Reasonable Doubt: Uncertainty in Education, Science and Law

This is the ms for:
Gardner-Medwin AR (2011) Proc. British Academy 171, 465-483. Ch.17 in *Evidence, Inference and Enquiry* (Ed. Twining W, Dawid P & Vasilaki D) OUP/British Academy
ISBN13: 9780197264843 ISBN10: 0197264840

Abstract

The use of evidence to resolve uncertainties is key to many endeavours, most conspicuously science and law. Despite this, the logic of uncertainty is seldom taught explicitly, and often seems misunderstood. Traditional educational practice even fails to encourage students to identify uncertainty when they express knowledge, though mark schemes that reward the identification of reliable and uncertain responses have long been shown to encourage more insightful understanding. In our information-rich society the ability to identify uncertainty is often more important than the possession of knowledge itself.

In both science and law there are fundamentally different kinds of uncertainty at issue. There is uncertainty whether a particular hypothesis is correct, and there is uncertainty about observable data that may be generated if a particular hypothesis is correct. Both are expressed in terms of probabilities. Each has its own domain of application and its own logic, but the inter-relationship is complex and sometimes misunderstood. Hypothesis probabilities are always open to error through possible failure to take account of realistic alternatives, while the proper inferences that can be drawn from data probabilities (often in the context of significance testing) are quite limited and easily over-interpreted.

When considering these two kinds of probability in a court of law it is possible to interpret the phrase 'reasonable doubt' in different ways. It can be seen as addressing data uncertainty: whether such incriminating evidence might with reasonable probability arise to confront an innocent person. Or (the more conventional view) it can be seen as some sort of threshold level on the probability that the defendant is guilty (a hypothesis probability). Each typically involves elements of subjective judgement, but fewer issues and uncertainties arise when considering the data probability and it is argued that this is often the more critical and proper issue for a jury to address. This has particular repercussions for cases involving identification of a suspect through trawl of a DNA or other database.

.....

Uncertainty and misconception in educational assessment

As a university teacher (in medical science) I have tried to raise students' awareness of uncertainties in their own knowledge. In conventional educational assessment students are often motivated to hide uncertainties. This is perverse, because it is obvious that it is a good thing to be able to distinguish between reliable and unreliable aspects of one's knowledge. Students should in fact be rewarded for acknowledging uncertainties that they have. By rewarding them, for example with simple but carefully designed mark schemes, they can be encouraged to reflect on the nature of any doubts, and on the evidence that they may be able to bring to bear to resolve these doubts. Instead of doing this, conventional marking usually encourages students to bluff their way through: it treats lucky guesses in the same way as well justified knowledge, and firm misconceptions as no worse than acknowledged ignorance. The result is a "go-for-it" culture in which decisions may be taken in the light of a marginal or superficial preference for some option, with little further thought.

The dangers of such behaviour are very obvious in a field such as medicine, where lives can be lost through reluctance to acknowledge uncertainty. Even in ordinary discourse however, it is a sign of weakness to shy from proper awareness and acknowledgement of one's uncertainties. It is disturbing, when talking to students familiar with essay writing – where you might think the discussion of uncertainties would be paramount – that they often say they would not normally mention doubts about a fact or argument they are tempted to include. Anyone who has marked exam essays knows that in practice a qualification, perhaps a

question mark in the margin, will make little difference to one's judgement if the point made is correct, while it makes a big difference if uncertainty is acknowledged about something incorrect.

It may at first seem paradoxical to reward someone for acknowledging uncertainty. Perhaps it runs counter to society's modern guidelines for "*How to get on in ….* " *politics, business*, and maybe sometimes even *science* or *law*. But the rewards for caution in the face of uncertainty are fundamental to biological survival. Children and young animals learn these lessons often through games, and it is no coincidence that the relevant mathematics comes under the heading 'Game Theory' (von Neumann & Morgenstern, 1944). Of course certainty also brings its own rewards - just as long as it is certainty about something that is correct. To pursue a cricketing analogy, the decision to try to hit a ball for six is fine, so long as one is pretty sure one can succeed: the price of failure tends to be high. If one is uncertain of success, then a more modest stroke will on average be better rewarded. Exactly the same principle is employed in the certainty-based mark scheme we employ for self-tests at UCL and Imperial College in London (Gardner-Medwin 1995, 2006). Acknowledging uncertainty (<67% estimated probability of being correct) gives 1 mark for a correct answer and 0 if incorrect. Claiming a high level of certainty (>80%) gives 3 marks or -6, while an intermediate level gives 2 marks or -2. This is a 'proper' or motivating mark scheme, in the sense that a student always expects to gain by giving an accurate indication of how reliable he/she thinks the chosen answer is (Dawid, 1986). With this scheme there is no benefit to be gained by the student trying to 'play the system': pretending to be either more or less confident than he/she really is. It is remarkable how readily students take to this scheme, appreciating the good sense and utilitarian value of settling for modest rewards and penalties when unsure.

