary scene are submitted to a laboratory along with clothing from a suspect. The laboratory finds glass fragments on the clothing and makes the comparison. The two samples are found to agree in all measured properties.

To calculate discriminating power in our hypothetical case we would consider glass samples generally and ask, ‘How often would two randomly selected glass samples *differ,* given the test criteria?’ To answer this question we need to make a survey of glass samples and (conceptually) pick out pairs at random. We count up the pairs which match and the pairs which do not and then calculate the chance of picking a *non-matching* pair at random. A simple calculation of discriminating power for qualitative characteristics of known frequency is shown in Table 4.2.

Table 4.2 —An example of a calculation of discriminating power

Mutually exclusive exhaustive outcomes	Frequency	Probability of joint occurrence in a pair
A	a	a^2
B	b	b^2
C	c	c^2
Totals	1.0	$a^2 + b^2 + c^2$

Notes:

(i) Discriminating power = $1 - (a^2 + b^2 + c^2)$.

(ii) The frequencies of each possible type are noted, and, assuming two independent random individuals from a population, the joint probability of a coincidental correspondence is given by the sum of the squares of the individual frequencies. This value has been referred to as PM, the probability of a match (Jones (1972), see also Chapter 2, section 6, for further discussion of this.) Discriminating power is then given by (1—PM).

For our present purposes we want to interpret the strength of the observed match. It is clear that the discriminating power of our test criteria has some bearing on this problem: the more discriminating a test is, the more significant will be the match in properties. We can also use discriminating power in a statement regarding the evidence, such as:

> ‘In the present case I applied a specific test protocol and found matching properties. Using this test I would expect only one in *N* pairs of randomly selected glass samples to match (where *N* is equal to {1/(1-DP)} and DP is the discriminating power, a number between 0 and 1).’

This statement is useful, but in any specific case it has serious limitations. Note that the actual properties of the evidence—the values that match—are not considered. We may have a very common or a very rare type of glass *in this particular case.* Discriminating power, which depends on the test criteria as applied to a population, is insensitive to this and gives, in effect, an 'average probability' of finding a match when using the test criteria. In actual cases this average probability will sometimes underestimate evidential value and sometimes overestimate it. Tippett *et al.* (1968) as well as Gaudette & Keeping (1974) have used discriminating power to estimate a general frequency of occurrence for the 'average' case. This practice has considerable potential for error. Gaudette (1978) cautions:

> 'The probability results previously found are average figures that apply to [samples] of average commoness. They should not be applied blindly to all cases'. 'Thus, for most [samples] of about average commonness, the probability figures provide a reasonable general estimate

Arguments have been made, however, that such average probabilities are useful and appropriate in circumstances where more specific data are unavailable and where the expert's experience indicates that the case materials are neither extremely common nor extremely rare. The danger is in indiscriminate or naive application (which, unfortunately, is not uncommon). The values and shortcomings of the method have been debated by a number of authors (Stoney (1984a), Gaudette (1978), Aitken (1987a), Aitken & Robertson (1987), Barnett & Ogle (1982), Gaudette (1982)). A further, more disturbing, variation on this theme, follows.

4.3.2 The double coincidence fallacy

As noted above in section 4.3.1, discriminating power or general match probabilities are not sensitive to the commonness of the type of material found in any particular case. Thus, if we had a very rare or very common glass type in our hypothetical case this very relevant information would be ignored.

A seemingly appropriate response is to consider the chance of a coincidental match of two items having the specific properties found in the case. For our glass samples we would ask, 'How rare would it be to select two random glass samples and have them share this set of properties?' Let F be this set of properties (note that 'F' will be used throughout this chapter to denote the laboratory findings), occurring with relative probability $P(\mathrm{F})$ among glass samples. Assuming independence, the joint probability of randomly selecting two samples with properties F would be given by the trivial product:

$$P(\mathrm{F}) \times P(\mathrm{F}) \tag{4.1}$$

The faults of this method are apparent when one recalls that the role of evidence is to help assess whether or not a particular individual, or suspect, is associated with a particular criminal act. We offer as self-evident that evidence implicating a suspect's involvement must be judged in relationship to how often randomly selected persons would be so implicated. If we consider a number of suspects, it is clear that we could judge each on the basis of how strongly they could be associated with the crime scene materials. The crime scene materials thus define a standard set of properties against which all possible offenders would be evaluated. (Later, when likelihood ratios are considered, this viewpoint is not essential and the terms ‘crime scene anchoring’ and ‘suspect anchoring’ are introduced to contrast methods where materials from one source or the other are taken as known.) Once a crime is committed this set is constant regardless of who becomes a suspect, or whether one is ever located, Since the circumstances defined by the crime are known and provide our basis for evaluating suspects, they cannot be considered variable after an association is made. In evaluating the significance of matching transfer evidence, we must remember that only one half of this match is relevant to the suspect; the other half is merely a result of which house was burgled, who was killed, or whatever the circumstances of the crime may be.

In this sense, the rarity of the crime scene materials themselves is not directly relevant to the evaluation of transfer evidence. To be sure, if very rare materials are at issue, the match would be highly significant. This significance, however, comes from the fact that the corresponding materials found associated with the suspect are rare, not from the fact that the crime scene materials are rare. This leads us to consideration of *Correspondence Frequencies.*

4.4 CORRESPONDENCE FREQUENCIES

Recognizing the fundamentally different roles of crime scene and suspect materials forces closer examination of our problem. Instead of considering the match itself as an event, we will now accept the crime scene materials as given and examine the rarity of correspondence to these materials.

4.4.1 Crime scene materials and suspect-related materials

When the identity of one item is given and we assess the rarity of corresponding items, the distinction between crime scene materials and suspect-related materials becomes very important. In general there will be differences in the form and occurrence of these materials. Previously we spoke loosely about the frequency of a match between two glass samples. In fact these two samples are of different forms. The crime scene glass is from a broken window and is *known* to be window glass. The sample from the suspect is in the form of glass fragments, and we have no such assurance. The fragments could be, for example, from a broken container. This difference is important and must be closely examined.

4.4.2 Bulk-source form and transferred-particle form of materials

Most often transfer evidence results when small particles are shed, fractured, or otherwise removed from a larger, bulk, source. Fracture of a glass item results in glass fragments; shedding of textiles produces loose fibres; cuts in our bodies result in blood stains. It is therefore useful to recognize two general forms of transfer evidence, the *source or bulk form* (bulk-source form) of the material and the *transferred-particle form* of the material. (The use of the terms ‘control’ for bulk-source form and 'recovered' for the transferred-particle form are sometimes encountered. Here, we specifically avoid these terms, as well as the terms 'questioned' and 'known'. The recovery of a jumper of unknown origin from a crime scene should amply illustrate the hazardous ambiguity of these terms.) Some examples of bulk-source and transferred-particle forms of common transfer evidence types are given in Table 4.3.

Table 4.3 —Forms and occurrence of common types of transfer evidence

Evidence	Bulk-source form	Transferred-particle form	Transfer event	Population options
Blood	Person	Bloodstain	Bleeding	People, bloodstains
Glass or paint	Glass item, painted surface	Glass or paint fragments	Breakage	Glass or painted items; locations, residences, fragments, fragments on people or clothing
Fibres	Textiles, clothing items	Fibres	Shedding	Textiles or clothing items, locations, residences, individual fibres, fibres on people or clothing
Hair	Person	Hairs	Shedding	People, locations, residences, individual hairs, hairs on people or clothing
Soil	Topsoil	Heterogeneous minerals, building and biological materials	Soiling	Sites, locations, residences, soil on people or clothing

Note: Transfer evidence frequently involves comparing bulk, composite sources with shed or fractured particles. Bulk and particle forms of common transfer evidence materials are given. The transfer event is that which causes the transfer or separation of the material from the bulk-source form. Proper assessment of the rarity of a particular material requires definition of a relevant population. Generally these populations could be bulk-source forms, transferred-particle forms, or the occurrence of each of these in association with people, their clothing, or their residences.

In our hypothetical case the crime scene material has the bulk-source form (broken window) and the suspect-related material has the transferred-particle form (glass fragments). Sometimes the opposite is true. If fibres from an offender's clothing are left at a crime scene, for example, the transferred-particle form is the crime scene material. The corresponding bulk-source form would then be the suspect's clothing.

4.4.3 Applying correspondence frequencies

Applying the correspondence frequency method to our hypothetical glass case, we pose the following question:

> 'Given the crime scene materials, what is the frequency of random occurrence of corresponding suspect-related materials?'

Or more specifically

> 'Given the properties of the broken crime scene window, what is the frequency of random occurrence of corresponding glass fragments?'

To estimate this frequency we need to measure properties on a population of glass samples and determine how often the properties correspond to the crime scene materials. Note that we are interested in the frequency of occurrence of this set of properties among *glass fragments* (the transferred-particle form), NOT in the frequency of these properties among windows (the bulk-source form).

Suppose we have extensive data on the properties of window glass. Can we assume the same frequencies of properties among glass fragments? The assumption would seem valid if all glass fragments come from broken windows. Glass fragments, however, also come from containers, headlamps, etc. Frequencies of properties in the two populations may therefore be expected to differ. Properties found only rarely in window glass might be quite commonly found in glass fragments, and vice versa.

4.4.4 Need for form-specific population data

Unless explicitly justified to do otherwise, we must use a *form-specific population* with the form being that of the suspect-related material. Because the suspect-related material can be in either form we will generally need two distinct population databases: in some cases we need a database for the transferred-particle form (as in our glass example); in other cases we would need a database for the bulk-source form. An example of this would be for headlamp glass properties, where glass fragments are left at the scene of a hit-and-run accident. The suspect-related material is the headlamp, a bulk source.

There are three mechanisms that can cause differences in bulk-source and transferred-particle form population frequencies. One of these is, as in our glass case, when a variety of bulk source types (windows, bottles, etc.) contribute to the particle population (glass fragments). Differences also arise when some types of bulk sources are more likely to produce or transfer the particle form. Frequencies of wool and polyester jumpers might be equal, for example, but the wool fibres would be more likely to be shed and transferred. This would result in a greater proportion of wool fibres in the particle population. A third cause for population differences is when there is variation in the stability of the transferred particle form. Once detached from their bulk source synthetic fibres would, for example, resist degradation longer than, say, cotton fibres. This mechanism would favour greater frequencies of synthetic fibres in the particle population.

For some types of transfer evidence no population differences are expected between the bulk-form and transferred-particle form. Bloodstain evidence is a good example. If type A blood is found with a frequency of 0.40 in a city's population, it is hard to imagine why the proportion of type A blood would be different in shed blood (bloodstains). For blood, then, we need not consider the distinction between bulk-form and particle-form populations. Furthermore, blood typing is a natural focus for quantitative interpretation because databases exist and assumptions of independence of various bloodgroup systems can be fully justified. This simplifies evidential interpretation considerably, but at the same time it masks the depth of the issues when blood is used as an example.

4.4.5 Need for environment-specific population data

Just as the *form* of a material can affect population frequencies, so can the *environment* in which the material is found. We might consider the population of glass fragments found on the street, for example, and contrast this with the population of glass fragments found on clothing. Since our suspect-related glass fragments were found on clothing, this population would be of most direct interest. A survey of street fragments would probably not be useful because there is reason to expect systematic differences between these two populations. Street fragments will undoubtedly have a greater proportion of automobile headlamp glass. Urban streets will have much bottle glass. Glass on clothing comes either from contact with loose fragments on some surface or as a result of being close to a breaking object. Our speculations regarding differences in the population frequencies can be replaced only by comparing environment-specific surveys and determining if any differences are significant (Walsh & Buckleton 1986). When possible, however, we would like population data for the specific form and environment in which the suspect-related material was found.

A variety of environment-specific population data has been published, most, as in our example, concerned with incidence of particle-form of materials on clothing. These include glass and paint fragments on clothing (Pearson *et al*. (1971), Harrison *et al*. (1985)), glass in footwear (Davis & DeHaan (1977)), loose fibres on clothing (Fong & Inami (1986)) and bloodstains on clothing (Briggs (1978), Owen & Smalldon (1975)). When the suspect-related material is the fragmentary form on clothing, these surveys are relevant and useful for estimating correspondence frequencies. If, however, the suspect-related material is in the bulk form, surveys of this type are not directly relevant. Instead we would need population data for the bulk-source form.

Environment-specific population data are useful in two distinct ways which we must not confuse. For the present we are interested in estimating relative frequencies of certain particle characteristics when the particle is found in a specific environment. The other way environment-specific data can be used is to estimate the frequency with which the particle form is found *in the environment.* Thus in the first instance we might wish to estimate the frequency with which paint chips, found on clothing, were red (the members of the population being paint chips of various colours). In the second instance we might wish to estimate the frequency of finding red paint chips (or paint chips, generally) on clothing. Here the members of the population are items of clothing, not paint chips. Both types of data are useful to help interpret transfer evidence. We will discuss the second use later in section 5.

4.4.6 Erroneous use of correspondence frequencies

There are two common errors that must be avoided when using correspondence frequencies. Each of these involves failure to address, properly, the initial question;

> 'Given the crime-scene material, what is the frequency of encountering corresponding suspect-related material?'

The first type of error follows from defining an inappropriately narrow field of possible suspect-related materials. Once a match in properties is observed there is a tendency to focus one's attention on the particular type of suspect-related material that is found *in the case,* instead of *any* suspect-related material that would correspond to the crime scene material. Most often this occurs when the suspect-related material is in the bulk form. Objects in their bulk form have a set of defining traits and an unambiguous identity. Frequencies of these specific items can be estimated by either direct surveys or through manufacturing and distribution data. The problem is that other items *capable of producing the particle-form crime scene material* are also possible corresponding suspect-related items.

Consider a case where yellow fibres are found on a murder victim and where a suspect is developed who has a yellow jumper composed of matching fibres. There is a tendency to ask, 'What is the frequency of encountering these yellow jumpers?' In fact, of course, any item of clothing capable of shedding the type of yellow fibres found at the crime scene would be a possible source and should therefore be included in our frequency estimate. This more broadly conceived frequency, although harder to estimate, is necessary in order to evaluate the evidence properly.

A particularly current example of this error was noted by Lander (1989) with respect to DNA bloodtyping as applied in the trial of *People v. Castro.* DNA blood types are observable as bands positioned in a column according to their molecular size. In comparing a suspect's bands to those from a crime scene stain their positions were first matched. An average of the two measured positions was then taken as the best estimator of molecular size, and a range of possible values was determined. The population frequency of this range for the average was then estimated from a database. As pointed out by Lander, this procedure fails in that the range of band positions that *would be matched* to the crime scene sample is not considered. The population frequency of this range in matchable band positions is required, not the frequency of the best molecular size estimate.

The second type of error in applying correspondence frequencies results from inverting the roles of the suspect-related material and the crime scene material. This occurs if, instead of asking,

> 'Given the crime scene material, what is the frequency of corresponding suspect-related material?'

we ask

> 'Given the suspect-related material, what is the frequency of corresponding crime scene material?'

Earlier, we discussed the need for different databases for the bulk-source form and transferred-particle form of a material, and, since this form will usually differ between the suspect and crime scene material, the need for a clear distinction between the above questions should be clear.

Evett (1984) has referred to the first question as 'scene-anchored' and the second type of question as 'suspect-anchored'. Anchoring refers to the portion of the data that is taken as known. For *correspondence frequencies* only the scene-anchored question is appropriate for the evaluation of suspects. (Later we will see that this distinction is not as important when applying Bayesian methods.) As noted above, evidential value must be judged in relationship to how randomly selected persons might be implicated by the evidence. The crime scene materials are taken as known and define a set of properties against which all suspects would be evaluated.

In the yellow fibre/jumper case above we commit this error if we estimate the frequency of fibres that would correspond to the suspect's sweater. In the glass fragment/broken window case the error would be made if we sought the frequency of windows that would correspond to the fragments found on the suspect. In each case we are estimating the frequency of *crime scenes* (or crime-related materials) with which the suspect-related material would correspond.

The seriousness of the distinction between the two questions above is amply illustrated by the following murder case (Kuo 1982). A young woman was missing after going on a boat ride with her boyfriend, and he became the prime suspect. Bloodstains were found on a suspect's boat, but since no body was found there was no known blood of the woman to type and compare with the bloodstains. The boyfriend was accused of murder and, in an effort to link the bloodstain to the missing woman, the woman's parents were bloodtyped. This is a type of parentage testing. It was found that the parents could have had a child with the type shown in the bloodstains. In the evaluation of this evidence we could ask either of these two questions:

> 'How common are *couples* that could have had a child with the bloodstain type?'

or

> 'how common are *bloodstain types* that could have come from a child of this couple?'

The first question is suspect-anchored because the bloodstains, found in association with the suspect, are taken as known. The missing woman's parents are the actual crime scene material since they are known to be associated with the missing woman. The second question is scene-anchored and is the appropriate question to ask. In two reported cases of this type (Kuo (1982), Ogino & Gregonis (1981)), the answers to the two competing questions were found to be dramatically different, as shown in Table 4.4. A full discussion of these cases appears elsewhere (Stoney 1984b). What is of importance here is that substantial interpretive errors can occur when correspondence frequencies are imprudently constructed.

Table 4.4 —Summary of statistics in two missing victim cases

Statistic	Case 1	Case 2	Description
Frequency of corresponding couples	0.006	0.26	Probability that a randomly selected couple could pass the *stain* type
Frequency of corresponding stains	0.34	0.024	Probability that a randomly selected type could be passed by the *couple*
Likelihood ratio	130	44	Odds on this *couple* passing the *stain* type relative to a randomly selected *couple*
Reciprocal of the likelihood ratio	0.008	0.023	Type frequency which would generate the observed likelihood ratio in a case where the victim's type is known

Note: Neither of the two correspondence frequencies accurately assesses the evidential value. Note the widely divergent values for the correspondence frequencies in these two cases. When the likelihood ratio is used the evidential value is accurately assessed and is found to be very close to the lower of the two correspondence frequencies in these two cases.

4.4.7 Limitation of correspondence frequencies

Correspondence frequencies cannot accurately evaluate evidence where the characteristics of the crime scene material are uncertain. The 'missing person bloodtyping' cases are an excellent illustration of this and will be discussed in the following section. A simpler example, following Evett's two-trace problem (Evett 1987), will suffice here. Consider a case where a crime has been committed and it is known that two men were involved, both of whom were injured and left bloodstains at the scene of the crime. (This may seem unrealistic, but directly analogous evidence is present in a case of multiple rape.) Later, the investigator develops a single suspect. There is no information about the second man, and the suspect denies involvement. Bloodtyping shows that one of the scene stains is type y_1 and the other is type $y_{2.}$ The suspect is type y_1. Suppose the relative frequencies of types $y_1$and y_2 are 0.01 and 0.20, respectively. In this case the correspondence frequency approach would ask for the frequency of a corresponding blood group. A corresponding bloodgroup would be found with a frequency of $0.01 + 0.20 = 0.21$. This would be the result regardless of whether the suspect was of type y_1 or y_2. The measure of evidential value is thus insensitive to whether or not the correspondence is with the rarer or more common phenotype. Even if there were a very rare blood type occurring in both the suspect and in one of the crime scene stains, the correspondence frequency would remain large owing to the more common stain type. We could imagine an extreme case where finding an extremely rare blood type in one of the stains was virtual proof of its origin. The correspondence frequency approach would be insensitive to this and would recognize *no difference* in evidential value with the circumstance where a suspect had the more common blood type. To generalize, any time *multiple crime scene items* could be matched to a suspect this weakness in the correspondence frequency will be expressed. The missing person bloodtyping cases cited earlier (Stoney 1984b, and see Table 4.4) also provide a dramatic illustration of this weakness. In this instance, the crime scene ‘materials’ are the *parents* of the missing person, and there is a large set of possible offspring bloodtypes. A correspondence frequency would sum together the incidence of all possible types when assessing bloodstains found in association with a suspect.

4.5 THE BAYESIAN METHOD

The application of Bayesian methods has been well argued in this collective volume and elsewhere (Evett 1987, Finkelstein & Fairley 1970, Lindley 1977, Gettinby 1984, Fienberg & Schervish 1986, Evett *et al.* 1987, Evett 1986). Key aspects of this argument are, firstly, that in simple cases the evidential value given by Bayesian methods is identical with the more intuitive correspondence frequencies, and, secondly, that in more complex cases the mathematics of Bayes’ Theorem must be accepted in order to retain this intuitive link. Our purpose here is to illustrate these two aspects.

4.5.1 Equivalency of the Bayesian approach in simple cases

Consider first a simple bloodstain case where an offender in a burglary case leaves a bloodstain at the scene of a crime. Some time later, as a result of information received, an investigator locates a man who, he suspects, is the person who committed the crime. The suspect denies involvement and provides a blood sample (It is assumed that sufficient time has elapsed since the crime for the cut which caused the bleeding to have healed so that the scientific examination is confined to blood typing.) Both the suspect and the bloodstain are type y_1. A survey has shown that the relative frequency of this phenotype in the population is q_1 (that is, $100q_1\%$ of the population are type y_1).

The investigator is interested in establishing the evidence relating to the proposition:

C: the suspect was the offender (that is, the man who committed the crime).

To evaluate the evidence it is necessary to consider the alternative proposition:

$\bar{C}$: the suspect was not the offender.

Denote the background information as I. Then, before the bloodtyping is done, we have the prior odds of the suspect being the offender:

$$O(C|I).$$

Denoting the scientific findings (bloodtyping) by F we require to update the odds to:

$$O(C|I,F).$$

This is done by means of Bayes' Theorem:

$$O(C|I,F) = \frac{P(F|C,I)}{P(F|\bar{C},I)} \cdot O(C|I). \tag{4.2}$$

Now the class of questions which the scientist should address to evaluate his evidence is defined. Whereas the investigator and, perhaps later, a court of law are concerned with the probability of C, given the findings, F, and I, the scientist must consider the probability of the findings, given both the alternatives C and $\bar{C}$. The ratio of these two probabilities is the likelihood ratio (LR):

$$LR = \frac{P(F|C,I)}{P(F|\bar{C},I)}. \tag{4.3}$$

In what follows, I is omitted purely for the sake of brevity. However, it is essential to remember that the findings can be evaluated only in the light of this background information. Occasionally, this background information will so influence the problem as to render the other, quantitative, data insignificant. For our simple case the likelihood ratio is evaluated as follows.

The scientific findings are a correspondence in bloodtype y_1 between a suspect and the crime scene stain. The probability of such a correspondence under C is simply the probability that the crime scene stain will be type y_1 since if the suspect is the offender then his type must correspond. The numerator is thus:

$$P(\mathrm{F/C}) = 1\times q_1.$$

Under $\bar{\mathrm{C}}$ the suspect is random with respect to the crime scene stain, and the probability of a correspondence of type y_1 is $q_1 \times q_1$. The likelihood ratio is therefore given by:

$$\frac{P(\mathrm{F}|\mathrm{C})}{P(\mathrm{F}|\bar{\mathrm{C}})} = \frac{q_1}{q_1 \times q_1} = \frac{1}{q_1}. \tag{4.4}$$

This is not a surprising result, and it conforms with what most forensic scientists consider intuitively reasonable. The rarer the blood type, the greater the likelihood ratio and hence the evidential value of the correspondence between the stain and the suspect's blood type. As a likelihood ratio, the value of the evidence would be summarized as: the evidence is $(1/q_1)$ times more likely if the suspect committed the crime than if he did not. More conventionally expressed, the evidential value would be summarized as: one in every $1/q_1$ individuals would have this type or, equivalently, the probability of a person selected at random exhibiting this blood type is q_1.

The role of the background information I can be illustrated quite simply. If there is an eyewitness who says that the person who committed the crime was of a particular ethnic group, then in order to evaluate the denominator it is necessary to use the relative frequency of the blood type among that ethnic group. Although obvious in this case, the role of the background information in determining the probability of the findings under different propositions is not always so clear, but the logical methodology embodied in the Bayesian approach is a valuable aid.

4.5.2 Factoring of the likelihood ratio, using conditional probabilities

An alternative approach to the likelihood ratio construction just presented involves the factorization of the ratio by the use of conditional probabilities. Instead of viewing the scientific findings as 'a correspondence in bloodtype y_1', we consider the findings F as two elements (F_1, F_2):

F_1: the bloodstain is type y_1,
F_2: the suspect is type y_1.

Using the multiplication law for probability:

$$\frac{P(\mathrm{F}|\mathrm{C})}{P(\mathrm{F}|\bar{\mathrm{C}})} = \frac{P(\mathrm{F}_2|\mathrm{C})}{P(\mathrm{F}_2|\bar{\mathrm{C}})} \times \frac{P(\mathrm{F}_1|\mathrm{F}_2,\mathrm{C})}{P(\mathrm{F}_1|\mathrm{F}_2,\bar{\mathrm{C}})}. \tag{4.5}$$

We can considerably simplify the right-hand side of this expression. $P(\mathrm{F}_2|\mathrm{C})$ and $P(\mathrm{F}_2|\bar{\mathrm{C}})$ denote the probability that the suspect is type y_1, given C and $\bar{\mathrm{C}}$, respectively. But the suspect's blood type was determined long before the event of interest and is independent of whether or not he committed the crime. It follows that $P(\mathrm{F}_2|\mathrm{C}) = P(\mathrm{F}_2|\bar{\mathrm{C}})$ and the first ratio on the right-hand side becomes one. In the denominator of the second ratio, knowledge of F_2 is irrelevant to F_1if the suspect had nothing to do with the crime, so (4.5) simplifies to:

$$\frac{P(F|C)}{P(F|\bar{C})} = \frac{P(F_1|F_2,C)}{P(F_1|\bar{C})}. \tag{4.6}$$

Consider first the numerator of this expression. This is the probability that the stain at the scene would be type y_1. This probability is simply one, that is, $p(F_1|F_2,C) = 1$. To evaluate the denominator it is necessary to visualize what happened if $\bar{C}$ were true. In that case, some other person left the stain at the scene. In the absence of any information to the contrary, then the unknown person is, as far as the blood type is concerned, a person selected at random from the population. It follows that $P(F_1|\bar{C}) = q_1$, the relative frequency of blood type y_1. So:

$$\frac{P(F|C)}{P(F|\bar{C})} = \frac{1}{q_1}, \tag{4.7}$$

which is equivalent to the result in the preceding section.

The simplification applied here is frequently justifiable on the basis that the suspect's characteristics are not dependent on whether or not he committeed the crime, but this should not be casually assumed. Some crime scenes would be very likely to transfer materials to an offender, thus helping to define his characteristics (in the sense of material on his clothing). If later identified as a suspect, the presence of this transferred material is certainly not independent of C. When in doubt the findings regarding both the suspect and the crime scene must be jointly considered.

4.5.3 Meaningful assessment in more complex cases

4.5.3.1 The two-bloodstain problem

The case example given earlier with two bloodstains left at a crime scene provides an excellent illustration of the applicability of the Bayes/likelihood ratio approach. This is a simple variation of Evett's (1987) two-trace problem.

Recall that two bloodstains were left at a crime scene by two offenders, and that these bloodtypes occurred with frequencies of $y_1 = 0.01$ and $y_2 = 0.20$ in the population. The typing result is that the suspect is type y1. Using likelihood ratios we can proceed using (4.6), where F_2 denotes the evidence of the suspect's bloodtype and F_1 denotes:

F_1: one bloodstain is type y_1,
one bloodstain is type y_2.

For the likelihood ratio we need to compute $P(F_1|F_2,C)$ and $P(F_1|\bar{C})$. The first term, the numerator, is the probability that one of the stains was y1 and the other was y_2, given that the suspect was one of the two men. In this case it is certain that one of the stains would be y_1, and there is no alternative other than to infer that the other stain came from the other man. In the absence of other information we regard this man as someone selected at random from the population, so the probability that the other stain would be y_2 is q_2. There appears to be no reasons why the two sets of stains should not be regarded as independent, so:

$$P(F_1|F_2,C) = 1 \times q_2 = q_2. \tag{4.8}$$

If $\bar{C}$ is true, then the crime was committed by two unknown men who can be regarded as men selected at random from the population. The denominator, then, is the probability that when two people are selected from the population one is type $y_1$and the other is type y_2. There are two ways of picking two men with this combination of phenotypes: the first could be y_1 and the second y_2, and vice versa. It follows that:

$$P(F_1|\bar{C}) = 2 \times q_1 \times q_2. \tag{4.9}$$

From equations (4.8) and (4.9) then:

$$\frac{P(F|C)}{P(F|\bar{C})} = \frac{1}{2q_1}. \tag{4.10}$$

An alternative way of viewing the same case circumstances follows from (4.5) directly, without the simplification using conditional probabilities. We can view the case findings as a correspondence between suspect and crime scene in type y_1 along with (random) scene blood types of y_1 and y_2. Under C the scene blood types would have a frequency of occurrence of $2q_1q_2$ and the correspondence in y_1 would have a frequency of 1/2 (that is, half of the time we would have the offender with type y_1 if we had one of the two offenders). Under $\bar{C}$, the scene blood types would again have frequency $2q_1q_2$ and the suspect's type would have the random frequency of y_1. The likelihood ratio is therefore:

$$\frac{P(F|C)}{p(F|\bar{C})} = \frac{(2q_1q_2)(\frac{1}{2})}{(2q_1q_2)(q_1)} = \frac{1}{2q_1}, \tag{4.11}$$

agreeing with the previous result.

This result is simple and reasonable. One would expect the evidential value in this case to be less than in the corresponding single-stain case; the fact that the likelihood ratio is halved is reasonable. Furthermore, the earlier objections to the correspondence frequency for this case are overcome. Recall that in our example there were frequencies of 0.01 and 0.20 for the two stains. The correspondence frequency was the sum of these and it did not matter if the rarer or more common stain was actually found in the case. In the extreme, where a very rare type was seen in both the suspect and one of the crime scene stains, this result was obviously unacceptable. The likelihood ratio, however, is responsive to this situation in a logical way. The measure depends only on the frequency of the corresponding bloodstain. The other bloodstain represents essentially a second chance at matching, but its population frequency is not ultimately relevant to evaluating this suspect.

4.5.3.2 Missing person bloodtyping

A second example where the likelihood ratio is necessary is in the missing person bloodtyping case considered earlier in connection with correspondence frequencies. This was the case where, since no body was found, a missing person's parents were typed in order to get blood type information for comparison with bloodstains found associated with a suspect (see Table 4.4). The appropriate (scene-anchored) correspondence frequency was the frequency of bloodstain types that could have come from a child of the couple. This correspondence frequency has a weakness directly analogous to the 'two-bloodstain' problem just discussed. Looking at the missing person's parents, a variety of bloodstain types are possible for the children. Some of these types may be quite common, others quite rare. In the case, however, only one stain type is involved. The correspondence frequency approach sums all the possible stain type frequencies, mixing the rare together with the common. A likelihood ratio approach overcomes this weakness. We would calculate the ratio:

$$\frac{P\text{ (this stain type given it is from one of the parent's offspring)}}{P\text{ (this stain type given it is from a random person)}}.$$

A full discussion of this case is given in Stoney (1984).

4.5.4 Suspect- and scene-anchoring in likelihood ratios

Another feature of likelihood ratios is that we need not distinguish between the suspect-anchored and scene-anchored perspectives. One direct consequence of Bayes' Theorem is the equality:

$$\frac{P(\mathrm{F}_1|\mathrm{F}_2)}{P(\mathrm{F}_1)} = \frac{P(\mathrm{F}_2|\mathrm{F}_1)}{P(\mathrm{F}_2)}. \tag{4.12}$$

For the missing person bloodtyping case, where correspondence frequencies were troublesome, there is now equivalency between the two possible likelihood ratios. L_1 (below) takes the scene-anchored approach by taking parents as given and asking for the probability that these parental phenotypes would produce a child that would have the types seen in the bloodstain that was found on the suspect:

$$\mathrm{L}_1 = \frac{\begin{array}{c}P\text{ (this set of parents would produce an off-}\\ \text{spring with these phenotypes)}\end{array}}{\begin{array}{c}P\text{ (a random set of parents would produce an}\\ \text{offspring with these phenotypes)}\end{array}}.$$

The alternative form of the likelihood ratio (L_2, below) uses the suspect-anchored approach by taking the bloodstain found on the suspect as given and asking for the probability that a person with these phenotypes would have had parents with the phenotypes seen in the missing person's parents.

$$L_2 = \frac{P\,(\text{this phenotype would have come from a set of parents with these types})}{P\,(\text{this phenotype would have come from a randomly selected set of parental phenotypes})}.$$

This latter somewhat laborious calculation is unnecessary because of Bayes' Theorem, and the distinction between suspect- and scene-anchoring is lost.

4.5.5 Multiple crime scene materials and their relevance

The two-bloodstain and missing person bloodtyping cases discussed earlier are both examples where the crime scene material allows a variety of possible incriminating characteristics for suspects. When a suspect is found with one specific characteristic, the likelihood ratio allowed meaningful assessment of the evidential value. In this section we will consider more closely how multiple crime objects and uncertainty regarding their relevance complicate the assessment of evidential value.

4.5.5.1 A comparison of the two-bloodstain and missing person bloodtyping cases

In the two-bloodstain case the variety (pair) of incriminating blood types arose from two distinct traces left at the crime scene, both of which were left by offenders. Contrast this with the missing person bloodtyping case where the parental blood types defined a field of possibilities for the blood type of the missing person. These possible types carried (unequal) probabilities of occurrence based on the principles of genetics. The random frequencies of occurrence in the population at large were also unequal. Essentially, because only the parental blood types were available as predictors of the missing person's type, there was ambiguity in our knowledge of this type.

We can make the two-bloodstain case more like the missing person case if we relax the assumption that the two bloodstains both came from offenders, and instead assume that there is only one offender. We still find two different blood types at the crime scene, but now only one of them is from the offender and, importantly, we don't know which blood type this is. The other blood type has nothing to do with the crime.

Note that, although this case example may appear contrived and unrealistic, identical considerations occur with many other types of physical evidence. There might be a variety of shoe prints on a rural path, with only one set being from the offender. Alternatively, with fibre evidence, for example, there is always a variety of different shed fibre types at a crime scene. Some of these types may have come from the offender, but some certainly did not. In the broadest view, the different kinds of evidence themselves represent multiple chances to link any given suspect with the crime scene, and the more evidence that is collected, the greater the chance of spurious incrimination.

In our modified two-bloodstain case we now have some doubt as to the (single) offender's blood type. The type is either y_1 or y_2, but one of these two types is known to be from someone other than the offender. If we assume that the two types are *equally likely* to be the offender's, then the likelihood ratio remains unchanged. If, however, there is an *a priori* probability r that the stain showing type y_1 (and corresponding to the suspect) is, in fact, from the offender, then the likelihood ratio becomes, from (4.5):

$$\frac{P(\mathrm{F}|\mathrm{C})}{P(\mathrm{F}|\overline{\mathrm{C}})} = \frac{(2q_1q_2)(r)}{(2q_1q_2)(q_1)} = \frac{r}{q_1}. \tag{4.13}$$

Most importantly, we now find ourselves examining the question of whether crime scene materials are definite predictors of some property of the offender, or whether some doubt exists regarding this connection. When such doubt exists, it must be incorporated into the likelihood ratio.

Generally speaking, when there are discrete, multiple occurrences of one form of evidence (such as blood types or shoeprints), we need to estimate the probability that the matching crime scene material does, in fact, represent traits of the offender. As we become less certain that the crime scene material is actually from the offender, the evidential value decreases according to (4.13).

4.5.5.2 Relevance of crime scene materials

During collection of crime scene materials investigators seek traces that may have come from the offender. Often these are transferred materials in their particle form. Depending on the type of material, its location, condition, and abundance the chances that it has a true connection with the offender may range from very likely to practically nil. A true connection means simply that the trace material came from the offender and is, therefore, a valid predictor of some traits of the offender. Crime scene material that has this true connection is said to be *relevant* in that it is relevant to the evaluation of suspects as possible offenders.

Relevance can be supported in many ways, but it generally arises by relating the material to the known or probable actions of the offender at the crime scene. A crime transpires through certain actions occurring at a certain time. When the types of material found at the crime scene, their location, condition, or abundance can be reliably attributed to the actions of the offender, then relevance can be reasonably assured. To illustrate how this connection can be established it is useful to consider the variety of common transfer evidence types shown in Table 4.3.

For some types of transferred material, such as blood or semen, there can be a clear connection with the crime. Bleeding or ejaculation results from a specific event. The location and condition of the fluids can make their connection to the crime virtually certain. Glass and paint fragments also result from a specific event (breakage) which can often be associated in time and place with the crime event. Other types of transfer evidence, such as hairs and fibres, can be deposited (shed) through more passive processes. Still, the abundance or location of these materials may allow a reasonably certain connection to be inferred. Examples are when pubic hairs are found in a rape victim's bed, or when abundant fibres of one type are found on a victim when a struggle must have occurred.

Occasionally relevance can be shown by more novel means. An example is the fibre evidence found in one series of murder cases (Deadman, 1984a,b). Here, the comparison of fibres from a series of cases established that some fibre types repeatedly occurred. These fibres could then be assumed to be relevant and therefore become valid predictors of fibre sources in the offender's environment.

