Statistical analysis of STR data

I.W.Evett[1] and J.S.Buckleton[2]

1. Forensic Science Service, Gooch St North, Birmingham, B5 6QQ, UK

2. ESR: Forensic, Mt. Albert, Private Bag 92-021, Auckland, New Zealand.

INTRODUCTION

The debate on the statistics of forensic DNA profiling has been dominated by restriction fragment length polymorphism (RFLP) data. However, conventional population genetic analyses require unequivocal identification of homozygotes and that is not, in general, possible with RFLP profiles. This has added to the air of confusion which has characterised the debate.

Fortunately, with short tandem repeat (STR) data the problem of identifying true homozygotes is minimised and this means that it has been easier to apply conventional methods of statistical analysis. This, in turn, has brought into sharper focus an issue that has troubled the authors for some time and which we discuss. In short, we question the relevance of conventional independence testing to the forensic use of profiling systems.

Classical wisdom would have it that the multiplication inherent in most methods of interpretation is only valid if the assumptions of independence are proved. We maintain that, not only is it not possible to prove the assumptions true, but they are certainly false. The essential issue is not to determine whether independence exists (it does not) but rather to assess the practical consequences of the departure from independence.

INDEPENDENCE TESTING

The basic forensic evidence transfer problem is as follows. A crime has been committed by a man who left a body fluid stain at the scene. The investigator has arrested a man whom he suspects of being the offender. A DNA profile of the crime sample is found to be indistinguishable from that of a sample provided by the suspect. How is this result to be interpreted?

Assuming - as is generally the case - that there is no question of a close relative of the suspect being involved then it is necessary to consider two competing explanations for the evidence:

C: the stain was left by the suspect
$\overline{C}$: the stain was left by an unknown man

Then it is necessary to consider the probability of the evidence given each of these hypotheses and the ratio of these two probabilities - the *likelihood ratio* (LR) - can in this situation be shown to reduce to $1/f$, where f is the frequency of the observed genotype among members of the population to which the 'unknown man' credibly belongs. This must be estimated using data from a sample of people from the relevant population.

The STR profile is a multilocus genotype and the simplest way to estimate its frequency is to multiply together the constituent allele frequency estimates both within and between loci: the recent debate has been dominated by questions about this procedure. The issue, of course, is independence and conventional wisdom calls for tests of independence. Within locus this is a test

for Hardy-Weinberg proportions - often called Hardy-Weinberg equilibrium (HWE). Between locus it is a test for linkage equilibrium (LE) proportions.

Classical independence testing revolves around the concept of a *null hypothesis*. This is the postulation of a state of affairs sufficiently simple for a test statistic to be devised with a known sampling distribution. For the within-locus test, the null hypothesis will be that the genotype frequencies are identical to the Hardy-Weinberg proportions. Put another way, it is that perfect independence exists in the population under consideration. The expectations under this hypothesis are then compared with functions of the sample data by means of a test statistic which itself is compared with a probability distribution. The area underneath the probability distribution corresponding to values of the statistic more extreme than that actually observed is called a *p*-value. Small *p*-values cast doubt on the validity of the null hypothesis. The word *significant* has been associated with the *p*-value 0.05 and the idea of "statistical significance" is something which has been included in training of most scientists.

There are various problems with this approach. Not least of these is that we are testing a hypothesis which is *certainly not true*: the conditions for HWE and LE cannot exist in real human populations. The alternative hypothesis, which is given scant attention, is that the null hypothesis is false. This is hardly an interesting hypothesis for two reasons. First, we know that it is true. Second it is so vague that, although it is true, it does not tell us how to proceed with forensic casework. The null hypothesis is a precise state of independence. The alternative hypothesis is everything else!

So the essence of this is that we are testing the truth of a precise hypothesis which we know to be false against a vague hypothesis which we know to be true. Scientists believe this to be the right way because their statistical education has taught them that it is necessary for scientists to look at their problems as classical statisticians do. It's not necessary at all - indeed, forcing practical scientific problems into an artificial theoretical framework inevitably causes confusion. This is one of the reasons why most scientists hate Statistics.