Uncertainty about the expression of knowledge or understanding is an example of uncertainty about ideas. Much of educational testing revolves around ideas that are generally accepted (at least within an agreed framework) as definitely either true or false, though students with partial knowledge may assign probabilities less than 1 and those with serious misconceptions may assign very low probabilities to what is correct. There are of course many elements to educational experience where there is no certainty and the task is to generate or discuss ideas that are neither known nor agreed to be objectively correct or incorrect. Here we come closer to the issues that arise in scientific research and in courts of law.

Uncertainty about ideas: the driving force in science

Scientists work with two fundamentally different kinds of uncertainty, though they don't always distinguish them very clearly. The first is an intrinsically subjective probability: *How certain is it that a particular idea or hypothesis is correct?* This is obviously fundamental within science, but the answers to such questions are always subjective in the sense that they may differ even between well informed and intelligent individuals. By gaining additional evidence one may shift such subjective probabilities in ways that all agree are rational and occasionally even quantifiable. Indeed, as Popper (1959) pointed out, one can sometimes demonstrate that an idea is certainly wrong, though never that it is certainly right as a general truth. Different levels of scepticism and uncertainty amongst individuals, about what is correct, may never be eliminated. Nevertheless, there are of course great swathes of scientific conclusions that are regarded, by more or less all who have studied them, as essentially certain. It is the power of science that the accumulation of evidence, given enough time and effort, tends to become overwhelming on one side or the other of any argument. However, evidence does not always diminish uncertainty, and even firmly embedded ideas can occasionally be challenged or overturned, even by simple pieces of evidence. Uncertainty about ideas is the motivation for progress in science, despite its nebulous and largely unquantifiable nature. It drives the testing of predictions and the devising of alternative theories and experimental challenges.

Darwin's theory of natural selection is a prime example of an idea that was conceived with a huge element of uncertainty, not least in Darwin's mind. Darwin's great success was that he

amassed a vast amount of evidence in its favour before going public: enough evidence to diminish his own uncertainty and to shake the conviction of many sceptics. There never was (and never will be) a time when Darwin's theory is totally certain, beyond any meaningful threshold. Of course, nearly every scientist believes, as I certainly do, in this theory. Darwin's idea is as firmly embedded as any in science, thanks to the progressive elucidation of biological mechanisms and the accumulation of data that are all (so far as I know) consistent with the core elements and predictions of this theory. But there can be no certainty that new ideas and observations could not change it. A scientist's natural response to the question "Is Darwin's theory beyond reasonable doubt?" is to say "What doubt? Let's see if we can devise a test for any ideas about how it may be wrong". Scientists love coherent scepticism, just so long as the ideas are testable. The history of science teaches them to regard dogmatic beliefs as pointless and arrogant. The strength of the chief tenets of science is not that they are certain, but that to overturn them would require not only observations inconsistent with them, but also a new explanation for all the evidence that has hitherto appeared to support them. Darwin achieved this in relation to the notion of divine creation, as did Einstein with classical mechanics. Scientists seldom of course achieve such dramatic revolutionary impact; but even when the pursuit of uncertainty doesn't overturn or strengthen existing ideas, it often throws light on facets that were previously unclear.

Uncertainty about expected data

The second kind of uncertainty is conceptually more straightforward and quantitative, though often underpinned by complex mathematics. This is uncertainty about the data that will arise in an experiment, given that the relevant mechanisms and principles are either fully understood or fully defined by a hypothetical model. Statistical analysis of such uncertainty is part of any scientific training (and the bane of many students!). It gives a framework for evaluation of new results, since a novel claim is of little value if the novelty or interest of the evidence could quite likely have arisen with conventional or uninteresting assumptions. The commonest format is that of 'significance testing', initiated by Fisher (1925). The assumptions on which a significance test is based are described as a 'null hypothesis' Ho. A point of potential interest is encapsulated in a statistic from the observations, quantifying a difference from the mean or median value expected on Ho. The outcome of the test is the probability ('P-value') based on Ho that, with the procedures adopted in the research, at least as large a value would be observed.

Since the use and misuse of significance tests and P-values has a huge and controversial literature (see for example very approachable reviews by Royall (1997) and Senn (2003)), I shall present here just a personal perspective on some of these issues. A fundamental point, clear I hope in the description above, is that a P-value is a probability of observing data, conditional on a specific hypothesis. It does not express uncertainty about hypotheses. P-values are hugely valuable in science because they quantify the degree to which data are consistent with some conventional or postulated idea. Perhaps the most useful role they perform is when the P-value is large and the result is 'not-significant': this means that the point of interest in the data would be quite likely to arise even if Ho is true, so there is little reason to pay attention to this aspect of the data as any kind of challenge to Ho. One can conclude this quite straightforwardly, without any consideration of alternative hypotheses or the rather nebulous business, discussed in the last section, of addressing the probabilities that hypotheses may be correct.