When there are no compelling circumstances indicating the relevance of crime scene materials, extreme caution is necessary in interpreting correspondence, and, as shown above, this uncertainty of relevance must be factored into the likelihood ratio. This evaluation may become quite complex. Suppose, for example, that vacuum sweepings are taken from a crime scene in order to look for fibres that match a particular suspect's clothing. Of the great many fibres collected, suppose several fibre types match. There is a tendency for analysts to focus only on these matching fibres when interpreting the transfer evidence, rather than to consider the full range of potentially incriminating fibre types. Essentially, the fibres *that match the particular suspect* are assumed to be relevant simply because a match has been made with this particular suspect's clothing. The point is often lost that with another suspect, or a randomly selected person, a different set of fibres from the sweepings could be incriminating. It is the overall potential for random incrimination that needs to be evaluated—as conceptually and practically difficult as this may be.

4.5.6 Multiple suspect-related materials

Laying aside single-valued characteristics of individuals, such as blood types, we now consider forms of transfer evidence where one individual or source can show variety in the expression of the transferred particle form. There are three general circumstances that will be discussed:

□ heterogeneous suspect-related materials,
□ multiple discrete suspect-related bulk sources,
□ uncertainties regarding the properties of a suspect-related bulk source.

4.5.6.1 Heterogeneous suspect-related materials

Heterogeneous suspect-related materials can be in either the bulk-source form or the transferred-particle form. A person's hair is a good example of a heterogeneous bulk source in suspect-related material. On the person's (suspect's) head there is great variety in the characteristics shown by single hairs. Individual hairs found at a crime scene may or may not fit into the range of hair varieties seen in a particular suspect. If they do fit into this range, they may still be of a rare or common variety in relation to the variation found on the suspect's head. There will also be an incidence, which may be rare or common, among the population of shed hairs. Application of the likelihood ratio in this instance is straightforward. If we assume the relevance of the crime scene hair, the numerator becomes the probability that a hair with the characteristics of the crime scene hair would be shed by the suspect. A probability model might estimate the probability that a hair, drawn at random from a suspect's head, would have these characteristics, to be one in 50. (That is, one in 50 of the suspect's head hairs would be expected to show the characteristics seen in the crime scene hair.) The denominator of the likelihood ratio is the probability that the crime scene hair characteristics would be encountered randomly.

This hair example is one where the heterogeneous suspect-related material is a bulk source. A good transferred-particle example is provided by a hypothetical case proposed by Aitken (1987b) and by Buckleton & Evett (1991). In Aitken's case a burglary occurs where access to the property is gained through breaking a window. A suspect is developed who is employed as a glazier and on whose clothing are found a large number of glass fragments. Some portion of these fragments do match the crime scene window. This particular suspect, because of the large number of glass fragments, has an abnormally high probability of being associated with (any particular) broken window.

Buckleton & Evett's (1991) case is an assault involving close contact between a victim and the assailant. The victim is wearing clothing composed of two fibre types. A suspect is later developed, wearing clothing as described by the victim. On his clothing are found twenty transferred-fibre types, including the two types that comprise the victim's clothing. In this case the multiple types of fibres on the suspect's clothing represent additional chances for random matching of crime scene materials. Buckleton & Evett derive a likelihood ratio for this case by using the following quotient to account for the multiplicity of fibre types.

$$\frac{P_{18}}{{}^{20}P_2 \times P_{20}}.$$

The numerator is the probability that a person from the population will have eighteen different transferred-fibre types on him. These represent the non-matching fibres in the case circumstances, under the hypothesis that the suspect is the offender. The denominator is the number of ways of choosing two items from twenty while paying attention to the order $\{{}^{20}P_2 = 20!/18!\}$ multiplied by the probability that a person from the population will have twenty different groups of foreign fibres on him. A full discussion is given in Buckleton & Evett (1991).

4.5.6.2 Multiple discrete suspect-related bulk sources

The second general circumstance where multiple suspect-related materials are encountered is where one individual possesses a variety of objects, any one of which could have contributed particulate materials to the crime scene. A suspect's wardrobe is a good example. If a suspect is apprehended very shortly after a crime has occurred, such that no change of clothing or visit home is reasonable, we need not consider the suspect's wardrobe clothing as possible sources of direct fibre transfer. If, however, the timing of suspect apprehension allows for a change of clothing it may be necessary to select garments from the suspect's wardrobe. Based on prior examination of crime scene materials, there may be a specific fibre type that is being sought (for example, a jumper composed of yellow polyester fibres), or there may be a much more crude effort to cross-match a large set of shed fibres from the scene with any textile in the suspect's environment.

Two extreme approaches to the search of the suspect's wardrobe demonstrate that how the transfer evidence is developed is of critical importance in assessing its evidential value. Suppose first that a particular fibre type is sought among the suspect's clothing and that a match is found. If we assume (a) that the fibre found at the crime scene is relevant (that is, it came from the offender who was wearing a source of these fibres) and further assume (b) that if the true offender's wardrobe were searched the source would always be found (that is, that the item of clothing would not be discarded), then the numerator of the likelihood ratio is one: we would find a match of this fibre type with probability 1 under the hypothesis C that the suspect was the offender. If there is doubt about either of these assumptions the numerator is less than one and the evidential value is decreased accordingly. The denominator of the likelihood ratio is the probability that, if a random person's wardrobe were searched for this fibre type, a matching source would be found. This probability will depend on the frequency of alternative sources for these particular fibres among a population of *individuals.*

Now suppose, secondly, that the fibre match is discovered after an extensive comparison beween a large set of individual shed fibres from the crime scene with the full wardrobe of the suspect. A fibre match now has considerably more opportunity to occur by chance alone. First we must consider the question of relevance. We cannot assume that all shed fibres collected from the crime scene are relevant. We can reasonably assume that a few of the fibre types might be relevant, giving a probability *r* for any particular fibre, and that, under C, the true offender would definitely match (that is, with probability one) one of the fibre types from the scene. This gives a value of *r* for the numerator of the likelihood ratio. In the denominator we must, analogous to Buckleton & Evett's fibre case, account for the variety of suspect-related materials that could result in a match. Each of the *n items of clothing* in the suspect's wardrobe represents a chance for a random match with the fibre types from the scene.

4.5.6.2 Uncertainties regarding the properties of a suspect-related bulk source

The third circumstance where multiple suspect-related materials occur is where the properties of the suspect-related source are open to some doubt. Consider the following simple example. In a hit-and-run case a suspect is known to have owned a particular make and model of car, but the car itself has since been destroyed and is now unavailable for sampling. The particular make and model that the suspect owned was painted with two different formulations during the model production, one quite distinct from that found at the crime scene and the other identical to it. There is now uncertainty regarding the actual paint type that was present on the suspect's car. This uncertainty must then be reflected in the numerator of the likelihood ratio.

4.5.7 Special issues arising in crime scene to offender transfers

A few of the examples in the previous sections have involved transfer from bulk sources at the crime scene to the offender. In this section we will look a bit more closely at these situations. The probabilities of encountering trace evidence materials on clothing taken at random are of primary interest. How often, for example, is blood or glass encountered on clothing if such clothing is collected at random? How, if at all, should this information be incorporated into an assessment of evidential value? The following example will serve as an illustration (Evett 1990).

A person has been killed by stabbing, and, as a result of information received, the investigator has a suspect. Scientific examination of the suspect's jacket reveals bloodstaining of type y_1. The victim is type y_1 but the suspect is some other type. Assume the offender is known to be a man. There are two alternatives:

C: the suspect is the offender, the man who stabbed the victim,

$\bar{\mathrm{C}}$: the suspect is not the offender.

The forensic science evidence can be summarized as follows:

F_1: the victim is type y_1.

F_2: the suspect's jacket is bloodstained with type y_1.

(Note that, as in (4.5), F_1 is still used for findings associated with the scene and F_2 for findings associated with the suspect.)

The multiplication law for conditional probabilities can be applied in order to evaluate the likelihood ratio, though it is convenient to express the laws in a different fashion from (4.5):

$$\frac{P(\mathrm{F}|\mathrm{C})}{P(\mathrm{F}|\bar{\mathrm{C}})} = \frac{P(\mathrm{F}_1|\mathrm{C})}{P(\mathrm{F}_1|\bar{\mathrm{C}})} \times \frac{P(\mathrm{F}_2|\mathrm{F}_1,\mathrm{C})}{P(\mathrm{F}_2|\mathrm{F}_1,\bar{\mathrm{C}})}. \tag{4.14}$$

This form is more suitable because it is straightforward to argue that $P(\mathrm{F}_1|\mathrm{C}) = P(\mathrm{F}_1|\bar{\mathrm{C}})$; the victim's blood type is quite independent of whether or not this particular suspect did the stabbing.

In the denominator of the second ratio, if $\bar{\mathrm{C}}$ is the case then the findings on the suspect are quite independent of the victim's blood type. The likelihood ratio $\{P(\mathrm{F}|\mathrm{C})/P(\mathrm{F}|\bar{\mathrm{C}})\}$ (denoted by LR) then becomes:

$$\mathrm{LR} = \frac{P(\mathrm{F}_2|\mathrm{F}_1,\mathrm{C})}{P(\mathrm{F}_2|\bar{\mathrm{C}})} \;. \tag{4.15}$$

The evaluation of the denominator will depend on whether or not the suspect offers an explanation for the presence of the bloodstaining. If we assume that no explanation is offered then we can take the view that, as far as the blood evidence is concerned, the suspect is a person selected at random from the population. This would not be the case if, for example, one of the reasons for picking out the suspect was that he had a bloodstained jacket: furthermore, if the suspect had an occupation or lifestyle which meant that he was more likely than the average person to acquire human bloodstains then a different view would be needed. However, we assume that neither of these alternatives applies in the present example. It is necessary to evaluate two probabilities:

- the probability that a man selected at random would have bloodstaining of a type different from his own on his jacket (let this probability be designated b),
- the probability that bloodstaining, found on the clothing of a man selected at random who is not himself $y_{1,}$ would be of type y_1 (let this probability be designated q'_1). In other words, this is the probability that bloodstaining of type y_1 would be found, given that bloodstaining of a type different from that of the selected man had been found.

The terminology takes no account of the extent of the staining, an important aspect of the evidence to which we will return later. The notation q'_1 is used to emphasize that it is necessary in this context to consider a different probability distribution from simply the random bloodgroup frequencies: whereas q'_1 might be close in value to q_1, q_1 and q'_1 are not the same, an issue that has been considered by Gettinby (1984). The multiplication law for probabilities then gives:

$$P(\mathrm{F}_2|\overline{\mathrm{C}}) = b \times q'_1 . \tag{4.16}$$

The evaluation of the numerator is a little more complicated because, strictly speaking, there are two alternatives for which we must cater:

Either: no blood was transferred, and the suspect already had bloodstaining of type y_1 on his jacket,
Or: blood was transferred, no bloodstaining having previously been present.

For simplicity it is assumed that complicated alternatives such as a person having bloodstaining from two or more sources on his clothing have negligible probabilities. Let *t* denote the probability that blood is transferred from the victim to the assailant. Then the numerator of the likelihood ratio is derived by adding together the probabilities of the above two mutually exclusive events:

$$P(\mathrm{F}_2|\mathrm{F}_1,\mathrm{C}) = \{(1-t) \times b \times q'_1\} + \{t \times (1-b)\} . \tag{4.17}$$

So, from (4.15) and (4.16):

$$\frac{P(\mathrm{F}|\mathrm{C})}{P(\mathrm{F}|\overline{\mathrm{C}})} = (1-t) + \frac{t(1-b)}{(b \times q'_1)} . \tag{4.18}$$

When the likelihood ratio $\{t(1 - b)/(b \times q'_1)\}$ is appreciably greater than one this expression can be simplified by approximation to:

$$\frac{P(\mathrm{F}|\mathrm{C})}{p(\mathrm{F}|\overline{\mathrm{C}})} = \frac{t(1-b)}{(b \times q'_1)} . \tag{4.19}$$

This evaluation is a simplification, and readers interested in a rigorous approach are referred to a more detailed treatment of such transfer problems by Evett (1984).

Briggs (1978) has described an examination of the clothing of 122 men, 4 of whom had bloodstains which were not their own blood. Three of the men had clothing which was lightly bloodstained, the fourth had clothing which was heavily bloodstained. Although the sample could not be considered to be representative of the population of men in general, Briggs argued convincingly that the nature of the sampled population is such that these figures are probably overestimates of the probabilities of finding such staining on a man selected at random from the general population.

Envisage a hypothetical case in which a person has been killed by stabbing. The victim is bloodtype AB. A suspect has been found and his jacket is found to be heavily stained with blood of type AB—the suspect himself is not type AB. We can estimate the variables b, q'_1, and t as follows. Briggs' study indicates that 1/122 is a reasonable estimate for b. In the absence of published information the distribution of ABO types among stains on clothing, but bearing in mind the relative frequency of AB among individuals, we can take 0.1 as a conservative value. Estimation of t requires the skilled judgement of the expert. Assume that, as a result of an appraisal of the crime, the expert makes a conservative estimate of 0.5 for t, the probability that blood would be transferred to the offender at the crime scene. Given these values, the likelihood ratio (4.19) is somewhere in excess of 600.

This analysis has been possible only because of Briggs' data, because it has been assumed (a) that the distribution of blood types among stains on clothing is the same as that among individuals in the population and (b) that the expert would be prepared to estimate transfer probabilities. If background information of this kind is not available, or such assumptions are invalid, then it is necessary for the scientist to consider a pair of alternatives which are further removed from the ultimate deliberations of the court, but that generally involve much less speculation, such as:

S: the bloodstain came from the victim,

$\bar{S}$: the bloodstain came from some other person.

In this situation the likelihood ratio is simply $1/q_1$. Because the scientist does not take account of the presence of the bloodstain when deriving this likelihood ratio, this aspect of the evaluation must be considered separately by the court.

4.6 DISCUSSION

Computational methods and hypothetical cases have been presented. It remains to place these in a more realistic context and to discuss the limitations which will be imposed on any quantitative method when it is placed within such a subjective discipline as the assessment of evidence in a court of law. Two major topics will be discussed: the limitations on the value of transfer evidence within a case context, and the reality of evidential development in case investigations.

4.6.1 Limitations on the value of transfer evidence within the case context

Any correspondence of transfer evidence has a fundamental limitation on what it can prove within a case context. Suppose we know, with certainty, that the correspondence is ‘true’. This will prove some fact which could variously be the identity of a person, a contact between two objects, the time of a contact, or perhaps even a contact between two specific objects at a specific time. In any evaluation of transfer evidence it is critically important to recognize this limitation explicitly at the beginning of the evaluation (Gaudette 1986). Transfer evidence can prove identity without contact, contact without time, and time or contact without identity. Note, however, that, often, the abundance of similar mass-manufactured items remains as a limitation on proof of specific contact. Combined with these fundamental limitations are uncertainties that follow from limitations on analytical precision, analytical error, the lack of relevant data, and from a lack of complete interpretive methodology.

Analytical imprecision limits the ability to differentiate samples that are, in fact, different and from different sources. This means that we will find some correspondences that are not true. This is acceptable so long as we recognize that our methods have this limitation and we are cautious in the conclusions reached regarding alternative sources.

Analytical error, or simply incorrect analytical results, remains as a limitation on all laboratory work. Operator and instrument error must be minimized by proper training, quality control, and proficiency testing.

There is a lack of relevant data in the forensic sciences. In this chapter we have treated the subject as if databases existed allowing estimation of, for example, the relative frequencies of shed fibres on clothing, the percentage of hairs on a person’s head with a given appearance, and the frequency of encountering automobiles with specific types of paint. Although there are some databases available which are extremely valuable in assisting with the formation of expert opinion, they are not sufficiently specific to be applied in the ways presented in this chapter. (Some very good attempts at this have been made, however. See, for example, Evett *et al.* 1987.) The final column of Table 4.3, labelled ‘population options’, provides some insight into this difficulty. For fibre evidence each of the following databases of relative frequencies would be useful:

(i) specific textile or clothing items,
(ii) items that are sources of a particular fibre type,
(iii) locations or residences with either (i) or (ii),
(iv) persons with sources of (i), (ii), or (iii),
(v) individual fibre types in the loose, shed state,
(vi) same as (v) but on clothing, in automobiles, in residences, or in association with individuals.

These frequencies will certainly vary with time, geographical location, and the particular analytical methodology that is applied. Furthermore, much other information is needed to use these fully, including details of fibre transfer and loss, persistence of fibres on various substrates, and joint probabilities of occurrence of fibre blends.

It should be apparent from the above that a comprehensive quantitative treatment of fibre transfer evidence will be elusive because there cannot be a completely reliable and relevant database. This does not mean, however, that a comprehensive understanding and valid assessment of evidential value cannot be attained. Rather, it requires that we appreciate the limitations of our quantitative methods and blend these with the subjective realities and limitations of casework.

The final limitation on the interpretation of transfer evidence is the lack of a complete interpretative methodology. Even if we had ideal population data, our ability to put these data to their purpose would be elusive. The simple methods presented in this chapter, though representing encouraging progress and better understanding, cannot effectively incorporate the range of issues that arise in actual casework as transfer evidence is discovered. The discussion which follows conveys some aspects of this difficulty.

4.6.2 The reality of evidential development in case investigations

After materials are analysed by a crime laboratory, transfer evidence emerges as a pair of items that show a correspondence. Often the evaluation focuses rather superficially on the rarity of such a correspondence by chance. This is a *post hoc* view of transfer evidence which can ignore essential features of how the evidence came to be or how it was previously used. The key element of transfer evidence is not this *post hoc* correspondence in properties, but rather the ability to infer contact. This inference requires that the correspondence be put into perspective. In this section we will examine the process that results in the production of transfer evidence, beginning with the investigator's crime scene examination.

At the crime scene transfer evidence is gathered while asking four questions which help guide the search:

□ where was there contact between the offender and the scene?
□ what was the nature of this contact?
□ a what transfers can be expected from the offender to the scene?
□ what transfers can be expected from the scene to the offender?

Based on consideration of these questions the investigator will collect materials that are likely to have been left at the scene by the offender along with samples of material that may have been transferred to the offender.

The materials initially collected at the scene are those which are first available to incriminate a suspect. There also exists a potential for incriminating evidence against a person selected at random. Of particular interest in this set, therefore, are materials that definitely have been transferred to the scene by the offender and that can be used unambiguously to exclude suspects who lack the bulk-source form of the material. These materials then have exclusionary value. (Pubic hair transfers from the offender in a rape case are an example.)

Along with other investigative leads, the materials from the crime scene that predict features of the offender are often used to help locate or screen possible suspects. The use of evidence for investigative screening of suspects is in conflict with its subsequent use to evaluate the suspect. That is, if we define and select a suspect based on the existence of a set of properties, the existence of this set of properties can no longer be viewed as an independent event. The evidence is far more meaningful if it can be used to evaluate a suspect after he is selected by other criteria. When used to screen suspects, transfer evidence essentially becomes part of the prior probability referred to in section 5, rather than part of the likelihood ratio itself.

Once a suspect is selected (by whatever criteria) he can then be evaluated further by looking for more detailed transfer evidence. A set of questions similar to those for the crime scene is useful as a guide for the collection of this evidence:

- □ is there a source of any materials that appeared to be left at the crime scene by the offender?
- □ do any materials from the crime scene appear to have been transferred to the suspect?
- □ are there any materials not previously recognized that may have been transferred from this suspect to the crime scene?

In response to the last question the investigator may need to visit the crime scene again to collect newly-recognized material that may have been transferred from the suspect to the crime scene. Depending on the case circumstances and investigative priorities a more or less intensive search for corresponding transfer evidence may be made. As we have seen, the extent of this search directly influences the construction of the likelihood ratio.

As the investigation proceeds two other types of samples gain in importance: control samples and alibi samples. Control samples are necessary whenever transfer evidence is used to support an inference of contact. Simply put, one must verify that the crime scene itself is not a possible source of materials believed to have been left by the offender, and conversely one must verify that the suspect’s own environment is not a possible source for materials supposedly coming from the crime scene. Alibi samples are very similar, but come about through a suspect’s alternative explanation for the origin of transferred materials that apparently link him to the crime. Procurement and comparison of control and alibi samples can modify evidential value considerably. An alternative control source can make otherwise strong transfer evidence worthless, while a false alibi explanation can make a stronger case for guilt.

The evidence development process outlined above is interactive. A selection of evidence is initially made from the crime scene on the basis that it is believed to be relevant. Subsequent developments and control tests modify this. Development of a particular suspect suggests other types of transfer and leads to re-evaluation and perhaps re-examination of the crime scene. Subsequent interviews and analysis modify the case circumstances further.

This subjective interaction and evaluation during the development of a case cannot be ignored when evaluating resulting transfer evidence that supports contact, and this limits the value of a generalized treatment. One possible approach to these difficulties is the use of influence diagrams as proposed by Aitken & Gammerman (1989).

4.7 SUMMARY

Consideration of quantitative approaches to transfer evidence has revealed a number of important aspects that will be summarized in this section.

(i) *Need for a broad case perspective*
There is a need to view transfer evidence within a broad case context so that information bearing on its interpretation will be recognized and incorporated into the analysis. Most importantly it is inappropriate and misleading to consider only that small portion of trace materials that result in incriminating evidence. The full range of materials collected and examined must be known so that the effects of multiple crime scene and suspect-related materials will not be overlooked.

(ii) *Need to focus on the inference of contact*
The ultimate goal of transfer evidence is the inference of contact between two specific objects at a specific time. Concentration on this goal, as opposed to that of finding a match, emphasizes the limitations of our evidence and allows a proper assessment of its role in proving contact. This perspective demands an assessment of the relevance of crime scene materials to this contact as reflected in their form, location, condition, and abundance.

(iii) *Databases needed for transfer evidence evaluation*
For the evaluation of transfer evidence databases must be developed which apply individually to bulk-source and transferred-particle form of materials. Furthermore, each of these must be evaluated for specific environments in which the materials are found.

(iv) *Limited application of match frequencies*
Match frequencies have limited application in evidential evaluation and considerable potential for erroneous application. They are valuable for assessing the efficiency of analytical methods in the differentiation of samples and for describing the average probabilities for evidential types that defy more precise definition of individual sample rarity.

(v) *Construction of correspondence frequencies*
Proper construction of correspondence frequencies requires a scene-anchored approach that takes the crime scene materials as given and seeks the frequency of correspondence of suspect-related materials. These corresponding materials must be recognized to include forms and varieties other than the specific type found in connection with the suspect.

(vi) *Limitation of correspondence frequencies*
Correspondence frequencies cannot fully evaluate transfer evidence when uncertainty exists in either the characteristics that would be transferred to the offender or whether such transfer would, in fact, occur. In these circumstances Bayesian methods must be used.

(vii) *Role of Bayes' likelihood ratio*
The Bayes' likelihood ratio is a generalization of the more intuitive correspondence frequency approach to evidence evaluation. It provides a theoretical framework within which specific quantitative data find their proper place. This approach requires recognition of auxiliary issues, and, because we are unable to address many of these, the subjective

environment into which all quantitative evidential assessments must fit becomes more apparent.

REFERENCES

Aitken, C.G.G. (1987a) Average probabilities—in the absence of frequency data is there anything better? Presented at the 11th International Association of Forensic Sciences, Vancouver, Canada. *Journal of the Canadian Society of Forensic Science* **20** 287.

Aitken, C.G.G. (1987b) Personal communication.

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

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

Barnett, P.D. & Ogle, R.R. (1982) Probabilities and human hair comparison. *Journal of Forensic Sciences* **27** 272–278.

Briggs, T.J. (1978) The probative value of bloodstains on clothing. *Medicine, Science and the Law* **18** 79–83.

Buckleton, J.S. & Evett, I.W. (1991) Aspects of the Bayesian interpretation of fibre evidence. *Journal of the Forensic Science Society* (in press).

Davis, R.J. & DeHaan, J.D. (1977) A survey of men's footwear. *Journal of the Forensic Science Society* **17** 271–285.

Deadman, H.A. (1984a) Fiber evidence in the Wayne Williams trial (part 1), *FBI Law Enforcement Bulletin,* March, 13–20.

Deadman, H.A. (1984b) Fiber evidence in the Wayne Williams trial (conclusion), *FBI Law Enforcement Bulletin,* May, 10–19.

Evett, I.W. (1984) A quantitative theory for interpreting transfer evidence in criminal cases. *Applied Statistics* **33** 25–32.

Evett, I.W. (1986) A Bayesian approach to the problem of interpreting glass evidence in forensic casework. *Journal of the Forensic Science Society* **26** 3–18.

Evett, I.W. (1987) On meaningful questions: A two-trace problem. *Journal of the Forensic Science Society* **27** 375–381.

Evett, I.W. (1990) The theory of intepreting scientific transfer evidence. In: Maehley, A. & Williams, R.L. (eds.) *Forensic Science Progress* **4**. Springer Verlag, Berlin, 141–180.

Evett, I.W., Cage, P.E., & Aitken, C.G.G. (1987) Evaluation of the likelihood ratio for fibre transfer evidence in criminal cases. *Applied Statistics* **36** 174–180.

Fienberg, S.E. & Schervish, M.J. (1986) The relevance of Bayesian inference for the presentation of evidence and for legal decisionmaking. *Boston University Law Review* **66** 771–789.

Finkelstein, M.O. & Fairley, W.B. (1970) A Bayesian approach to identification evidence. *Harvard Law Review* **83** 489.

Fong, W. & Inami, S.H. (1986) Results of a study to determine the probability of chance match occurrences between fibers known to be from different sources. *Journal of Forensic Sciences* **31** 65–72.

Gaudette, B.D. (1978) Some further thoughts on probabilities and human hair comparisons. *Journal of Forensic Sciences* **23** 758–763.

Gaudette, B.D. (1982) A supplementary discussion of probabilities and human hair comparisons. *Journal of Forensic Sciences* **27** 279–289.

Gaudette, B.D. (1986) Evaluation of associative physical evidence. *Journal of the Forensic Science Society* **26** 163–167.

Gaudette, B.D. & Keeping, E.S. (1974) An attempt at determining probabilities in human scalp hair comparisons. *Journal of Forensic Sciences* **19** 599–606.

Gettinby, G. (1984) An empirical approach to estimating the probability of innocently acquiring bloodstains of different ABO groups on clothing. *Journal of the Forensic Science Society* **24** 221–227.

Harrison, P.H., Lambert, J.A., & Zoro, J.A. (1985) A survey of glass fragments recovered from clothing of persons suspected of involvement in crime. *Forensic Science International* **27** 171–187.

Jones, D.A. (1972) Blood samples: Probability of discrimination. *Journal of the Forensic Science Society* **12** 355–359.

Kuo, M.C. (1982) Linking a bloodstain to a missing person by genetic inheritance. *Journal of Forensic Sciences* **27** 438–444.

Lander, E.S. (1989) DNA fingerprinting on trial. *Nature* **339** 501–505.

Lindley, D.V. (1977) A problem in forensic science. *Biometrika* **64** 207–213.

Ogino, C. & Gregonis, D.J. (1981) Indirect blood typing of a victim's blood using parentage testing. Presentation before the *California Association of Criminalists, 57th Semi-annual Seminar*, *Pasadena, California, USA*.

Owen, G.W. & Smalldon, K.W. (1975) Blood and semen stains on outer clothing and shoes not related to crime: Report of a survey using presumptive tests. *Journal of Forensic Sciences* **20** 391–403.

Pearson, E.F., May, R.W., & Dabbs, M.D.G. (1971) Glass and paint fragments found in men's outer clothing. Report of a survey. *Journal of Forensic Sciences* **16** 283–300.

Smalldon, K.W. & Moffat, A.C. (1973) The calculation of discriminating power for a series of correlated attributes. *Journal of the Forensic Science Society* **13** 291–295.

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

Stoney, D.A. (1984b) Statistics applicable to the inference of a victim's bloodtype from familial testing. *Journal of the Forensic Science Society* **24** 9–22.

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

Walsh, K.A.J. & Buckleton, J.S. (1986) On the problem of assessing the evidential value of glass fragments embedded in footware. *Journal of the Forensic Science Society* **26** 55–60.

5

Applications of statistics to particular areas of forensic science

Edited by G.Gettinby, Department of Statistics and Modelling Science, University of Strathclyde, Glasgow, UK

This chapter presents contributions by various authors on the implementation of statistical methods to problems in forensic science. With the exception of the final discussion on the use of post mortem data to predict time of death, the contributions examine the interpretation of transfer evidence in courts of law. Such evidence frequently arises in the search for proof of innocence or guilt in criminal investigations where materials such as blood, soil, or hair have been transmitted to or from a crime locus. Transfer problems also relate to claims of paternity and to the discharge of firearms.

The first contribution presents a gentle discussion on evidence associated with blood transfer and establishes many of the principles commonly encountered with the interpretation of forensic evidence. A gentle introduction to the interpretation of DNA profiles, commonly measured from samples of blood, continues the discussion of blood evidence, and this is more specifically examined in the discussion of paternity evidence. After blood, the analysis of soil data is probably the most rigorous in statistical terms. Difficult and controversial problems arise in the interpretation of hair and ballistics data, largely because the data sets are small and the information they contain subjective.

The future for the application of statistical models within forensic science clearly lies toward more extensive collection and analysis of relevant data, and educational and delivery systems which will enable the courts to implement findings. Computers will greatly assist with these tasks, but there will still remain the need for a greater consensus among experts if statistical models are not to be rebuked or found unsafe.

EVIDENCE FROM BLOOD

G.A.F.Seber†

With the current technological advances in serology (the analysis of sera), blood can now be used more effectively as evidence in criminal proceedings and paternity suits.

† Department of Mathematics and Statistics, Auckland University, Auckland, New Zealand

Human blood is made up of about 60% fluid plasma and 40% solid substances. The solids are mainly red blood cells (erythrocytes), platelets (thrombocytes), and white blood cells (leukocytes). There are a large number of substances in the blood which are determined genetically and which are generally independent (Grunbaum *et al.* 1978). By independence we mean that the occurrence of a particular gene in one substance is statistically independent of the occurrence or non-occurrence of a particular gene in another substance. These substances can be grouped under four headings. Firstly there are the red cell antigens on the surface of the red cell membrane. These form part of the immunological system and examples are the ABO, MNSs, Rh (Rhesus), Kidd, Duffy, and Kell systems. The second group are the red cell enzymes such as PGM (phosphoglucomatase) and EAP (erythrocyte acid phosphotase, also abbreviated as acP). The third groups are the serum proteins and include Hp (haptoglobin), Gc (group specific component), GPT (glutamate pyruvate transaminase), and the immunoglobins Gm and Km. The fourth, more recently discovered system, is the white cell antigens or HLA (human leukocyte antigen) system.

To demonstrate some of the basic genetic ideas we consider the ABO system as an example. This system has three genes or alleles which we denote by a, b, and o — some authors use capital letters, which can cause notational confusion. Each parent passes one ABO gene to their child, so that we have a pair of genes called a genotype. These take the form aa, bb, ao, bo etc. where the order of the letters does not matter. Thus ab means either a from the mother and b from the father, or vice versa. However, as a and b are dominant genes and o is recessive, the usual detection tests do not detect gene o, only the presence or absence of a and b. This means that we cannot distinguish between aa and ao, and what is detected is the so-called phenotype, A say. Here A means that a is present but b is absent and includes the genotypes aa and ao. Similarly, phenotype B includes both bb and bo. When a and b are both present, the phenotype AB is the same as the genotype ab, while if a and b are both absent, the phenotype O is the same as the genotype oo. With better detection procedures, the allele a can be subdivided into a_1 and a_2, with a_1 dominant with respect to both a_2 and o, and a_2 dominant with respect to o. The phenotype A_1 then includes the genotypes a_1a_1, a_1a_2 and a_1o, while phenotype A_2 includes the genotypes a_2a_2 and a_2o. Procedures for determining the phenotypes of the various systems are described in a legal context by Reisner & Bolk (1982). In general, the more alleles that can be detected in a given blood system, the more useful is the system in forensic work. Recently a technique called isoelectric focusing (IEF) has led to finding more alleles in some of the systems. For example, the two genes in the PGM system labelled as 1 and 2 have been further subdivided into + 1, -1, + 2, -2. Some technical details about the IEF method are given in Murch & Budowle (1986).

Referring again to the ABO system, we assume that the probabilities that a parent passes on one of the three genes, a, b, and o to a child are *p, q,* and *r* respectively, with $p + q + r = 1$. If the population is in equilibrium, which can be achieved by random mating (the so-called Hardy-Weinberg law), the proportions of the various phenotypes in the population can be readily calculated. For example, the phenotype A consists of genotypes aa and ao. The probability of genotype aa will be $p^{2,}$ the probabilities p for each parent multiplying as the father and mother can be regarded as independent donors of the gene a. Similarly, the probability of genotype ao will be $pr + rp$ or $2pr$ as either the father or the mother can contribute the a gene. The probability of getting phenotype A is then the sum of the probabilities of the two genotypes, namely $p^2 + 2pr$. Similarly the probability of phenotype AB is $2pq$. If we know the phenotypes of a large random sample of people from a population, we can then use certain statistical techniques to obtain estimates of the gene probabilities and, by substitution, estimates of the phenotype probabilities just described. Standard deviations (standard errors) of these phenotype estimates are also available (cf. Greenwood & Seber 1990).

We have mentioned just the ABO system. However, there are over 60 such systems known, but these are not all used for forensic investigations, for a number of reasons. First, for many of the systems, antisera for phenotype testing are not readily available to all laboratories. Only very large laboratories can afford to make their own reagents, so that most laboratories rely on commercial products. This raises the question of quality control and the accuracy of results. Second, even if all the antisera were available, the time and cost for doing all the tests would be prohibitive. Many of the tests require a high degree of technical skill so that there is the added problem of checking work. Third, the systems that can be used will depend on the quantity and the quality of the blood sample available. In paternity testing, blood samples are taken from the mother, child, and alleged father, and the phenotypes for each blood system are determined and compared. In this case about 12 to 20 systems are typically used. However, with bloodstains, there is a limited amount of blood available, and the blood will have begun to age. Blood systems applicable here are those (i) which can be readily extracted from objects such as clothing, (ii) which can be analysed using small samples, and (iii) whose phenotypes are fairly stable when allowed to age. About six or more systems are currently used for bloodstains though the list is growing as the technology improves. Aging studies play an important role here (e.g. Kwan & Pallos 1980).

With bloodstain evidence we have the typical situation where blood has been transferred to an object or another person. For example, there may be blood on the suspect's clothing which could have come from the victim, or blood on the victim which could have come from the suspect. In either case, the phenotypes for several blood systems are determined by using samples from the victim, bloodstain, and the suspect(s). If the bloodstain on the suspect does not match the victim, then the suspect is excluded by this evidence. On the other hand, if such a match occurs, then various probabilities of such an occurrence can be calculated. However, in associating probabilities with 'transfer' evidence in general, it is important that the right question is asked, otherwise conflicting probabilities can be calculated (e.g. Stoney 1984a). This point was discussed in detail by Stoney (1984b) who suggested that the proper question is 'Given the crime object, what is the probability of encountering a corresponding suspect object?' Here the crime object is that defined by the offence and the suspect object is that connected with the accused. Buckleton *et al.* (1987) identify two basic situations.

In situation A, the victim's blood is transferred to the offender or some object related to the offender so that the blood of the victim is the crime object and the bloodstain on the suspect is the suspect object. The question then becomes ‘Given the blood phenotypes of the crime object (victim’s blood), what is the probability of encountering the suspect object (a bloodstain with phenotypes corresponding to the victim’s blood)?’ In situation B it is the offender’s blood which is transferred to the victim or some object at the scene of the crime. The bloodstain made at the scene of the crime now becomes the crime object and the blood of the suspect becomes the suspect object. The question is again ‘Given the blood phenotypes of the crime object (bloodstain), what is the probability of encountering the suspect object (persons with blood phenotypes corresponding to the bloodstain)?’

Before looking at these two situations, we shall introduce some probabilities which play a central role in forensic science. The first is the probability Q of non-discrimination or matching (cf. Selvin *et al*. 1979, 1983). This is the probability that two randomly selected people have matching blood phenotypes. For example: let Q_{ABO} be this probability with respect to just the ABO system. Then

$$\begin{aligned} Q_{\mathrm{ABO}} &= P(\text{matching phenotypes}) \\ &= P(\text{both A}) + P(\text{both B}) + P(\text{both AB}) + P(\text{both O}) \\ &= P_{\mathrm{A}}^2 + P_{\mathrm{B}}^2 + P_{\mathrm{AB}}^2 + P_{\mathrm{O}}^2 \\ &= (p^2 + 2pr)^2 + (q^2 + 2qr)^2 + (2pq)^2 + r^4, \end{aligned} \tag{1}$$

which can be evaluated using estimates of p, q, and r. Here phenotypes of the two people multiply, for example $P(\text{both A}) = P_{\mathrm{A}}P_{\mathrm{A}}$, as the people may be regarded as being independent. If m blood systems are used for typing, and Qj is the probability of non-discrimination for the jth system, then, as the systems are statistically independent, the probability Q that all m systems match is the product

$$Q = Q_1 Q_2 \ldots Q_{\mathrm{m}}.$$

The probability 1-Q is called the probability of discrimination, and it is the probability that two randomly chosen individuals do not match for at least one phenotype in m phenotype comparisons. It is also called the discriminatory or discriminating power, see section 4.3.1.