Let's see what happens when we do play the significance testers' game. We carry out a test and let us assume that we accept that a particular *p*-value - typically 0.05 -is "significant". Then one of two things can happen: either the result is significant or it isn't. If it is significant then we reject the null hypothesis - hardly a surprising result because we knew it wasn't true anyway. If it's not significant then we are told - correctly - that we haven't proved the null hypothesis (but then we didn't want to prove something that we knew was false) and if we'd taken a bigger sample, or used a different test then we might have seen a "significant" result. It is a game that we can't win.

This last argument invokes the concept of "power" which says, basically, that the more powerful the test (or the bigger the database), the more likely it is to detect departures from independence. The detection comes through the *p*-value which, whatever the power of the test, tends to be used as a measure of the extent of the departure from independence: a small *p*-value is taken to imply an appreciable departure from independence.

Directing attention to possible meanings of "appreciable" is a fruitless pursuit. Our concern is with the use of the typing system in forensic casework and our question is "if we use this system in casework then to what extent may the results mislead a court?". The *p*-value is of no use here because in no case is it designed to provide any sort of measure of the practical consequences of using the independence model.

A DNA analysis is undertaken for the purpose assisting a criminal justice system in its objectives of excluding the innocent and incriminating the guilty. The greater the power of the analysis to assist the achievement of these ends, the greater the utility of the technique. But there are two possibilities which have negative utility: excluding the guilty and incriminating the innocent. The latter is rightly regarded with abhorrence and for this reason it has come to dominate the DNA debate.

If we set aside any discussion of laboratory error then an STR comparison has a theoretically zero chance of excluding the true offender. The chance of incriminating an innocent person is that of a chance match between two people. With any new technique, forensic scientists know that it is necessary to estimate PM - the probability of a match between two unrelated people - or (1-PM) (Jones, 1972), called the discriminating power. For the STR quadruplex of VWA, THO1, F13A1 and FES, PM is about 1 in 10,000 in Caucasians. Here is our first measure which contributes to an assessment of forensic utility.

CONSERVATISM

It has become accepted, and the authors must accept some of the blame, that wherever there is some doubt the 'most conservative' answer should be given in the sense that the most conservative answer is that which most favours the defendant. We now maintain that this principle is wrong and that its application leads to serious misconceptions.

There appear to be three main ideas underlying the wish to be conservative. The first, which is laudable, is an attempt to save the falsely accused suspect from conviction. The second is a search for some kind of "comfort factor" for the forensic scientist. The third is that the assumptions underlying the calculation may not be correct in the case at hand. We consider these three in turn.

Falsely accused suspect

The first of these ideas has strong intuitive appeal but appears to have no foundation in logic. It says, effectively, "we know that our technique can lead to false inclusions so we will revise our LR's downward". This is a matter of policy rather than science and adopting a deliberately conservative policy means that in every case in which the suspect is truly the offender we will understate the strength of the DNA evidence. But the policy has only negligible impact on cases in which the suspect is not the offender because in 99.99% of such cases the suspect is excluded and the LR is at its absolute minimum of zero. Safety for innocent suspects lies in very discriminating techniques, both technological and inferential.

It must be recognised that a policy of understating the evidence applies equally to the truly culpable and the truly innocent. What do we really think we have achieved? Moderate understatement of the evidence will provide only a small effect in assisting the falsely accused suspect. Perhaps we intend a massive understatement of the evidence. This renders DNA evidence essentially useless to the courts. Is this useful?

Comfort factor

The principle of conservatism, though we are not aware of its having been formally stated anywhere, seems to take the form "any plausible explanation of the evidence that most favours the defendant is to be preferred". It is a fallacy to believe that this line can make it more comfortable in court. There is a never ending line of "could be's" for creating less and less plausible situations that favour the defendant still more. Whatever the scientist does to be conservative it will always

be possible for defence, perfectly legitimately, to find another scientist who will argue for a more conservative interpretation. The only way to be uniformly conservative is to report an exclusion in every case, a LR of zero is unequivocally the lowest LR that can be reported.

For a scientist to claim in evidence that his LR is conservative in that it is in some sense a lower limit of some "true LR" is to invite a cross-examination along the lines "could the true answer be as small as ...." or "how can you be sure that your LR is smaller than the true value". There is no true LR - every calculation exists within a framework of assumptions.

Validity of the assumptions

It is not possible to make any assessment of evidence without making assumptions and, as far as possible these should be made clear to the court. We suggest the following.