A hazard in the interpretation of P-values arises from terminology often used to express conclusions based on them. Low P-values are often said to justify "rejection of H_0" at a particular significance level. This is only true in a very restricted formal sense, when a P-value is used as a decision criterion between two actions that would be appropriate if there were solid reasons to accept or reject Ho. For example one might decide to ignore the data because it is consistent with Ho, or start researching alternative hypotheses if the data would be surprising on the basis of Ho. If there are alternative hypotheses already formulated, with some

basis for assigning them relative probabilities, then a result with a low P-value doesn't necessarily even argue against Ho, let alone justify rejecting it. An interesting result may be so extreme that it is very unlikely on Ho, but it may be even less likely on whatever alternative hypotheses are considered plausible. As an example, suppose that 10 tosses of a coin yield 9 heads and one tail. Does this result, surprisingly far from 50:50 (P-value=2%) argue for rejection of a hypothesis (H_o) that this is a fair coin? If the only alternatives seem to be that a coin might have 2 heads or 2 tails, then the data clearly shows both of these false, supporting Ho. However, if one envisages the possibility that a coin might be biassed so as to be more likely to land one way, and considers this plausible, then the data lend more support to this idea than to Ho. Whether one ends up considering Ho more or less likely than at the outset depends on the initial probabilities one assigns to any of these (and possibly other) ideas - one's prior probabilities. In science, especially at the frontiers, the alternatives to a well formulated hypothesis are often simply a matter of speculation. In that situation all one can sensibly say about data with a low P-value is that the parameter of interest is surprisingly far from expectation on the basis of Ho (or similar hypotheses) and that other hypotheses, not necessarily very plausible or even thought of, could have rendered it less surprising.

The borderline between uncertainty about hypotheses and data

A useful extension of significance tests is the calculation of so-called 'confidence limits': a range of null hypotheses that would yield P-values above a specified level for a result of interest in the data - in other words, they would make the result unsurprising. This practice helps to clarify how useful the observations are: how well they serve to discriminate between different values of a parameter within a hypothetical model; but it still says nothing about the probability that one should assign to the correctness of any of these models. The commonly used terminology is again, in my view, misleading because the term '95% confidence limits' for a parameter in a model can easily be understood to mean that there is a 95% probability that the true parameter lies within these limits. This is simply not so, as is evident if you think about an experiment that tries to measure an effect that you consider to be almost certainly non-existent (perhaps extrasensory perception or homeopathy). Whatever the protocol for the experiment, you should expect on 5% of occasions that a result will be obtained for which a nil effect is outside the calculated confidence limits. If you happen to experience one of these results it would obviously be inappropriate to adopt a belief that there is a 95% probability that the postulated effect is real. Perhaps 'consonance limits' (Kempthorne & Folks, 1971) would better convey the true meaning, which is a range of models for which the data would be relatively commonplace - falling within the 95% most likely outcomes.

In general it is important to make a category distinction between the probability that data will arise and the probability that a hypothesis about the generation of data is correct. Sometimes, however, there is a simple relationship between the two. For example, an experiment may estimate (by sample measurements) the mean value of a parameter that is considered at the outset equally likely to take any of a wide range of values: perhaps some characteristic of a newly encountered substance, object or species. Then (with a few, often justifiable, assumptions[1]) the posterior probabilities of different hypotheses about the true mean are the same as the data probabilities that would be calculated for sample means, on the hypothesis that the true mean is what was actually measured. Returning to terminology, this means that when no hypothesis initially has preferred status, the limits cautiously described above as '95% consonance limits' are truly '95% confidence limits' - they identify a range that one can be fairly confident includes the true mean (with 95% probability based on only these data). Examples like this where one can convert directly from a data probability to a hypothesis probability are not rare, but they lead to confusion if their limitations are not properly taken into account.

[1] For example, measurements that are normally distributed random variables with mean μ, where the prior probability distribution for μ is essentially uniform over a range much larger than the measurement standard deviation.

When alternative hypotheses are well defined and do have special status in an experiment, for example rival causal explanations of the data, then updating of probabilities for the alternatives requires Bayesian analysis (Lindley, 1972). As we saw above (in the coin example) attempts to make such inferences using P-values can be quite misleading since an unlikely result on a hypothesis Ho may actually be evidence in favour of Ho. Bayesian analysis involves identifying prior probabilities for the various hypotheses and updating these in the light of the data, using likelihood ratios (the relative probabilities that the data would be observed on the different hypotheses). The technicalities need not concern us here, but the procedure is demonstrably successful in situations where alternative hypotheses have clear prior probabilities. Even when this is not the case, and prior probabilities are little more than guesses, sufficient data can sometimes lead to highly reliable inferences (see for example, MacKay, 2003). The main reservation is that conclusions can be wholly misleading if the correct hypothesis has been omitted from consideration or assigned an inappropriate probability. If this is a major risk, as it can of course be at the frontiers of research, then the more limited logic by which significance tests challenge individual hypotheses can be a more comfortable basis for scientific progress.

An important application of Bayesian analysis is in medical diagnosis. Here one starts with fairly objective probabilities that different hypotheses might be true, for example that a person with a persistent cough may have each of a number of different diseases. Then these probabilities are updated on the basis of data from the patient's history and tests carried out, using information about how frequently the data would arise in the population for each of the conditions. This process may simply occur in the doctor's mind, leading eventually either to confidence about specific diagnoses or to residual uncertainty. Alternatively, a Bayesian computer algorithm may process the data to generate explicit posterior probabilities. Which is better is debatable, given the possible unreliability of the doctor's compilation of inferences set against his or her ability to take into account subtle aspects of the data ignored by a computer algorithm. But there is no disputing the appropriateness of this strategy to resolve medical uncertainties and to help make decisions.