We now show that Q_{ABO} can be regarded as the average probability that there is a match with the ABO system. Since $P_{\mathrm{A}} + P_{\mathrm{B}} + P_{\mathrm{AB}} + P_{\mathrm{O}}=1$, we can see from (1) that

$$\begin{aligned} Q_{\mathrm{ABO}} &= \frac{(NP_{\mathrm{A}})P_{\mathrm{A}} + \ldots + (NP_{\mathrm{O}})P_{\mathrm{O}}}{NP_{\mathrm{A}} + \ldots + NP_{\mathrm{O}}} \\ &= \frac{N_{\mathrm{A}}P_{\mathrm{A}} + \,.\,.\, + N_{\mathrm{O}}N_{\mathrm{O}}}{N_{\mathrm{A}} + \ldots + N_{\mathrm{O}}} \end{aligned}$$

where $N_{\mathrm{A}}(= NP_{\mathrm{A}})$ can be interpreted as the number of potential victims of phenotype A which a random suspect would match with probability P_{A}. Thus Q_{ABO} is the average probability of matching with respect to the whole population of potential victims. Although this probability will not help with individual cases (Aitken 1987 and section 2.6), it does assess how well the ABO system fares, on average. For example, Selvin *et al.* (1983) used gene probability estimates from Grunbaum *et al.* (1980) to calculate $Q_{\mathrm{ABO}} = 0.381$ and $Q_{\mathrm{A1A2BO}} = 0.320$. Since we want Q to be as small as possible, we see that by subdividing the allele a into a_1 and a_2 we get a substantial reduction in Q. In addition to seeing what happens when further alleles are introduced, we can also assess the effect on Q of adding a new blood group to the system. Some systems are substantially less useful in terms of excluding innocent individuals than others. Selvin gives $Q = 0.825$ for the Kell system and $Q = 0.3753$ for the Kidd system, so that the latter is more useful. It can be shown that the probability of non-discrimination for a k-allele system cannot be less than $d = (2k - 1)/k^3$ (Selvin *et al.* 1983). For a codominant system, that is one in which the alleles are equally dominant, the minimum occurs when the gene probabilities are all equal to 1/k. Blood systems with a Q value close to this minimum are good ones to use. For example, when $k = 2$, $d = 0.375$. Cost and time are also important factors in assessing the usefulness of a particular system (Laux & Wurster 1983).

One other probability of interest is the probability of concordance; also known as the probability of coincidence. This is the probability that a randomly selected person matches a given set of phenotypes. For example, if the given set of phenotypes for the ABO, PGM, and the Hp systems are A, PGM1, and Hp1 respectively, then a randomly selected person will have these phenotypes with probability

$$c = P_{\mathrm{A}} P_{\mathrm{PGM1}}\, P_{\mathrm{Hp1}}. \tag{2}$$

In general, if P_j is the phenotype probability for the jth system ($j = 1,2,\ldots, m$), then the probability of coincidence is

$$c = P_1\, P_2 \ldots\, P_m.$$

It is assumed in the above discussion that all phenotypes are correctly typed. However, there is always the possibility of laboratory error. Selvin & Grunbaum (1987) demonstrated that such errors have little effect on the probability of non-discrimination but can have a substantial and misleading effect on the probability of concordance. Fortunately most errors give the benefit of the doubt to the suspect in that blood groups which actually match tend to be declared as not matching.

Referring to situation A above, we can answer the appropriate question by finding the probability of getting a match (which we shall call event E, or the evidence) given the blood phenotypes of the victim and suspect. We consider this probability for the cases when the suspect is guilty (G) and not guilty ($\overline{\mathrm{G}}$). These probabilities will depend on the phenotypes of the suspect. If the suspect has the same phenotypes as the stain, then the stain could have come from the suspect. Now suppose that we have a large population of N people. Let Np be the number of bloodstains of which a proportion a are self stains and 1 - a are non-self stains. Then, if the chance of acquiring a bloodstain from any source does not depend on the phenotypes of the source, we can work out the number of bloodstains in the population which match those of the bloodstain evidence. There are two cases.

Firstly, if the suspect has different phenotypes from the stain, then of the Np$(1 - \alpha)$ people that have non-self stains, Np$(1 - \alpha)c$ will match: c is given by equation (2). Hence

$$P(\text{evidence}|\text{innocent}) = P(\text{E}|\overline{\text{G}}) \ .$$

Secondly, if the suspect has the same phenotypes, then the bloodstain will be either a self stain or a non-self stain, hence

$$\begin{aligned} P(\text{E}|\overline{\text{G}}) &= P(\text{self stain}) + P(\text{non-self matching stain}) \\ &= p\alpha + p(1 - \alpha)c. \end{aligned}$$

Gettinby (1984) discussed these probabilities for just the ABO system and expressed a in terms of other parameters. He gave an example where a could be estimated. For the case where the suspect is guilty, the above probabilities will depend on how sure we are that the victim's blood was transferred to the suspect. If we are sure of this evidence, then the probabilities are both 1.

In situation B, suppose we are sure that the blood at the scene of the crime was left by the person who committed the crime. The evidence E is again a match, but this time between the phenotypes of the bloodstain and those of the suspect. Then $P(\text{E}|\overline{\text{G}}) = c$, being the probability that a random person has matching phenotypes and $P(\text{E}|\text{G}) = 1$. We note the role played by the coincidence factor c in the above discussion. Usually it is the probability that is presented in court. This is appropriate in situation B but not in situation A. However, in situation A, if c is very small then $p(1 - \alpha)c$ will be even smaller.

What is the effect of having the blood evidence? Before the blood evidence is considered, suppose that the prior odds are

$$\frac{P(\text{guilty})}{P(\text{innocent})} = \frac{P(\text{G}|\text{I})}{P(\overline{\text{G}}|\text{I})} \ ,$$

where I is the information or evidence up to that stage. Then, after the blood evidence is given, the posterior odds become (Evett 1983)

$$\begin{aligned} \frac{P(\text{G}|\text{E,I})}{P(\overline{\text{G}}|\text{E,I})}) &= \frac{P(\text{G \& E}|\text{I})/P(\text{E}|\text{I})}{P(\overline{\text{G}}\text{ \& E}|\text{I})/P(\text{E}|\text{I})} \\ &= \frac{P(\text{E}|\text{G,I})\ P(\text{G}|\text{I})}{P(\text{E}|\overline{\text{G}}\text{,I})\ P(\overline{\text{G}}|\text{I})} \\ &= L\cdot\frac{P(\text{G}|\text{I})}{P(\overline{\text{G}}|\text{I})} \ , \text{say.} \end{aligned}$$

If the blood evidence E is independent of the previous evidence I, then I drops out of the conditional probabilities in L and, in situation B, for example, we have

$$L = \frac{P(\mathrm{E}|\mathrm{G})}{P(\mathrm{E}|\overline{\mathrm{G}})} = \frac{1}{c} ,$$

the factor which converts prior odds into posterior odds.

In concluding this section, it should be noted that care is needed in calculating c. If the race is known, then c must be calculated by using the gene estimates for that particular race. However, if the race is unknown, then the phenotypes must be calculated for a total (mixed race) population. This is not so easy. For a further discussion of this question see Walsh & Buckleton (1988).

References

Aitken, C.G.G. (1987) The use of statistics in forensic science, *Journal of the Forensic Science Society* **27** 113–115.

Buckleton, J.S., Walsh, K.A.J., Seber, G.A.F., & Woodfield, D.G. (1987) A stratified approach to the compilation of blood group frequency surveys, *Journal of the Forensic Science Society* **27** 103–112.

Evett, I.W. (1983) What is the probability that this blood came from that person? A meaningful question? *Journal of the Forensic Science Society* **23** 35–39.

Gettinby, G. (1984) An empirical approach to estimating the probability of innocently acquiring bloodstains of different ABO groups on clothing, *Journal of the Forensic Science Society* **24** 221–227.

Greenwood, S.R. & Seber, G.A.F. (1990) Standard errors of blood phenotype estimates. Submitted for publication.

Grunbaum, B.W., Selvin, S., Pace, N., & Black, D.M. (1978) Frequency distribution and discrimination probability of twelve protein genetic variants in human blood as functions of race, sex, and age, *Journal of Forensic Sciences* **23** 577–587.

Grunbaum, B.W., Selvin, S., Myhre, B.A. & Pace, N. (1980) Distribution of gene frequencies and discrimination probabilities for 22 human blood genetic systems in four racial groups, *Journal of Forensic Sciences* **25** 428–444.

Kwan, Q.Y. & Pallos, A. (1980) Detectability of selected genetic markers in dried blood on aging, *Journal of Forensic Sciences* **25** 479–498.

Laux, D.L. & Wurster, J.W. (1983) The relative indices of efficiency for selected methods of bloodstain analysis, *Journal of Forensic Sciences* **28** 1000–1003.

Murch, R.S. & Budowle, B. (1986) Applications of isoelectric focusing in forensic serology, *Journal of Forensic Sciences* **31** 869–880.

Reisner, E.G. & Bolk, T.A. (1982) A laymen's guide to the use of blood group analysis in paternity testing, *Journal of Family Law* **20** 657–675.

Selvin, S. & Grunbaum, B.W. (1987) Genetic marker determination in evidence blood stains: the effect of classification errors on probability of non-discrimination and probability of concordance, *Journal of the Forensic Science Society* **27** 57–63.

Selvin, S., Black, D.M., Grunbaum, B.W., & Pace, N. (1979) Racial classifications based on blood group protein systems, *Journal of Forensic Sciences* **24** 376–383.

Selvin, S., Grunbaum, B.W., & Myhre, B.A. (1983) The probability of non-discrimination or likelihood of guilt of an accused: Criminal identification, *Journal of the Forensic Science Society* **23** 27–33.

Stoney, D.A. (1984a) Statistics applicable to the inference of a victim's blood type from familial testing, *Journal of the Forensic Science Society* **24** 9–22.

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

Walsh, K.A.J. & Buckleton, J.S. (1988) A discussion of the law of mutual independence and its application of blood group frequency data, *Journal of the Forensic Science Society* **28** 95–98.

DNA FINGERPRINTING

R.N.Curnow†

DNA fingerprinting, more correctly DNA profiling, has rapidly become an important tool for the forensic scientist. The initial enthusiasm, following the discovery of highly variable numbers of repeated sequences in certain positions on human chromosomes (Jeffreys 1987), has been succeeded by more critical questioning of some of the biological and sampling assumptions underlying its use.

DNA profiling is used in the determination of paternity. Since paternity determination is discussed elsewhere in this chapter by Berry, I shall concentrate on its use in crimes of murder or rape where a sample, generally blood or semen, has been recovered from the scene of crime and is known to be from the person responsible. We can simplify the biological background underlying the use of multi-locus probes by saying that there are, at a number of positions on our chromosomes, a variable number of repeats of a particular relatively short sequence of DNA. Using a restriction enzyme that cuts the chromosomes at specific positions and a radioactive multi-locus DNA probe, these repeated sequences can be removed from a blood or semen sample and separated on an electrophoretic gel according to size, essentially the number of repeats, and X-ray filmed. The separated fragments, called bands, can be seen in Fig. 5.1.

The banding pattern for the scene of crime sample can then be compared with the pattern in a sample from a suspect. If the scene of crime sample contains a band not present in the sample from the suspect then, assuming no contamination of the scene of crime sample and no errors in the laboratory analyses, the suspect is clearly not the criminal. This is called exclusion. The advantage of DNA profiling over other methods is that we can often show that the chance of a random innocent individual showing the same degree of match as that between the suspect and the scene of crime sample is very small indeed, and therefore that the suspect is almost certainly the criminal. This process is called inclusion.

† Department of Applied Statistics, University of Reading, Reading, UK

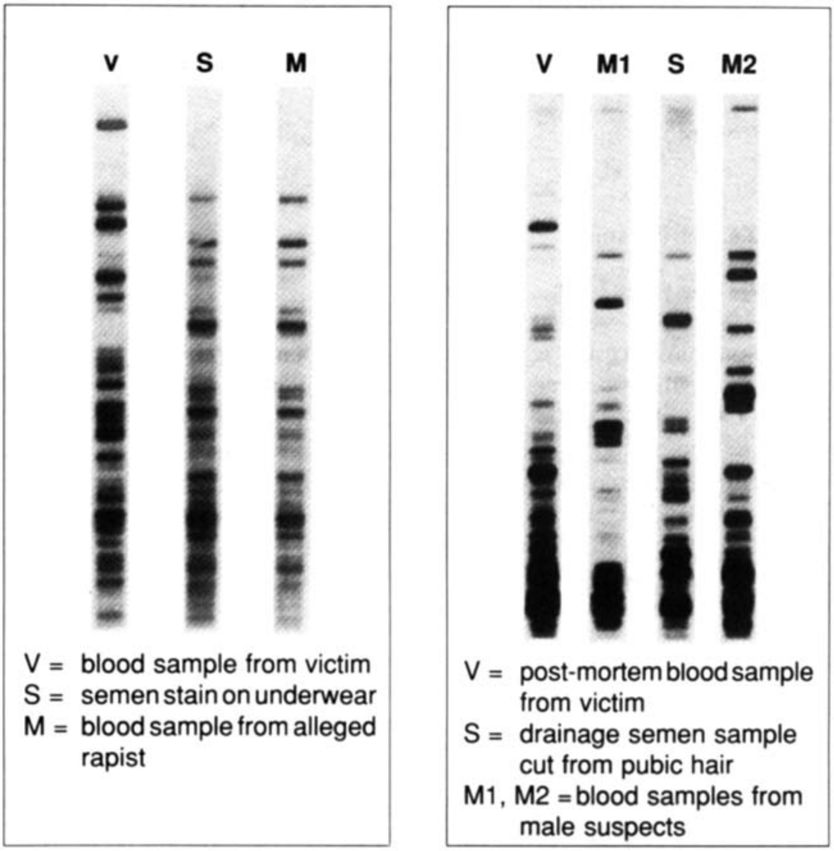

V
S
M
V = blood sample from victim
S = semen stain on underwear
M = blood sample from alleged rapist
V
M1
S
M2
V = post-mortem blood sample from victim
S = drainage semen sample cut from pubic hair
M1, M2 = blood samples from male suspects

Fig. 5.1 —DNA profiles.

Samples from a scene of crime are often in an unsatisfactory state for accurate analysis. The absence of a band may be significant, but, alternatively, may be due to the degraded nature of the sample. The number of observed bands in both the sample from the scene of crime and the sample from the suspect is therefore the main evidence, from the DNA profile, concerning the suspect's guilt. Suspects lacking any of the bands in a scene of crime sample will have been eliminated, and so the number of such shared bands will simply be the number of bands present in the sample from the scene of crime. To calculate the probability that a random innocent person has these bands we have to make a number of assumptions. Empirical evidence suggests that each position on the chromosomes generally provides at most one band within the window of the gel and that only one position will provide a band of given size. Further, there is evidence that there is very little association between the occurrences of the repeated sequences at different positions on an individual's chromosomes. In genetic terms they are in linkage equilibrium. The frequency with which any position provides a band on the gel has been found to be at most 0.26. We use this upper limit since it is conservative in the sense of being helpful to the suspect. With all these assumptions we can calculate the probability that a random person has a particular set of n bands as 0.26". Therefore the likelihood ratio

$$\frac{\text{Probability } (n \text{ shared bands}|\text{suspect is guilty})}{\text{Probability } (n \text{ shared bands}|\text{suspect is not guilty})} = \frac{1}{(0.26)^n} = 0.26^{-n} \;.$$

As an example, for ten shared bands,

$$0.26^{-10} = 1.4 \times 10^{-6} = 1/710000$$

This ratio is sometimes converted into a statement that the odds on guilt are 710000 to 1. This is true only in the very unlikely situation that the odds prior to the DNA profiling are 1 to 1, that is, that all the other information, prior and evidential, is neutral with respect to the guilt or otherwise of the suspect.

In the calculation above, the probability that the suspect has all n bands when he is not guilty takes no account of the total number of bands the suspect has. The probability should be conditioned on this number. If the total possible number that could appear in the window of the electrophoretic gel is N, then the probability that a random member of the population has the n particular bands, given he has a total of s bands, is

$$\frac{{}^{N-n}C_{s-n}}{{}^{N}C_{s}}$$

where ${}^{b}C_{a}$ is the combinatorial expression for $b!/(a!(b - a)!)$. The value of N, from empirical evidence, is about $N = 44$. The probabilities for random individuals with various total numbers of bands, s, now depend on the substituted value for N rather than on the assumed value for the frequency of occurrence of a band of 0.26, and are as follows for $N = 44$ and $n = 10$ shared bands:

s	Probability
10	4.0×10^{-9}
20	7.4×10^{-5}
30	1.2×10^{-2}
44	1

These probabilities should be used rather than the value 0.26^{10}. which is their average, averaging over a binomial distribution for s with $N = 44$ and $p = 0.26$.

Evett *et al.* (1989) have suggested that the degradation of the scene of crime sample could be modelled by assuming that the probability of any band present in the criminal being identified in the scene of crime sample takes some value m, depending on the state of the sample from the scene of crime. The likelihood ratio is then

$$\frac{m^n(1-m)^{s-n}}{(0.26m)^n(1-0.26m)^{N-n}} = \frac{0.26^{-n} \times (1-m)^{s-n}}{(1-0.26m)^{N-n}}$$

where *s* and *N* are as defined before. Evett *et al.* (1989) show that the likelihood ratio is increased by taking account of the suspect's bands not present in the scene of crime sample unless *m* is sufficiently large that the presence of bands in the suspect not in the scene of crime sample reduces the strength of the evidence against him. Clearly if degradation is thought to be slight and the suspect has a number of bands not in the scene of crime sample, then, in the interests of the suspect, this additional calculation with a number of possible values for the likely level of degradation should be made. The level of degradation does vary according to band size, with the smaller bands being less likely to degrade, so that any small bands in the suspect not in the scene of crime sample should be given careful consideration.

Multi-locus probes are now being replaced by single-locus probes that, as their name implies, determine the size of the bands at a single specific position on a chromosome. At such a position both chromosomes of the pair of chromosomes provide a band that will appear within the window of the electrophoretic gel, although they may both be of the same size, in which case only a single band will appear. The number of possible band sizes at the locus is very large, with each particular size being rare. Suspects can clearly be eliminated from the enquiry if their band size or band sizes are not the same as those in the sample from the scene of crime. There are always errors of measurement and variations in mobility both within and between gels. At the same time there may be a number of band sizes close to, but not the same as, those in the sample from the scene of the crime. Currently, somewhat arbitrary decisions, based on estimated standard deviations and correlations of errors, are made about whether an apparent match is a real match or not. Work is in progress on the more appropriate likelihood ratio approach in which the numerator is the probability density of the extent of the lack of an exact match of the two bands, assuming that this lack of match is solely due to normally distributed errors of measurement and variations in mobility with known variance and covariance values. The denominator is calculated as the probability of a random member of the population having two bands in the observed positions on the gel relative to those in the scene of crime sample. This probability is estimated by sampling pairs of band sizes from an appropriately smoothed empirical distribution of band size in the population, and then inserting these probabilities in a formula which allows, as in the numerator, for errors of measurement and variations in mobility. This approach should answer some of the criticisms raised after a recent court case (Lander 1989).

The most worrying assumption underlying both multi- and single-locus probes is that of a single population with a single value for the frequency of each band, and, for the multi-locus probes, a lack of association between the occurrences of different band sizes in the same individual. Geographically isolated populations and ethnic minority populations may differ, through natural selection and genetic drift, in both the frequency of the different band sizes and also in associations between band size occurrences. There is a need to analyse all the data that are now becoming available and to organize special studies of small populations to assess the importance of these assumptions.

In summary, DNA profiling provides a method that is much more powerful than previous typing systems based on, for example the ABO blood groups. Even with the typing of eight red cell enzyme systems as well as the ABO system, the average probability of being able to exclude an innocent person is only 0.995 (Sensabaugh 1981) compared with a probability of 1 - 0.26^{-10}if a multi-locus probe reveals ten bands in the scene of crime sample. The main gain, however, is that, through the relatively greater variety of particular band sizes or combinations of band sizes, DNA profiling leads to very low probabilities of a profile match unless the suspect is in fact guilty. The average value of this probability using the eight red cell enzymes systems and the ABO blood group is 0.014, compared with 0.26^{-10} from a multi-locus probe identifying ten bands in the scene of crime sample. Further work is needed to improve the analysis and interpretation of the single-locus profiles and also to check the numerous assumptions made, particularly that we are sampling from a single population and so the frequencies of the various band sizes and the associations between them can be calculated from amalgamated data and then applied to the profiles of individuals from any section of the population.

References

Evett, I.W., Werrett, D.J., & Smith, A.M.F. (1989) Probabilistic analysis of DNA profiles, *Journal of the Forensic Science Society* **2** 191–196.

Jeffreys, A.J. (1987) Highly variable minisatellites and DNA fingerprints, *Bioche-mical Society Transactions* **15** 309–317.

Lander, E.S. (1989) DNA fingerprinting on trial, *Nature* **339** 501–505.

Sensabaugh, G.F. (1981) Uses of polymorphic red cell enzymes in forensic sciences, *Clinics in Haematology* **10** 185–207.

PROBABILITY OF PATERNITY

Donald A.Berry†

Cases of disputed paternity are frequently resolved by using genetic testing. Sometimes these cases are decided by courts, but increasingly they are settled out of court once the genetic testing results become known. The current standard genetic system in the United States is human leukocyte antigen (HLA). Many laboratories also use red cell and other blood groups, and not all use HLA. Some labs (Gjertson *et al.* 1988, Berry 1991) now use DNA restriction fragment length polymorphisms (RFLPs) in paternity testing. Earlier in this chapter, in the section on DNA fingerprinting, RFLP was simply referred to as fingerprinting or profiling. The standard laboratory report includes a ‘probability of paternity’. The main purpose of this section is to explain how this probability is calculated. It will examine the assumptions required to obtain this probability, and describe how uncertainty might best be communicated to a court in a paternity case.

Consider ABO blood type as an example. Like other blood groups and HLA systems, this is a discrete system in that it has a fixed number of distinct genotypes. These are the homozygous genotypes AA, BB, and OO and the heterozygous genotypes AB, AO, and BO. Since A and B are codominant genes and O is

† School of Statistics, University of Minnesota, Minneapolis, USA

recessive, laboratories can distinguish only four phenotypes: A (the union of genotypes AA and AO), B (the union of BB and BO), AB, and O(= OO). (Modern methods can distinguish two kinds of A allelles, but I will not consider this possibility.)

Consider this scenario. A woman claims that a particular man is the father of her child. The woman's blood type is determined to be A and the child's is found to be B. If she is indeed the mother then both mother and child must be heterozygous, and the child's paternal gene must be B. So all type A and type O males are excluded from paternity; a man with one of these types could be the father only if there has been a laboratory error or a gene mutation.

Suppose the putative father is type AB. He has a B gene and so he is not excluded. There is an unfortunate tendency for people to view such non-exclusion as irrelevant. To see that it is positive evidence of paternity consider a much more extreme case. Suppose the child has a paternal trait that is so rare that at most one other person in the World could have it. (I admit that there may be no trait that is *this* rare.) Then a putative father who is not excluded is obviously the father. Many men have B genes. So non-exclusion is far from conclusive in the case at hand. The problem is to quantify non-exclusionary evidence. As indicated in Chapter 1 of this volume, a standard way to quantify uncertainty is by using probability. And the standard tool for incorporating evidence using probability is Bayes' Theorem.

It is convenient to decompose the evidence into two pieces, E and E´. Evidence E consists of all non-genetic background information in the case, including, I will assume, conclusive testimony that the 'mother' is in fact the mother, and the following genetic information: the ABO system phenotypes of the mother and the putative father and the frequency distribution of ABO phenotypes in some reference population, one containing the true father if he differs from the putative father. (Deciding on a reference population is obviously problematic—I will return to this issue below.)

Evidence E´ is the ABO blood type of the child. Let F stand for the event that the putative father is the real father, and let $\bar{F}$ be its complement. Assuming the condition E throughout, Bayes' Theorem says

$$P(\mathrm{F}|\mathrm{E}',\mathrm{E}) = \frac{P(\mathrm{E}'|\mathrm{F},\mathrm{E}).\ \ P(\mathrm{F}|\mathrm{E})}{P(\mathrm{E}'|\mathrm{E})}$$

According to the 'law of total probability', the denominator is

$$P(\mathrm{E}'|\mathrm{E}) = P(\mathrm{E}'|\mathrm{F},\mathrm{E}).\ P(\mathrm{F}|\mathrm{E})) + P(\mathrm{E}'|\bar{F},\mathrm{E}).P(\bar{F}|\mathrm{E}).$$

(See section 1.15 and the extension of the conversation.)

The factor P(E´|F,E), which occurs in both numerator and denominator of Bayes' Theorem, is the probability that a child of this woman and the putative father would be type B. Conditioning on F means that the only relevant phenotypes contained in E are those of the mother and putative father, assumed now to be the father. For the child to be type B the mother must pass on her O gene (she must be heterozygous since she is assumed to be the mother under E) and the father must pass on his B.

According to Mendelian genetics, each occurs with probability 1/2, and the probability they both occur is $(1/2)^2$; so $P(\mathrm{E}'|\mathrm{F},\mathrm{E}) = 1/4$.

The factor $P(\mathrm{E}'|\overline{\boldsymbol{F}},\mathrm{E})$ is the probability that a child of this woman would be type B if the putative father is not the father. Again the mother must pass on her O gene, which she does with probability 1/2. An assumption is now required concerning who the father is if he is not the putative father. All blood banks and other laboratories that calculate probabilities of paternity assume that, given $\overline{\boldsymbol{F}}$, the true father is a male who has been selected randomly (a so-called 'random man') from the reference population mentioned above.

The probability that a randomly selected male from this population has a B gene and passes on that gene to an offspring is the same as the probability of obtaining a B gene when selecting randomly from the corresponding population of *genes*. The latter is just the proportion of B genes in the population (which is distinct from the proportion of people with B *blood type*—the latter being a phenotype and not a genotype). This proportion depends on the population. Among Caucasians it is about 9 per cent. So if E refers to a Caucasian population, $P(\mathrm{E}'|\overline{\boldsymbol{F}},\mathrm{E})$ is about 0.09/2, or 4.5 per cent.

The remaining factor required to apply Bayes' Theorem is $P(\mathrm{F}|\mathrm{E})$. This is the prior probability of paternity, prior in the sense that it is not conditioned on the child's phenotype. I will discuss this quantity at greater length below. For now, I will make the assumption of Essen-Möller (1938) that the prior probability is 1/2. In the case at hand this gives

$$P(\mathrm{F}|\mathrm{E}',\mathrm{E}) = \frac{(1/4)(1/2)}{(1/4)(1/2) + 0.045(1/2)} = 84.7\% \ .$$

So the putative father's probability of paternity has jumped from 50% prior to E to about 85% posterior to E. This jump is rather large because B genes are moderately rare in the population.

There are many genetic systems. These include HLA, DNA RFLPs, and any other systems whose phenotypes can be identified and their frequencies in the reference population measured. An approach for incorporating the evidence contained in these systems is to apply Bayes' Theorem repeatedly, each time using the posterior probability for the previous system as the prior probability for the next. In the example case above, the second genetic system introduced would use 0.847 as the prior probability. The final probability is that which is posterior to the last system introduced. Another approach is to calculate $P(\mathrm{E}'|\mathrm{F},\mathrm{E})$ and $P(\mathrm{E}'|\overline{\boldsymbol{F}},\mathrm{E})$ by multiplying the corresponding probabilities for the individual systems and then applying Bayes' Theorem just once, using prior probability 1/2. These two approaches are equivalent and assume that the various genetic systems are independent.

The standard way most laboratories (in the United States, at least) present genetic calculations to courts is via the Bayes' factor or gene system index

$$\mathrm{GSI} = \frac{P(\mathrm{E}'|\overline{\mathrm{F}},\mathrm{E})}{P(\mathrm{E}'|\mathrm{F},\mathrm{E})} \ .$$

When there are multiple gene systems the various GSIs are multiplied to give the paternity index, or PI. (Such multiplication also assumes that the various gene systems are independent.) Bayes' Theorem then says

$$P(\mathrm{F}|\mathrm{E}',\mathrm{E}) = \frac{P(\mathrm{F}|\mathrm{E})}{P(\mathrm{F}|\mathrm{E}) + [\mathrm{PI}]\ P(\bar{\mathrm{F}}|\mathrm{E})}\ ,$$

In terms of odds,

$$\frac{P(\bar{\mathrm{F}}|\mathrm{E}',\mathrm{E})}{P(\mathrm{F}|\mathrm{E}',\mathrm{E})} = \mathrm{PI}\ \frac{P(\bar{\mathrm{F}}|\mathrm{E})}{P(\mathrm{F}|\mathrm{E})}$$

which says that the posterior odds ratio equals the paternity index times the prior odds ratio.

The above description is the essence of the technique used by blood banks and other laboratories to calculate probabilities of paternity. The basic idea is correct, but there are some important caveats and essential modifications. I will discuss these below.

Prior probability. For a laboratory to assume a prior probability of 1/2, or any other particular value, is misleading at best. The assumption has the effect of pre-empting the roles of the judge and jury. The resulting posterior probability is assuredly not the probability of paternity. Indeed, applying the Essen-Möller version of Bayes' Theorem to several putative fathers could well result in no probability distribution at all since the sum may be greater than 100%. Taking the above case as an example, having two putative fathers who are both type AB would give a total percentage of 169.4%.

The prior probability should be judged by each juror based on E, the background evidence in the case. Particularly relevant parts of E include testimony concerning whether sexual intercourse took place between the mother and the putative father and between the mother and other men, the timing and frequency of such intercourse, testimony concerning the fertility of the possible fathers, the time interval during which conception might have taken place (based on the woman's menstrual cycle and the child's birthday), and testimony concerning the use of birth control methods and the reliability of such methods. Probability assessment is a standard requirement whenever Bayes' Theorem is used. One approach (DeGroot 1970) to probability assessment is to reason in terms of betting odds or lotteries by comparing the event F with events of known (or assumed to be known) probabilities, such as outcomes when rolling a die. While this method can be effective, and has no serious competitors, it may meet with resistance from jurors who are offended by comparisons with gambling. A simpler approach is to assume that jurors are naturally blessed with reasonable conceptions of probability. This is clearly wishful thinking, but, especially under rules that disallow questions from jurors during a trial, no approach is perfect.

However jurors' prior probabilities are derived, laboratories should provide them with something equivalent to a table showing how the genetic evidence in the case at hand converts prior probabilities into posterior probabilities, also providing help in using such a table. An abbreviated table for the example case considered above looks like this:

Prior probability	0	0.100	0.250	0.500	0.750	0.900	1
Posterior probability	0	0.382	0.649	0.847	0.943	0.980	1

An appealing alternative is to present this relationship in reverse, saying, for example, that to have a posterior probability of paternity of 50% implies a prior probability of about 15%.

A graph can be used to show the relationship between prior and posterior. Fig. 5.2 shows this relationship for the above example.

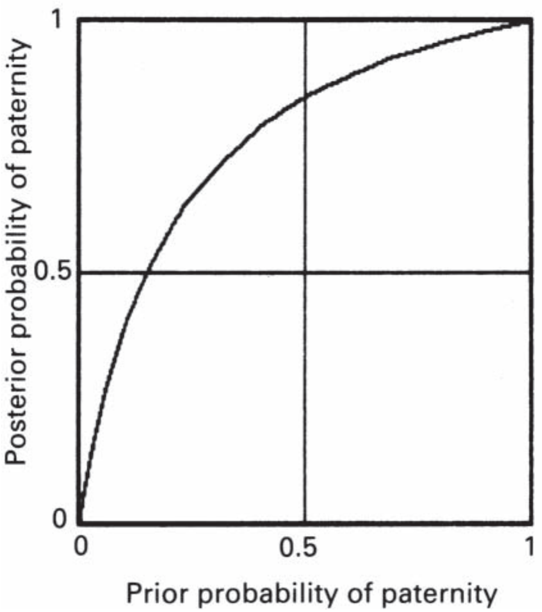


Fig. 5.2 —Relationship between posterior and prior probabilities.

Independence of genetic systems. An issue of importance in cases of disputed paternity, and in forensic science generally, is multiplying probabilities of several pieces of evidence together. Such multiplication assumes that the individual pieces are statistically independent. This assumption is often made but seldom verified. This statement certainly applies to paternity testing. For ways to check for independence, and for related discussion, see Berry & Geisser (1986).

Reference population. The above development assumes a known reference population. As indicated above, the reference population is irrelevant if the putative father is the true father. Therefore, the correct reference population is that of the true father assuming he is different from the putative father. Since this is unlikely to be known, laboratories use a convenient population (usually on the basis of race) containing the putative father! This makes no sense at all. It introduces a bias that can be large, though its direction is not clear. In the absence of testimony concerning the race of the true father, laboratories should calculate probabilities for each possible race. These probabilities should be presented to the court along with an explanation that makes it clear that the one used should be the race of the true father.

Sampling variability. The phenotype distribution in the reference population is not perfectly known (unless there is a specific set of men known to contain the true father and their phenotypes are known). Rather, samples are collected from the various populations. (These are always samples of convenience and are not random; this may introduce a bias.) Laboratories assume that the sample frequencies are the same as the population frequencies. This is obviously incorrect; it is more nearly true for larger samples—for random samples of size n the sampling error is inversely proportional to $\sqrt{n}$. This source of uncertainty could be incorporated into Bayes' Theorem, but laboratories have not done so; doing so could either increase or decrease the posterior probability.

Laboratory errors. Laboratory test results may be wrong. It is important to know the likelihood of errors, but learning about the frequency of such laboratory errors is very difficult. Some errors come to light when, for example, the child's mother is excluded. (Such exclusions are not always errors: there have been paternity cases in which the 'mother' was excluded on the basis of genetic testing and further investigation revealed that the child and another had been exchanged inadvertently in the hospital nursery!) But many errors are never uncovered. One possibility is routinely to run duplicate tests to estimate error rates. But two different tests carried out at the same laboratory cannot be independent; they may not be independent even if carried out at different laboratories. Replicate testing underestimates the actual error. But such testing should be carried out routinely to establish at least rough estimates of the laboratory's reliability.

The possibility of laboratory errors could easily be incorporated into Bayes' Theorem; again, laboratories have not done so. The direction of this oversight is predictable: for non-excluded males, including the possibility of laboratory error would tend to decrease the posterior probability of paternity. The size of this decrease obviously depends on the likelihood of errors.

Further reading. This section is a brief introduction to the use of probability in cases of disputed paternity. Papers by Berry & Geisser (1986), Aickin & Kaye (1982), Ellman & Kaye (1979), Salmon & Salmon (1980), Schatkin (1984), Sussman & Gilja (1981), and Wiener (1976) provide more information on the issues touched upon here.

References

Aickin, M. & Kaye, D. (1982) Some mathematical and legal considerations in using serological tests to prove paternity. In: R.H.Walker (ed.) *Inclusion probabilities in parentage testing,* American Association of Blood Banks, Arlington, VA, pp. 155–168.

Berry, D.A. (1991) Inferences using DNA profiling in forensic identification and paternity cases, *Statistical Science* **6** 175–205.

Berry, D.A. & Geisser, S. (1986) Inference in cases of disputed paternity. In: M.H. DeGroot, S.E. Fienberg and J.B. Kadane (eds.) *Statistics and the law,* Wiley, New York, 353–390.

DeGroot, M.H. (1970) *Optimal statistical decisions,* McGraw-Hill, New York, Chapter 6.

Ellman, I.M. & Kaye, D. (1979) Probabilities and proof; Can HLA and blood group testing prove paternity? *New York University Law School Journal* **54** 1131–1162.

Essen-Möller, E. (1938) Die Biesweiskraft der Ahnlichkeit im Vater Schaftsnach-weis; Theoretische Grundlagen, *Mitt. Anthorop, Ges. (Wein)* **68** 598.

Gjertson, D.W., Mickey, M.R., Hopfield, J., Takenouchi, T. & Terasaki, P.I. (1988) Calculation of probability of paternity using DNA sequences, *American Journal of Human Genetics* **43** 860–869.

Salmon, D. & Salmon, C. (1980) Blood groups and genetic markers; polymorphisms and probability of paternity, *Transfusion* **20** 684–694.

Schatkin, S.B. (1984) *Disputed paternity proceedings,* Vols I and II, 4th ed., rev., Matthew Bender, New York.

Sussman, L.N. & Gilja, B.K. (1981) Blood grouping tests for paternity and nonpaternity, *New York State Journal of Medicine* **81** 343–346.

Wiener, A.S. (1976) Likelihood of parentage. In: L.M.Seideman (ed.), *Paternity testing by blood grouping,* 2nd ed. Charles C.Thomas, Springfield, pp. 124–131.

FORENSIC IDENTIFICATION OF SOIL, USING STATISTICAL MODELS

G.Gettinby†

Introduction

Soil is formed from the aggregation of weathered rock particles and is important in forensic studies where it may be transmitted to or from the scene of a crime. Typically, a soil sample may be recovered from the property of a suspect, such as clothing, footwear, vehicles, etc., and the identity of its origin sought. In particular, the recovered soil sample may have to be compared with a control soil sample obtained from the scene of a crime.