"The LR is my best assessment of the strength of the evidence. In making my calculations I have made assumptions and approximations in ways which I do not believe to be prejudicial to the defendant. My methods have been tested in experiments on actual data and these have included simulations of cases in which the offender and the defendant are different unrelated people. These experiments indicate that there is a small chance of my quoting a LR as big as the one I have given if the defendant is not the offender. From such experiments I estimate that on average this chance is about 1 in ....." The experiments described later shows how such an answer may be given.

TWO PRACTICAL ILLUSTRATIONS

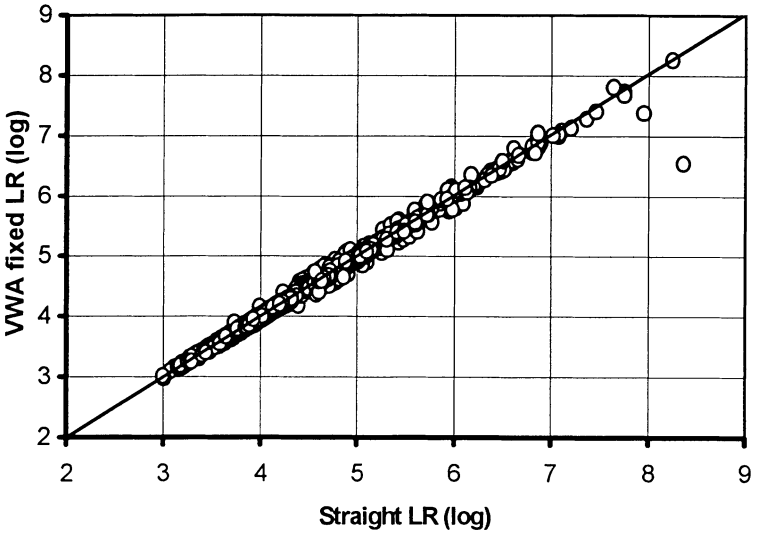
First we consider the recent paper by Drozd et al (1994) which reports a survey of the VWA genotypes in 200 British Caucasians. The authors say "Before such a polymorphic system is used, checks are made to ensure that the chosen system is in Hardy-Weinberg equilibrium.." They duly carried out a goodness of fit test which compared the observed genotype frequencies with those expected from Hardy-Weinberg proportions. They found the result to be: "very significant ($P<0.01$), suggesting that the sampled population may not be in Hardy-Weinberg equilibrium." That's an unduly cautious statement - the population is most certainly not in Hardy-Weinberg equilibrium - it couldn't possibly be, because the required conditions cannot be met in real human populations.

The authors showed that the high value of the goodness of fit test statistic was caused by differences between observed and expected frequencies for the 17/17 and 16/17 genotypes. They saw 22 17/17's against an expected number of 14 and 8 16/17's against an expected 19. Is this cause for concern?

We have taken a file of 1660 British Caucasians typed at four loci which is an amalgamation of data collected by the FSS, MPFSL and Strathclyde laboratories. We have calculated the four locus genotype frequency for each sample in the combined database using the allele frequencies multiplied together for three of the loci (THO1, F13A1 and FES) but using two different methods for VWA: (a) multiplying allele frequencies and (b) using the actual observed genotype frequency. Figure 1 compares the LR's based on the two methods.

It would be difficult to expect better correspondence. The vigilant observer will point to the extreme right point which shows a two order of magnitude difference between the two calculations. This point corresponds to a single observation of genotype (12,20) and has nothing to do with the 16/17 'problem'. The independence model for VWA represents no practical disadvantage.

Figure 1.



Next, here is an example from the analysis of quadruplex data on 1400 Caucasians as described in Evett, Gill, Scranage and Weir (1995). The exact test was used, determining *p*-values by shuffling the databases. The single locus tests for F13A1 and FES gave moderate *p*-values, but the two locus composite test gave a *p*-value of 0.001 - "highly significant" according to conventional wisdom. Does this result mean that we must abandon our independence model?

One way of looking deeper at the effect is to consider the individual allele specific disequilibria Weir (1990). This can be done by comparing the observed frequency of the co-occurrence of each two locus allele combination with that expected from the assumption of between-locus independence. When this is done it is found that the effect can be localised to three specific combinations as shown in the following table.