Uncertainty in criminal trials

The uncertainties in a criminal trial have superficial similarities to the problems of medical diagnosis. The endpoint is a decision that turns graded uncertainties into specific actions: whether to convict a defendant, or apply treatment for a specific diagnosis. The relative utilities of right and wrong decisions may in part determine what level of doubt is acceptable for a final decision. In medicine this uncertainty is normally explicit, since the patient usually makes the final decision about whether to accept treatment, and must therefore be informed of doubts and likely outcomes. In law, the criterion for conviction ('beyond reasonable doubt') is substantially open to a jury's interpretation, with only partial knowledge of the utilities involved since a sentence for conviction may only be determined at a later stage by a judge. But these are not profound differences.

Legal cases often have many layers of complexity, for example concerning facts, motives, identification, intention and witness credibility. Final decisions may hinge on a complex synthesis of many types of doubt (Cohen, 1977; Anderson et al., 2005), and it is far from clear that mathematical approaches are helpful in coming to appropriate conclusions. Alternative conceptions of probability (for example Cohen's inductive probability based on qualitative eliminative reasoning) may sometimes be more rational (or at least more manageable for a jury) when deciding whether facts are "beyond reasonable doubt" in a trial. Such issues are much debated in the legal literature (see e.g. Tillers & Gottfried, 2006) with cogent arguments from many perspectives. However, the debate that follows naturally from the discussion here on the nature of uncertainties in science is not about whether probabilities should be quantified but whether we should be dealing with uncertainties about hypotheses or about data. Ideally, of course, one would like court decisions to be based on certainty rather than uncertainty, but this is rarely the case. If the evidence does support certainty, then concluding that defendant D definitely committed crime C (an assertion about a hypothesis probability) is equivalent to

concluding that the evidence could not have arisen if D were innocent (an assertion about a data probability): each conclusion follows from the other. But if uncertainties are present, these become different questions.

I have argued elsewhere (Gardner-Medwin, 2005) that it can be rational for a jury to come to the conclusion that the hypothesis of guilt is very probably correct, while acquitting on the grounds that the evidence could with reasonable likelihood have arisen for an innocent person. It may at first seem that these issues are opposite sides of the same coin, which would make my assertion either paradoxical or merely a matter of setting different thresholds for judging the two probabilities. I will first set out a caricature example, based on cases that first drew my interest (as discussed by Dawid, 2002) to show that this is not the case.

Suppose multiple infants in a family have died in circumstances consistent with either sudden infant death syndrome (SIDS: a rare medical condition that leaves no specific signs post mortem) or infanticide by the mother. This may reasonably lead to suspicion of crime, and indeed cases have come to court with little more in the way of pertinent evidence. It is obviously relevant to ask questions analogous to those underlying tests of significance: "How likely is it that such evidence would arise in an innocent family with comparable genetic, medical and socio-economic background, and how often would such cases be expected to arise in an innocent population?" Such questions cannot be answered with precision, but competent experts should be able to give reasonable ranges for the answers. Unfortunately the supposedly expert testimony in recent UK cases was not competent, raising serious concerns about how expert testimony should be validated (Royal Statistical Society, 2002); but this does not detract from the relevance of proper answers. The conclusion might be that the risk of the deaths that have brought this mother to court arising in her family without crime was at least a probability P, and that similar multiple deaths in families with no greater risk factors for multiple SIDS might be expected to arise somewhere in the UK once every X months. A jury might reasonably decide to acquit simply on the basis of such testimony, justifying their decision by saying that if juries convict on such evidence they might be responsible for convicting innocent mothers at the rate of one every X months in the UK, which they deem unacceptable. The acceptability criterion is of course a subjective matter: how many false convictions of this sort might be acceptable in a given population per year, decade or century in the interests of justice. But it is no more subjective than what constitutes "reasonable doubt" on any other basis.

An alternative approach is to decide such a case based on the probability that the hypothesis of guilt is correct, i.e. that this defendant is guilty within the meaning of the law. This view is often implicit in the legal literature and the media. It was clearly expressed in the letter from the Royal Statistical Society (2002) to the Lord Chancellor in the UK concerning the multiple SIDS case of R vs. Sally Clark, and was analysed in relation to this case in more detail by Dawid (2002). To quote the letter: "Two deaths by murder may well be even more unlikely [than two deaths by SIDS]. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation." The implication (made explicit by Dawid (2002) in Section 2.3 of his paper) is that a jury should use evidence about the incidence of this type of murder in the population, and convict if this is sufficiently greater than that of SIDS, taking account of all the known circumstances. This is certainly a rational way to infer probability of guilt in light of the evidence, but it has uncomfortable consequences when used as the basis for conviction or acquittal. In particular, it means that a defendant who would be acquitted on the grounds that the evidence is reasonably consistent with innocence (the argument of the last paragraph) might find herself convicted because sufficiently many other people have committed the crime of which she is accused (Gardner-Medwin, 2005). This seems ethically improper and probably unacceptable in law, because the voluntary criminal acts and intentions of people in other legal cases cannot reasonably be used as an argument to establish that a particular defendant has broken the law. Such evidence might be admissible to demonstrate, in unusual cases, that people are indeed sometimes capable of surprising