Soil offers several features for comparison. The chemical composition of the soil may be examined for the presence of phosphates, ammonia, salt, pH (Dudley 1976a) etc. The physical condition of the soil may be used to assess colour (Dudley 1975a),

† Department of Statistics and Modelling Science, University of Strathclyde, Glasgow, UK

density (Chaperlin & Howarth, 1983), etc. In addition, most soils contain biological materials in different amounts. For operational purposes a procedure for comparing soil samples needs to be practical, inexpensive, accurate, and most of all it should be suitable for use with small quantities of soil. Microscopic examination can be useful in revealing pollen grains, decomposed plants, etc. Neutron activation analysis and atomic absorption methods can be used for the measurement of trace elements (Hoffman *et al.* 1969). However, mineral particles remain stable, and so one of the most widely used comparisons is particle size distribution (Dudley 1975b, Dudley & Smalldon 1978b). In particular, sieving and Coulter Counter methods can be used to obtain good estimates of the volume per unit weight of soils over a wide range of particle size diameters. This section provides descriptions of two complementary statistical methods for the matching of soil samples, using particle sizes and other attributes.

Elliptical contour model

Dudley & Smalldon (1978a) carried out an examination of 18 duplicate soil samples collected in Berkshire, England. Inspection of the silt particle size distributions indicated that soils from different locations could not always be easily distinguished. Moreover, the use of conventional two-sample statistical tests to compare the distributions led to duplicate soil samples being identified as different at conventional levels of significance. This suggests that such tests were too unreliable for operational use. Instead a model was proposed which used two new variables to characterize duplicate soil samples. These variables were summary quantities constructed from the particle size distribution of duplicate soil samples. The first variable was derived from the volumes per unit weight of the duplicate samples, by dividing the difference of the volumes by the sum of the volumes to give an index of variability (IV). The second variable was calculated as the maximum difference (MD) between the cumulative probability distributions of the particle sizes of duplicate soils, a quantity commonly used in Kolmogorov-Smirnov two-sample tests.

A simple procedure to characterize the variation between soil samples from the same locus would be to construct 95% or 99% confidence levels for the IV or MD variables using their respective means μ_{iv} and μ_{md}, and respective variances σ^2_{iv} and σ^2_{md}. Estimates of these parameters are available from the collected duplicate soil samples. Since IV and MD are not independent, the univariate confidence intervals can be improved upon by using a confidence contour constructed from the bivariate distribution of IV and MD. If IV and MD can be assumed to have a bivariate normal distribution with correlation coefficient p, obtaining the confidence contours can be simplified by using the transformation:

$$X = \text{IV Cos}\theta + \text{MD Sin}\theta,$$
$$Y = - \text{ IV Sine}\theta + \text{MDCos}\theta,$$

where $\tan 2\theta = 2\rho\sigma_{iv}\sigma_{md}/(\sigma^2_{iv} + \sigma^2_{md})$.

The transformed variables X and Y are independent and have normal distributions with mean 0 and respective variances

$$2\sigma_x^2 = \sigma_{iv}^2 + \sigma_{md}^2 + \{(\sigma_{iv}^2 - \sigma_{md}^2)^2 + 4\rho^2\sigma_{iv}^2\sigma_{md}^2\}^{0.5} \ ,$$
$$2\sigma_y^2 = \sigma_{iv}^2 + \sigma_{md}^2 - \{(\sigma_{iv}^2 - \sigma_{md}^2)^2 + 4\rho^2\sigma_{iv}^2\sigma_{md}^2\}^{0.5} \ .$$

It follows that

$$Z = X^2/\sigma_x^2 + Y^2/\sigma_y^2$$

has a λ^2 distribution with two degrees of freedom. Contours for the 95% and 99% levels can be sketched using the ellipse

$$X^2/\sigma_x^2 + Y^2/\sigma_y^2 = c \ ,$$

where c is obtained from λ^2 tables, using

$$P[\lambda^2 > c] = 0.05$$

and

$$P[\lambda^2 > c] = 0.01$$

respectively.

Finally, the transformation

$$\begin{aligned} \mathrm{IV} &= X \,\mathrm{Cos}\theta - Y \,\mathrm{Sin}\theta \\ \mathrm{MD} &= X \,\mathrm{Sin}\theta + Y \,\mathrm{Cos}\theta \end{aligned}$$

rotates the contours to provide the required confidence contours in the IV and MD space. For operational use, control and recovered soil samples which generate IV and MD values within a contour can be classified as similar, and those which fall outside can be classified as different.

Validation. The power of the model to discriminate between soil samples from different origins is important. Using confidence intervals constructed from parameter estimates of means, variances, and correlation coefficients obtained from the 18 duplicate soil samples, Dudley & Smalldon were able to assess the performance of the model. A total of 153 comparisons of those pairs of soil samples which were not duplicates indicated that the model, used with the 95% confidence contour, correctly identified 97% of the comparisons as different. This compared favourably with the use of a univariate confidence interval for IV alone, which gave only an 86% discriminating power.

Observations. The elliptical contour model of Dudley & Smalldon remains an important method for the comparison of soils by using particle size distributions. Its success depends on the availability of sufficient duplicate soil samples which are representative of the geographical area in which the model is to be applied. Extension of the method to a wider geographical domain would require the construction of a new database, containing records of duplicate soil sample analysis, for the production of relevant elliptical contours. The method does make the assumption that the IV and MD variables obtained from duplicate soil samples have a bivariate normal distribution to simplify obtaining confidence contours. If this is not true, modern computer simulation methods could be employed to produce simulated confidence contours from empirical data.

Similarity probability model

Wanogho *et at.* (1985a) reported a method for the comparison of soils in close proximity. The procedure arose from an experimental study in which a ploughed field was divided into 100 cells each of area 3 × 3 m^2. By examining the variability of the soil between and within cells, using 6 replicate soil samples from each cell and taking measurements on up to 7 variables, it was possible accurately to predict which cell a soil sample came from. The procedure is relevant to forensic investigations where the origin of a recovered soil sample within a control area is to be identified.

The control area is divided into *c* cells of equal area. Each cell is classified according to its row and column location. From within each cell *n* replicate soil aliquots are collected. Laboratory analyses are undertaken on each aliquot to obtain the particle size distribution. The particle size distribution provides estimates of the percentage of the total aliquot volume to fall within particle class sizes. If the particle size distribution consists of *p* classes this gives measurements on each of *p* variables $X_1,\ldots,X_p$. In addition, a measurement is obtained for X_{p+1}, t_{he} percentage organic matter of the aliquot. The identification of soil from different cells is based upon these p +1 variables. Two-way analysis of variance is undertaken on each variable, using row and column locations as factors. This provides a screening test. If significant row × column interactions are indicated, the analysis can proceed and the residual mean square,s^2_{Xj}, used as a measure of within cell variation. If no significant differences are obtained the variable has little value in discriminating between soil from different cells.

A recovered soil sample, known or suspected of coming from the control area, is divided into *m* aliquots, and similar laboratory measurements are obtained on variables $X_1,\ldots,\mathrm{X}_p +{}_1$. For variable X_j, $j = 1$ to $p + 1$, the recovered data are compared with the data from control cell *i* by using the *z* statistic

$$z_i = \frac{\bar{x}-\bar{y}}{s_{X_j}(1/n+1/m)^{0.5}}$$

where $\bar{x}$ is the mean of the X_j data for the n control aliquots, and $\bar{y}$ is the mean of the X_j data for the *m* recovered aliquots.

The probability of obtaining such an extreme value as z_{i} is calculated by using

$$P_j(z_i) = 2\Phi(-z_i)$$

where $\Phi(-z_i)$ is the area under the standard normal distribution between $-\infty$ and $-z_i$. The most likely cell of origin of the recovered soil sample based on a single variable X_j is cell i with the largest similarity probability, that is,

$$\max\,\{P_j(z_i)\} \qquad \text{i} = 1,\ldots,c\ .$$

The most likely cell of origin of the recovered soil sample based on all variables $X_1,\ldots,X_{p+1}$ is cell i corresponding to

$$\max\{P_1(z_i) \times \ldots \times P_{p+1}(z_i)\} \qquad \text{i} = 1,\ldots,\text{c}\ .$$

Instead of choosing the control cell associated with the maximum likelihood, the control cells may be ranked in order of similarity with the recovered soil sample. The method is also sufficiently flexible to include any other variables, such as median particle size, which may enhance accurate identification of origin.

Validation. In the initial study, out of 20 recovered soil samples, chosen at random from the 100 control cells, the cells of origin of 19 samples were correctly identified. Control and recovered soil samples were each divided into six 1.5 g aliquots, and the variable percentage of particles in class interval 90–250 μm (X_1), median particle size (X_2), and percentage organic matter (X_3), provided best identification.

In a subsequent study (Wanogho *et al.* 1985b), the similarity probability model was used to predict the origins of 14 recovered soil samples within 26 control samples. Only 10 of the recovered samples originated from the 26 control samples, and the correct cell of origin was identified in each case. On the remaining of 4 recovered samples, 2 were wrongly predicted to have originated from the control set. These 2 samples came from a location close to the identified control cells. In these comparisons only three 0.5g aliquots were analysed from each soil sample, and the predictions were based on the variables: percentages of particles in class intervals =63 μm (X_1), 63–90 μm (X_2), 90–250 μm (X_3), 250–500 μm (X_4), 500 μm–1 mm (X_5), median particle size (X_6), and percentage organic matter (X_7).

Observations. The success of the similarity probability model at predicting the correct identity of soils taken from within a control area arises because the control area is well defined in terms of the variables being used for prediction. With many problems of transfer in forensic science, such information is not obtainable, or the cost of obtaining the information is prohibitive. The method is easy to apply, and once the control data are available the model can be routinely implemented on microcomputers, using stored commands in the MINITAB statistical software. The method does make assumptions about the independence of variables in order to justify multiplication of similarity probabilities from different variables. In practice, this assumption is unlikely to be entirely true. Nevertheless, the method gives satisfactory results under operational conditions and performs well with small quantities of soil.

Discussion

The models described show that it is possible to use statistical methods for the forensic identification of soils. The elliptical contour model is suitable for use in situations where a recovered soil sample is to be compared with any control sample, to decide whether the two samples exhibit particle size distributions typical of duplicate samples. The method will work well, provided that the population of duplicate soil samples has been adequately characterized by using empirical data. In contrast, the similarity probability model would be of particular value in forensic studies where a suspect may admit to being within the vicinity of a crime locus. Soil recovered from the suspect's belongings could be confidently pinpointed to a particular locus. Like most statistical models in forensic science, both models suffer from inadequate testing under operational conditions. In the case of the ellipitical contour model the construction of the contours may not have been properly understood, and with the similarity probability model the generation of control data would be expensive and could be justified only in exceptional circumstances. In addition, neither method has been presented in a form suitable for operational use. Microcomputer software for the implementation of the methods would promote greater awareness, critical assessment, and adoption within the forensic community.

The methods highlight the importance of particle size distribution. The measurement of particle sizes has traditionally depended on tedious sieving methods (Dudley 1976b) which were routinely used for particle sizes with diameters in excess of 63 μm. The wet sieving method is assumed to be more useful than the dry method because it provides a better estimate of the soil particle sizes owing to lack of agglomeration (Wanogho *et al.* 1987a, b). Using this procedure, soil samples as small as 0.5 g can be successfully analysed. The sub-63 μm particles can account for up to one third of clay soil samples. Recently, laser diffraction and Coulter Counter methods (Wanogho *et al.* 1987c) have shown that the sub-63 μm fractions can be analysed quickly and effortlessly. It is therefore possible that with efficient laboratory methods for processing soil samples and computer storage, analysis, and delivery systems, statistical models like those described will become increasingly important for the forensic identification of soil.

References

Chaperlin, K. & Howarth, P.J. (1983) Soil comparison by the density gradient method—a review and evaluation, *Forensic Science International* **23** 161–177.

Dudley, R.J. (1975a) The use of colour in the discrimination between soils, *Journal of the Forensic Science Society* **15** 209–218.

Dudley, R.J. (1975b) Techniques for soil comparison in forensic science, *International Microfilm Journal of Legal Medicine* 10No. 4 Paper 65.

Dudley, R.J. (1976a) A simple method for determining the pH of small soil samples and its use in forensic science, *Journal of the Forensic Science Society* 16 21–27.

Dudley, R.J. (1976b) The particle size analysis of soils and its use in forensic science. — The determination of particle soil size distributions within the silt and sand fractions, *Journal of the Forensic Science Society* **16** 219–229.

Dudley, R.J. & Smalldon, K.W. (1978a) The comparison of distributional shapes with particular reference to a problem in forensic science, *International Statistical Review* **46** 53–63.

Dudley, R.J. & Smalldon, K.W. (1978b) The evaluation of methods for soil analysis under simulated scenes of crime conditions, *Forensic Science International* **12** 49–60.

Dudley, R.J. & Smalldon, K.W. (1978c) The objective comparison of the particle size distribution in soils with particular reference to the sand fraction, *Medicine Science and the Law* **18** 278–282.

Hoffman, C.M., Brunelle, R.L., & Snow, K.B. (1969) Forensic comparisons of soils by neutron activation and atomic absorption analysis, *The Journal of Criminal Law, Criminology and Police Science* **60** 395–401.

Wanogho, S., Gettinby, G., Caddy, B., & Robertson, J. (1985a) A statistical method for assessing soil comparisons, *Journal of Forensic Sciences* **30** 864–872.

Wanogho, S., Gettinby, G., Caddy, B., & Robertson, J. (1985b) A statistical procedure for the forensic evaluation of soil particles. In: P.J.Lloyd, (ed.) *Particle size analysis.* John Wiley & Sons, 105–112.

Wanogho, S., Gettinby, G., Caddy, B. & Robertson, J. (1987a) Some factors affecting soil sieve analysis in forensic science: 1 Dry Sieving, *Forensic Science International* **33** 129–137.

Wanogho, S., Gettinby, G., Caddy, B., & Robertson, J. (1987b) Some factors affecting soil sieve analysis in forensic science: 2 Wet Sieving, *Forensic Science International* **33** 139–147.

Wanogho, S., Gettinby, G., Caddy, B. & Robertson, J. (1987c) Particle size distribution analysis of soils using laser diffraction, *Forensic Science International* **33** 117–128.

HUMAN HAIRS

C.G.G.Aitken† and J.Robertson‡

Introduction

Hair analysis is difficult. Much of this difficulty arises from the lack of discrete, numerical, and quantifiable features in (at least) human hairs. Robertson (1982) has discussed the use of qualitative and quantitative features in the examination of hairs. The consensus of opinion amongst hair examiners would seem to be that the currently available numerical features are of very limited value (Robertson & Aitken 1986). The discussion here relates mainly to the microscopic analysis of human head and pubic hairs because these are the types most often encountered in casework. Attempts to individualize these hairs rely on detailed microscopic examination.

† Department of Mathematics and Statistics, The University of Edinburgh, Edinburgh, UK
‡ Australian Federal Police, Canberra, Australia

However, analysis of the microscopic features of hairs has been criticized as too subjective. This is perhaps a fair criticism, but it is not a reason for the dismissal of microscopic hair analysis. Rather, it should act as a spur for the improvement of the methods of data collection from such analyses.

Non-microscopic approaches

Useful information can be obtained by enzyme typing of cellular material comprising the sheath of actively growing anagen hairs roots (Gambel *et al.* 1987). The disappointing fact is that most hairs recovered in casework have roots in the catagen or telogen growth phase in which sheath material is not present. Telogen hairs are hairs at the end of their growth cycle which are naturally shed. As most hairs transferred in contact are not direct transfers from the head of the donor but rather are indirect from the clothing worn by the donor (Gaudette & Tessarolo 1987) it is then not surprising that it is the naturally shed telogen hairs which are found as extraneous hairs. A second point worth consideration is that the information content to be gained from the analysis of sheath material for blood group substances is of quite a low order. One in 4 or 1 in 10 of the population might have the blood group present in the sheath material, and this would be considered good evidence. An experienced hair examiner will achieve much higher levels of discrimination than these ‘odds’ by microscopic examination. Recent advances in the extraction and analysis of minute amounts of DNA from hairs (Higuchi *et al.* 1988) hold the promise of a much higher level of discrimination. At the present time it appears that an active growing hair root may be required, although not the cellular sheath material. This would again limit its application in casework.

Other non-microscopic approaches to the analysis of hairs have been tried, and, in particular, elemental analysis of hairs was thought to hold considerable promise for a numerical, objective approach to hair analysis. Unfortunately, this early optimism has not been fulfilled. This subject has recently been reviewed by one of us (Robertson 1987).

Thus, as Robertson concluded in his 1982 paper, microscopic examination remains the pivotal technique for the examination and comparison of hairs. The remainder of this section deals with the application of statistics to the analysis of information obtained by microscopic examination.

Statistical approaches

In other sections of this chapter, as far as possible, statistics have been used to provide a direct measure of the value of evidence by use of the likelihood ratio. This approach is possible since frequency data and measurements have been available on characteristics of interest. This use of statistics is not possible with hairs since very few measurements of characteristics of interest are made. Those which are, tend to be discretized to form what is known as an ordinal variable, for example, the medullary index may be discretized as less than 0.167, 0.167–0.250, greater than 0.250, and the three categories have a natural ordering. With other possible characteristics of interest, such as colour, pigment density, and medulla, it is possible to record on a nominal scale, but these characteristics have no natural ordering, though there may already be some argument about this. For example, pigment density may be recorded as absent, light, medium, heavy, or opaque, which, arguably, has a natural ordering. Statistical analysis, using discrete multivariate analysis, would, theoretically, be possible on data of these types if records were kept on the frequencies with which each characteristic occurred. However, the number of variables and the number of possible responses for each variable mean that an unpractically large data set would be required for any meaningful statistical analysis. Also, the analyses would be required with a considerable investment of time and resources. At current rates of hair examination the data sets could take many years to compile, even if all the problems of ensuring consistency of recording between laboratories could be solved. These problems are not trivial. For example, a part of a hair which is one colour to one examiner may be recorded as a different colour by a different examiner. Pigment density which is light to one examiner may be medium to another. Indeed the same examiner may record the same feature in a different way from week to week (if not day to day). A reference book giving standardized examples for each possible category for each feature would be invaluable for the development of a consistent recording scheme. The use of such a reference book could slow down the analysis of hairs but with the benefit of improving the standard of consistency and reproducibility. Of course, as an examiner became more experienced the requirement to consult the photographic reference would reduce. A hair examination form, which goes some way to meeting these needs, has been developed after consideration of the questionnaires distributed by the present authors (Aitken & Robertson 1986, Robertson & Aitken 1986). A report, by Verma *et al.* (1987), also describes an attempt to gather data which would be useful for this type of analysis. Their record sheet allowed for information on twelve variables to be recorded at each of thirty points along the length of a hair as well as general information from another eleven variables, in the Introduction to their report, they stated that

> '(t)his proposed research is one (such) attempt to gain a better understanding of the quantitative aspect of hair individualization by addressing the issue of population frequencies, as an essential parameter, for a more meaningful estimate of inter- and intra-individual variations. The objective status of the individualization potential of hair evidence can thus be improved and defined.'

Notice the reference to population frequencies. The importance of such data has been emphasized in Chapter 2.

It is the absence of such data that is restricting the potential for this type of statistical treatment for hair analysis. Other attempts at the development of a system for the characterization of human head hairs have been described by Shaffer (1982) and Strauss (1983).

The potential for the use of computerized data entry, accumulation and treatment is obvious but has so far not attracted the interest of computer 'buffs'. The use of a computer in this way would not, of course, solve any of the problems of reproducibility or the accuracy of microscopic data, but it would make it much easier to record features at numerous points along the hair shaft, giving more detailed information than is acceptable for manual handling.

One approach to overcoming the subjective judgement of the examiner/observer would be to use image analysis techniques. Quite sophisticated equipment is now available with computer software back-up for the analysis of the recorded data. A digitized image of the longitudinal cross-section could be obtained and the data could be easily collected by this method. The analysis problems would be considerable since the amount of information in one image (128 × 128 pixels, say) would be large, and many images would be needed in order to develop reasonable statistical models. Thus, whilst this approach might offer the promise of producing the type of data which could be subjected to statistical analysis, its application is, at present, not even a remote possibility.

If the scenario for objective data production is so poor, what other approaches may be viable? One way of assessing the value of hair examination would be to test the ability of the analyst to discriminate hairs and to assign unknown hairs to a range of possible sources.

Gaudette has described such a series of experiments (Gaudette & Keeping 1974, Gaudette 1978). These experiments relied on many pairwise comparisons of hairs, known to be from different people, and the recording of the proportion of the total number of comparisons in which it was not possible to distinguish the different sources. The proportion calculated is known as an average probability. Statistics has a small role in the derivation of the results from Gaudette's experiments, and this has been described in Chapter 2 in the section on average probabilities. However, it should be emphasized that Gaudette's results do not refer to measurements or counts of particular characteristics on hairs directly. Rather, they refer to counts of comparisons made on pairs of hairs where the comparisons are made, using the subjective criteria which are not measurable or necessarily reproducible between different hair examiners.

It is not within the scope of this chapter to examine in detail the arguments which have appeared in the literature surrounding the approach of Gaudette (see for example, Barnett & Ogle 1982, Aitken & Robertson 1987). Suffice it to say that much of this discussion has centred around the wording of the questions concerning the value of hair evidence which Gaudette attempts to answer. This is important since the nature of the question asked determines the appropriate form of statistical analysis, and, as Aitken & Robertson (1987) have shown, this can affect the probability figures which Gaudette derived in his papers for the value of the evidence relating to both head and pubic hairs. The justification for the use of a Bayesian approach remains paramount, however, and is unaffected by the conflicting views of those who have entered into a dialogue with Gaudette.

Whether the estimates Gaudette has produced are meaningful or not depends on factors beyond pure statistical considerations. It is our view that, at best, his estimates apply only to experienced hair examiners in the Royal Canadian Mounted Police laboratory system. Before they can have more universal application, similar studies require to be done in other systems. These have not been forthcoming. Gaudette is to be applauded for raising the level of awareness and discussion of this important issue.

One of the difficulties in the application of the Bayesian approach is of course providing (in any practical sense) reasonable estimates (or odds) for the factors included in the equation. In this regard, one should consider not only questions regarding the origin of a hair, based on analytical factors, but arguably the very finding of a hair may have significance.

Thus, the possibility of transfer, secondary or otherwise, of human head hairs may be relevant. For example, Gaudette & Tessarolo (1987) described a series of experiments in which they studied the effects of various factors on the possibility of hair transfer. In the Introduction, they referred to a case (Anon. 1985) in which a hair was transferred from a suspect to a victim's body by way of the investigating officer's jacket. This caused a murder inquiry to take an improper course. The conclusions from Gaudette & Tessarolo's study included a warning that 'in statistically evaluating hair evidence, each case must be evaluated on its own merits'. This is a perfectly legitimate warning in that the statistical evaluation referred to is that based on average probabilities, and these probabilities refer to a general description of the value of the evidence type not to a particular case.

Further studies by Somerset (1985) have looked at the transfer and persistence of human scalp hairs, and a preliminary study of the transfer of pubic hair during acts of sexual intercourse has been carried out (M. Keating, personal communication). In this latter study it was shown that pubic hairs are not always transferred during even close intimate contact.

Although most of the hairs recovered during the searching of items are probably shed naturally (telogen hairs), in some cases the finding of pulled hairs (anagen hairs) (although see King *et al.* (1982) for a fuller discussion of what is not a black and white situation) is surely of significance. For example, a victim may pull hairs from an attacker. The presence of 'pulled' hairs which 'match' a suspect is surely of greater significance than 'shed' hairs? The difficulty lies in placing a numerical value of the significance to be attached to this finding. However, simply because it is difficult, should we run away from confronting the issue?

Aitken & Robertson (1987) discussed two approaches toward improving the objectivity of hair analysis. The first was to extend the method of pairwise comparisons between hairs from different sources to include comparisons between hairs known to be from the same source. The proportion of such comparisons in which the two hairs were found to be indistinguishable should be noted. It is expected that this proportion would be close to 1. This proportion would enable an estimate of the probability of the evidence of association, given the hairs came from the same source, to be made. This probability could then be compared with the probability of the evidence of association determined from the pairwise comparisons made between hairs from different sources.

The second approach suggested was the collection of frequency data, using the hair examination form mentioned earlier in this paper. Use of this form, together with a central collection of data through a computer network and a microcomputer on the investigator's bench, would enable a practical method of collection and analysis to be developed. Only when data have been collected can investigations be made of the distribution of characteristics between individuals, between hairs from the same individual and between different parts of the same hair from the same individual.

Conclusion

In this short section we have described the two main approaches which may be used in attempting to evaluate the worth of hair as an evidence material. One approach relies on attempting to evaluate the ability of the analyst to assign correctly unknown hairs to a source; the other, on the ability of the analyst to describe the individual features providing data which can then be subjected to the appropriate statistical test. This in itself is not as straightforward as it seems, because it is quite probable that the features recorded are not independent of each other (for example, pigment features) and there are the previously discussed problems with obtaining reliable numerical data. These two approaches need not be mutually exclusive, and perhaps the damaging aspect of some of the debate which has taken place in this area is the unstated assumption that one approach is right and the other wrong.

At present, statistics is of limited value in attempting to evaluate the evidential value of the finding of a hair or hairs. Improvements lie not in the hands of the statistician but in those of hair examiners. They need to carry out the basic research and data collection. The prognosis is not promising. Those of us who are interested in hair examination and believe it to have value are also the people who are too busy doing casework to carry out the relevant research. However, too much scientific research proves to be less useful than it needs to be through inadequate experimental design which then limits statistical treatment. As in most areas there needs to be a team approach with the statistician participating in the design stage.

Ideally, we want to be able to say that the hair evidence is so many times more likely if the suspect were at the scene of the crime than if he were not. At present, our knowledge falls well short of this ideal.

References

Aitken, C.G.G. & Robertson, J. (1986) The value of microscopic features in the examination of human head hairs: statistical analysis of questionnaire returns, *Journal of Forensic Sciences* **31** 546–562.

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

Anonymous (1985) The case of the adroit agent, *Forensic Science Digest* **11** 15–20.

Barnett, P.D. & Ogle, R.R. (1982) Probabilities and human hair comparison, *Journal of Forensic Sciences* **27** 272–278.

Gambel, B.A., Budowle, B., & Terrell, L. (1987) Glyoxalase 1 typing and phosphoglucomutase—1 subtyping of a single hair, *Journal of Forensic Sciences* **32** 1175–1181.

Gaudette, B.D. (1978) Some further thoughts on probabilities and human hair comparison, *Journal of Forensic Sciences* **23** 758–763.

Gaudette, B.D. & Keeping, E.S. (1974) An attempt at determining probabilities in human scalp hair comparison, *Journal of Forensic Sciences* **19** 599–606.

Gaudette, B.D. & Tessarolo, A.A. (1987) Secondary transfer of human scalp hair, *Journal of Forensic Sciences* **32** 1241–1253.

Higuchi, R., von Beroldingen, C.H., Sensabaugh, G.F., & Erlich, H.A. (1988) DNA typing from single hairs, *Nature* **323** 543–546.

King, L.A., Wigmore, R., & Tiribell, J.M. (1982) The morphology and occurrence of human hair sheath cells, *Journal of the Forensic Science Society* **22** 267–269.

Robertson, J. (1982) An appraisal of the use of microscopic data in the examination of human head hair, *Journal of the Forensic Science Society* **22** 390–395.

Robertson, J. (1987) Trace elements in human hair—a significant forensic future, *Trends in Analytical Chemistry* **6** 65–69.

Robertson, J. & Aitken, C.G.G. (1986) The value of microscopic features in the examination of human head hairs: analysis of comments contained in questionnaire returns, *Journal of Forensic Sciences* **31** 563–573.

Shaffer, S.A. (1982) A protocol for the examination of hair evidence, *The Microscope* **30** 151–161.

Somerset, H. (1985) The persistence of human head hairs on clothing, MSc. thesis, University of Strathclyde.

Strauss, M.A.T. (1983) Forensic characterization of human hair I, *The Microscope* **34** 15–29.

Verma, M.S., Reed, P.A., & Kobremski, T.E. (1987) *Quantifying the individualization of hair source through an application of computer assisted syntax-directed pattern reecognition technique,* part I. Colorado Department of Public Safety.

STATISTICS IN FORENSIC BALLISTICS

W.F.Rowe†

The most basic examinations in the field of forensic ballistics (more properly called firearms examination) are directed at determining whether bullets or cartridges found at the scene of a crime were fired from a particular firearm. To do this, the forensic firearms examiner test fires rounds from the suspect weapon and, using a comparison microscope, compares the markings on the resulting test-fired bullets and cartridges with those on the bullets and cartridges from the crime scene (often referred to as the questioned bullets and cartridges). The markings made by a firearm on a fired bullet consist principally of land and groove impressions made by the rifling of the barrel. The lands are the raised spiral ridges of the rifling that grip the bullet and give it a stabilizing rotation as it traverses the barrel; the grooves are the recesses between the lands. Minute imperfections on the land and groove surfaces engrave microscopic scratches (variously called striations or striae) on the surfaces of fired bullets. If the patterns of striations on the surfaces of two bullets match, then the firearms examiner concludes that the same weapon fired both bullets. The firearms examiner may reach the same conclusion in the case of

† The Graduate School of Arts and Sciences, The George Washington University, Washington DC, USA

cartridges if he can match markings such as firing pin impressions, breechblock markings (as called bolt face signatures), and extractor and ejector marks appearing on two cartridges. For more detailed information on firearms examinations the interested reader may refer to a number of textbooks or reviews (Hatcher *et al.* 1957, Davis 1958, Ceccaldi 1962, Mathews 1962, Berg 1977, and Rowe 1988).

Unlike forensic serology where statistics have been applied routinely for many years, firearm examination is a discipline in which statistics play a limited role at the present time. Consider the most basic of all examinations in forensic ballistics: the matching of striations in land and groove impressions on the surfaces of fired bullets. When a firearms examiner matches patterns of striations, using a comparison microscope, he has no definite guidelines as to how many striations in a pattern must match before he reaches a positive conclusion (viz. that both test-fired bullet and questioned bullets were fired from the same weapon). Despite decades of acceptance by law enforcement agencies and courts of law, positive conclusions rest primarily on examiners' personal experiences and on a very limited number of experimental studies in which bullets were fired through weapons that were sequentially rifled with the same rifling tools (Hatcher *et al.* 1957 and Austin 1969). These studies showed that, even when new, different weapons produced different patterns of striations. Some authorities (Hatcher *et al.* 1957 and Goddard 1927) have asserted that these differences arose from the wearing of the rifling tools used to rifle the weapons. Booker (1980) has disputed this explanation, pointing out that rifling tools may be used to cut thousands of barrels without showing significant wear.

Many of the early experimental and theoretical studies on striation matching have been reviewed by Thomas (1967) and Dougherty (1969). For the most part the early experimental work arose out of the investigation of particular cases, and consequently lacked generality, while the theoretical studies were founded on untested assumptions about the random nature of striations on bullets or in toolmarks. In 1959 Biasotti (1959) finally published a statistical examination of striations and striation patterns on a number of fired bullets. This researcher fired bullets through both new and used Smith and Wesson .38 Special revolvers. Prior to firing, each bullet was marked to indicate its orientation in the cylinder. (This ensured that in subsequent microscopic examinations corresponding land and groove marks could be compared.) Biasotti compared bullets fired from the same pistols with one another, as well as with bullets fired from different weapons. He recorded the total number of striations in each land and groove mark, the total number of matching striations for each pair of bullets compared, and the frequency of series of consecutively matching striations. He found that bullets fired from the same weapon had 21–24% (for full metal jacket bullets) or 36–38% (for lead alloy bullets) of their striations matching. When bullets fired from different weapons were compared, as many as 15–20% of the striations on the pair of bullets could be matched. This illustrates the limited value of simply matching striations without considering whether or not they are part of a pattern. Biasotti found that even when using a very liberal interpretation of consecutive matching striations, 3–4 consecutive matching striations were very uncommon when bullets fired from different weapons were compared; on the other hand, when bullets fired from the same weapon were considered, patterns of up to fifteen consecutive striations could be matched.

From these results it appears that a pattern of five or more matching consecutive striations virtually assures that two bullets were fired from the same weapon.

Booker (1980) has strongly criticized the work of Biasotti, in particular for Biasotti's varying of his criteria for defining matching striations, depending on whether the bullets being compared were fired by the same weapon or not. Booker has also pointed out that if the number of fine striations present is large, the probability of a coincidental matching of consecutive striations is appreciable. To illustrate his point, Booker prepared two toolmarks, using a tool whose edge had been ground with 400 mesh abrasive particles (approximate diameter 40 μm). Five examiners were given photographs of the toolmarks and were asked to determine the maximum number of consecutive matching striations when the photographs were compared in mismatched configurations. The examiners all reported seven or more consecutive matching striations.

The most recent studies of this question are those of Uchiyama (1975, 1988) and Gardner (1978). Both of these workers developed similar probability models of striation matching. The presentation below follows that of Gardner.

Consider two bullets: Bullet 1 having n_1 striations and Bullet 2 having n_2 striations. We can assume without loss of generality that n_1 is less than or equal to n_2. When these two bullets are compared, n striations match to within a distance d. Because a striation on Bullet 2 will match a corresponding striation on Bullet 1 equally well if it falls anywhere within a distance d on either side of the striation on Bullet 1, we can imagine that the surfaces of the bullets being compared are divided into m equal intervals of width $2d$. Now calculate P_1, the probability that one and only one striation on Bullet 1 matches a striation on Bullet 2:

$$P_1 = {}^{n_1}\mathrm{C}_1 \left[\frac{n_2}{m}\right] \left(\frac{m-n_2}{m-1}\right) \left[\left(\frac{m-n_2-1}{m-2}\right) \cdots \left(\frac{m-n_2-n_1+2}{m-n_1+1}\right)\right]$$

where ${}^{n}\mathrm{C}_r$ is the combinatorial expression for $\dfrac{n!}{r!\,(n-r)!}$.

Similarly, the probability P_2 that two and only two striations on Bullet 1 match those on Bullet 2 is given by

$$P_2 = {}^{n_1}\mathrm{C}_2 \left[\left(\frac{n_2}{m}\right)\left(\frac{n_2-1}{m-1}\right)\right]\left[\left(\frac{m-n_2}{m-2}\right) \left(\frac{m-n_2-1}{m-3}\right) \cdots \left(\frac{m-n_2-n_1+3}{m-n_1+1}\right)\right]$$

Generally, the probability P_k that k striations on Bullet 1 will match those on Bullet 2 is given by

$$P_k = {}^{n_1}\mathrm{C}_k \left[\left(\frac{n_2}{m}\right) \cdots \left(\frac{n_2-k+1}{m-k+1}\right)\right]\left[\left(\frac{m-n_2}{m-k}\right) \cdots \left(\frac{m-n_2-n_1+k+1}{m-n_1+1}\right)\right] .$$

In combinatorial notation this becomes

$$P_k = \frac{{}^{n_2}\mathrm{C}_k \; {}^{m-n_2}\mathrm{C}_{n_1-k}}{{}^{m}\mathrm{C}_{n_1}} \; .$$

The probability *P(P)* that just as good a match could be obtained at random is given by

$$P(P) = 1 - \sum_{k=0}^{n-1} P_k \; .$$

A fundamental requirement for the application of probability models to the problem of striation matching is the demonstration that the striations in the land and groove marks are randomly distributed. Booker (1980) has suggested that these striations, in fact, may not be randomly distributed. However, the experimental results reported by both Uchiyama and Gardner are generally consistent with striations being distributed randomly according to a uniform distribution.

Uchiyama calculated *P(P)* for both right phase and wrong phase comparisons of full metal jacket .45 rounds and .22 long rifle rounds. In a right phase comparison land marks made by the same land are compared, while in a wrong phase comparison land marks made by different lands are compared. For the .45 rounds, the right phase comparison yielded *P(P)* values between $4.62 \times 10^{-8}\%$ and 35.49%, while the wrong phase comparisons gave *P(P)* values between 0.385% and 97.15%. The right phase comparisons of the .22 calibre bullets gave *P(P)* values between $2.36 \times 10^{-8}\%$ and 6.35%; the wrong phase comparisons gave values of *P(P)* between 4.56% and 96.05%. Davis (1958) has pointed out that wrong phase matches of striations on bullets fired from the same weapon are common. That being the case, it is difficult to see what useful conclusions can be drawn from this research. Certainly quite different *P(P)* values would have resulted had Uchiyama attempted to match striations on bullets fired from different weapons.

Uchiyama found that *P(P)* could be greatly reduced by including matching patterns of consecutive striations in the probability calculations. The criticisms that Booker directed at the work of Biasotti apply with equal force here: when the land and groove impressions contain a large number of fine striations, finding patterns of consecutive matching striations on bullets fired from different weapons may not be particularly difficult.