F13A1	FES	Observed	Expected
3	11	88	78.8
3	12	26.5	38.1
15	11	13.5	19.7

These are the observed and expected numbers of times that individuals carry the given pair of alleles. For two of the combinations the observed number is *smaller* than expected so, using the expected number from multiplying across the two loci is actually in the defendant's favour. For the first combination of the three the observed frequency is 12% larger than that expected: is this cause for concern? We do not believe that such a difference would make any discernible impact on the impression created on a jury.

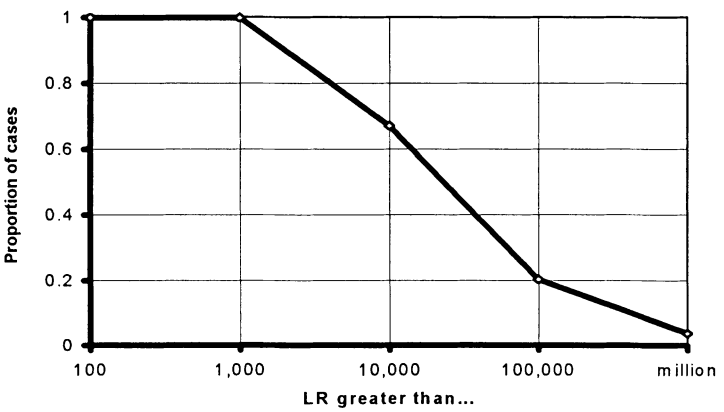
TIPPETT EXPERIMENTS

It is not normally recognised by those who have recently entered the DNA debate that the forensic community has been thinking about issues of discrimination and evidential value for many

years. In particular, an imaginative experiment was carried out by a team headed by Tippett (1968) to study the power of paint comparisons and similar methods were later used by Gaudette and Keeping (1974) for hair examination. These researchers introduced the concept of between-source comparisons as a way of investigating the forensic utility of a technique and their methods form the basis of our preferred means for evaluating DNA techniques as already described in papers such as Evett, Scranage and Pinchin (1992).

Given a file of STR data for a combination of loci, the following experiments can contribute to assessing the forensic utility of the method. For illustration we use the file of 1660 Caucasians that we described earlier. The first step is to estimate the performance of the method in cases in which the suspect is truly the person who left the crime stain. For STR's, where the matching stage is trivial (RFLP's require duplicate analyses for this stage) this is simply done by going through the file and calculating the LR for each individual using the chosen method of calculation. In this instance we have simply used allele frequencies estimated from the file itself and multiplied within and between loci. Our one concession to allowing for sampling effects is to use a default minimum frequency of 0.01. The distribution of LR's from this experiment are summarised as shown in the Figure 2.

Figure 2

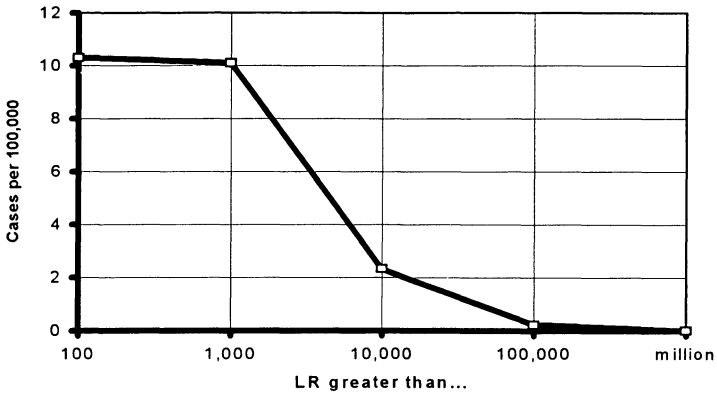


It has an unusual design so it is worth explaining it carefully. The *x* axis is in terms of LR and it is decremental in that the *y* co-ordinate is the proportion of those cases (in which the suspect truly is the person who left the crime sample) in which the LR can be expected to *exceed* the value on the *x* axis. See, for example, that in all cases the LR will be greater than 100 and that in nearly all cases it will exceed 1,000. In slightly over 65% of cases the LR will exceed 10,000 and in 20% of cases it will exceed 100,000.