behaviour (for example, a mother killing her own children). But a liberal society would not be content, I think, with legal practice based on a utilitarian principle that if crime is rife then the law should convict with lower standards of evidence, so as to suppress crime at the cost of imprisoning the innocent (Gardner-Medwin, 2005). This would be the action of a totalitarian state. An enlightened jury must be prepared to acquit if the evidence could plausibly have arisen without guilt, however likely it may seem that the defendant is guilty on statistical grounds. If common crimes go unpunished as a result, this needs to be rectified by improving the quality of evidence rather than by lowering the threshold for conviction.

Uncertainty in trials with evidence of a definite crime

Trials are of course usually more complex and less amenable to quantitative analysis than the caricature discussed above. Commonly there is clear evidence of a crime, and amongst the many hypotheses that might account for this evidence there are two of primary concern: those put forward by the prosecution and defence. The prosecution case entails guilt on the part of the defendant, while that of the defence may not identify at all who is guilty, merely offering alternative explanation for evidence brought against the defendant and adding further evidence that may tend to prove innocence. Both the prosecution and defence hypotheses are typically composite in the sense that there are unknown elements for which there may at best be a reasonable set of assignable probabilities (for example that the victim either walked or got a lift from A to B). The situation is somewhat similar to comparison of two scientific hypotheses, where the relative likelihood of the data arising through the alternative explanations can provide a rational basis for increasing belief in one or the other. However, as we saw above, there can be fundamental problems in this process.

The first problem is the need for prior probabilities, from which to generate posterior probabilities in the light of the evidence. With scientific theories we saw how there may simply be no way of resolving differences of opinion about such priors. Similarly, in court the priors may depend on such dubiously subjective factors as a juror's expectation on being summoned to take part, suppositions about the rigour of the prosecution services, or how the defendant is dressed in court. Aspects of the background history of the defendant, which may or may not be allowed to emerge in court, may reasonably be considered relevant to prior probabilities and thereby to the posterior probability of guilt, but constraints and debate often centre round the fairness of admitting such evidence and the legal balance between potential prejudicial and probative value (see e.g. Roberts & Zuckerman, 2004: Chs. 4, 11).

A second problem about the use of posterior probabilities arises from the composite and incomplete nature of the hypotheses under consideration. The defence hypothesis seldom specifies who actually committed the crime. Anyone familiar with detective fiction (and probably, though I do not have experience, detective fact) knows that a prosecution case can look overwhelmingly stronger than a defence case up to the point when a new idea or emerging fact suddenly makes plausible a hitherto unconsidered suspect, motive or opportunity. In an ideal world such a development would have arisen and been investigated before trial, but miscarriages of justice can occur if this does not happen. This is analogous to the simple coin tossing example above, where the same evidence was seen either to support or negate the hypothesis that a coin is fair, depending on details of the alternative hypotheses that may have been thought of, and whether these were assigned significant prior probabilities. Miscarriages of justice can very easily occur if the true criminal is absent from consideration or seems above suspicion.

These problems concerning probability of guilt are analogous to the problems that make scientists shy from debate about the probability that a particular hypothesis is true. The probability of guilt cannot however be avoided altogether, because it would obviously be unreasonable to convict someone without somehow concluding that there is a high probability of guilt. But we have seen both ethical concerns and logical problems about treating this as a well-founded sufficient criterion for conviction. Hence, as in science, it is also constructive to

focus on the more limited but in some respects more direct inferences that may be drawn from data probabilities. In court, a relevant probability is whether an innocent person could have been brought to court to face at least the weight of incriminating evidence that has been seen for the defendant. This can be a substantial concern where there is a high profile crime, intense police effort and the opportunity to trawl a wide population: it becomes quite possible that an innocent suspect may be found against whom a convincing case can be made. The murder in London of Jill Dando and the subsequent conviction and acquittal of Barry George come to mind. A jury must be prepared to recognise that a seemingly strong case can often be made against a person innocent of the crime at issue.

In court there is not such a clear distinction to be made between hypothesis and data probabilities as there is in science. This is because hypotheses and explanations of data in court are not the well defined stochastic models that are the usual basis for calculation of data probabilities in science. There are subjective elements to the probability that evidence might arise without guilt, for example how likely it is that a particular scenario would develop or that witnesses may lie. Uncertainty about whether evidence might arise for an innocent defendant is therefore in some respects just as subjective and arguable as is uncertainty about the hypothesis of guilt. However, it involves fewer priors and can be judged within a more limited framework. This framework focuses largely on the things that could happen to innocent people, rather than on the behaviour of criminals. An incidental advantage (though hardly a reason in itself for preferring this approach) is that such scenarios may be easier for a jury to envisage.