Even without using patterns of consecutive matching striations, Gardner obtained much higher values of *P(P)* than Uchiyama; however, his results cannot be applied to firearms examinations as they are conventionally performed. Gardner carried out his examinations of bullet surfaces by using a scanning electron microscope (SEM) The backscattered electron signal was digitized and then averaged over 700 to 1000 scans in the direction of the striations to suppress surface anomalies such as craters. The averaged signals were subjected to parabolic smoothing, converted to first derivatives, and then reduced to histograms. Subsequent probability calculations were performed on the data in this reduced form.

The work of Biasotti, Uchiyama, and Gardner is open to criticism on the ground that bullets fired from only a limited number of weapons were examined. Biasotti examined bullets fired from twenty-four .38 Special Smith and Wesson revolvers; Uchiyama examined bullets fired from a .45 Colt National Match semiautomatic pistol and a .22 Woodsman semi-automatic pistol; and Gardner's bullets came from four .38 Special Smith and Wesson revolvers. Given that there are a vast number of firearms currently available, rifled by at least five different methods, extrapolation of their results to all firearms is a leap of faith. Even granting that the striations in the land and groove marks are randomly distributed and that very small values of $P(P)$ have been obtained in a very limited number of experimental comparisons, the question of the uniqueness of the pattern of striations produced by a particular firearm remains unanswered. For the sake of argument, suppose that the four weapons used in Gardner's study were rifled with four different tools (as in fact was probably the case). In such a case, it would be unreasonable to expect much similarity in the patterns of striations on bullets fired from different weapons. On the other hand, there could be thousands of weapons rifled with each of the four rifling tools whose barrels contain identical or very similar imperfections and produce similar markings on the bullets fired through them.

This is probably not the case. Even if the rifled barrel of a weapon were microscopically identical to every other weapon of the same make and model when new, the patterns of wear and corrosion within the barrel would soon set it apart from all others, including those manufactured at the same time with the same machinery. In the cases of firing pins, breechblocks, extractors, and ejectors, the manufacture of each of these components of a firearm involves such hand operations as filing, milling, or turning on a lathe (Berg 1977). It is simply inconceivable that the patterns of tooling marks resulting from such production methods could be duplicated on the components of different weapons.

Firearms examiners are frequently called upon to aid in the reconstruction of shooting incidents by estimating the range from which a firearm was fired into a target (frequently a human being). The range of fire of a handgun or rifle is usually determined from the distribution of gunshot residue (burned and unburned powder, lead fouling, and primer residue) on the target around the bullet hole. The usual approach is for the examiner to try to duplicate the pattern of gunshot residue on the target by firing the weapon at different ranges into clean cloth targets, using ammunition from the same lot as that believed to have been used to fire the original gunshot residue pattern. Krishnan (1967, 1974) has explored the use of the distribution of antimony and lead at certain distances from the bullet hole to determine the range of fire. Those two elements (along with barium) come from the cartridge primers. Krishnan tried to fit his experimental data with a number of different functions, and found that the logarithms of the antimony and lead concentrations were linear functions of the logarithm of the range of fire. Unfortunately, Krishnan did not attempt to test this method of estimating the range of fire experimentally.

The range of fire of shotgun may also be estimated from the distribution of gunshot residue, provided that the muzzle was close enough to the target for such residue to reach it. In addition, the distribution of pellets in the shotgun pellet pattern may be used to estimate the range of fire. For ranges of fire beyond several feet, little detectable gunshot residue will be deposited on the target, and the size of the shotgun pellet pattern becomes the only basis for the estimation of the range of fire.

The problem of determining the range from which a shotgun pellet pattern was fired is one that seems to be amenable to the application of statistics. The range of fire may be determined by using the 'rule of thumb' that a shotgun pellet pattern spreads approximately 1 inch for every yard the pellets travel down range. While this is frequently a good approximation, some combinations of shotgun and ammunition may significantly depart from it. Another approach is to test fire pellet patterns, using the shotgun believed to have fired the pattern whose range of fire is to be determined as well as the same ammunition (preferably from the same batch and lot). The firearms examiner fires patterns at various ranges until he obtains a pellet pattern the same size and density as the questioned pattern. Although such an approach seems rather unsophisticated, it is capable of a remarkable degree of accuracy. Rowe (1988) has reported a blind study in which questioned patterns were fired at randomly selected ranges and a neophyte examiner was required to determine the range of fire of each pattern based on a visual comparison of the questioned pellet patterns with his own test-fired patterns. The ten randomly selected ranges varied from 6 to 41 feet. The average error was 1.5 feet and the average percentage error was 6.8%.

The sizes of shotgun pellet patterns may be measured in a variety of ways. Some of these are given below:

If D_H is the horizontal dispersion, D_V is the vertical dispersion, and N is the number of pellets, we can define

$$\begin{aligned} \langle D\rangle &= (D_\mathrm{H} + D_\mathrm{v})/2\ , \\ D &= (D_\mathrm{H} D_\mathrm{V})^{1/2}, \\ A &= D_\mathrm{H} D_\mathrm{V}, \\ d &= N/d_H D_\mathrm{V}, \end{aligned}$$

If x_i and y_i are Cartesian coordinates of the ith pellet hole in the pattern, we can define

$$S = \left\{ \sum_{i=1}^{N} [(x_i - \bar{x})^2 + (y_i - \bar{y})^2] \right\}^{1/2} \ ,$$

where

$$\bar{x} = \sum_{i=1}^{N} x_i/N \ ,$$

$$\bar{y} = \sum_{i=1}^{N} y_i/N \ .$$

Jauhari and his co-workers (1973, 1974) determined D_H, D_V, $<D>$, D, A, and d for pellet patterns fired with a .410 smoothbore musket. D_H, D_V, $<D>$, and D were approximately linear functions of the range of fire, while A and d were not. Mattoo & Nabar (1969), who first proposed the use of S, found that this quantity was an approximate linear function of range when it was applied to 00 buckshot patterns. Some examiners have also used the radius R of the smallest circle that will just enclose the shotgun pellet pattern, as a convenient measure of size of a pellet pattern.

The choice of a particular measurement of pellet pattern size is largely a matter of convenience. Wray *et al.* (1983) measured A, D, R, and S for a series of seventy-two 00 buckshot patterns fired at ranges from 5.2 to 10.7 m (17 to 35 feet). D, R, and S were all found to be linear functions of range, while A was a quadratic function. The computation of S is very time-consuming and therefore has little to recommend it for practical work.

Heaney & Rowe (1983) have suggested the application of linear regression to the problem of range-of-fire estimation. The methods of linear regression are familiar to most scientists, and most computer statistics software packages contain simple regression analysis programs. Most important of all, regression analysis provides a way of calculating the confidence limits of the range-of-fire estimate. Two other methods for determining confidence intervals have been proposed by Jauhari *et al.* (1972). In the first, no assumptions are made concerning the distribution of the sizes of the pellet patterns fired at a particular range (that is, no assumption of a normal distribution is necessary); the lower distribution-free confidence limits are obtained by interpolation between consecutive highest values of the measure of pellet pattern size, while the higher limits are obtained by interpolation between consecutive values of the lowest values of the measure of pellet pattern size. Normal-distribution confidence intervals may be calculated under the assumption that the test-fired patterns at each range are random samples of normally distributed populations. In that case, a fraction gamma of the intervals from

$$<\text{pattern size}> - ks \text{ to } <\text{pattern size}> + ks$$

should contain at least $100P\%$ of the population. Gamma is the confidence coefficient, P is the coverage, $<$ pattern size $>$ is the mean of the measure of pellet pattern size, s is the standard deviation of the pellet pattern size, and k is a numerical factor depending on gamma and P.

Rowe & Hanson (1985) carried out a blind study in which the ranges of fire of five 00 buckshot patterns and five No. 4 buckshot patterns were estimated by using regression equations. The regression equations were computed from thirty-six 00 buckshot patterns and thirty-six No. 4 buckshot patterns fired at ranges from 3.0 to 15.2m (10 to 50 feet). Because the standard deviations of the pellet pattern sizes increased with range, a weighted linear regression was performed on the pellet pattern sizes. Both the distribution-free and Normal-distribution confidence intervals were calculated and compared with the confidence intervals calculated by the methods of regression analysis. Several problems immediately emerged: for the No. 4 buckshot patterns it was necessary to carry out linear regressions on the highest and lowest values of pellet pattern size for each range of fire in order to obtain interpolation functions for the calculation of the distribution-free confidence intervals; the distribution-free confidence interval for one of the 00 buckshot patterns excluded the actual range of fire; and no upper Normal-distribution confidence limit could be calculated from the experimental data at all. The actual ranges of fire for the ten pellet patterns all fell within the 99% confidence intervals calculated by regression analysis.

A major problem with the application of regression analysis to the estimation of the range of fire of a shotgun pellet pattern is the large number of test shots needed to establish the regression line of pattern size versus range. At least twenty shots fired with the same lot of ammunition would be required to obtain useful results. Frequently the firearms examiner does not have a sufficient number of shotshells from the same lot of ammunition as that used to fire the questioned pellet pattern to undertake regression analysis. Fann *et al.* (1986) have examined this problem and proposed that because different brands of ammunition should have similar (although probably not identical) ballistic behaviour, a useful range-of-fire estimation could be obtained by test firing a different brand of ammunition, using the sizes of these test-fired pellet patterns to determine the regression equation, scaling the sizes of the questioned patterns, and calculating the estimated range of fire from the scaled pattern sizes using the regression equation for the other brand of ammunition. Fann *et al.* (1986) tested this hypothesis in a blind study, using questioned patterns fired with two different brands of 00 buckshot cartridges at randomly selected ranges. Sufficient test shots were also fired with each brand to establish regression lines for both. Scaling factors for the questioned pellet patterns were calculated as the ratios of the mean pattern sizes for the pellet patterns test-fired at a range of 15.2m (50 feet). The true ranges of fire all fell within the 99% confidence intervals of the estimated ranges of fire calculated from the scaled pattern sizes, despite the fact that the ballistic performance of the two brands of ammunition was sufficiently different that the sizes of the pellet patterns fired at 15.2 m (50 feet) were different at the 95% confidence level.

When only a portion of a shotgun pellet pattern is available, the firearms examiner can no longer rely on the size of the pattern in estimating the range of fire. A study by Nag & Lahiri (1986) attempted to address this problem. These workers argued from theoretical grounds that the distribution of pellets in shotgun pellet patterns ought to follow a Normal distribution centred on the geometrical centre of the pattern. Based on this model, the authors then calculated the number of pellets that should fall within 10-inch and 20-inch squares at various ranges of fire. The results agreed well with data published three decades ago by Burrard (1960). It should be noted that this study contained a number of flaws: the experimental test of the Gaussian model was indirect and less stringent than a direct examination of the distributions of pellets within test-fired patterns; moreover, changes in the manufacture of shotshells over the last thirty years make the reliance on Burrard's data suspect. Bhattacharyya & Sengupta (1989), on the other hand, have advanced a model of shotgun pellet distribution based on a Maxwell distribution of the radial components of the pellet's velocities. These researchers also presented experimental data in good agreement with their model. Unfortunately for proponents of both the Gaussian and Maxwell distribution models, Boyer *et al.* (1989) examined the distributions of 00 buckshot and No. 2 birdshot in patterns fired at a variety of ranges. These workers found that the distributions were in general not Gaussian; moreover, at the longer ranges the

No. 2 shot patterns exhibited bimodal distributions,which also contradicts the Maxwell distribution model. Obviously, the problem of estimating the range of fire from a partial pellet pattern offers opportunities for further research.

References

Austin, P.F. (1969) The identification of bullets fired from high-velocity rifles with consecutively rifled micro-groove barrels. *5th Triennial Meeting of the International Association of Forensic Sciences, Toronto.*

Berg, S.O. (1977) The forensic ballistics laboratory. In: C.G. Tedeschi, W.G. Eckert, & L.G. Tedeschi (eds) *Forensic medicine: a study in trauma and environmental hazard,* Vol. 1, W.B. Saunders Company, Philadelphia, PA.

Bhattacharyya, C. & Sengupta, P.K. (1989) Shotgun pellet dispersion in a Maxwellian model, *Forensic Science International* **41** 205–217.

Biasotti, A.A. (1959) A statistical study of the individual characteristics of fired bullets, *Journal of Forensic Sciences* **4** 34–50.

Booker, J.L. (1980) Examination of the badly damaged bullet, *Journal of the Forensic Science Society* **20** 153–162.

Boyer, D.A., Marshall, L.D., Trzicak, L.M., & Rowe, W.F. (1989) Experimental evaluation of the distribution of the pellets in shotgun pellet patterns, *Forensic Science International* **42** 51–59.

Burrard, G. (1960) *The modern shotgun,* Vol. III, Herbert Jenkins, London.

Ceccaldi, P.F. (1962) Examination of firearms and ammunition. In: F.Lundquist (ed) *Methods of forensic science,* Vol. 1, Interscience Publishers, New York.

Davis, J.E. (1958) *An introduction to tool marks, firearms and the striagraph.* Charles C.Thomas, Springfield, IL.

Dougherty, P.M. (1969) Report on two early United States firearm identification cases, *Journal of Forensic Sciences* **14** 453–59.

Fann, C.H., Ritter, W.A., Watts, R.H., & Rowe, W.F. (1986) Regression analysis applied to shotgun range-of-fire estimations: Results of a blind study, *Journal of Forensic Sciences* **31** 840–854.

Gardner, G.Y. (1978) Computer identification of bullets, *IEEE Transactions on Systems, Man, and Cybernetics* **8** 69–76.

Goddard, C.H. (1927) Who did the shooting? *Popular Science Monthly,* November, 21–22, 171, 173.

Hatcher, J.S., Jury, F.J., & Weller, J. (1957) *Firearms investigation, identification and evidence.* Stackpole Books, Harrisburg, PA.

Heaney, K.D. & Rowe, W.F. (1983) The application of linear regression to range-of-fire estimates based on the spread of shotgun pellet patterns, *Journal of Forensic Sciences* **28** 433–436.

Jauhari, M. (1973) Functional relationships between the range of firing and the various characteristics of a pellet pattern, *Journal of the Indian Academy of Forensic Sciences* **12** 5–9.

Jauhari, M., Chatterjee, S.M., & Ghosh, P.K. (1972) Statistical treatment of pellet dispersion data for estimating range of firing, *Journal of Forensic Sciences* **17** 141–149.

Jauhari, M., Chatterjee, S.M. & Ghosh, P.K. (1974) A comparative study of six parameters for estimating range of firing from pellet dispersion, *Journal of the Indian Academy of Forensic Sciences* **13** 17–24.

Krishnan, S.S. (1967) Firing distance determination by neutron activation analysis, *Journal of Forensic Sciences* **12** 471–483.

Krishnan, S.S. (1974) Firing distance determination by atomic absorption spectrophotometry, *Journal of Forensic Sciences* **19** 351–356.

Mathews, J.H. (1962) *Firearms identification,* Vol. 1, Charles C.Thomas, Springfield, IL.

Mattoo, B.N. & Nabar, B.S. (1969) Evaluation of effective shot dispersion in buckshot patterns, *Journal of Forensic Sciences* **14** 263–269.

Nag, N.K. & Lahiri, A. (1986) An evaluation of distribution of pellets due to shotgun discharge, *Forensic Science International* **32** 151–159.

Rowe, W.F. (1988) Firearms identification. In: Richard Saferstein (ed) *Forensic science handbook,* Volume II, Prentice Hall, Englewood Cliffs, NJ.

Rowe, W.F. & Hanson, S.R. (1985) Range-of-fire estimates from regression analysis applied to the spreads of shotgun pellet patterns: Results of a blind study, *Forensic Science International* **28** 239–250.

Thomas, F. (1967) Comments on the discovery of striation matching and on early contributions to forensic firearms identification, *Journal of Forensic Sciences* **12** 1–7.

Uchiyama, Tsuneo (1975) A criterion for land mark identification. *Reports of the National Research Institute of Police Science* **28** 179–187.

Uchiyama, Tsuneo (1988) A criterion for land mark identification, *AFTE Journal* **20** 3 236–251.

Wray, J.L., McNeil, J.E., & Rowe, W.F. (1983) Comparison of methods for estimating range of fire based on the spread of buckshot patterns, *Journal of Forensic Sciences* **28** 846–857.

THE USE OF STATISTICS IN THE INVESTIGATION OF POST MORTEM INTERVAL ESTIMATION FROM BODY TEMPERATURE

J.C.Wright† and **M.A.Green**‡

Introduction

For many fields of science, data collection and analysis may be carried out in the laboratory, and the analysis may be based on a set of known formula. The use of statistics in those investigations is typically as a means of supporting or qualifying a result. In the use of cadaveric temperature as a means of post mortem interval

† Digital Exploration Ltd., East Grinstead, UK

‡ Department of Forensic Medicine, St. James’s University Teaching Hospital, Leeds, UK

estimation, most pathologists recognize that the factors affecting the result are numerous and not well understood. There are many empirical formulas describing the relationship between time since death and temperature. Some of these have a sound theoretical basis. All of them make only basic attempts to take into account any of the unknown factors (Marshall & Hoare 1962, Marshall 1962a, Henssge 1979, 1981, Fiddes & Patten 1958, de Saram *et al.* 1955, Brown & Marshall 1974, Joseph & Schikele 1970).

The lack of any true theory forces the statistical analysis of data to be the means of deriving and improving an empirical relationship between temperature and time since death, rather than supporting an existing theory. It is thus necessary to analyse the data from a very detached standpoint, and to use statistical methods which do not detract from the fact that no known relationship exists.

Empirical formula and premises

At the department of Forensic Medicine, University of Leeds, a bank of data was collected which, for about 100 coroner's cases, was a record of cadaveric temperature over the range 2 to 23 hours post mortem. These data were assumed to obey an empirical formula:

$$\text{Post mortem interval} = F\ (\text{temperature}\ldots)$$

where F describes some function, and the dots indicate an acceptance that there are several other factors. The actual formula used is not of importance, except to note that the one used was chosen originally for its mathematical simplicity (Green & Wright 1985a, b).

Note that, strictly speaking, the above equation is inverted. The temperature is actually a function of the time. However, the equation more accurately reflects the aim that the post mortem interval may be calculated from the temperature.

The raw data were compared with the result derived from the empirical equation and the results were expressed as a percentage error, δ. For example, if the empirical formula indicates a post mortem interval of 9.0 hours, and the known post mortem interval is 10.0 hours, then δ is -10%. The experimental error in time and temperature measurement for the data collection was negligible (typically <3%), and so the calculated error is assumed to be due to the failure of the empirical formula to indicate accurately the true post mortem interval. This is an important point since, included in this error, is the effect of all the factors not taken into account in the formula.

As the formula is totally empirical, the following assumptions are made. The mean percentage error δ for the formula is 0. This simply means that the formula will give, on average, the correct result over many cases. Secondly, the distribution of the error is a Normal one. That is, the errors given by the formula *IF APPLIED TO A LARGE NUMBER OF CASES* will follow the Gaussian distribution.

The ensuing statistical analysis took the approach that the derived data, that is, the percentage error δ from the formula, were a sample of all the possible data. Thus, theory relating to small samples within a data set was applied.

Sampled data significance

The most common, and perhaps the simplest, test of the significance of a sample of data is the *'t'* test. The *t* test was used to assess the significance of the errors recorded.

There are other methods of assessing the significance of data such as those analysed here, where the number of samples is low. One of these is the chi-squared test, which is generally more applicable to situations where discrete events are recorded, for example, as an analysis of the frequency of numbers resulting from the throw of a die.

The equations shown below should be familiar to the reader: they are repeated here simply to introduce the symbols used for variables.

For large amounts of data which are expected to follow a Normal (Gaussian) distribution, the standard deviation *s* (see below) is a measure of the spread of the data. For Normal distributions, it is expected that approximately 95% of all the data will fall within the range

$$\bar{x} \pm 2s,$$

and that approximately 99.5% of all the data will fall within the range

$$\bar{x} \pm 3s,$$

where $\bar{x}$ is the mean value of all the data.

Where only a small number of data have been collected, the scatter of the results may cause more than 5% of the data to fall outside the range

$$\bar{x} \pm 2s$$

Conversely, and just as importantly, it is possible that more than 95% of the data may fall within the same range. Either of these cases would indicate that the assumption of a Normal distribution might not be valid.

The *t* test of the significance of a data set is an attempt to determine whether the scatter of the results might be due to the low volume of data, or if the scatter is due to an incorrect hypothesis or formula.

For a random sample of *N* measurements, having an expected mean μ and a measured mean $\bar{x}$, the statistic *t* is

$$\text{deviation } t = \frac{(\bar{x} - \mu)\sqrt{N}}{s}$$

where *s*, the standard deviation of the data, is

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{(N-1)}$$

(For the calculation of *s* where a large number of data *(N)* are available, the divisor used is *N* not *N* - 1, on the basis that almost the same result is obtained.)

Use of the *t* statistic

As has been noted, the original empirical formula was chosen from a mathematical viewpoint. It led to a form of equation which was not only mathematically simple, but also easily altered in order to improve its accuracy. The application of the *t* test offered a means of assessing the accuracy of the empirical formula, and then altering the formula to improve its accuracy. This becomes a classic case of making the theory fit the data.

The results of the original application of the empirical formula to the derived errors data are shown in Table 1 and Fig. 5.3

Table 1 —Analysis of data collected using the original formula

Time	*N*	Mean	SD
2	3	−6.3	13.1
3	16	9.0	29.2
4	34	17.3	38.7
5	47	23.1	37.9
6	54	18.1	32.2
7	62	17.1	28.4
8	64	15.7	22.8
9	66	12.5	21.5
10	66	12.8	20.9
11	66	6.2	18.3
12	66	2.8	17.5
13	66	0.0	17.0
14	61	−0.3	15.8
15	58	−1.4	18.5
16	56	−1.7	19.7
17	51	−3.8	13.3
18	44	−5.5	13.2
19	37	−4.1	15.7
20	30	−7.8	18.1
21	20	−5.7	16.7
22	14	−8.9	14.6
23	5	−8.6	17.2

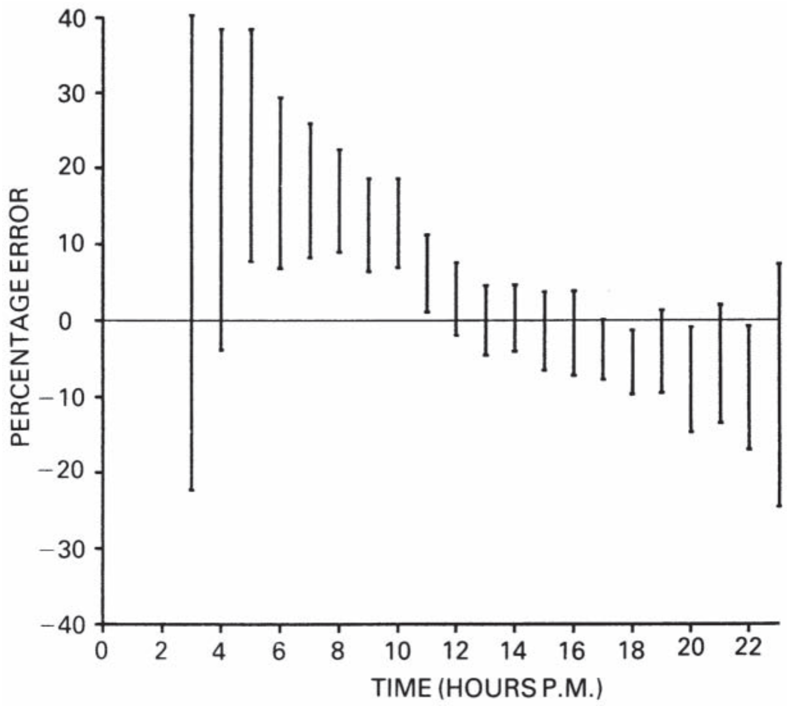


Fig. 5.3 —The 95 percentile ranges for the data of Table 1, original formula.

For each data set, at hourly intervals, the percentage error was calculated:

$$\text{Percentage time error } \delta = \frac{\text{formula time} - \text{known time}}{\text{known time}} \times 100$$

The number of cases available *(N)* at the true post mortem interval (Time) and the mean of δ (Mean) are listed. Also shown is the standard deviation (SD).

The *t* test range is shown in the bar graph. The bars indicate the range of error that would be expected for a fiducial limit of 95%. That is, for each of the sets of data at any one time, the relevant value of the *t* score for significance at the 5% level was used to find the upper and lower limits from the equations:

$$\bar{x} \pm \frac{s}{\sqrt{N}} \times t \ .$$

Note that the choice of the 95 percentile was based on the fact that for the majority of times, 20 or more cases were available for analysis. When using *t* score tables, remember that *N* data provide *N* - 1 degrees of freedom. For example, for 19 data, the value of *t* for the 0.05 significance level is approximately 2.10.

There are two immediate points to make about the ranges. The first is that the range of error is smallest where the number of cases *(N)* is largest. This is to be expected for two reasons. The first is that conceptually and statistically, we expect the results to improve as more data are collected. Morevoer, as *N* increases, the *t* test of the significance of the data improves, that is, we approach a better approximation to having sampled the whole data set. This underlies the meaning of the *t* test. It indicates how much the deviation is due to the data, how much the deviation is due to a lack of data, and, highly important in this case, it is a measure of the inability of the empirical formula to take into account any of the unknown factors.

The results of Fig. 5.3 show dramatically the failure of the first version of the empirical formula. For most of the time range recorded and analysed, the range of error does not include 0. That is, for this version of the formula, we would expect most cases to give a poor result. The *t* test has shown therefore that the formula is incorrect for our data. The next step was to change the formula according to the errors indicated by the first pass.

Formula revision

The requirement of the formula revision was that it should match the data as much as possible. In simple terms, since the mean error at, for example, 10 hours post mortem, is equal to +12.8%, we might decide to adjust the equation to give a value 12.8% lower at this time. Similarly, we could adjust the equation to give a value 6.3% higher at 2 hours post mortem. These percentages are, after all, the errors indicated in Table 1. This simple approach, however, ignores some of the information furnished by the statistical analysis. In particular, calculate, for time 10 hours, the *t* value:

$$\begin{aligned} t &= \frac{(\bar{x} - \mu) \times \sqrt{N}}{s} \\ &= 4.98 \quad . \end{aligned}$$

μ, the expected mean, is 0 since we are analysing percentage error. Referring to a set of *t* score tables, it is seen that the probability of getting such a *t* score for 66 cases is practically zero. Indeed, the above value is unlikely to appear in any set of *t* tables. For time 2 hours, however, the *t* score is 0.83. The probability of such a value occurring is about 0.5. In other words, the errors measured at 10 hours indicate a gross failure of the formula. Using statistical methods, they prove the formula to be in error. At 2 hours, because of the lack of data, it is quite reasonable to get such errors recorded.

When the formula was corrected, the probability of the *t* score was taken into account, and correction was not made where no statistical evidence suggested a need.

The results from the revised formula are shown in Table 2 and Fig. 5.4. More data were available for this second pass, and the effects of this are indicated by marginally lower standard deviations for the percentage error δ.

Table 2 —Analysis of data collected using the corrected formula

Time	*N*	Mean	SD
2	9	−12.1	32.4
3	33	−2.7	29.1
4	57	−2.4	36.3
5	64	1.7	32.2
6	66	2.2	29.2
7	78	4.7	28.8
8	78	2.3	21.8
9	78	2.5	22.1
10	75	2.2	20.8
11	74	−0.5	18.5
12	71	−1.9	16.3
13	70	−1.6	13.9
14	70	−2.0	14.5
15	68	−2.6	12.9
16	66	−0.9	13.9
17	64	−1.0	12.2
18	61	−0.3	12.8
19	57	−1.5	15.4
20	47	−0.5	13.5
21	36	−2.9	14.7
22	26	−0.9	13.8
23	16	−2.2	15.3

It should be noted that the results shown for the revised formula include a final revision to the formula which was not statistically derived. This extra revision ensured that the results of the use of the formula were consistent with the data at times between multiples of an hour. In effect, the formula was mathematically ‘smoothed’ so that its failure to give the ‘right’ result was consistent.

Thus, the new version of the formula is shown to give results that, for the 95% interval estimate, always include 0 in the error range.

Conclusions

A simple statistic, measuring the validity or significance of the results of a series of experiments, has been used to improve the performance of the data analysis. The same statistic is used to indicate the residual error when the improved analysis is used.

There is nothing to preclude further iterations of the adjustment to the formula. However, it is worth noting that the work done indicates the extent of errors to be expected by a forensic pathologist when using cadaveric temperature as a means of estimating the post mortem interval.

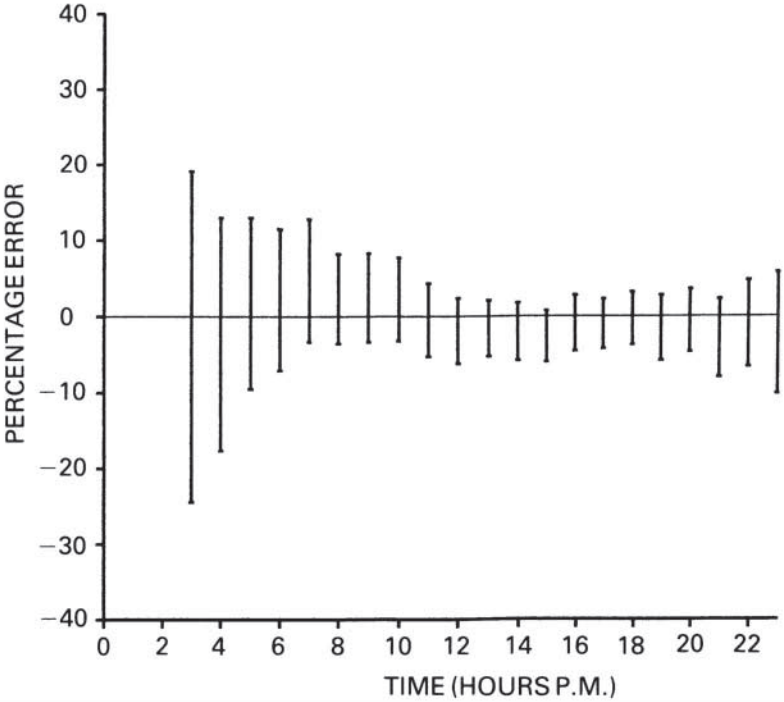


Fig. 5.4 —The 95 percentile ranges for the data of Table 2, corrected formula.

The data presented here were collected under relatively well controlled conditions. A pathologist at the site of a crime can, at best, use only the empirical formula, together with the statistics indicating the significance of its results, to arrive at an estimated range of times between which it is likely that the death occurred.

References

Brown, A. & Marshall, T.K. (1974) Body temperature as a means of estimating the time of death, *Forensic Science* **4** 125–133.

de Saram, G.S.W., Webster, G., & Kathigamatamby, N. (1955) Post mortem temperature and time of death, *J. Crim. Law Criminol. Pol. Sci.* **46** 562–577.

Fiddes, F.S. & Patten, T.D. (1958) A percentage method for representing the fall in body temperature after death, *Journal of Forensic Medicine* **5** 2–15.

Green, M.A. & Wright, J.C. (1985a) Post-mortem interval estimation from body temperature data only, *Forensic Science International* **28** 35–46.

Green, M.A. & Wright, J.C. (1985b) The theoretical aspects of the time dependent Z equation as a means of post-mortem interval estimation using body temperature data only, *Forensic Science International* **28** 53–62.

Henssge, C. (1979) Precision of estimating the time of death by mathematical expression of rectal body cooling, *Z.Rechtsmedizin* **83** 49–67.

Henssge, C. (1981) Estimation of death-time by computing the rectal body cooling under various cooling conditions, *Z.Rechtsmedizin* **87** 147–178.

Joseph, A.E.A. & Schikele, E. (1970) A general method for assessing factors controlling post-mortem cooling, *Journal of Forensic Sciences* **15** 364–391.

Marshall, T.K. (1962a) Estimating the time of death, *Journal of Forensic Sciences* **7** 189–210.

Marshall, T.K. (1962b) Estimating the time of death, *Journal of Forensic Sciences* **7** 211–230.

Marshall, T.K. & Hoare, F.E. (1962) Estimating the time of death, *Journal of Forensic Sciences* **7** 56–81.

6

Knowledge-based systems

John Buckleton and **Kevan Walsh**
Department of Scientific and Industrial Research, Auckland, New Zealand

6.1 KNOWLEDGE-BASED SYSTEMS

6.1.1 Definition

Defining a knowledge-based system, or the related term, expert system, is not entirely straightforward. If you picked up a standard computer text on artificial intelligence it might say of expert systems that ‘an expert system is a rule-based AI (artificial intelligence) application program for doing a task which requires expertise’ (Charniak & McDermott 1985). They (may) have some method of approximate reasoning often based on probabilities, fuzzy logic, or certainty factors for deriving a result even from imperfect data (Forsyth & Nayor 1986), although very forcible arguments have been advanced that only the first of these is necessary (Lindley 1987). We prefer the term ‘knowledge-based system’ partly because the term ‘expert’ has a special meaning in the forensic science context but also because an expert system is often meant to acquire knowledge incrementally, that is, to learn, whereas a knowledge-based system need not.

Knowledge-based systems are usually considered as a subset of artificial intelligence, the ‘clever tricks department of computer science’ (Forsyth & Nayor 1986). They may sometimes emulate human thought in that they can deal with uncertain or incomplete information, or deal with other aspects not usually considered the domain of the computer. The concept of knowledge-based systems is not new, having gained some acceptance in the nineteen-seventies. One of the best known expert systems is MYCIN, a system designed to diagnose and then prescribe treatment for an infectious disease, particularly a bacterial infection of the blood. Application of these systems to forensic science problems, however, is relatively new.

Most readers will be familar with the concept of a computer program where an algorithm is used to compute an answer, using a database. The knowledge-based system can be viewed as an evolutionary development on this theme, whereby an ‘inference engine’ is used to make an intelligent decision with regard to a knowledge base. The knowledge base will not only contain information or data but also rules by which decisions can be made. Whereas a program might only give an answer, a knowledge-based system should give advice or make a decision and be able to explain its reasoning.

We do not, however, want to be constrained by definitions of knowledge-based systems which are too narrowly constructed. If we were to do so, then many of our later examples would not qualify. The exact nature of the programming is of less importance to the operational forensic scientist than is the concept of what is possible with these systems. We will therefore include in this discussion any computer-based system that embodies knowledge within itself. Using this pragmatic concept we see knowledge-based systems as attempting to extend what can be done with a computer from the domain of mathematical algorithms and numerical processing into the areas of judgement, uncertainty, rules of thumb and, as a key element, an explanation of the reasons behind a result.

One of the reasons that there has been an increase in interest in knowledge-based systems in forensic science is the availability of relatively cheap software that can run on personal computers. The cost of starting has therefore come down such that it is affordable for forensic science use. There still is a very considerable cost in time for developing these systems.

6.1.2 The special nature of a knowledge-based system

The ability of a knowledge-based system to take on functions which were previously the domain of human expertise is achieved predominantly by the use of rule-based programming. The concept of a rule-based program might be foreign to some forensic scientists, but there is no need for this. An example of a rule might be, for instance, in a case of disputed paternity—‘if the alleged father has none of the genes present in the child then he is excluded as the true father of the child’. These rules could be generated in any field. For instance, a laboratory manager might have a set of rules for setting casework priority. These might be something like—‘assign high priority if the case is a homicide and all the samples are well packaged. Assign low priority if the case is a non-fatal traffic incident or if the samples are not correctly packaged’. These rules would continue until the manager’s system for setting priority was fully described. The assembly of a number of such rules comprises a body of knowledge often known as a knowledge base, and is the core of a knowledge-based system. This in fact corresponds in part to the way that human experts ‘think’, although it is less than clear exactly how human experts do think. The object of building such as system would, in the case of the fictional casework priority setting system, be that the receptionist could assign priority by using the manager’s rules without further reference to the manager. The system would embody the knowledge of the manager and be available to others.

6.1.3 Knowledge acquisition

There are two ways to acquire the knowledge inherent in a knowledge-based system. The first involves formalizing the knowledge of human experts into a set of rules that can be used in the building of the system. Obviously this requires the time and effort of experts who are often busy or uninterested in the project. This is not surprising;

the expertise that they possess is their personal property and probably took a great deal of time and effort to acquire. In many cases they may be reluctant, understandably, to hand this over.

The second method of knowledge acquisition is an aspect of artificial intelligence known as rule induction or machine learning. In this application an expert system shell is used to generate rules directly from the data without the need for the presence of a human expert. This was partly the programmer's response to the difficult task of getting busy and valuable experts to lay down their rules of operation in a computer-compatible form. The authors are aware of two instances of this in forensic science, both of which will be discussed later.

The authors feel that if the system is useful there is little point in debating whether proper use of rule-based logic has been made. In our experience of forensic science 'expert systems' (some of which would not qualify under a strict definition) we must say that no forensic scientist (but many a computer specialist) has asked us whether the system is goal driven with opportunistic forward chaining and how it handles uncertain information. Such questions we will leave to the specialists.

6.1.4 The reasons for using a knowledge-based system

The philosophy of a knowledge-based system as used in forensic science is markedly different from its equivalent in the commercial world. People advocating the use of knowledge-based systems generally highlight the main advantages in commerical terms and point to the fact that with such a system lay persons can perform 'nearly as well' as experts, which in turn results in obvious financial or competitive benefits. Commercially, then, one of the main criteria for choosing an expert system application is that expertise is rare and valuable.