But the concerns, such as they are, seem never to be about cases in which the suspect truly is the offender. They relate to cases in which the suspect is truly *not* the offender. The between-person experiment is carried out to assess the performance of the method in cases in which the suspect is not the person who left the crime stain. A file of 1660 gives us 1660.1659/2 = 1,376,970 comparisons; put another way, the simulation of about 1.4 million cases in which the offender and suspect are different people. Figure 3 summarises the output from that experiment. Note that the horizontal axis is the same as that in the previous figure. The *y* axis is now, however, in cases per 100,000. So, a fortuitous match will occur in about 1 case in 10,000, which tells us the average discriminating power of the technique. Further, we can see that a fortuitous match and a LR of

10,000 can be expected in about 1 case in 50,000. It is our view that this diagram sets the forensic value of this technique in context.

Figure 3



It can be shown that the number of large LR's in the between-person experiment is robust to dependence, whether artificially induced by mixing of populations or by creating populations with known disequilibria. It would appear that some substantial departure from the model is required before the performance of the independence model starts to produce a large excess of big LR's. So it might be argued by proponents of the significance testing that this analysis has low power to detect dependence effects. This is true: but this is entirely the point - small or moderate departures from independence have negligible impact on the forensic value of the technique.

Other methods - such as the exact test - will detect the departures from perfect independence *long before these departures have any practical consequence*. The Tippett style experiment leads to an estimate of the magnitude of the *effect* of the departure.

It has been the custom for forensic scientists to tell courts about the length of their experience, how many cases they have done, how many comparisons they have made. The between-person comparison experiment is an extension of this tradition. One of the authors is a firearms examiner and one is a former document examiner: each of us can cite the hundreds of comparisons we have made within our respective evidence types. Most of these comparisons were in cases so the true state of affairs was unknown. With the between-person comparisons we can go much further - we can cite millions of comparisons (Lambert et al, 1995), not made personally, of course, but on our behalf using the power of modern computers. In these instances we know the true state of affairs - we set the experiment up this way. We can assess our model really quite rigorously. Yet the DNA evidence we give is no stronger in essence than the qualitative opinions we would offer in our other fields.

DISCUSSION

We maintain that the model of perfect independence performs adequately in the context we have been discussing though we accept that it is for each court to judge its reliability. This judgement can be usefully informed by a description of the results of Tippett experiments.

We further maintain that classical hypothesis testing has only a small part to play in assessing reliability. We do not advocate the abandonment of testing for disequilibria but contend that the results do not address the questions of practical impact, and provide no guidance on how to proceed after testing.

We also maintain that comfort in the witness box should be based, not on deliberately diluting the evidence, but on the strength which comes from depth of understanding of the underlying issues, full experimentation, and comprehensive training. It is the function of the scientist to aim for the best assessment of the evidential strength within the circumstances of the case, supported with a plausible explanation of the reliability of that assessment.

REFERENCES

Brookfield JFY (1995) The effect of relatedness on likelihood ratios and the use of conservative estimates. Genetica: 13-19.

Drozd MA, Archard L, Lincoln PJ, Morling N, Nelleman LJ, Phillips C, Soteriou B, Syndercombe Court D (1994) An investigation of the HUMVWA31A locus in British Caucasians. For Sci Int 69: 161-170.

Evett IW, Gill PD, Scranage JK and Weir BS (1995) Establishing the robustness of STR statistics for forensic applications. Am J Hum Gen (Submitted)

Evett IW, Scranage JK and Pinchin R (1992) An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. Am J Hum Gen 52: 498-505.

Gaudette BD and Keeping ES (1974) An attempt at determining probabilities in human scalp hair comparison. J For Sci 19: 599-606.

Jones DA (1972) Blood samples: probability of discrimination. J For Sci Soc 12: 355-359.

Lambert JA, Scranage JK and Evett IW (1995). Large scale database experiments to assess the significance of matching DNA profiles. Int J Leg Med (In press).

Tippett CF, Emerson VJ, Fereday MJ, Lawton F and Lampert SM (1968) The evidential value of the comparison of paint flakes from sources other than vehicles. J For Sci Soc 8: 61-65.

Weir BS (1990) Genetic Data Analysis, 1st edn. Sinauer, Sunderland, MA.