Selection of suspects: the DNA database controversy

A hazard familiar to statisticians is the potential use of data twice over, once to select a hypothesis of interest (in this case, to pick a suspect from a population) and then again as evidence that the hypothesis (guilt of this suspect) is true. There can be an element of this in police protocols (as in the SIDS/murder cases, probably the Barry George case and many cases involving DNA matching). The procedures to avoid logical errors due to double use of data are not always appreciated by scientists, so it is unsurprising that juries may have difficulty handling the issue correctly. They may even not be aware that a suspect came to attention through a trawl of police records rather than through a connection with the case, since the fact that a suspect had a police record may be considered prejudicial in court (see e.g. Kaye, 2009 for examples and court rulings in relation to DNA testing). There are diverse views and heated debate in legal and statistical circles about how to handle evidence that serves these dual roles. For example, the American National Research Council Committee on DNA Forensic Science (1996), Stockmarr (1999) and Devlin (2000) advocate downgrading the weight of evidence when the suspect is identified by a single match in a database trawl, while Balding (2002, 2005), Dawid (2002), Kaye (2009) and others claim that the use of a database does not diminish the evidence - indeed strengthens it, albeit usually only slightly, by ruling out those in the database who test negative. The differences can correspond to large factors in the estimated probability of a false conviction: factors comparable (as shown below) with the number of potential suspects. As Dawid (2002) points out, differences between statistical approaches to uncertainty must rarely have such serious potential consequences for those affected.

The issues can be somewhat clarified by considering the safety of convictions - how likely the procedures are to incriminate innocent persons. Suppose that the person who left DNA at a crime scene is considered certain to be the true perpetrator of a crime (TP), and that the technical probability of the sample matching a random person other than the TP is a very small number (p), with no possibility that a match would fail to be evident if the TP is tested. Suppose there is a prime suspect who is estimated on the basis of non-genetic evidence to have a probability s=50% of being guilty (strong suspicion, but certainly not enough on its own to convict). If the DNA profile of this suspect is tested and matches the crime sample, then the probability that this match is false is p/(1+p), or almost exactly p. Conviction on this basis has a small estimated probability p of being a miscarriage of justice. The risk is greater if the

evidence against the prime suspect is weaker ($s<0.5$)[2], though with only a small number (M) of possible suspects this would make little difference for a test that confirms a prime suspect[2].

Cases involving trawl of a DNA database are more complex. Suppose there are initially no prime suspects based on non-genetic evidence, but a large number N of potential suspects (perhaps all the males of a plausible age within a city on a given day). Suppose a DNA database contains profiles for a random subset including D of these possible suspects, and that a search of the database has revealed exactly one of these as matching the DNA at the crime scene. There is at the outset a probability D/N that the database includes the true perpetrator (TP), and therefore a probability $(D/N)(1-p)^{D-1}$ that he would give rise to a unique match, taking account of the probability $(1-p)^{D-1}$ that no additional match arises by chance for any of the (D-1) innocent persons in the database. But there is also a probability $(1-D/N)pD(1-p)^{D-1}$ that the TP is not in the database and that exactly one of the innocent persons in the database does match. Given the fact that a unique match has been found, the probability that this is the TP is $1/(1+p(N-D))$. There is a complementary probability (approximately $p(N-D)$ if this is $<<1$) that this is a false match, typically much greater than the probability of a false match when a single test is carried out on a prime suspect[3]. The factor by which this risk is increased compared with the illustrative example in the last paragraph is (N-D), the number of potential suspects outside the database - potentially 100,000 or more in some cases (Kaye 2009)[4]. Of course a match found in a DNA trawl leads to investigation of the suspect and further non-genetic evidence. In some cases this may lead to confident conviction (e.g. if the suspect turns out to match a CCTV image from the crime scene) or elimination or acquittal if the match is clearly false (e.g. if there is a definite alibi). But DNA evidence often appears to the jury to be the strongest evidence in a case and conviction following a DNA trawl may even result when all other evidence seems to weigh in the defendant's favour (Donnelly, 2005; Kaye, 2009). It is crucial in such cases that a jury should recognise that the DNA data comes from a trawl and carries a much greater risk of false incrimination than if it were confirming suspicion of a prime suspect. Without this, such trials must be considered unsafe.

How are we to reconcile this conclusion with the seemingly contrary Bayesian argument presented by Dawid (2002) and Kaye (2009) that the order in which DNA and non-genetic data are obtained is immaterial? They argue that since a trawl eliminates suspects as well as identifying one who matches, when combined with other evidence it will rationally lead to a probability of guilt as great or greater than would be inferred if the same person had been tested as the prime suspect. This argument is illustrated in Fig. 1 for the simple case considered above, where there is strong non-genetic evidence (E) against a defendant, either emerging at the outset to justify confirmatory DNA testing of the defendant as a prime suspect or else emerging after the suspect has been identified in a trawl. This shows how the odds on the suspect's guilt are accumulated to similar outcomes, simply in a different order. This near equivalence of outcomes has been used to challenge the notion that juries need to be informed of the special facts and statistical considerations when suspect identification occurs through trawls or multiple testing (Balding 2002, 2005; Kaye, 2009).