Forensic science is different; the major objective of any system designed to aid the scientist is of course to make available to him the most reliable methods and background information. There can be no question of a system that would dictate an 'answer' to a lay person or anyone else for that matter. The legal system in most Western countries has taken a humanist approach. The decision making is done by the judge or jury, and it is expected of forensic witnesses that they deliver their own opinion based on their study and experience, certainly not the opinion of another individual, be it a person or computer system, knowledge-based or otherwise. For these reasons it is even more vital in forensic science that the system can explain the result and the logic behind it. Any forensic system that fails this criterion is less than worthless and possibly dangerous.

It is self evident to say that a knowledge-based system must be both feasible and useful. Applied to forensic science, however, this leads to a paradox, for two reasons —in some circumstances its use may be neither feasible nor useful. The most feasible system would be in fields where the knowledge is complete, exact, and well formalized. In fields such as this there is little need for a system since the subject is well understood and the methodology is well founded. In other fields the decision-making process is too complex or the knowledge is incomplete. This has been described as the Goldilocks syndrome (De Monchy *et al.* 1988): some tasks are too soft, some are too hard, and some are just right. A system in a field that is well understood still has the potential to relieve people of mundane jobs, but given the difficulty overcoming hearsay rules this aspect will be limited in forensic science. Herein lies the second paradox. Some of the most useful applications of knowledge-based systems must be in aiding the development or forming of expert opinion for court purposes. Expert opinion in many fields of forensic science relies on deductive reasoning involving many factors and drawing on vast databases, surveys, and other knowledge. Such a process is the ideal task for a knowledge-based system to attempt to duplicate. However, it becames a pointless exercise if no court will accept the findings of a computer *per se* or by virtue of the hearsay rule. This is discussed later in more detail.

The decision to implement knowledge-based systems in a forensic science laboratory could be made for a number of reasons, but often these systems will be viewed as tools with which to maximize laboratory efficiency. An identification of the most appropriate applications is then necessary. One of the first questions that should be asked is 'do our present personnel handle this field adequately, or effectively?'. If the expertise is well established and freely available and the experts are easily replaced, then there is little need for a computerized system. Likewise, if the field is well understood then it should be handled as well by a human expert. If the process simply involves calculations then a computer may be useful, but there is probably no point in using a knowledge-based system. In other areas it can be readily acknowledged that human expertise is far superior to computerized expertise. One such area is image recognition, where what is a fairly simple task for even non-experts is an extremely difficult task to duplicate by computer. Many forensic science methods involve some kind of image comparison, such as chromatographic identification of chemicals, shoeprint and toolmark comparisons, and fingerprint identification.

Attempts to computerize these functions have met with varied success, and where they are employed it is to handle the data or for the information-searching side rather than to perform the image recognition aspects. They help to eliminate subjectivity, surely, which can be treacherous in our field.

In forensic science a knowledge-based system would be well suited to a field of expertise which is large and complex. In such fields expertise is difficult to obtain as it requires a great deal of effort in training to become familiar with all aspects. This expertise becomes difficult to replace, and consistency among applications and among experts may be lacking. Formalizing the expertise in the form of a knowledge-based system not only enables consistency to be achieved but also guarantees continuity of expertise if, for example, an expert leaves the laboratory.

A field of expertise that is wanting in some aspect of its knowledge is also suited to the knowledge-based system. Such a field may have been poorly studied with incomplete and informal knowledge often containing uncertainties or contradictions. Except in extreme cases these factors can be handled by rule-based logic, and the resulting system, whilst markedly more difficult to build, will be very useful. At the very least, attempts to implement a knowledge-based system can, by virtue of the difficulty in system construction, identify those areas where data acquisition is required.

A major advantage associated with the implementation of a knowledge-based system is that it will be a force for consistency. Also, this consistency is coldly applied and takes no account of pressure or prejudice. It is difficult to envisage a computer being swayed by the emotive aspects of a case or making a mistake because of time pressure.

6.2 LEGAL CONSIDERATIONS

The potential exists in forensic science for knowledge systems to be used to assess the significance of evidence where the analyst has been involved only in the gathering of experimental data or in the input of certain parameters. For example, in a toxicological case, the analyst may prepare post mortem samples, inject the sample into a chromatographic system, and enter into the computer various case-specific details. The knowledge-based system, if linked to a gas chromatograph with mass selective detector, for example, could detect a chemical, identify it, and come to some decision as to its influence regarding the cause of death. Although the laboratory may have confidence in such a system the courts will probably not share this confidence and will either reject it outright or investigate it thoroughly. The problem (in this example) is that the witness is only a technician, and the true expert is the computerized system. Strictly, the computerized system should be the witness.

It is the legal constraints of forensic work that prevent, or at least restrain, the implementation of some knowledge-based systems in forensic science. Of particular difficulty are those legal systems that incorporate rules requiring the exclusion of evidence in certain circumstances.

The following discussion on the legal aspects relies heavily on points made by a prominent New Zealand judge (Prichard 1985), and the following quoted passages are taken directly from this reference.

The English law of evidence comprises one principle with five exceptions to that principle, known as the exclusionary rules, and a number of exceptions to the exceptions.

The basic principle is that all relevant evidence is admissible.

‘The five exclusionary rules are:

(1) The rule against receiving evidence of the witness’ opinion.
(2) The rule against hearsay.
(3) Privileged communications.
(4) The rule against hearing evidence of bad character, disposition or propensity.
(5) The Judge’s discretion, in criminal trials, to exclude prejudical evidence in order to ensure a fair trial.

The opinions of experts are received as an exception to the first exclusionary rule —but only if the following conditions are satisfied:

(1) The subject must be one in which competence to reach an informed opinion can only be acquired by a specified course of study or experience.
(2) The witness must be appropriately qualified.

(3) The witness must state all the facts on which he bases his opinion.
(4) The opinion must be fairly based on prove facts, using methods and principles which are recognised by scientists as being reliable.

'The fourth condition requires *(inter alia)* that each of the facts on which the opinion is based must be proved by evidence which is not excluded by the rule against hearsay.

'The hearsay rule is that the Court will not receive, as evidence of the truth of what is asserted, evidence of any assertions not made in Court by the witness who is testifying. In short, statements made out of Court are not admissible evidence if tendered as proof of the truth of those statements. Thus the opinion of an expert witness is inadmissible if the witness founds his opinion on any data (or any fact) to which he cannot testify from his own observation unless evidence of the accuracy of the data (or the existence of the fact) is given in to the Court by a witness who has the required knowledge from his own observation.

'In the case of computer-generated evidence the rule against hearsay requires that the Court be furnished with first-hand evidence of all data supplied to the computer. Data which are hearsay are not converted into admissible evidence by being reduced to writing or fed into a computer.

'However there are numerous exceptions to the strictness of the hearsay rule. One important exception applies only to the opinion evidence of expert witnesses. Expert witnesses are entitled to refer to and rely on information contained in text books and scientific publications in their own field of expertise.

'This relaxation of the hearsay rule is justified on two grounds:

(1) Because no scientist can know from personal observation more than a minute fraction of the data which he must use every day as working truths.
(2) Because an expert can be expected to know which text books and publications within his own field of expertise can be regarded as reliable sources of information.

'The fourth condition also requires that when a computer is called in aid by an expert witness there must be evidence as to the programming of the computer. And there must be evidence that the computer has been tested and found to be functioning correctly. A scientist who makes use of a computer may or may not be able to give evidence of these matters from his own knowledge. If he cannot, an additional witness or witnesses must be called to supply the evidence. (There is a restricted category of instruments which are presumed by the Courts to be trustworthy in the absence of proof to the contrary. The instruments in that category include clocks, thermometers, speedometers, etc. There is no such presumption in favour of computers).

'It is obvious that the foregoing conditions cannot be readily satisfied by computer-generated conclusions if the computer employs a so-called 'expert' or knowledge-based system.

'It is arguable that reference to a body of computer-stored knowledge is not distinguishable in principle from reference by an expert witness to the information embodied in text books and scientific publications. Provided that the Court is informed of the authoritative sources of all the computer's store of knowledge, is given an explanation of the rules distilled from these sources, and is told how these rules are applied by the computer to the facts of the instant case, a case could be made out for the proposition that the evidence is not based on inadmissible hearsay and is derived by a reliable method.

'However, there is a further difficulty. The principle underlying the law of evidence is that the Court will act only on evidence which is given by witnesses who are sworn to give truthful evidence and whose evidence is tested in open Court.

'A statement by an expert witness that he holds a certain opinion is a statement of fact. The weight to be attached to that statement will be measured, in part, by the validity of the reasoning by which the opinion is reached and, in part, by the experience, standing, and reputation of the witness.

'When the witness is a forensic scientist these are matters which can be explored and tested by cross-examination. But the same can hardly be said of a computer, which is incapable of a moral judgment, is not concerned to preserve its reputation, is not on oath, and cannot be questioned in cross-examination.

'In the case of a computer using a knowledge-based system a persuasive explanation of the computer's function would be essential and would inevitably involve a considerable body of technical evidence. The witness or witnesses who give that evidence would of course be subjected to cross-examination designed to demonstrate that the rules from which the computer output is derived are inadequate, not authoritative, or in conflict with other sources of knowledge and/or that the computer programme is otherwise not infallible.

'Whether the Courts will regard the cross-examination of witnesses who give evidence of the programming of knowledge-based systems as an adequate substitute for the cross-examination of a forensic scientist on his oath remains to be seen. It should not be thought that Judges are unreceptive of the benefits of modern technology. In time the position may be reached when the output of at least some knowledge-based systems which have been tried and tested in the forensic crucible will be accepted by the Courts as *prima facie* evidence without further proof.

'But there are manifest difficulties, and it is certain that defence counsel in criminal proceedings will be strenuous in their resistance to the admission of evidence which, in their eyes, is tantamount to opinion evidence emanating from a machine which cannot be cross-examined.

'As we see it, knowledge-based systems will have their place as tools for the assistance of forensic scientists; but it seems unlikely that computer-generated evidence using such systems will find acceptance as admissible evidence in Courts where the English legal system prevails.'

The use of a knowledge-based system to form an opinion which is then presented in court, without the scrutiny of a human expert, could obviously conflict with some of the exclusionary rules, but a well-designed system may overcome some of these constraints. The knowledge-based system should not function like a ‘black box’, which is the hallmark of most programs. Rather, the knowledge-based system should explain why a decision was made or advice was given and on what information it based the result. Such considerations should help to satisfy the need for the witness to state the basic facts upon which the opinion was based. The question of whether the inference is justified should have been established early during the setting up of the system, possibly by publication of the inference procedure followed by international scrutiny. There are obvious problems if the inference procedure upon which the opinion is based has not been proven, but the problem is the same whether a knowledge-based system is used or not.

Of course, knowledge-based systems are not the only computer-based generators of evidence for court purposes, and therefore this problem has been addressed before. However, there exists with knowledge-based systems the potential to draw conclusions which may not be possible by the analyst, alone or with very little human intervention (for example, see the discussion on Rule Induction).

The use of systems involving the interpretation of gas chromatography and mass-spectra has been contested in American Courts. In a heroin importation case (*Scientific Sleuthing Newsletter* 1983) where the scientific witness was unable to describe the program doing the mass-spectral searching and matching, the appeal Court ruled that the evidence was still admissible and that the issue was one of the weight to be given to the evidence. The expert was permitted to give his opinion that the substance was heroin, and the lower Court conviction was upheld.

However, in a case involving gas chromatography in conjunction with an integrator and a computer, a Court of Appeal held that the analyst was not qualified to give evidence that the computer was correctly programmed and maintained (*Holt v Auckland City Council* 1980).

It is important to realize then that the programmer would probably be expected to testify, at least to set the precedent. Problems may be encountered where a commercial package is used and the underlying programming is inaccessible or when the programmer has a proprietary interest in the program and will not disclose the program in discovery or at the trial. An example of such refusal during an actual trial is a patent violation case in the United States (Riesel 1986).

Aside from the legal problems likely to be encountered, the advantages to be obtained from use of knowledge-based systems are enormous. The training and knowledge-acquisition period of a forensic expert is lengthy, difficult, and expensive. The prospect of new forensic scientists, technicians, or clerks using the knowledge of fully fledged experts must be alluring to forensic managers. The choice of personnel involved can be tailored to the ultimate need of the task. If, ultimately, the output will be required for court, then the user must be fully conversant with the system and the system must be compliant to the needs of the scientist. Conversely, a clerk could use a knowledge-based system to set casework priority with little knowledge of the underlying decision processes. The consistency that could be obtained by universal use of a knowledge-based system is a factor that also must strongly recommend their use.

6.3 EXAMPLES OF USE IN FORENSIC SCIENCE

6.3.1 Introduction

The advent of knowledge-based systems was met with much publicity and excitement. To an extent, they were built up to do more than they can deliver. It is true that well designed systems can reason under uncertainty and can in some ways emulate the human thought process, but achieving these goals is a substantial task, and the methods available are in their infancy. Computer programmers find this difficult, and some of the early enthusiasm has evaporated.

Forensic science is just beginning to be exposed to knowledge-based systems. It is important that they are not oversold. Most of the current forensic science systems are pilot projects or early prototypes, and a realistic view must be taken of them. If we accept that the concept of a knowledge-based system in forensic science is totally different from other applications because of the legal and other constraints, then there would be no point in discussing non-forensic knowledge-based systems. We propose, therefore, to discuss the two instances of rule induction and the three forensic knowledge-based systems with which we have first-hand experience either as part of the team building them or because they have been made available to us for examination. The use of knowledge-based systems in forensic science is in its infancy, and the following examples represent the majority of forensic science expert systems in existence.

Not all of this work has been published, and this limits what can be said.

6.3.2 Early developments in forensic science

Several workers at the Home Office Central Research and Support Establishment in England were amongst the first to recognize that knowledge-based systems would have a place in forensic science. Lambert and Evett were involved in the interpretation of glass evidence in the late seventies and early eighties, and it was these workers who saw an application for a knowledge-based system in this field.

Experimental work was undertaken in conjunction with Queen Mary College with the object of producing a working prototype for the accumulation of forensic glass data and expertise that could be used to demonstrate the technology. This work resulted in a report in 1983 (Efstathiou *et al.*) which contained some salient points which are as valid today as they were then:

> ‘Expert systems can be expensive to install because of the complexity of knowledge elicitation . . . . Experts are busy and in demand, so it is difficult to train them to express their knowledge in a computer-compatible form’.
>
> ‘… the organization was concerned to improve the performance of their experts within the sensitive domain of providing evidence that would be used in Court. It was important that the experts perform in a way that is fair and consistent and could be seen by the jury to be so.’
>
> ‘… The experts found that the expert system provided a good way of expressing their knowledge and ensuring consistency and thoroughness amongst experts.’

Even though this work was described by the authors as a pilot project and these comments were written very much at the beginning of forensic involvement in knowledge-based systems, it must be said that they are as pertinent today as they were, although ‘a lot of silicon has flowed under the bridge since then’ (Forsyth & Nayor 1986). The goals of consistency and thoroughness are very worthwhile goals in forensic science.

The next development in the field of glass evidence was the investigation by Evett and Spiehler of the use of rule induction for the classification of glass samples as window or non-window, float or non-float, from elemental analysis (by scanning electron microscopy) and refractive index (Evett & Spiehler 1988). The object was to produce a set of rules by which glass samples could be classified from analytical data. These rules were then built into a rule-based system.

Much of the following discussion is taken from this work.

Simply stated, rule induction packages seek to form rules from the data themselves rather than by asking human experts. As such, this constitutes an aspect of artificial intelligence often called machine learning. The package used was BEAGLE (Bionic Evolutionary Algorithm Generating Logical Expressions, Warm Boot Ltd., Nottingham, UK). In contrast to some packages which attempt to model the way in which human beings acquire knowledge, BEAGLE, in some ways, mimics the process of natural evolution.

BEAGLE requires data and a goal. In this case the analytical data were obtained from samples known to be window or non-window glass, float or non-float glass. The goal was of the type, ‘find rules for deciding whether or not a sample is window glass’.

A human analysis would quickly show that magnesium levels tend to be higher in window than in non-window glass. A simple rule might, therefore, be that ‘if magnesium levels are greater than zero then the sample is a window’. Such a simple rule would misclassify all those non-window glass samples which had measurable magnesium content. Refining the rule might, therefore, be worthwhile. One possibility is ‘if the magnesium level is greater than zero and the aluminium level is less then 1.5% then the sample is from a window’. This rule could be checked and its performance in correctly classifying samples compared with the first rule. This cycle of rule modification and subsequent assessment of performance can be repeated.

BEAGLE follows this sort of procedure, generating rules and assessing them. The poorly performing rules are deleted and the survivors are modified.

The highly successful experiment by Evett and Spiehler using BEAGLE produced a set of rules that outperformed both nearest neighbour and discriminant analysis in classifying unseen glass samples. Out often trials the BEAGLE rules gave one incorrect response, whereas discriminant analysis gave three errors and nearest neighbour gave two errors. BEAGLE is inexpensive and warrants serious consideration in interpretation problems.

A salutary word of caution from the manual is included. ‘There is no magic in BEAGLE, so it cannot find patterns that do not exist. It may be that you are expecting to predict the price of gold in London from the number of motor accidents in New Zealand (so to speak). This example is somewhat far-fetched, but it is all too easy to collect data that are available rather than data that are relevant’.

6.3.3 CAGE: Computer assistance for glass evidence

Of the knowledge-based systems to be discussed this one is closest to the interpretation issues that form the theme of this book. It is also the one of which the authors have the most experience. We will, therefore, discuss it in more depth.

CAGE is a knowledge-based system written at the Home Office by Buckleton, Pinchin, & Evett (1990) with copious help from operational staff. Its purpose is to assist the forensic glass examiner in interpreting forensic glass evidence. The concepts used are mostly forward chaining and deterministic, in that it steps forward in a logical chain towards the solution. As such, the system must be described as simple in that it does not deal with uncertain or incomplete data. It is written in CRYSTAL (Intelligent Environments Ltd) with some arithmetic aspects performed externally in TURBO PASCAL.

The knowledge contained within CAGE is derived from three sources: the Bayesian theory for interpreting glass cases (Evett & Buckleton 1990) developed at the Home Office, surveys regarding the random incidence of glass on clothing (McQuillan 1988, personal communication), and the ability to select the appropriate data elements in the case in question. The system also advises on any pertinent decisions or judgements required in the interpretation, and offers advice relating to the reporting of the result. There is also a feature to assist in explaining the logic and reasoning behind the decisions.

CAGE was built with the assistance of a large number of operational staff, and has been tested on them. A quote from the assistant director of one of the participating laboratories sums up the project concisely: The value of the system lies in the way it draws the reporting officer's attention to the relevant point of issue and demands that they use their informed judgement in a structured manner to reach a proper conclusion.'

It is too early to say whether this system will find a place in the routine of forensic glass examiners, but early indications are positive.

As CAGE is the only one of the systems to be discussed whose purpose relates to the interpretation of transfer evidence, certain points relating to this, but generally applicable to knowledge-based systems, should be made here.

One of the key tasks of the forensic scientist is to offer rational advice about an unknown event founded on evalutions of the accumulated evidence and based on his background knowledge. In the context of glass evidence the unknown event is often 'the suspect is the person who broke the window'. The evidence will be glass fragments on the clothing of the suspect indistinguishable in physical properties from the broken window at the scene of the crime. The background knowledge of the forensic glass examiner will relate to occurrence of glass on clothing, the frequency of glass of this 'type', and some feel for the probability of glass transfer and retention in the circumstances of the crime.

The key 'breakthrough' in this context has been the growing acceptance of Bayesian inference as the framework within which to evaluate the evidence. There is nothing new about Bayesian inference: the ideas of the Reverend Thomas Bayes were published in the eighteenth century; see Chapters 1 and 3 for a more detailed discussion of these ideas. It is, however, only in the nineteen eighties that there has been widespread acceptance of his methods in forensic science circles. For this the forensic community must thank those 'pure' statisticians who took an interest in forensic science matters and a dedicated few forensic scientists who realized the worth of these initiatives.

In the context of glass evidence Bayes' Theorem directs us to consider the ratio of two probabilities: the probability of the evidence if the suspect is the person who broke the window and the probability of the evidence if the suspect is not the person who broke the window. The ratio of these is called the likelihood ratio or Bayes factor and is the key to expressing the evidential value that we are seeking. The theoretical background to this argument is given in Chapter 3.

Inspecting first the denominator, we are led to consider the probability of glass fragments of this type being present if the suspect is not the person who broke the window. Typically, forensic scientists use the concept here of a random man; that is, if the suspect is not the person who broke the window then we will consider him to be a person selected at random from the population. (See Chapter 2 for a discussion of the meaning of 'population'.) We will therefore, in our interpretation, need to understand the nature of glass fragments on the clothing of persons selected at random from the population. In CAGE this is provided by a survey of persons unconnected with crime conducted by McQuillan at the Northern Ireland Forensic Science Laboratory in Belfast. These data are held within the program and may be selected in a form to correspond to the specific needs of the case in question. In addition, graphical displays are made to assist the scientist in judging for himself whether intervention in some or all of the terms is warranted. This is consistent with the policy that the forensic scientist must dominate the system and not let the system dominate the forensic scientist.

Turning to the numerator of the likelihood ratio we are faced with a more difficult task. Here we are directed to consider the probability of glass fragments of this type being present if the suspect is the person who broke the window. This relates to the probability that glass fragments will be transferred to the clothing during the breakage, retained, and subsequently found by the examiner. Obviously this will be very different in every case, depending on the nature of the clothing, the elapsed interval since commission of the crime, the packaging of the clothing, and a myriad of other factors.

A review of the literature and unpublished studies produced only two works which were considered relevant. These were unpublished transfer experiments done at the Metropolitan Police Forensic Science Laboratory and studies of the retention of glass on clothing done at the Home Office Central Research and Support Establishment (Pounds & Smalldon 1978). These data were judged to be scant, and therefore considerable efforts were made to improve the quality of advice offered at this point.

After some consideration of possible mathematical models it was considered that the 'best' approach to take was based on the premise that the body of experience possessed collectively by the operational glass examiners might represent the 'best' knowledge available on transfer. It was, therefore, decided to collect the opinions of forensic glass examiners and add these progressively to the knowledge of the system. Taking this approach, the system was then able to advise on transfer based on the scant literature material available and on the opinions of other examiners in cases similar to the one is question.

It must be obvious to the reader that this would tend to favour consistency of assessment of transfer probabilities. This should generally be beneficial. Obviously, if the opinions stored are erroneous then the advice will be erroneous. In such a situation the system would be a force for consistent error. Again, there is no magic in this knowledge-based system. In the words of Spiehler *et al.* (1988) The more expertise (knowledge) that is entered into the expert system, the better it can evaluate data and solve problems. Conversely, if little expertise or faulty or incoherent knowledge is entered into the system, it will likely yield poor results.'

Having assessed the numerator and the denominator of the likelihood ratio by a combination of objective survey data and subjective opinion, the system then supplies computational power to perform the arithmetic.

A final module attempts to associate verbal equivalents to the numerical result. An instance of this is that a likelihood ratio of between 1000 and 10000 is translated into the statement 'the evidence very strongly supports the suggestion that the suspect is the person who broke the window'. No doubt this aspect of the system will be contentious and will cause some debate. This is not the first such attempt at a verbal scale for evidence (Brown & Cropp 1987, Aitken 1988). It has the advantage of being firmly anchored in a logical framework (Bayesian inference). In any case, debate will be beneficial and the system can be modified if a different scale is desired. The full scale, as far as it extends, is presented in Table 1.

Table 1

Likelihood ratio	Verbal equivalent
$1 < \mathrm{LR} < 10$	The evidence slightly supports the suggestion that . . .
$10 < \mathrm{LR} < 100$	The evidence supports the suggestion that . . .
$100 < \mathrm{LR} < 1000$	The evidence strongly supports the suggestion that . . .
$1000 < \mathrm{LR} < 10000$	The evidence very strongly supports the suggestion that . . .

A likelihood ratio of 1 represents inconclusive evidence, and forensic scientists may wish to describe evidence with a likelihood ratio greater than 1, but not much so, as inconclusive also, in order to be conservative. This would depend on the confidence with which the estimates involved in the calculations were held.

At the request of the scientist, the system will display the most significant, in the forensic sense, terms (objective and subjective) used in the analysis and will demonstrate the arithmetic pathway to the result. This very key feature was not built by using the intrinsic 'why' function of CRYSTAL, which displays the rules assessed and the results of that assessment, but rather was built specifically to assist with the explanation of the mathematics. It has already been mentioned that one of the key requirements of a knowledge-based system is that it can explain itself. This is even more essential in forensic science than in other applications. The authors intended the scientist to reach this stage and critically reappraise all significant terms so that the scientist can be confident that an appropriate and (if necessary) conservative assessment has been made.

It was recognized at an early stage that one of the contributions that knowledge-based systems could make to forensic science is in promoting consistency. Many surveys have shown (e.g., Kind *et al.* 1979, Lawton *et al.* 1988) that forensic scientists are inconsistent both in their choice of words and in the interpretation of case data *per se.* Equally there have been a number of calls for consistency (Brown & Cropp 1987, Kind *et al.* 1979). These attempts at promoting consistency have not been successful, and this is regrettable. Associating phrases with numbers to express the evidential value of transfer evidence is a difficult and dangerous task. However, if agreement of sorts could be reached, then associating such a scale with a knowledge-based system must make a contribution towards consistency of reporting.

The following is a quote from the paper on CAGE (Buckleton *et al.* 1990).

> 'In fact it must be realized that all evidence exists somewhere on the scale between conclusive exclusion and conclusive identification, and it is not totally in the realm of fiction to imagine one international reporting scale for all evidence whether it be glass, fibres, classical serology, or DNA profiling. The path to such consistency must be widespread discussion, but a system such as this (CAGE) might function as a catalyst.'

This system, if ever used extensively, must assist in obtaining consistency in the interpretation. This has already been discussed under the assessment of transfer probabilities, but there are many other aspects in which more consistency could be advanced. Not the least of these would be the acceptance of the Bayesian framework itself.

It is accepted amongst knowledge engineers, that is, professionals who build knowledge-based systems, that in the process of encapsulating knowledge into an expert system the expert will be forced to formalize his own rules for decision making. This may turn out to be the first time that the expert has had to do this, and it will be educational for him to do so.

In the CAGE project this process resulted in clearly identifying those areas in which the knowledge of the experts or from the literature was weak or absent. In particular, the area of probabilities of transfer and persistance of glass on clothing were identified as requiring attention. Such revelations are not in any way detrimental, but rather, very constructive. They can be used to point clearly the way for further development. This aspect was widely accepted by those persons testing the CAGE system, and, we believe, by managers in charge of the Home Office research effort.

This beneficial aspect is not peculiar to CAGE but is an expected benefit from the construction of any knowledge-based system.

6.3.4 Toxicology of amitriptyline

Described here is a study undertaken at the Home Office by Spiehler *et al.* (1988). This is one of the two instances of rule induction in forensic science of which the authors are aware. We have had the pleasure of viewing one of the systems produced from the machine-generated rules.

The project was designed to study the applicability of commercially available expert system shells to interpretation in forensic toxicology. Expert system shells possess the inference mechanism of a full expert system, but lack the knowledge base. To use them one must supply the knowledge base.

The goal of this work was to advise on dose, time since ingestion, and effect. It was proposed to undertake rule induction on the raw data rather than to build a system from 'human' rules.

The toxicology of amitriptyline was selected as the trial field since conventional pharmacokinetic models had not proved useful. In many ways this was a stringent test. As has been mentioned earlier, there is no magic in rule induction, and if no pattern exists in the data then no software package can find one. The failure of conventional methods therefore made this a difficult and worthwhile test.

Table 2 —Inference of time since ingestion: BEAGLE

Target time since ingestion	Success rate
<7 hr	92%
>7 hr and <18 hr	88%
<18 hr	88%
>18 hr	92%
<7 hr or >18 hr	92%

Rule induction (or machine learning) was undertaken, using the expert system shell Expert 4 and BEAGLE (previously described). Fatal amitriptyline case data were obtained from the available literature and from the Registry of Human Toxicology of the Home Office Central Research and Support Establishment. Both of the shells were able to derive rules and associated probabilities from these data for use in a rule-based expert system advisor. Two sets of rules were generated, one to advise on time since ingestion and one to advise on the probability of the amitriptyline overdose being directly responsible for a fatality. The Expert 4 shell was also used to advise on dose. The accuracy of these rules is shown in Tables 2 and 3. More complete data appear in the original paper.

It is interesting to compare the machine-induced rules with human opinion on the subject. Again in the words of Spiehler *et al.* (1988), 'both expert systems, using quite different models of knowledge and diagnosis and different statistical measures, have confirmed the hypothesis proposed by Bailey *et al.* (1978, 1979, 1980),...that a parent/metabolite ratio greater than 3 suggested a short survival time, and, based on the pharmacokinetics of amitriptyline, that the ratio would decline with increasing survival time.'

Table 3 —Inference of fatal response: BEAGLE. Performance of rules

Rule	Rules considered one at a time			
	True positives	False positives	False negatives	True negatives
1	38	2	1	12
2	28	1	11	13
3	28	0	11	14

Another finding of machine-induced logic was 'that tissue (heart and liver) amitriptyline/nortriptyline ratios and tissue nortriptyline concentrations are the most useful toxicological data for interpretation of time since ingestion'.

These findings and the resulting knowledge-based system are valuable *per se.* This is, however, not the end of the benefit from this study. As is typical for the creation of a knowledge-based system, Spiehler *et al.* were able to make general comment as to how future data could be collected. '...(T)o capture the knowledge base of interest in forensic toxicology, toxicology registries must document the targets (dose, time, and response), contain determinations with high information value, and must have been maintained to reflect the occurrence of cases in the population. By suggesting the information value of various analytical tests, expert systems can be useful in the design of registry database collections and in the management of laboratory protocols.'

It was also found '...that heart or liver amitriptyline and nortriptyline determinatons contain more information about the critical decision levels for case interpretation than other measurements. This would suggest that these analyses are more cost-effective than others'.

The architects of these systems do, however, offer self-criticism of the ability of their systems to explain how and why conclusions were reached. Work is continuing in this key field.

Dr K.Bedford and members of his staff from our laboratory kindly consented to comment on the system. The following comments are taken from a set of trials where the system's interpretation was compared to that already arrived at by the human experts. Dr Bedford felt that amitriptyline was a moderately difficult drug to interpret, citing that it had a good, easily quantifiable metabolite and had been extensively studied. Complicating factors, such as interpreting cases involving chronic users and the effect of other drugs, were mentioned. In a number of tests the system agreed with Dr Bedford's interpretation and also those of his staff, who varied in experience. Comments were also made on the friendliness of the system. From comments made, it was clear that Dr Bedford was applying a set of rules not dissimilar to the ones described above; he also had with him literature to which he referred when making his decisions (Stead & Moffat 1983). The system does not appear to handle the two complicating factors mentioned above, the former of which is not thought to be a problem with amitriptyline. For Dr Bedford, who has considerable experience, the system was unnecessary, although the potential as a ‘second person to talk to’ was apparent.

Other members of the staff valued the advice and support offered by the system. The ‘second person to talk to’ aspect was mentioned, and it may be that our examiners would have more confidence in their own interpretation after discussing it with the system. It was also mentioned that the system was quicker than looking up the texts, which are based on previous cases much as this system is.

If Dr Bedford is using a set of ‘rules of thumb’, it would be very interesting to build a system based upon these and to compare its performance to the system based on the machine-induced rules.

Spiehler *et al.* described their project as a pilot study, and it is very successful in this context. Comments arising from users are that they would like to see it expanded to a larger variety of drugs and to handle multi-drug cases, especially those involving alcohol. They would certainly find a use for systems of this type.

6.3.5 TICTAC

TICTAC is a system designed to assist in the identification of a tablet from a physical description. It contains a large database and a search program written in BASIC. This program is top-ended, that is, it communicates with the user by using either Xi PLUS or CRYSTAL. The project is under the direction of M.D.Ossleton of the Home Office Central Research and Support Establishment, whilst much of the programming has been done by J.Ramsey of St Georges Hospital Medical School, London. This system is not a good example of rule-based logic. The two rule-based languages are used primarily for their splendid interfacing with the user, whilst the bulk of the work is done by the BASIC program and the bulk of the knowledge is in the database rather than in the rules. It is included here, nonetheless, because it is quite clearly a system that incorporates a very considerable body of knowledge and is of enormous utility in forensic science.

The authors have had the pleasure of using TICTAC, but could not see the programming, nor has this material been published. We therefore wish to thank M. D.Ossleton and J.Ramsey for allowing us to use this material.

Because of these facts our comment must, necessarily, be restricted to a description of the capabilities of the system. The database is now quite considerable with over 6000 records. The primary function of TICTAC is to identify a tablet from its physical description. Details are therefore elicited from the user regarding whether the tablet is marked, the dose form, the availability, whether it is scored, the tablet coating, the plan shape, the elevation shape, colour type (mottled/layered/ uniform/etc.), and the colour. It is possible to reply ‘undecided’ to these questions. The Xi PLUS version achieves this dialogue with the use of very attractive overlay menus that one of the authors (J.Buckleton) was able to drive with minimal instruction, despite not being a drug analyst. The CRYSTAL version, whilst also friendly, did not present such an attractive appearance. The authors believe that this should not be viewed as a criticism of the CRYSTAL package.

The system then attempts to identify the tablet from these data. Obviously, this may result in a unique solution or in multiple possibilities. To assist further in the analysis the system possesses knowledge relating to product name, product licence number, and active ingredient, all of which can be made available to the user. In addition, company details, such as name, address, telephone number, and manufacturer's code, are available.

Further knowledge is available under the heading 'Drug search table'. This option allows the user to input analytical results such as retention index on OV1, retention index on poly A-103 and UV maximum, and then search the database for a drug, or drugs, fitting the analytical data.

6.3.6 Forensic serology advisory system

The authors were allowed to use this system, sent to the Home Office by J. Rynearson who works at the California Criminalistics Institute. We could not, however, see the programming; the system came as a package written in two languages. There was available, however, associated program documentation giving detailed description of the purpose and function of the system. The following write-up is taken from that documentation and from our experience with the system. This work is not as yet published in full, but a report was presented at the Tenth Australian International Forensic Science Symposium in Brisbane in 1988 (Rynearson *et al.* 1988). This will restrict what we can say about the system. The authors wish to thank Mr Rynearson for the opportunity of using this system. In many ways this system is an exemplar for knowledge-based systems and the methods for building them. Accordingly, some discussion will be made of the approach followed by Mr Rynearson.

The forensic serology advisory system has the following purposes:

> 'To provide the analyst an advisory tool offering recommendations for technical analysis of bloodstains in criminal investigations. (To) integrate the existing written and unwritten expertise dealing with storage conditions instead of the usual alphabetical list. (To provide) access to technical information which will assist in the selection of viable markers for examination combined with the subtle precautions in the performance of actual analysis. (To perform the) calculation of simple frequency distributions of genetic markers within a population of predetermined racial background.'

The system presents a very attractive interface to the user and was easily used by the authors without instruction. To simplify it excessively, there are two primary functions, 'characterize' and 'discriminate'. We will discuss firstly the 'discriminate' function which is intended to advise on appropriate tests to perform to discriminate between the suspect and the victim, having regard for the quality and quantity of the stain.

The user is asked to input the whole blood types of the suspect and the victim and the typing systems and methodology available at the laboratory. Next the stain is described. Information such as stain age, colour, odour, packaging, size, temperature of storage, substrate (that is, whether it is on cloth etc.), time available for the analysis, and the reaction to initial tests (screening/ species/ etc.), are input. The strength of reaction to initial tests is evaluated for compatibility with the history of the stain. The markers in which the suspect and the victim differ are then determined.

This list is then modified to select those markers which are the most discriminating and which are still viable, given the history of the stain. If the analysis is required in a short time, then the quickest methods are selected. If the quantity of stain is small, then those markers and methods consuming the least sample are recommended.

The second option of the system is the ‘characterize’ function. This does not require the whole blood types of the suspect and the victim but rather seeks to characterize the stain fully. The factors of viable markers, the most discriminating markers, the quantity of stain and the time available for analysis, are considered as before in the light of knowledge about the methods and marker systems.

The recommendations are accompanied by an explanation of the reasons why a particular system was recommended to the user and provides cautions and special areas of emphasis.

The combined frequency of occurrence of the blood markers is calculated for a particular race.

The whole of this system is supported by a reference system which directs the user to the original literature.

Mr Rynearson has used the term ‘electronic reference book’ with regard to this system, a term which the authors feel is most apt and demonstrates one of the uses of a knowledge-based system.

To describe the knowledge inherent in this system without seeing the programming is dangerous, but it must be apparent to the reader that a considerable body of knowledge is encapsulated in the system. This knowledge must relate to the viability of various systems under different stain histories, the quantity required for particular analyses, the time for particular analyses, and the discrimination of the various marker systems. All of this is assessed in view of the methodology available at a particular laboratory.