---

[2] If the prior probability of guilt of the matching suspect is s, based on non-genetic evidence and elimination of any previously tested suspects, then the probability that this match is false is approximately $p(1-s)/s$ (assuming this itself is small). Even if s for the prime suspect is considered unquantifiable, but there are at most M suspects who are at all plausible, s must rationally be taken as at least 1/M on the basis that his probability of guilt must be considered at least as high as each of the others. This sets an upper bound on the probability of a false match as $p(M-1)$.

[3] A different derivation of the same conclusion is given by Song et al. (2009).

[4] Note that a larger and more relevant database (larger D) reduces the risk that a unique match will be a false match. This can be seen as an argument in favour of maintaining large DNA databases. Large databases would also increase the frequency with which database use would generate suspects for investigation. However, there might be legitimate concerns that inappropriate handling in court of the statistical and technical issues surrounding database use might lead to more false convictions.

I would wholly accept this argument if the collection and evaluation of non-genetic evidence were rigorous, quantitative, free of selection bias and uncertainties, and derived from investigation of all possible avenues. But this is unrealistic, sometimes even for evidence backed by science. The reality is that much evidence depends on hunches and qualitative argument, and thorough investigation has often been restricted to just one or a few of the possible suspects. The degree of suspicion appropriate for a prime suspect is often highly uncertain, and the discovery of a positive DNA match provides strong vindication of the potentially uncertain arguments that led to the status as prime suspect. When a suspect is instead identified by DNA trawl, the arguments constructed around non-genetic evidence obtained after identification will lack that test of authenticity and must retain all their caveats and uncertainties. Contrast for example, two situations[5]. In (A), S becomes a suspect because an unreliable informant asserts "S did it", S is DNA tested and proves to match the crime scene. In (B), S is suspected because trawl of a database reveals that he is a DNA match and then the informant adds his assertion "S did it". In neither case will the informant's bland assertion carry significant weight, but it is essential that a jury should know which scenario applies to the case, because the probability that the procedure would produce a DNA match incriminating an innocent person is much greater in (B) than (A), by a factor equal to the number of plausible suspects within the database. A worthwhile approach is for a jury to consider in trawl cases whether, if subsequent non-genetic evidence had been established at the outset, it would have justified the defendant being tested as the prime suspect. Only then is the probability of false incrimination by DNA as low as the random match probability p. In case (B) above this consideration would highlight the crucial importance of knowing whether the informant's post-trawl assertion was made with or without knowledge that S was by then a suspect.

My thesis here is that courts need to address the probability that police procedures (including trawls for suspects) could have produced evidence against an innocent person at least as incriminating as that presented. This is a form of data probability and is how the court should ultimately judge the safety of a conviction. It is a judgement that takes account of more than just facts linking the defendant to the case - but so it should. The defendant is in court due to operation of these procedures, which must therefore be under scrutiny just as much as the defendant himself. A judgement that focuses only on the probability that the specific defendant is guilty will necessarily be deficient. It is essentially trying to apportion probability between all possible suspects, only one of whom is fully investigated and presented, while the true culprit and nature of the crime may not even have entered consideration. There is a strong analogy here to the difficulty of judging how likely it is that a particular scientific hypothesis is correct: this is always uncertain because we simply never know if we have had the luck, insight or imagination to consider the correct explanation of the data we observe.

Conclusion

Acknowledgement and characterisation of uncertainties are key elements in the application of knowledge and evidence. A lack of awareness of the nature or extent of uncertainty can lead to confusion and inappropriate decisions in areas of education, science and law. My focus here has been on a distinction that cuts across the major debates that often emerge in statistics (about the definition of probability in frequentist or Bayesian terms) and in law (about the merits of quantitative or qualitative approaches to decision making in court). The distinction I am concerned with is between probabilities assigned to hypotheses and probabilities for the observation of data, conditional on a specific hypothesis. There is not a choice between addressing just one or the other: each has its own logic and its domains of rigorous application in both science and law. Both are central to the process of establishing knowledge, which itself consists of gradations of belief in hypotheses - justified, at least partially, by evidence from data. The link between hypothesis and data probabilities is seldom rigorous in either science or

[5] I am grateful to D.Kaye for suggesting consideration of this scenario.

law. There are grey areas where conclusions may be matters of opinion. Science in general gets round this problem by relying on the indefinite accumulation of evidence that will ultimately swamp all but the most extreme differences of opinion. Criminal law relies more on the moderating effect of a jury decision to generate an outcome that is generally accepted as reasonable. In the legal context however, I argue that the usual focus on probability that the defendant is guilty (a hypothesis probability) is inadequate. Clear benefits arise if a verdict is considered to be ultimately constrained by a data probability: how likely is it that such incriminating evidence could have arisen for an innocent person?

Acknowledgements

I am grateful to D. Kaye and D. Balding and to the editors of this volume for constructive comments.