Some but not all of the serologists to whom the authors have shown the system have suggested that the bulk of this decision making is relatively straightforward and that, therefore, a machine is not required. However, this will, to an extent, depend on the number of systems in use by the laboratory. The greater the number of systems, the more complicated the decision process becomes. This is particularly so in relation to ageing effects on the viability of the stain for grouping. The authors feel that this system should be a factor in developing a rational and consistent approach to the selection of tests for serology, and would like to see it tested against human operators. It may be that human decision making could, in some cases, be seen to be erratic, inconsistent, not optimal, or *ad hoc.* This could be tested only by experiment.

The process used in building this system is, in many ways, typical for the building of a system from human experts (as opposed to rule induction). The initial prototype was built by using input from a limited number of experts and a literature review. This stage is often described as knowledge acquisition. This is followed by critical appraisal by a larger field of experts and leads to the incorporation of additional expertise. This can obviously be cycled until the system is developed to a level where its performance is adequate. At this point, if not earlier, the documentation must be produced and the users introduced to the system. This last stage should not be viewed lightly. Many system builders overrate the friendliness of their systems. If the users have initial bad experiences with the system, then these first reactions may taint all future use.

6.4 INITIAL STEPS

It is important not to fall into the trap of using an expert system for the sake of it. There must be a good reason for doing so, and most of the advantages and disadvantages have been discussed above. A system that no one uses because the task it carries out is too simple, or because it is too unfriendly, or one that fails to consider all that a true expert considers, or that is unable to be supported in court, wastes the time of all concerned. On the other hand, if you are just beginning to learn how to use knowledge-based systems, your first attempt to set an expert system should not be complex at all, lest the task become too discouragingly difficult. An excellent starting point is to acquire a relatively inexpensive expert building tool such as CRYSTAL and experiment with applications. Such a system can be run on inexpensive PCs. We have been running on IBM 8088 processors which are adequate for the amitryptyline toxicology system and the serology advisor, but the added speed of an 80286 or better is advantageous for CAGE.

ACKNOWLEDGEMENTS

The authors wish to thank those persons already mentioned who made their systems available to us. We would like to thank Mr Justice Prichard for reviewing and contributing to the text. There have been numerous people on whom we have tried various systems, and we wish to thank them for their comments.

REFERENCES

Aitken, C.G.G. (1988) Statements of probability. *Journal of the Forensic Science Society* **28** 329–330.

Bailey, D.N., Van Dyke, C., Langou, R.A., & Jatlow, P. (1978) Tricyclic antidepressants: Plasma levels and clinical findings in overdose. *American Journal of Psychiatry* **135** 1325–28.

Bailey, D.N. & Shaw, R.F. (1979) Tricyclic antidepressants: Interpretation of blood and tissue levels in fatal overdose. *Journal of Analytical Toxicology* **3** 43–46.

Bailey, D.N. & Shaw, R.F. (1980) Interpretation of blood and tissue concentrations in fatal self-ingested overdose involving amitriptyline: An update. *Journal of Analytical Toxicology* **4** 232–236.

Brown, G.A. & Cropp, P.L. (1987) Standardised nomenclature in forensic science. *Journal of the Forensic Science Society* **27** 393–399.

Buckleton, J.S., Pinchin, R., & Evett, I.W. (1990) Computerised assistance for glass evidence: An experimental knowledge based system for assisting in the interpretation of forensic glass examination. (In press.)

Charniak, E. & McDermott, D. (1985) *Introduction to artificial intelligence.* Addison-Wesley Publishing Company, Massachusetts.

De Monchy, A.R., Forster, A.R., Arrettig, R., Lan Le, & Deming, S.N. (1988) Expert systems for the analytical laboratory. *Analytical Chemistry* **60** (23) 1355–1361.

Efstathiou, H.J., Lambert, J., & Hepworth, N. (1983) Applications of expert systems to forensic science. Personal communication available from Mr J. Lambert, Home Office Forensic Science Laboratory, Sandbeck Way, Audby Lane, Wetherby, West Yorkshire LS22 4DN, England.

Evett, I.W. & Buckleton, J.S. (1990) The interpretation of glass evidence: A practical approach. *Journal of the Forensic Science Society* **30** 215–223.

Evett, I.W. & Spiehler, E.J. (1988) Rule induction in forensic science. In: Duffin, P. H. (ed) *Knowledge based systems applications in administrative government.* Ellis Horwood Limited, Chichester, UK.

Forsyth, R. & Nayor, C. (1986) *The hitch-hikers guide to artificial intelligence.* Chapman and Hall/Methuen. London, New York.

Holt v Auckland City Council (1980). *New Zealand Law Reports* **2** 124.

Kind, S.S., Wigmore, R., Whitehead, P.H., & Loxley, D.S. (1979) Terminology in forensic science. *Journal of the Forensic Science Society* **19** 189.

Lawton, M.E., Buckleton, J.S., & Walsh, K.A. J. (1988) An international survey of the reporting of hypothetical cases. *Journal of the Forensic Science Society* **28** 243–252.

Lindley, D.V. (1987) The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science* **2** (1) 17–24.

Pounds, C.A. & Smalldon, K.W. (1978) The distribution of glass fragments in front of a broken window and the transfer to individuals standing nearby. *Journal of the Forensic Science Society* **18** 197–203.

Prichard, Justice (1985) The presentation of expert evidence. Presented at a conference on *The expert witness—The interface between the expert and the law* at Auckland, New Zealand.

Riesel, D. (1986) Discovery and examination of scientific experts: *The Practical Lawyer* **32** (6) 59–91.

Rynearson, J., Goodyear, M. & Chisum, W.J. (1988) Artificial intelligence in the crime laboratory. In: *Proceedings of the Tenth Australian International Forensic Science Symposium, Brisbane, 23–27 May 1988,* p. 3212.

Scientific Sleuthing Newsletter (1983) Drugs loss of standard spectra no impediment to chemist's testimony nor is the lack of knowledge of computer program used in GC/MS. 7 (4). Item 2117 5. (also correspondence relating to the above: The origin of SSN material of 'where's the beef. (1984) *Scientific Sleuthing Newsletter* **8** (2) 3–5.)

Spiehler, V., Spiehler, E., & Ossleton, M.D. (1988) Application of expert systems analysis to interpretation of fatal cases involving amitriptyline. *Journal of Analytical Toxicology* **12** 216–224.

Stead, A.H. & Moffat, A.C. (1983) A collection of therapeutic, toxic and fatal blood drug concentrations in man. *Human Toxicology* **3** 437–464.

7

Quality assurance in the forensic laboratory

Thomas A.Kubic, M.S., J.D.
Detective Criminalist, Scientific Investigation Bureau, Nassau County Police Department, Mineola, New York, USA
and
JoAnn Buscaglia, M.S.
Deputy Director and Quality Assurance Officer, Thomas A.Kubic & Assoc. (TAKA), Forensic and Environmental Consultants, Northport, New York, USA

7.1 INTRODUCTION

To understand the role of statistics in a Quality Assurance System, a discussion of Quality Assurance concepts is beneficial.

Although all forensic laboratories are operating under, or with, some system which is designed to ensure the quality of the analytical product, not all have a formal, documented Quality Assurance System.

The word ‘quality’ will be used repeatedly as an adjective in this chapter, and its meaning should not be interpreted in such a way as to imply that laboratories that are not practising under a Quality Assurance Program are performing ‘poor’ work. Rather, as used here, it implies that the work being performed meets or exceeds some defined standard, which can be measured and whose performance can be monitored. It is implicitly assumed that there is a continuous effort to elevate the standard of performance.

The advanced technology that is becoming commonly available, even to smaller laboratories or laboratory systems, carries as a concomitant the necessity to guarantee, to a higher degree than ever before, the quality of the laboratory product. This is not only on account of the ideal to which all science should ascribe, that is, of the highest possible quality product, but also because of the impact that the work of a forensic laboratory normally has upon individuals and society. As the probative value of a forensic scientist’s evidence increases in such a way that the probability that a verdict will be consistent with that evidence becomes more certain, then the effort to ensure that the evidence is consistent with the true facts should likewise increase in order to ensure that an injustice does not take place. For example, errors in class characterization of a bullet or in classification of the blood on a victim as that from a major type A person would not be as damaging as an erroneous bullet-to-firearm match, or an improper match of the suspect's DNA to a stain found on the victim.

Moreover, as the technologies employed at forensic laboratories become more complex, it is often the case that subtle variations in either analytical methodology or results can be significant. It is therefore necessary to ensure tighter control of the analytical process and the interpretation of the generated data.

It should not be overlooked that, as caseload pressure increases, subdivision of tasks is likely to maximize production. Subsequently, it is unlikely that every portion of the processing of a case will be performed by a highly experienced individual. The laboratory system must make allowances for the development and maturation of the expertise of the more junior members of the laboratory staff.

Review of this total work product, whether it be by higher management, first line supervisors, or peers, is required, and a system that facilitates such a review should be operational. Having recognized that the implementation of a Quality Assurance System is not only desirable but necessary for scientific, ethical, and social reasons, definitions of the general terminology follow (Taylor 1985).

Quality assurance: A system of activities which provides the producer or user of a service the assurance that it meets defined standards of quality.

Quality assurance consists of two separate, but related, activities: quality control and quality assessment. Both must be operational and coordinated and should be evidenced by appropriate written or electronic documentation.

Quality control: The overall system of activities which controls the quality of the product or service so that it meets the needs of users. The aim is to provide quality that is satisfactory, adequate, dependable, and economic.

Quality assessment: The overall system of activities which provides assurance that the quality control activities are performed effectively. It entails a continuing evaluation of the performance of the production system and the quality of the products produced. The objective of quality control in a laboratory setting is to fine-tune the measurement process and maintain it in a desired state of stability and reproducibility. In a forensic laboratory, this assessment should include the opinions, based on the results of the measurement process, that are expressed by its 'experts'.

Once control is established the precision of the process may be defined, and biases can usually be identified. Hence, proper action can be taken to eliminate or compensate for them. Quality assessment provides the means of establishing the evidence that quality control has been accomplished.

A successful Quality Assurance Program does not just happen. It requires a well organized plan. It is not unusual for the uninitiated to believe that *implementation* of a Quality Program is the most difficult task. In reality, it is the *design* of a program that will be able to function and bring about the desired result that is the most difficult undertaking.

The cost of quality assurance is often mentioned as a reason for failing to adopt a far-reaching Quality Program. Some experts have estimated that quality assurance activities require 10–20% of an analyst's efforts, with a substantial part of the time centred on preventive maintenance and the analysis of check and reference samples

(Garfield 1984). The American Industrial Hygiene Association (AIHA), United States' Environmental Protection Agency (EPA), and United States' National Institute of Standards and Technology (NIST) all require a minimum of 10% re-analysis of quality control samples as part of their laboratory accreditation programs (EPA 1968, AIHA 1989, Berger *et al.* 1988). When the efforts of the QA/QC supervisor and the record-keeping, training, and quality review, etc. are considered, some laboratories maintain that about one third (1/3) of the time of all technical personnel can be ascribed to the maintenance of the quality program (Buscaglia 1991, Emma 1991). These costs can be quite complex and difficult to evaluate, but are important to consider. There will be some point at which additional gain in quality may not be justified by the increased cost (Wilcox 1977). If it is accepted that there is such a point of 'diminishing returns' for health-related laboratories, it must be recognized, no matter how unpalatable to some, that there are similar economic considerations for the forensic laboratories.

In the opinion of some authors (McCully & Lee 1980) '…a laboratory cannot afford not to have good quality assurance. The gain in not having to re-do (re-analyse), correct, throw-out, etc. will more than pay for the cost of the program.' They assert that the benefits to the laboratory of elevated credibility, increased expertise of the analysts, and improved morale of the staff are added dividends. An economic axiom might be 'It is cheaper to do it right the first time.'

7.2 DEVELOPMENT OF A QUALITY ASSURANCE PROGRAM

There are a number of components to a Quality Assurance Plan, or Program, as defined by various experts. In reality, all of the components are contained in one of three essential elements: Prevention, Assessment, and Correction. A number of the components are clearly assignable to one of the key element titles, while many can appear in two or all three.

Those elements that are preventive in nature include, but are not limited to, the design of the written Quality Program, along with the acquisition of the necessary instrumentation, standards, equipment, and adequate physical facilities. The adoption of standard operating procedures, good calibration, and measurement practices using validated methods all tend to prevent problems.

The assessment portion of the program may include determination of accuracy and precision of an accepted methodology with periodic evaluation by duplicate and check samples. Interlaboratory exchange samples, blind in-house quality control samples, and external proficiency testing all have a role in quality assessment. Where the service supplied may require expert conclusions or opinions based on the data, peer review of the results and conclusions can become an integral part of the assessment element. In the forensic laboratory, where court testimony may be the end product of a substantial amount of analytical effort, review by supervisory personnel of the intended, as well as the actual, testimony may be crucial to the quality of the laboratory product.

Correction, the last part of the triumvirate, consists of actions taken to determine the cause of quality defects and to eliminate the defects, where possible. After evaluation, procedures are implemented or practices altered so that proper functioning of the analytical system is restored. This can entail the servicing of malfunctioning instrumentation, recalibration, scheduling of more frequent calibrations, increasing the number of quality check samples, institution of more comprehensive control and blank sample tests, or re-evaluation of methods used for analyses. Retraining or more advanced training can often be part of the corrective element. Many individuals, including bench scientists, supervisors, and managers, tend to confuse disciplinary action with correction in quality assurance. This is unfortunate because it leads to defects in the product not being remedied because of the desire neither to accept nor to impose blame. Correction in this context means the resolution, to the extent possible, of any and all problems with the product for the benefit of all parties.

Participation in either a voluntary or mandated laboratory accredition program administered by independent parties, whose goal is the improvement of laboratories and the services rendered and who have no vested interest, can be of great assistance in elevating the product of a laboratory. Programs that are implemented by the laboratory’s own industry, although possibly not totally independent in action or view, can similarly be of tremendous value in upgrading the performance of associated laboratories.

7.3 PORTIONS OF A QUALITY ASSURANCE PROGRAM

Authorities delineate various components necessary to a functional Quality Assurance Program. The National Institute for Occupational Safety and Health (NIOSH, United States) has identified over twenty individual elements that, it believes, must be covered in a Quality Program (NIOSH 1990). The reader is referred to the *Bibliography,* as space will prevent detailed treatment of any of the topics and only passing mention of many.

It may appear trivial to mention that the written Quality Program should contain the following: a statement of the laboratory’s and Program’s objectives, plans for attaining a quality product, the laboratory’s organization, with the responsibilities of the staff delineated, and policy statements where appropriate.

The laboratory should function under a system of written standard operating procedures (SOP), which should contain information of the day-to-day operation of the laboratory. The Quality Control Program may be contained within the SOP Manual, or be referenced therein and found as a totally separate document. The day-to-day quality control records may be kept together in a common location or in various sites within the laboratory, when that is more convenient. In either case, their locations, descriptions of entries, and scheduling of entries must be found in the Quality Control Manual. Often, quality records such as control charts are duplicated and found both in the permanent records and at the sites where analytical measurements are performed.

If actual copies of the laboratory's standard methods of analysis and all revisions are not kept in the Quality Manual or SOP Manual, then, as a minimum, reference to their locations, updating and revision procedures, as well as dates of applicability, must be maintained in the central SOP files.

Another portion of the Quality Manual that is often overlooked, and which is part of the prevention element, is that of analyst training. The written program should contain minimum academic, educational, and experience standards for each different analytical expertise required. There may, and should indeed, be quite different standards for the molecular biologist or biochemist performing DNA typing, the forensic toxicologist, the document examiner, and the chemist performing arson accelerant and explosive residue identification.

Forensic laboratories tend to rely heavily on in-house training of their analytical staff. Such programs are highly valuable, but the agenda of all such programs should be found in a written training program, which contains a schedule for advancement, a system for monitoring progress, and a method to determine completion of the training with the attainment of examiner status.

Training does not terminate upon completion of the step-wise program, but is, rather, a continuing educational process. The policy portion of the Quality Program should be explicit concerning the laboratory management's views on advanced training, whether it is promulgated through the pursuance of advanced degrees, continuing education via specialized courses, seminar attendance, or professional advancement through library and laboratory research.

It is worth noting that true commitment to quality cannot exist without adequate in-house or easily accessible library facilities with texts and journals enabling scholarly research, as well as the possession of adequate reference materials consisting of actual physical standards and exemplars.

The entire technical staff of the laboratory is responsible for the Quality Program. Some may have a specific duty, for example preparing the weekly calibration of retention times for a certain gas chromatographic analysis, while others may be assigned a general duty, such as that of the laboratory director who is ultimately responsible for the quality of the laboratory's product.

It is prudent and often necessary for the laboratory director to delegate the design, implementation, and updating of the quality plan to an individual who, because of education, training, interest, and temperament is well suited to act as Quality Assurance Officer or Manager.

For the Quality Program to take hold and accomplish its enumerated goals, management must not only have faith in the Quality Assurance Officer and vest in him the responsibility for attaining quality production, but also give him the commensurate authority to achieve that end. The Quality Officer must have the support of management who must ratify his actions in all but the most unusual situations.

The record-keeping requirements for a forensic analytical laboratory should be more stringent than those of almost all other analytical laboratories, save those concerned with human life and health. Examination of a final report of the laboratory and supporting notes of the analyst should allow a competent forensic scientist to reconstruct the thinking and procedure of the original examiner so that, employing the data, the initial results and conclusions are verified. For forensic analyses, mere copies of the reports and hard data such as spectra and chromatograms may be insufficient. The original notes of the examiner with pertinent details as to his observations, approach, and even his thought process may be essential at a later date.

7.4 SAMPLE CONTROL

The often quoted dictum from the computing world of ‘garbage in, garbage out’ has an analogy for the analytical laboratory. An improper sample will almost universally result in false or unusable data, along with improper conclusions. That is not to say that imperfect samples cannot be employed when proper controls are available. But the resurrection to usefulness of the data from such improper samples is often difficult, if not impossible.

Forensic laboratories have been unable in many cases to make submitters realize that analyses of contaminated, undocumented, non-secured, co-mingled samples, or of those submitted without sufficient or proper controls, or without known exemplars, can only lead to disastrous results. Requests for services that are vague, insufficient in detail, or without proper background material should likewise be rejected by the laboratory. Forensic laboratories find themselves in the unenviable position of attempting to produce quality services with limited resources. This requires the screening of submissions, while political pressures from within and outside their host organization prefer the laboratory to accept everything and ‘Do what you can with it!’.

A written policy adopted by the laboratory’s agency dealing with the disposition of improperly submitted samples will support the laboratory’s position of refusing to accept such materials. Such policies that result in rejection of samples before their entry into the laboratory’s measurement system are a proper part of the Quality Program and can be found not only in environmental (EPA 1968) and health (NID A 1989) laboratories but also in some forensic science service laboratories (FBI 1990).

In law enforcement, there has been a continued trend to limit or do little or no investigation of crimes where there are not present at least one or more solvability factors. (Solvability factors are facts and circumstances that indicate that an investigation may lead to solution of the crime.) It would not be inappropriate for crime laboratories to promulgate similar policies for acceptance and processing of submitted physical evidence. If the facts indicate that analysis of the evidence is not likely to result in useful information, then the analysis of it should be postponed until the situation has changed. It may even be profitable for the policy to dictate that the evidence not even be accepted into the laboratory, but be safeguarded by the submitting entity until analysis would be fruitful. Naturally, such a policy would make exceptions for materials that need special preservation techniques (cryogenic freezing), or are accepted for training purposes or database construction, or which are processed as investigation aids.

The control of samples once accepted into the laboratory by proper logging, transfer, and evidence continuity procedures cannot be overemphasized. Who had it? For how long? Why? These are all questions that must be able to be answered.

Computer-generated sample bar coding and control can be invaluable in a Quality Program that deals with large volumes of similar samples such as found in forensic toxicology, or drug screening laboratories. Computerized laboratory management information systems are beginning to take their place in forensic laboratories. These systems allow managers to track sample location and evaluate the timeliness of the laboratory's services.

All the components that make up good measurement practice are contained within quality control. We will mention a number of topics, none in great detail, with which the Quality Assurance Officer should become acquainted, in order to be able to assist in the production of quality results. The authors again refer readers to the reference texts found in the *Bibliography* at the end of this chapter as being of great value in developing their mathematical skills and understanding of the application of statistical analyses to the treatment of data.

It is necessary that technically sound methods are employed, samples are valid, errors are evaluated, interpretations are correct, and the process is monitored to ensure a quality measurement process.

The use of validated methods ensures that the product of the measurement process will meet stated standards. In forensic laboratories, the term 'validated methods' is often used interchangeably with standard methods. In our view they are not the same. That is, a valid method may not meet the rigours of a step-by-step standard method of analysis. Although the term standard method seems to imply it is valid, this may not be the case. It is possible to adopt a 'standard method', perform all analyses by that method, without deviation, and attain results which lack the required accuracy or precision. However, methods that are adopted as 'standard' by regulatory agencies or professional associations have usually been validated. The method itself will usually have a statement regarding the conditions, sampling procedures, calibrations, concentrations, interferences, etc., under which one can expect the stated accuracy and precision (Eller 1984).

Members of a methods advisory committee for forensic science reported that the adoption of 'standard methods', as they are thought of in most situations, are not applicable to forensic science investigations because 'a basic requirement of a standard method is a standard sample. In reality, such samples rarely exist in the forensic sciences' (Gallo & Field 1980). The lack of close similarity of samples is one fact on which the forensic analyst can usually rely. However, this should not be confused with the establishment of the validity of a method.

Validation of methodology deals with the evaluation of the performance of a measurement process against the requirements for the analytical data. Careful consideration of the data requirement is essential to the establishment of firm criteria for performance. Put simply, what is it that we expect this method to do and with what accuracy and precision? During the validation of the process, the ruggedness of the methodology (that is, the effect caused by deviations from the procedure), is evaluated. Certainly methods that are too sensitive to sample variation may not be suitable for forensic laboratories.

Validation of methodology can be accomplished in a number of ways (Taylor 1987, p. 194). If reference samples are available that are similar in all respects to the test samples, the task is simple. The average of the results from the tests performed on the reference samples is compared with the accepted value for these reference samples. If the accepted value falls within an interval known as the confidence interval, constructed about the experimental value, then there is no reason to believe that the experimental value is inconsistent with the accepted value at the level of confidence appropriate for that interval. The confidence interval is constructed as follows.

Assume n measurements are taken on the test sample, denote these by the symbols $x_1, x_2, \ldots, x_n$, with x_i indicating the ith measurent. Let $\overline{x}$ denote the sample mean $(\{x_1 + x_2 + \ldots x_n\}/n)$ and let s denote the sample standard deviation,

$$s = \sqrt{\left\{\sum_{i=1}^{n} (x_i - \overline{x})^2/(n-1)\right\}} . \tag{7.1}$$

Denote the appropriate percentage point, a say, of the t-distribution, with n - 1 degrees of freedom, by $t_{(n-1)}(\alpha)$. Then, the 100(1-2α)% confidence interval for the true mean is

$$\overline{x} \pm t_{(n-1)}(\alpha)\, s/\sqrt{n} \tag{7.2}$$

(Note that this is a general statement concerning the size of the interval. As an example, let $\alpha = 0.025$. Then $(1-2\alpha) = 0.95$ and $100(1-2\alpha)\% = 95\%$.)

The number, n, of sample measurements should be sufficient to ensure that the confidence interval has reasonable precision. Seven is usually considered a minimum value for n in this application.

The reader is reminded that the t-distribution is employed rather than the Normal distribution because the standard deviation of the population is estimated from the sample.

When suitable reference materials are not readily available a number of other approaches are acceptable. One approach compares the results of the candidate method with those of one previously proven to be applicable and reliable. The test statistic given in (7.3) is used to determine if there is a reason to believe that the values generated by the two methods disagree for reasons other than by chance (Miller & Miller 1988, p. 55). Expression (7.3) is used when it can be reasonably assumed or is known that the two methods give rise to similar levels of variability in their results.

$$t = (\overline{x}_1 - \overline{x}_2)/[s_p\sqrt{\{(1/n_1) + (1/n_2)\}}] \tag{7.3}$$

where the pooled standard deviation s_p is given by

$$s_p = \sqrt{[\{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\}/(n_1 + n_2 - 2)]}$$

and s_1^2 and s_2^2 are the sample variances from samples 1 and 2, respectively.

Expression (7.3) provides a test statistic for the null hypothesis (H_0) that the two methods do not disagree in the values they provide, other than by chance. If this hypothesis is true then (7.3) has a t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

The null hypothesis can then be tested by referring the value for *t* calculated from (7.3) to tables of the *t*-distribution. If the calculated *t* value is not greater than the value taken from the *t* table at the stated significance and required degrees of freedom, then the null hypothesis is not rejected and there is no statistical reason to believe that the results of the two methods differ.

If the variances and standard deviations are not or cannot be assumed to be equal, then (7.4) is used (Natrella 1966, p. 3–27).

$$t = (\overline{x}_1 - \overline{x}_2)/\sqrt{}(v_1 + v_2) \qquad (7.4)$$

where $v_1 = s_1^2/n_1$ and $v_2 = s_2^2/n_2$ and the effective degrees of freedom (f) is given by (7.5):

$$f = ((v_1 + v_2)^2/[\{v_1^2/(n_1+1)\} + \{v_2^2/(n_2+1)\}]) - 2\ . \qquad (7.5)$$

If the calculated t value does not exceed the *t* variate for the required effective degrees of freedom, then the null hypothesis is not rejected and there is no statistical reason to believe the results generated from the evaluated method differ. It should be noted that the fact that the variance of one method is less than that of another may, by itself, be enough reason to decide that the first method is not the equivalent of the second.

Even if there are no previously accepted methods for the analysis, agreement of the results with any additional independent methods lends support to the validity of the method.

Another, though less satisfactory, approach of spiked or surrogate samples may be employed, but the analyst must be mindful of possible difficulties due to matrix effects. Additionally, split samples, prepared from actual test samples, are valuable to evaluate the precision of a method, but are unlikely to provide any information as to the presence or magnitude of bias.

Often the appropriateness of a given methodology can be inferred from the proven validity of the method for analogous materials. In regulatory matters, such as revenue raising (for example, tax on alcohol) or environmental protection (for example, PCBs in waste oil), the agency has a responsibility to specify valid methods for enforcement. Otherwise, nothing but chaos will result from conflicting claims (Taylor 1987, p. 194).

7.5 METHOD EVALUATION

Both descriptive statistics and statistical inference play an important role in evaluating errors and attempting to control the measurement process.

A perfunctory examination of the kinds of measurements made in a forensic laboratory could lead to the conclusion that most are qualitative and that quantitative measurements requiring estimation of precision and accuracy are the exceptions. The determination of toxic levels of poisons found in a deceased's organs, the amount of trace metals found on the hands of a person alleged to have discharged a firearm, and the level of purity or amount of pure controlled substance possessed are areas in which the evaluation of quantitative errors is recognized as important.

However, every qualitative test has within it a quantitative aspect. When a sample is tested for the presence of a given component and the result is negative, it is a general practice to report that 'the sample did not contain that substance' or that 'the sought material was not present'. This may not be, in fact, the true situation. The component may indeed be present but at a level too low to be detected by the method employed. A more proper statement would be that, if present, the material is below the level of detection (LOD) of the method, followed by a statement of the established LOD for this method.

Often decisions are made as to the similarity of two materials such as paint based on the qualitative absence of an element in one sample which is present in a small amount in the second. Might not the questioned element be truly present, but not detected?

Let us propose a null hypothesis that two samples originated from the same source (are the same). When we reject, or not, this null hypothesis based on a retention volume in gas chromatography, or on distance travelled relative to a solvent front (Rf) in thin layer chromatography, or on migration distance on an electrophoresis gel in DNA typing, we are making a decision based on an evaluation of some factor that has a definite quantitative aspect. The errors in evaluating this factor can and should be studied and treated statistically.

Although the probability of an accidental match will not be dealt with here, it should be apparent that the specificity (the proportion of true negatives that are detected) of any such testing method should be high. In probabilistic terms, the test should be negative when the sample is, in fact, negative (Rosner 1990, p. 55). Forensic scientists say this by stating that they wish to minimize false positives as much as practicable. In other words, they do not wish to get a positive result or association when the sample is truly negative. For them, avoidance of the Type II error (not rejecting the null hypothesis when it is false), that is, in this case, of deciding the samples come from the same source when they do not, is supreme.

Biases or errors that affect the accuracy of a measurement process are classified as either systematic or random. Systematic errors, whether they be absolute or relative, are compensated for by ensuring that measurement devices are properly calibrated. A prerequisite of calibration is that the measurement process must be in statistical control. Without statistical control a measurement cannot be logically said to be measuring anything at all. Statistical control is reached when the means of a large number of individual values approach a limiting value called a limiting mean, with a stable distribution described by a stable standard deviation. Such data have predictability, and the measurements have credibility as a degree of consistency has been attained (Taylor 1987, p. 13).

Typical calibration curves used in analytical measurements assume that there is a compositional uncertainty that is negligible. A sufficient number of measurements are made of each standard and the average response (Y) for each standard value X is plotted against the corresponding value of X. A straight line is drawn either by visually fitting the data or by calculating the slope and intercept by the method of least squares.

The latter method is preferred because it removes the bias of the graph drawer. Statistical analyses may be performed on the slope, the intercept, and points on the line for estimation of the quality of the fit and the consistency of calibration (Natrella 1966, Miller & Miller 1988, pp. 102–115).

The least squares linear fit or any similar nonlinear mathematical model fitting process can be very sensitive to outliers. The practitioner will be mindful to employ a valid statistical method like that of Dixon (Crow *et al.* 1960, p. 252) or others (Taylor 1987. pp. 33–36, Miller & Miller 1988, pp. 135–150.) to support the exclusion of a data point.

In some situations, where calibration is being performed near the Level of Detection (LOD) or Level of Quantitation (LOQ), the uncertainties of the measurements made on the standards cannot be considered insignificant. It is then not justifiable to imply simple regression, even with weighted errors. The assistance of those knowledgeable in handling such data may be required (Mandel 1984).

We have remarked on the difficulties of quantitation at trace levels, as well as mentioned the terms LOD and LOQ. We will now define these terms and discuss how they apply to the assurance of quality in the measurement process.

The level, or limit (which is often used interchangeably), of detection (LOD) can have a number of meanings depending on the context in which it is used. The authors prefer to think that there should be an ‘R’ placed in each of this type of acronym to represent ‘Reliable’, because there is a level of confidence inherent in each statement. Normally, that level is assumed to be 95%, but this may not be the case.

There is an instrument level of detection (IDL), or even that of a specific detector within the instrument, which is that level of signal that can be differentiated above the background ‘noise’. Do not forget the ‘R’ for reliable.

The method limit of detection (MDL) is the minimum amount of an analyte that a method can detect, regardless of its origin, and determination of its value is based on the variance of the analyzed samples.

The level of detection (LOD) is usually defined as that level that can be detected above the background of the sample matrix. A measured value becomes believable when it is greater than the uncertainty associated with it. The LOD is often arbitrarily defined as $3s_0$, where s_0 is the standard deviation of analyses performed on the blank, and it can be determined by extrapolation (Taylor, 1987, p. 79). An alternative is to determine the standard deviation for a reasonable number of replicates, at least 7, at a small but measurable level at about twice LOQ, and define this value of the standard deviation as s_0. Recent guidelines from public agencies suggest the definition of LOD should be given by:

$$\mathrm{Y}\text{-}\mathrm{Y}_0 = 3s_0, \tag{7.6}$$

where Y is the response and Y_0 is the response of the background (Miller & Miller 1988. p. 115). The usefulness of the definition is illustrated by Fig. 7.1.

The analyst is faced with two problems when making measurements near the LOD. If his (null) hypothesis (H_0) is that an analyte is present and the true state of fact is that it is present, he may commit a Type I error (reject H_0 when it is true) by failing to detect it, or a Type II error (not reject H_0 when it is false) by stating it is present when it is not. The analyst normally wishes the probability of both errors to be equal.

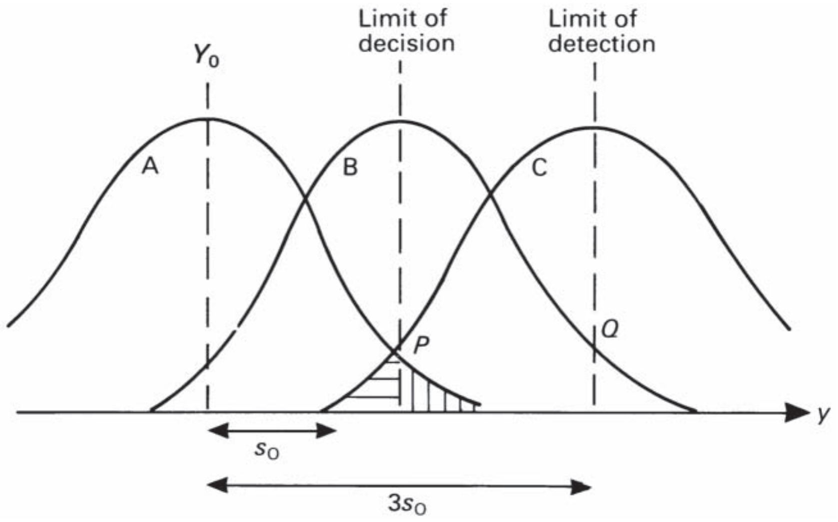


Fig. 7.1 —Curves to illustrate the variation of signal response Y_0 due to the blank (curve A), the variation of response of analyses of samples whose average response was P, the limit of decision (curve B), and of samples whose average response was Q, the limit of detection (curve C). The standard deviation of measurements on the blank is denoted by s_0.

In forensic laboratories for certain analyses, for example the presence of a controlled substance, the likelihood of committing the Type II error is minimized as much as is practicable.

In Fig. 7.1, curve or distribution A represents the variation of signal response Y_0 due to the blank and is represented here for simplicity as a Normal distribution. In reality, these data are often censored; that is, there is some minimum signal below which values cannot fall (sometimes zero). Point P is located at high values for distribution A, such that for values of Y greater than P it is decided that the analyte of interest is present. It could be argued that values greater than P are unlikely to be due to background effects only, although values less than P may still be obtained when the analyte of interest is absent. Distribution B represents a series of analyses of samples whose average response was P. These are normally distributed with half of the responses below P, and these are considered as being consistent with a blank, while half of the samples will be correctly reported as being positive. This solves the problem of false positives but not of false negatives, Type I errors. This point, P, sometimes called the level of decision, is unsatisfactory as a LOD (Level of Detection). Let us now identify point Q so that it is twice as far from Y_0 as P. Distribution C represents a series of analyses of samples whose average response was Q. It can be shown that (Q-Y_0) is 3.28 times s_0. At this point, the probability of each of the two original errors is 0.05. If (Q-Y_0) is $3s_0$ then the likelihood of either of the errors is about 7%. Many would consider this a reasonable and practical definition for LOD (Miller & Miller 1988, pp. 115–116.). Those analysts who wish to employ other than equal probabilities for Type I and II errors are referred to the literature (Mark & Workman 1989, Clayton *et al.* 1987) for suggested solutions.

The Level of Quantitation (LOQ) is the lowest level at which measurements become quantitatively meaningful, and it has been defined in a number of ways, such as two to five times LOD or $10s_0$. These definitions have not generally been accepted (Taylor 1987, p. 79). Most analysts, in place of the above, prefer to employ the point at which they feel the relative uncertainty is reasonable or a point at which it begins to approach a minimum value. Fig. 7.2 displays the relationship of the uncertainty of a measurement and LOD and LOQ.

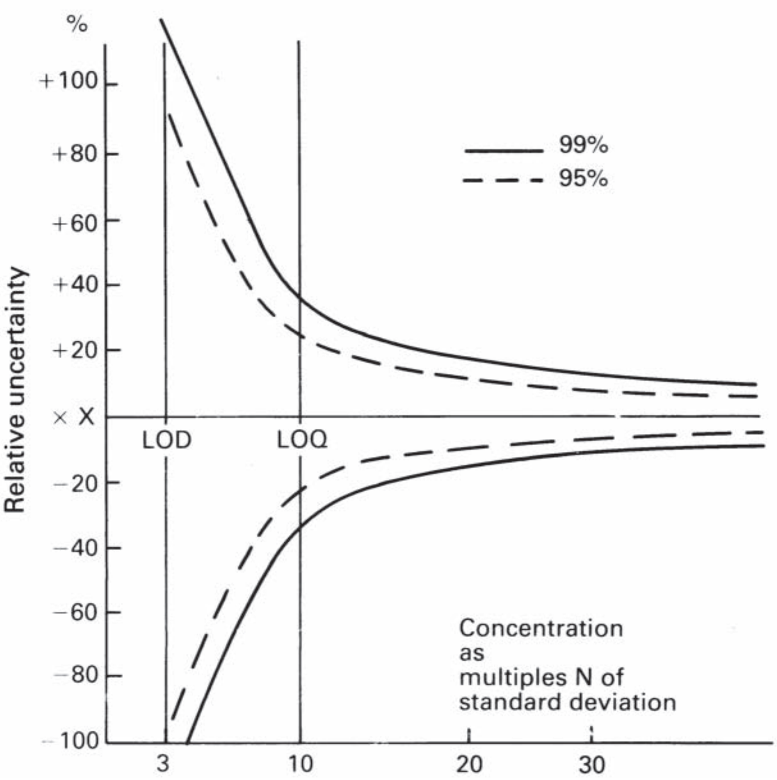


Fig. 7.2 —The relationship of the uncertainty of a measurement, expressed as a percentage, with the concentration of the analyte in multiples of the standard deviation of analyses conducted on the blank.