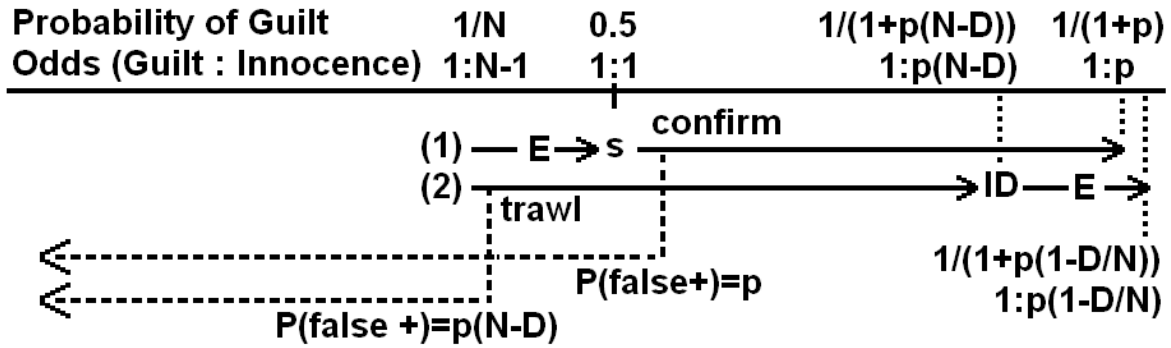

Fig. 1

Comparison of Bayesian treatments of probability of guilt. Arrows show successive changes in probability of guilt and associated odds, for a suspect identified (1) by non-genetic evidence (E) followed by a confirmatory DNA test, or (2) through trawl in a DNA database followed by investigation and discovery of E. The non-genetic evidence on its own leads to a probability of guilt (s) that is taken here to be 50%. N, D and p are explained in the text. Dashed arrows show approximate probabilities of a false DNA match being found, which if subsequently identified as false would lead (given the assumptions) to low probability of guilt and acquittal.

References

Anderson T, Schum D, Twining W (2005) Analysis of Evidence. 2nd edition. Cambridge University Press

Balding DJ (2002) The DNA database search controversy. Biometrics 58, 241-244

Balding DJ (2005) Weight-of-evidence for forensic DNA profiles. John Wiley, Chichester, UK

Cohen LJ (1977) The Probable and the Provable. Clarendon Press, Oxford

Dawid AP (1986). Probability Forecasting. Encyclopedia of Statistical Sciences vol. 7, edited by S. Kotz, N. L. Johnson and C. B. Read. Wiley-Interscience, 210–218.

Dawid AP (2002) Bayes's Theorem and weighing evidence by juries. Proc Brit Acad 113, 71-90

Devlin B (2000) The evidentiary value of a DNA database search. Biometrics 56, 1276

Donnelly P (2005) Appealing Statistics. Significance 2,46-48

Fisher RA (1925) Statistical Methods for Research Workers, Oliver & Boyd

Gardner-Medwin AR (1995) Confidence Assessment in the Teaching of Basic Science. ALT-J. Association for Learning Technology Journal 3, 80-85

Gardner-Medwin AR (2005) What probability should a jury address? Significance 2, 9-12. Available online at http://www.ucl.ac.uk/~ucgbarg/doubt.htm

Gardner-Medwin AR (2006) Confidence-Based Marking - towards deeper learning and better exams In : Innovative Assessment in Higher Education. Ed.: Bryan C and Clegg K. Routledge, Taylor and Francis Group, London

Kaye D (2009) Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls. North Carolina Law Review, 87,425-503

Kempthorne O, Folks L (1971) Probability, Statistics and Data Analysis. Iowa State University Press

Lindley DV (1972) Bayesian Statistics: A Review, Society for Industrial ad Applied Mathematics, Philadelphia

MacKay DJC (2003) Information Theory, Inference and Learning Algorithms. Cambridge University Press 628pp.

National Research Council Committee on DNA Forensic Science (1996) An update: the evaluation of DNA forensic DNA evidence. National Academy Press, Washington DC.

Popper K (1959) The Logic of Scientific Discovery. Hutchinson, London

Roberts P & Zuckerman A (2004) Criminal Evidence. Oxford University Press, Oxford

Royall RM (1997) Statistical Evidence: A likelihood paradigm. Chapman & Hall, London

Royal Statistical Society (2002) Letter from the President to the Lord Chancellor regarding the use of statistical evidence in court cases. http://www.rss.org.uk/statsandlaw

Senn S (2003) Dicing with Death: Chance, Risk and Health. Cambridge University Press

Song YS, Patil A, Murphy EE & Montgomery S (2009) Average probability that a "Cold Hit" in a DNA database search results in an erroneous attribution. J. Forensic. Sci. 54, 22-27

Stockmarr A (1999) Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. Biometrics 55, 671-677

Tillers P, Gottfried J (2006) A collateral attack on the legal maxim that proof beyond a reasonable doubt is unquantifiable. Law, Probability and Risk 5,135-157

von Neumann J, Morgenstern O (1944), Theory of games and economic behavior, Princeton University Press

Tony Gardner-Medwin , Emeritus Professor of Physiology at University College London, has a research background in neuroscience and neural computation. He has a long-standing interest in the improvement of student learning experience and student assessment by getting students to judge the reliability of arguments and knowledge (Certainty-Based Marking: https://tmedwin.net/cbm/selftests). He participated frequently in the UCL Evidence Programme. Home page: https://tmedwin.net/~ucgbarg .