The horizontal axis in Fig. 7.2 represents the concentration of analyte in multiples of the standard deviation s_0 of analyses conducted on the blank. The vertical axis represents a quantity known as the relative uncertainty. The value of the relative uncertainty is dependent on the level of confidence at which one is prepared to express one's uncertainty, and is related to the idea of confidence intervals. Values for two levels, 95% and 99%, of confidence are shown by the dashed and solid lines, respectively, in Fig. 7.2.

An example will show how relative uncertainty may be determined. Consider the 95% confidence level. Let y_0 be the measurement on the blank and y the measurement on the analyte with the standard deviation of both being s_0. Then, assuming independence, the variance of the difference $(y\text{-}y_0)$ is $2s_0^2$ with corresponding standard deviation $s_0\sqrt{2}$. Then, at the 95% level of confidence, one is 'uncertain' if the difference between y and y_0 is less than 2 standard deviations $(s_0\sqrt{2})$; that is, if

$$|y\text{-}y_0| \leq (2\sqrt{2})s_0.$$

(The value 2 is used as an approximation to 1.96, the appropriate percentage point of the standard Normal distribution for the 95% level of confidence.)

Now, assume $|y\text{-}y_0|$ is a multiple, N, of s_0 so that $|y\text{-}y_0| = Ns_0$. Thus, one is uncertain if $N \leq (2\sqrt{2})$. Relative uncertainty is measured by the reciprocal of this, $(2\sqrt{2})/N$, expressed as a percentage. As N increases, relative uncertainty decreases. When $N = 3$, the level of detection, $(2\sqrt{2})/N = 0.94$ or 94% . When $N = 10$, the level of quantitation, $(2\sqrt{2})/N = 0.28$ or 28%. These are both calculated assuming a 95% level of confidence. For a higher level of confidence, 99% say, the LOD, that is, the value at which relative uncertainty is 100%, will be correspondingly larger. The appropriate percentage point of the standard Normal distribution is 2.58; relative uncertainty is then $(2.58\sqrt{2})/N$, and for this to equal 1 we need to take $N = 3.6$. Thus, at the 95% level of confidence, the LOD is $3.6s_{0.}$

If the slope of the plot of response versus amount *(X)* on the calibration curve begins to decrease, then the point at which this happens is known as the Level of Linearity (LOL). The straight line begins to curve gently, and this becomes more severe as the values on the X (horizontal) axis increase. Scientists historically preferred to work from straight lines because it was easy to fit them mathematically to data points and extract values from them. With micro- or mini-computers and sophisticated software available in most laboratories, non-straight line calibrations can be routinely employed. The sensitivity, that is, the slope of the response curve, must remain reasonably steep or the ability to detect differences in the amount, *X,* will be severely diminished.

The reader is now asked to give some thought to the following. Assume two similar paint samples are being analysed for association by electron microbeam energy dispersive spectroscopy. Sample A, from the standard, exhibits what appears to be the presence of trace, near LOD, iron in the paint. The visual appearance of the peak on the CRT display is insufficient evidence that it is not due to random electronic instrument noise. The analyst tests the suspect peak with the following technique for determining if the peak could be due to random error in pulse counting. This is an IDL test. The peak is declared real if P, the net number of counts above the number b of background counts, is greater than $3\sqrt{b}$.

Using this technique, he correctly concludes that iron is present in A. Upon analysis of another sample, sample B, no iron is detected. Is the analyst on firm ground to state that the samples are not from the same source because of the detected presence of iron in A and not in B? The answer is 'no', unless the analyst could evaluate the variance of counts from samples of similar iron concentrations in matrices similar to A. If the net counts for iron in B fell outside the desired confidence limits for A, then the analyst would be justified in rejecting the hypothesis that A and B had a similar origin. But if the LOD for this analysis is inside this caluclated interval, he may not reject the possible common origin.

Please refer to Fig. 7.2 for an appreciation of the relative errors that are possible when measurements are performed at, or near, the LOD for an analyte.

7.6 QUANTITATIVE ANALYSIS

There are innumerable types of analyses performed in forensic laboratories where quantitative measurements are performed. The question often arises as to how much confidence the laboratory has in the results it is reporting. That is, do they believe the 'true value' is either close to, or, in reality, equal to the reported figure? The role of statistical analysis on the evaluation of how much faith or confidence the laboratory should have in the reported value of its measurement is undeniable.

Whenever repeated measurements are made on a given sample, or multiple samples are analysed from a population, the results usually exhibit variation in values. This variation is due principally to random errors in the measurement process that are not totally controllable, and it has, generally, a Normal distribution.

There are other continuous distributions that are of interest to the forensic scientist, such as the log-normal. This has been shown to describe the levels of acid phosphatase (ACP) found pre- and post-coital in the vagina (Sensabaugh 1979). A number of distributions of discrete data also play an important role in forensic science. The binomial distribution describes the probability of obtaining a certain number of outcomes of a particular type (A, say) in a sequence of independent trials in which, in each trial, there are only two possible outcomes (outcome A or outcome B, say) and the probability P_a (P_b) of obtaining type A (B) at any one trial remains constant from trial to trial ($P_a + P_b = 1$). The Poisson distribution may be considered as a special case of the binomial wherein the probability of outcome A, for example, is rare. The reader is referred to Natrella (1966), Crow *et al.* (1960), Rosner (1990) or any good statistical methods text for further information concerning this topic.

Many distributions that are difficult to handle mathematically can be transformed into a reasonable approximation to a Normal distribution, and statistical analyses based on the Normal distribution may then be used to analyse the data (Natrella 1966, pp. 20–1, 20–8). Alternatively, when a large number of observations are made from a population where the measurements of the characteristic of interest may not have a Normal distribution, the average results of these measurements will have a Normal distribution. This result is known as the Central Limit Theorem (Rosner 1990, p. 154), and it allows statistical analyses based on the Normal distribution to be used more often than may otherwise have been accepted. Both the binomial and Poisson distributions can be approximated for large sample numbers by Normal distributions in this way. In such situations, the approximate answers obtained from a Normal treatment of the data are sufficiently accurate that they are routinely employed (Rosner 1990, pp. 121–133, Barlow 1989).

Methods are also available that allow statistical evaluation of errors, no matter how the data are distributed, other than that some require that the data be distributed symmetrically. These methods, known as non-parametric methods, are valuable and are described in the text by Miller & Miller (1988 pp. 137–159) and others.

With that background dealt with, we can discuss some of the methods of evaluating confidence in our quantitative data.

The first important concept is that of a Confidence Interval. We could take a random sample of size *n* from a population of data, take measurements of some characteristic of interest, calculate the average $\bar{x}$, and compute a confidence interval for the true mean value, μ, of these measurements according to (7.7):

$$\bar{x} \pm z\sigma/\sqrt{n}. \tag{7.7}$$

This is said to contain μ with a given pre-specified level of confidence. This pre-specified level determines the value of *z*, and this may be found in tables of the standard Normal distribution. This assumes that σ, the true standard deviation, is known. An alternative interpretation of a confidence interval is that if we take a large number of samples, each of size *n,* then μ will be found in a proportion of intervals, corresponding to the appropriate probability, generated about each of the sample means ($\bar{x}$). Thus, given 100 95% confidence intervals, for example, it is expected that 95 of these will contain the true value of μ. Note, however, that for any particular interval, μ either is or is not contained therein. The width of a confidence interval reflects the precision with which the true mean is estimated; the narrower the interval for a given level of confidence, the more precise is the estimate. Note, however, that the magnitude of the interval is inversely proportional to the square root of the number of measurements *(n).* Thus, to halve the width (double the precision) it is necessary to take 4 times as many observations ($\sqrt{4} = 2$). If a 95% confidence interval is desired, this level of certainty is often referred to as a 'reasonable scientific certainty', and the value for *z* is 1.96 (often approximated by 2).

Unfortunately, it is seldom that the analyst is able to collect an entire population of data and determine the parameters μ, and σ. The analyst is forced to estimate them from data generated from a limited sample. The best, that is the least biased and least variable, estimate of μ is $\bar{x}$, the sample mean, while the unbiased estimate of σ, and the one used in the construction of confidence intervals for the mean, is s, the sample standard deviation, defined in (7.1).

When *s* is used to estimate σ, the test statistic, defined below, is not Normally distributed but has a *t*-distribution. When σ is estimated by *s*, the corresponding confidence interval is derived by replacing *z* and σ in (7.7) with *t* and *s* respectively; see (7.2). As mentioned, the *t* variate necessary to generate a given confidence interval, for example 95%, varies with the degrees of freedom (*n*-1), but can be approximated by the corresponding z variate when the sample size is large.

A practical application of the above is the case where an exhibit is analysed and found to contain 506 mg of a controlled substance. A greater penalty is assigned if more than 500 mg are present. Is the analyst correct in testifying, to a reasonable scientific certainty, that the exhibit contains 500 mg or more of the illicit substance?

A lower one-sided confidence interval may be constructed in a similar manner to the previously discussed two-sided interval by employing (7.8):

$$\overline{x} - t_{(n-1)}(\alpha)\, s/\sqrt{n}. \tag{7.8}$$

This gives a value, x_1 say, above which the true value would be expected to be found with the desired confidence. Suppose s in this example is 7 mg and n is 7 ($\sqrt{7}$ = 2.646); the appropriate percentage point of the t distribution for (n-1) = 6 degrees of freedom and α = 0.05 (which gives a 100(1-α)%, or 95% lower confidence limit) is 1.943, therefore

$$x_l = \overline{x} - \{1.943(7/2.646)\}$$

where $\overline{x}$ = 506 mg. The lower confidence limit, x_l, is then 500.86. The analyst is, therefore, justified, to a 'reasonable scientific certainty', in his testimony.

A second area in which statistics can be of assistance is the following. One has a large population of items which were sampled, and measurements of some characteristic of interest made, and $\overline{x}$ and s calculated. One might wish to make a statement as to the proportion of the population that is expected to be between certain limits with respect to the measureable characteristic. These are referred to as Statistical Tolerance Limits. These furnish limits between, above or below which, at selected confidence levels, we expect to find a prescribed proportion of the items of a population (Natrella 1966, pp. 2–13, 2–15).

Similarly to the one-sided confidence interval generated by (7.8), a tolerance limit x_1 can be calculated as

$$x_l = \overline{x} - ks. \tag{7.9}$$

The value for k is obtained for specified *n, p* (the proportion of the data to be included), and probability level, from appropriate tables (see, for example, Natrella 1966). Tables such as these may be used to provide statements of the following form: 'For a sample of size 10, the probability is 0.75 that 90% of the population will be greater than $\overline{x}$- 1.671s.'

It is important that the concepts above, and others, be understood by the analyst and quality assurance manager. Not only do they apply to straightforward quantitative analysis such as blood alcohol and drug purity, but they are also critical when association is to be established or rejected on the basis of quantitative measurements.

The standard deviations, s and *s,* of populations and samples, respectively, are often referred to in the calculation of various statistics. Formula (7.1) is the formula for the calculation of s for a set of measurements. Although it can be calculated, some experts feel uneasy about calculations of s based on fewer than five measurements. Taylor (1987) seemed to indicate that the minimum should be seven. There are a number of other useful methods for determining s (Mark & Workman 1987), two of which are described below.

The standard deviation can be estimated from a series of duplicate measurements. This series can be determined on the same sample on several occasions, or on duplicate measurements on several samples. If several different samples are employed, they should be similar enough in analyte and matrix, such that the required assumption, that the precision of measurement is essentially the same, is supportable. The standard deviation *s* can be calculated from (7.10) (Mark & Workman 1987):

$$s = \sqrt{\left\{\sum_{i=1}^{k} d_i^2/(2k)\right\}} \tag{7.10}$$

where *d* is the difference between a pair of duplicate measurements, *k* is the number of sets of duplicate measurements, and the summation is over *k*.

The standard deviation *s* may be related to the range *(R)* of a set of measurements. The range is the difference between the highest and lowest value of a set. The average range *($\overline{R}$)* may be calculated from a number *(k)* of sets of measurements. The estimate of *s* is obtained by employing (7.11)

$$s = \overline{R}/d_2^* \tag{7.11}$$

where d_2^* is obtained from appropriate tables (Taylor 1987, p. 265). Such tables also give estimates for the number of degrees of freedom which may be used in appropriate statistical tests. For example, given 5 sets of replicates with 4 replicates in each set, d_2^* is given as 2.10 and the value to use for degrees of freedom is 13.9.

7.7 SAMPLING

Many of the methods described here, as well as many others, require the sample evaluated to be random. In many cases it is not possible to prove the randomness of the samples, and it needs to be assumed (Rosner 1990, p. 43). A random sample is a selection of some members of a population in a manner such that each member of the sample is independently chosen and each member of the population has a non-zero probability of being chosen. Most samples are what are known as simple random samples wherein the probability of being selected is equal for each member of the population (see Chapter 2, section 2.2). This requires that the sample be the product of a random sampling plan. It should be noted that random does not mean haphazard. Random sampling requires a specific plan to ensure randomness. Plans designed to ensure randomness are routinely based on random number tables, and the reader desiring to develop such a plan is referred to further reading; see, for example, Rosner (1990, pp. 142–147), Crow *et al.* (1960, pp. 3, 83–85), Natrella (1966, Chapter 1), Duncan (1986).

A number of particular sampling problems are of particular concern to the forensic chemist. The first, the collection of a representative sample from a number of individual items in a population, involves random sampling and will not be treated here. Similarly, the methodology for obtaining a representative sample of a bulk powder, liquid, etc., by methods such as cone and quarter, are left for the reader to review in quality analytical chemistry texts.

Questions arise when a very large number of exhibits, 500 to 2000 ‘crack’ vials, for example, are submitted as one population, and the analysis of each vial is considered economically not feasible. The forensic scientist would like to select, randomly, a pre-ordained number of samples, analyse them, and, based on either all or almost all resulting in positive results, be able to state with a given probability a that a certain proportion of them are indeed positive. Frank *et al.* (1991) have approached this problem, employing hypergeometric probabilities, and have developed a satisfactory solution.

The above sampling problem, wherein the item measured results in a go or no-go, positive or negative, classification, is commonly referred to as a sampling by attributes. The military, in dealing with large sample volumes of items such as small arms ammunition, and civilian industries, in dealing with items such as fuses, have developed plans suitable for this type of testing.

The military employs Mil.STD-105D (USDOD 1963), while industry often adopts the slightly different counterpart ANSI/ASQC standard Z1.4-1981 (ASQC 1981) for monitoring a continuous process. Use of these plans which depend on the size of an inspected lot and an inspection level, allows for the prediction of a maximum percentage of ‘defective units’, at a given confidence level, in the population. This prediction is based on the number of ‘defective units’ detected in the sample. The reader is cautioned that, although 105D has been employed for drug lot evaluations (Springer & McClure 1988), the above methods are designed to guarantee an overall percentage of ‘acceptable’ items for continuous processes. Extrapolation to any particular lot is improper.

The testing for a certain characteristic, perfect or defective, positive or negative, is described by the binomal distribution. Here, for instance, we may wish to estimate the confidence interval about the true proportion *(P)* of positive items based on the proportion *(p)* of positive samples found in a sample of size *(n)*. Exact solutions for these intervals based on the binomial can be found in tables and charts in Natrella (1966, Tables A22–24), Crow *et al.* (1960, pp. 257–272), and Rosner (1990, pp. 543–544), for sample sizes up to thirty (30) and specific additional sample sizes up to one thousand (1000). For other sample sizes, the exact confidence interval for *P* can be determined by employing equations (Rosner 1990, p. 172), where the evaluation of expression (7.12) is required for various values of x, between 0 and n, inclusive,

$$\sum_{k=0}^{x} \binom{n}{k} p^{k}(1-p)^{n-k} \,. \tag{7.12}$$

This calculation is cumbersome, but is not exceedingly so with the aid of a computer. In most cases, as previously mentioned, when large sample sizes are employed, the evaluation of the intervals employing the Normal approximation is sufficiently accurate. Expressions enabling the calculation of the required sample size to estimate

P within stated error are available (Rosner 1990, pp. 173–175; Natrella 1966, pp. 7-2, 7-3).

7.8 ASSESSMENT AND CONTROL

It requires a program of quality assessment to maintain the quality level of a laboratory's product. The data generated by evaluating the measurement process via various assessment tests are employed to maintain control of the process and cause corrective action to be taken when necessary. The use of control charts is invaluable for tracking the performance of a measurement method.

A general assessment of the laboratory's quality can be made by outside agencies during accreditation inspections. However, an in-house, continuous assessment program must be maintained, using the analysis of a number of different types of quality control samples.

Calibration procedures and their results must be monitored, whether they are performed on a scheduled basis or oh a routine basis when an instrument is put into use. The charting of instrument performance characteristics such as sensitivity, linearity, and resolution can give an indication that replacement of parts (for example, a hollow cathode source in an A.A. spectrophotometer), or rejuvenation of a detector (for example, solid state, Li, Si, in EDS spectrometer), are, or will be, required in the near future. In chemical analysis requiring extraction techniques, a methods recovery rate should be monitored.

The control chart, referred to as an *X* (control) chart for single measurements or $\overline{X}$ (read as *X*-bar) (control) chart for means of measurements, is often employed to assess the above types of performance. Fig. 7.3 displays representative control charts of this type.

Fig. 7.3a could represent the sensitivity check of an atomic absorption spectrometer with a standard analyte concentration. The dashed lines indicate performance limits determined by a requirement of the analytical method. The positioning of a large number of data points above or below the dotted lines, either scattered or grouped, is used as an indication of lack of statistical control. Occurrence of most of the points below the central line, but within the dotted limit, is an indication of a bias. A continued trend by the points in a downward direction is an indication of gradual loss of sensitivity which could, in this case, be caused by source degradation.

Fig. 7.3b represents the type of *X* control chart that might be used to track the performance of the measurement of a calibration standard or a check sample. Performance is monitored about the expected or standard value with upper and lower warning limits (UWL, LWL) normally set at ± one standard deviation (s.d.). The respective control limits (UCL, LCL) are, in this case, set at 1.5 s.d. Performance limits are often set at 95% confidence, approximately 2 s.d., and occasionally, as high as 3 s.d. The occurrence of more than a few data points greater than the warning limits is a cause for concern, while more than the rare event of a point outside the control limit may be cause to cease the use of that measurement method or process until statistical control returns. The population s.d. is never known. It is replaced by an estimate based, in some way, on previous knowledge. No practical problems should be caused by this procedure except with very small data sets.

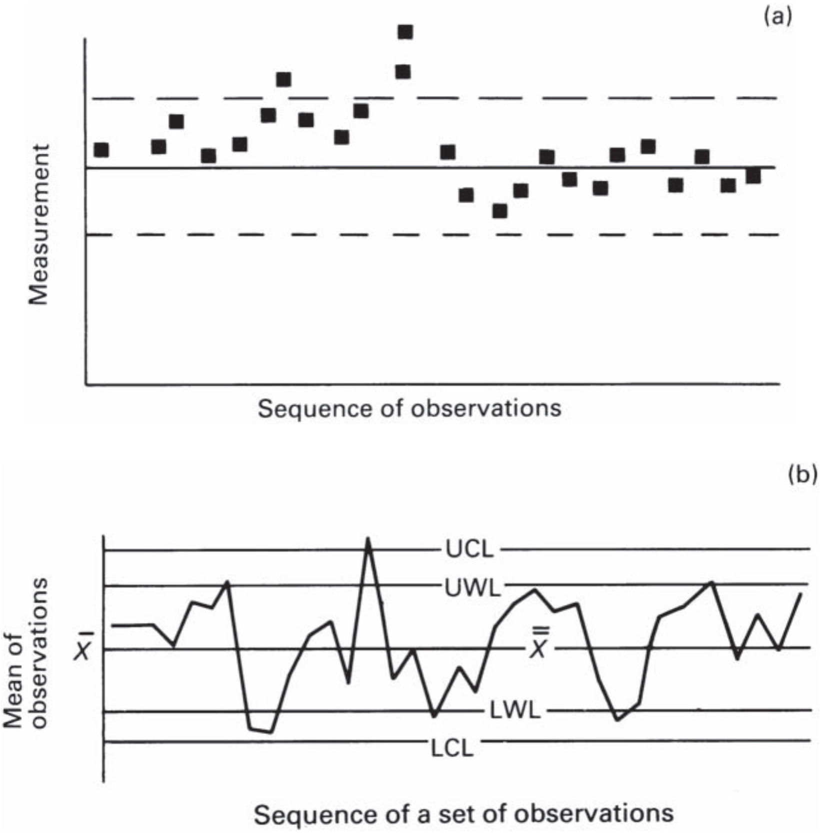


Fig. 7.3 —(a) A sample X control chart with performance limits indicated by broken lines. (b) A sample $\overline{X}$ control chart with the expected value ($\overline{X}$) and upper (U) and lower (L) control (C) and warning (W) limits (UCL, UWL, LWL and LCL) indicated.

An *R* control chart, which plots the measurement range and which has a minimim value of zero, is another useful control tool. It is similar to an $\overline{X}$ chart but represents, and is used to control, process variability. Certainly, one would expect a chart (*S*) based on the standard deviation, to be more valuable than the *R* chart, because the standard deviation is less biased. But the *S* distribution is not Normally distributed even when the measurement data are. This results in the *S* chart being more problematical to interpret. The popularity of the *R* chart results from its ease of interpretation and the fact that often, before $\overline{X}$ charts can be successfully employed, statistical control must be established. *R* charts aid in proving statistical control. Examples and detailed application of the use of the control charts mentioned above and others can be found in a number of texts (Duncan 1986, part 4, Taylor 1987, Chapter 14, Ryan 1989).

A number of different types of control samples should be employed to assess laboratory performance. Check samples are samples supplied to or by the individual analyst to monitor performance. The analyst is usually aware of the expected result.

Performance tracking by control charts can be maintained for each type of control sample listed, as well as the re-analyses of replicate samples by analyst, method, laboratory, and even instrument. All of the foregoing are recommended.

Proficiency samples are those submitted to the measurement process with the expected result unknown to the analyst. He is, however, aware that it is a proficiency test. Blind proficiency tests are, of course, the best type of control samples to employ, especially for a forensic laboratory, as they can be used to assess the entire laboratory process including receipt, transmission, communication, turnaround time, results, and interpretation.

It is recognized that the implementation of a ‘blind’ proficiency testing program can present almost unsurmountable problems for a quality control manager or laboratory director, but the information obtained from such a program can be the most useful.

Participation in a proficiency testing program managed by an agency outside the particular laboratory can also be invaluable in aiding in the assessment and improvement of quality. The results, when supplied, usually contain not only the expected values, as determined from reference laboratories, but also an indication of the performance of other laboratories and their methodologies. Error analysis of the submitted results also usually accompanies the returned report to the laboratory.

Replicate analysis of true case samples conducted by the same analyst, or by another, without knowledge that it is a replicate, can be a useful tool for quality assessment, especially of quantitative results. This type of evaluation, again, requires that statistical control has been established.

When the difference (X_1-X_2) between the two replicate measurements exceeds the test value calculated, one can assume that the difference is not due to chance alone, and one of the values is likely to be an error. The 2.77 constant in the equation

$$X_1\text{-}X_2 = 2.77s,$$

where s is the standard deviation, provides an upper limit that (X_1-X_2) is not expected to exceed more than 5% of the time (Duncan 1986, pp. 151–152, 1006–1007).

Interlaboratory collaborative testing studies are powerful tools for the validation of methods and assessment of laboratory biases and precision. The conduct of any collaborative test requires careful planning and execution (ASTM 1988, Youden & Steiner 1975, pp. 9–13, 27–36). If this is not done, not only may the data obtained from a significant effort be useless, but worse, the test may result in unwarranted and undesirable conclusions.

Treatment of the data can likewise be complex and confounding owing to all the origins of error that may be possible. The reader is referred to Youden & Steiner (1975, pp. 13–26, 36–63). A data analysis software package is available as a complement to the ASTM methodology (ASTM Interlaboratory Data Analyses Software, E691, available from American Society for Testing and Materials, Philadelphia, Pennsylvania, USA). Youden & Steiner (1975, pp. 99–104) have described a method using so-called two-way plots to evaluate the performance of laboratories, during which they analyse two samples of the same material of similar but not identical composition. An example of such a two-way plot appears as Fig. 7.4.

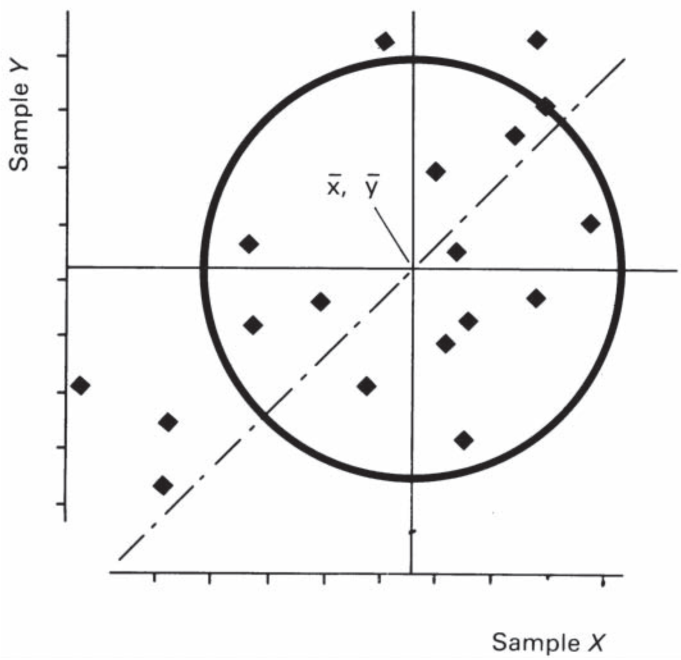


Fig. 7.4 —A two-way plot used in the evaluation of laboratory performance by plotting the results of analyses of two samples (*X* and *Y*) of the same material of similar but not identical composition. The horizontal and vertical lines represent the average values determined for samples *Y* and *X,* respectively. The broken line indicates the line *X* = *Y.* A circular confidence limit is indicated.

The horizontal and vertical lines that intersect represent the average values determined for samples *Y* and *X,* respectively; the units are arbitrary. Note that the horizontal, vertical, and 45° lines each divide the data points into approximately equal groups, but the data are not equally distributed between the four quadrants. There are more points in the upper right or lower left quadrants, indicating that systematic biases have a greater effect than random errors. Dispersion along the dotted 45° line indicates that laboratories are high or low on both samples, while dispersion orthogonal to this line is an indication of lack of agreement between results from a laboratory. If systematic errors were trivial, we would expect the points to tend to array as a circle. A circular limit about the mean point can be calculated at a desired confidence level. For further details on interpretation of this plot see Youden & Steiner (1975, p. 103).

Whenever one desires to evaluate interlaboratory performance, one should be mindful of the difficulties that will result if the collaborative samples are at trace quantities below LOQ and, worse, near LOD. A consistently increasing coefficient of variation (CV), or Relative Standard Deviation (RSD), given by (7.13), as concentration decreases, and represented by the classic horn-shaped confidence limits described by Horwitz *et al.* (1980), will make data interpretation problematical. The caveat ‘Make proficiency samples reasonable!’ should be followed.

$$CV = RSD = \frac{s}{\bar{x}} \ . \qquad (7.13)$$

7.9 CORRECTIVE ACTION

The assessment process may be successful in fixing the location of a specific problem, such as an incorrect standard, or may just point to a broad area where tighter control is needed, for example, sample control.

Generally, laboratory measurement errors are assignable to three general types: blunders, imprecision, and bias. Any technically significant problem identified should be investigated and, after the cause is minimized, if not eliminated, the result will be an overall higher quality measurement system.

Any of the following can lead to measurement blunders:

wrong sample,
improper method,
contamination,
wrong reading,
sampling problem,
miscalibration,
losses,
transposition and transcription errors,
lack of statistical control.

Some causes of bias that have been identified are:

interferences,
instrument warm-up,
matrix effects,
calibration problems,
improper controls or blanks,
samples losses,
operator bias.

The principal causes of lack of measurement precision are as follows:

instrument instability,
environmental effects,
operator ability,
reagent stability,
variation in blanks, standards, and controls,

unverified methods,
variable recovery rates,
failure to maintain statistical tolerances.

Record-keeping by all parties in the measurement process is critical to the location and correction of problems. It is equally important for analysts to retain data that they believe are erroneous. Decisions concerning rejection of outlying data need to be supported on sound technical or statistical bases. Even in a forensic laboratory, where an admission to the committing of an error may seem to be catastrophic, the documentation of the evaluation and resolution of measurement problems, with supporting records, is a necessity.

Every laboratory system is unique in the problems it encounters, the volume of measurements performed, and the types of clients' needs it services. Every laboratory director should be familiar with the general concepts outlined in this brief discussion, and, with the assistance of a competent quality assurance manager, should be able to elevate the quality of the product emanating from his laboratory and maintain it at a high level.

REFERENCES

AIHA, American Industrial Hygiene Association (1989) *Guidelines for laboratory accreditation,* American Industrial Hygiene Association Laboratory Accreditation Committee, Akron, Ohio, USA.

ASQC, American Society for Quality Control (1981) *Sampling procedures and tables for inspection by attributes,* ANSI/ASQC Standard Z1.4–1981, American Society for Quality Control, Milwaukee, Wisconsin, USA.

ASTM, American Society for Testing Materials (1988) *Standard practice for conducting an interlaboratory study to determine the precision of a test method,* STD E691–87, American Society for Testing Materials, Philadelphia, Pennysylvania, USA.

Barlow, R.J. (1989) *Statistics: a guide to the use of statistical methods in the physical sciences,* John Wiley & Sons, New York, USA, pp. 24–35.

Berger, H.W., Galowin, L.S., Horlick, J., Steel, E., & Verkouteren, J. (1988) *Bulk asbestos handbook: Operational and technical requirements of the laboratory accreditation program for bulk asbestos analysis,* U.S. Department of Commerce, National Bureau of Standards, Gaithersburg, Maryland, USA.

Buscaglia, J. (1991), Quality Assurance Supervisor, TAKA Asbestos Analytical Services, Northport, New York, USA, personal communication.

Clayton, C.A., Hines, J.W., & Elkins, P.D. (1987) Detection limits and specified assurance probabilities. *Analytical Chemistry* **59** 2506–2514.

Crow, E.L., Davis, F.A., & Maxfield, M.W. (1960) *Statistics manual.* Dover, New York, USA.

Duncan, A.J. (1986) *Quality control and industrial statistics.* 5th ed., Irwin, Homewood, Illinois, USA.

Eller, P.M. (ed.) (1984) *NIOSH manual of analytical methods.* 3rd ed., U.S. Government Printing Office, Washington, DC, USA.

Emma, T. (1991) Director of Laboratories, Hunter Environmental Laboratories, Waltham, Massachussets, USA, personal communication.

EPA (U.S. Environmental Protection Agency) (1968) *Test methods for evaluating solid waste.* Vol. 1A: laboratory manual physical/chemical methods, U.S. Government Printing Office, Washington, DC, USA.

FBI, U.S. Federal Bureau of Investigation Laboratory (1990) *Gunshot primer residue examination policy.* FBI Laboratory, Elemental and Metals Analysis Unit, Washington, DC, USA.

Frank, R.S., Hinkley, S.W., & Hoffman, C.G. (1991) Representative sampling of drug seizures in multiple containers. *Journal of Forensic Sciences* **36** 350.

Gallo, J.L. & Field, K.S. (1980) *Criminalistics methods of analysis feasibility study, final report.* Forensic Science Foundation, Colorado Springs, Colorado, USA.

Garfield, F.M. (1984) *Quality assurance principles for analytical laboratories.* Association of Official Analytical Chemists, Arlington, Virginia, USA, p. 4.

Horwitz, W., Kamps, L.R., & Boyer, K.W. (1980) Quality assurance in the analyses of foods for trace constituents. *Journal of the Association of Official Analytical Chemists* **63** 1344–1354.

Mandel, J. (1984) Fitting straight lines when both variables are subject to error. *J. Qual. Technol.* **16** 1–14.

Mark, H. & Workman, J. (1987) Alternative ways to calculate standard deviations. *Spectroscopy* **2** 38.

Mark, H. & Workman, J. (1989) One- and two-tailed tests. *Spectroscopy* **4** 52.

McCully, K.A. & Lee, J.G. (1980) Q.A. of sample analysis in the chemical laboratory. In: Garfield, F.M. *et al.* (eds.) *Optimizing chemical laboratory performance through the application of quality assurance principles.* Association of Official Analytical Chemists, Arlington, Virginia, USA. pp. 57–86.

Miller, J.C. & Miller, J.N. (1988) *Statistics for analytical chemistry.* 2nd ed., Ellis Horwood, Chichester, UK.

Natrella, M.G. (1963) *Experimental statistics, NBS Handbook* 91. National Bureau of Standards, Gaithersburg, Maryland, USA.

NIDA (U.S. National Institute of Drug Addiction) (1989) *Guidelines for quality assurance in personal drug abuse screening.* National Institute of Drug Addiction (NIDA) accreditation program, NIDA, Washington, DC, USA.

NIOSH (U.S. National Institute for Occupational Safety and Health) (1990) *NIOSH Specifications for industrial hygiene laboratory quality program requirements.* U.S. National Institute for Occupational Safety and Health, Cincinnati, Ohio, USA.

Rosner, B. (1990) *Fundamentals of biostatistics*. PSW-Kent, Boston, Massachussets, USA.

Ryan, T.P. (1989) *Statistical method for quality improvement.* Part II, John Wiley & Sons, New York, USA.

Sensabaugh, G.F. (1979) The quantitative acid phosphatase test, a statistical analysis of endogenous and postcoital acid phosphatase levels in the vagina. *Journal of Forensic Sciences* **24** 346–365.

Springer, J.A. & McClure, F.D. (1988) Statistical sampling approaches. *J. Assoc. Off. Anal. Chem.* **71** 246–250.

Taylor, J.K. (1985) *Principles of quality assurance of chemical measurements NBSIR85–3105*. National Bureau of Standards, Gaithersburg, Maryland, USA.

Taylor, J.K. (1987) *Quality assurance of chemical measurements.* Lewis Publishers, Chelsea, Michigan, USA.

USDOD (U.S. Department of Defense) (1963) *Sampling procedures and tables for inspection by attributes, Mil-STD-105D.* U.S. Government Printing Office, Washington, DC, USA.

Wilcox, K.R. (1977) Laboratory management. In: Inhorn, S.L. (ed.) *Quality assurance practices for health laboratories.* American Public Health Association, Washington, DC, USA. pp. 3–126.

Youden, W.J. & Steiner, E.H. (1975) *Statistical manual of the Association of Official Analytical Chemists.* Association of Official Analytical Chemists, Arlington, Virginia, USA.

BIBLIOGRAPHY

ASTM (1988) *Standards on precision and bias for various applications,* 3rd ed., American Society for Testing and Materials, Philadelphia, Pennysylvania, USA.

Crow, E.L. *et al.* (1960) *Statistical manual.* Dover Publications Inc., New York, USA.

Doudy, S. & Weardon, S. (1983) *Statistics for research.* John Wiley & Sons, New York, USA.

Duncan, A.J. (1986) *Quality control and industrial statistics.* 5th ed., Irwin, Homewood, Illinois, USA.

Dux, J.P. (1986) *Handbook of quality assurance for analytical chemistry laboratory.* Van Nostrand Reinhold Co., New York, USA.

Garfield, F.M. (1984) *Quality assurance principles for analytical laboratories.* Association of Official Chemists, Inc., Arlington, Virginia, USA.

Mason, R.L. *et al.* (1989) *Statistical design andanalysis of experiments.* John Wiley & Sons, New York, USA.

Miller, J.C. & Miller, J.N. (1988) *Statistics for analytical chemistry.* 2nd ed., Ellis Horwood, Chichester, UK.

Natrella, M.G. (1966) *Experimental statistics NBS handbook* 91. U.S. Government Printing Office, National Bureau of Standards, Washington, DC 20402, USA.

Rosner, B. (1990) *Fundamental of biostatistics.* 3rd ed., PWS-Kent Publishing Co., Boston, Massachussets, USA.

Ryan, T.P. (1989) *Statistical methods for quality improvement.* John Wiley & Sons, New York, USA.

Scott, W.W. (1939) *Standard methods of chemical analysis.* 5th ed., Van Nostrand Co., Inc., New York, USA.

Taylor, J.K. (1987) *Quality assurance of chemical measurements.* Lewis Publishers Inc, Chelsea, Michigan, USA.

Wernimont, G.T. & Spindly, W. (eds) (1985) *Use of statistics to develop and evaluate analytical methods.* Association of Official Analytical Chemists, Arlington, Virginia, USA.

Youden, W.J. & Steiner, E.H. (1975) *Statistical manual of the Association of Official Analytical Chemists.* Association of Official Analytical Chemists, Arlington, Virginia, USA.

Index

Note: Surnames beginning Mac,. Mc have been placed before all other entries beginning with ‘M’ or ‘m’.