Accounting for uncertainty in forensic calculations

Amy D. Anderson, Gary W. Beecham, Jr. and Bruce S. Weir
Program in Statistical Genetics
Department of Statistics
North Carolina State University

Outline

1. Sources of Uncertainty
2. Types of methods for dealing with uncertainty
3. Confidence Interval of the likelihood ratio associated with mixed stain DNA evidence

Statistics in a nutshell–the height example

Suppose I wanted to know the average height of all men in some particular group. I could choose a random sample of 100 random men, and use the sample mean as my estimate.

- If I repeated my actions and drew a different random sample of 100 men, I would obtain a different estimate.
- Hence, there is uncertainty in my estimate. This uncertainty can be summarized by reporting my result as a confidence interval or including a margin of error with my estimate.
- Note: As my sample size increases, I get a more and more refined estimate and the margin of error will decrease.

The allele frequency example

- The same thing happens with estimating allele frequencies–a database corresponds to a sample of alleles.
- If different individuals had been included in the database, our estimated allele frequencies would have been different.
- Since there is uncertainty in our allele frequency estimates, there will be uncertainty in quantities that are calculated using these estimates (e.g. match probabilities).

The effect of database size on the margin of error.

Here, I used a true allele frequency of 0.20 to show how database size and level of confidence effect the margin of error for an allele frequency at a single locus.

Database Size	Margin of Error for 95% CI	Margin of Error for 99% CI
50	0.111	0.146
100	0.078	0.103
500	0.035	0.046
1000	0.025	0.033
5000	0.011	0.015

A bigger issue...

- People don't all belong to one homogeneous group.
- Different subpopulations have different allele frequencies.
- Over the next few slides I will give an extreme example to show how the existence of subpopulations can effect our estimation of somebody's uniqueness.

Logical thought in a world with one population

- Imagine a world where people live in villages, but the heights of people are pretty much the same in all the villages.
- One day, a villager sees a ten-foot tall woman walking through town. Since everybody knows that women are rarely over six foot tall, the villager knows he has seen something extraordinary–probably the tallest person in the world.
- Years later, if news comes that a ten foot tall woman committed a crime, the villager gets to have the pleasure of telling his family about how that criminal once passed through their very own village.

Logical thought in a world with many subpopulations

- Imagine a world where people live in villages, and the heights of people are pretty much the same in most villages, but the people also know that, far up in the mountains there are some villages with very strange and different people.
- In this world, when a villager sees the ten foot tall woman, he has two thoughts—maybe this is the tallest person in the world, or maybe, way up in the mountains there is a whole village where the women are ten feet tall.

- Years later, when news comes that a ten foot tall woman committed a crime, the villager gets to have the pleasure of telling his family about how one of those giant women once passed through the village—and maybe it was even the very same criminal.

How this applies to match probabilities...

- A very tall woman is like a rare profile...
- Luckily, we have better information than the villagers in either of the two worlds. We know that there are many subpopulations out there, *and population geneticists have learned a lot about how allele frequencies within a subpopulation might compare to the overall average*.
- This is where the "theta correction" comes in. When we use formulae that include θ, we are taking into account that the defendant and the perpetrator might both belong to some subpopulation where allele frequencies might differ from the allele frequencies we see in the database.

More about θ

- The parameter θ is essentially a measure of how different we think the subpopulation might be from the population as a whole.

- Values of θ within the European population are usually estimated to be less than 0.01 at each marker, but, to be a bit conservative, we usually pretend that θ is known and has a value of 0.03.

- Higher values of θ indicate that the subpopulation can be more different from the overall population.

Proposed methods for dealing with uncertainty

- Confidence Intervals
  - Pro: Familiar from opinion polls.
  - Con: The probability statement is about the database, not about the allele frequency (or match probability).
- Bayesian probability intervals / Bayesian credible intervals / Bayesian highest posterior density regions
  - Pro: The probability statement is about the allele frequency or match probability. "With probability 0.95, the match probability is between 0.012 and 0.023."
  - Con: Requires the use of a "prior" distribution, which may or may not have a theoretical justification and may or may not have a great influence on the result.

- Just plugging in conservative estimates at every step (e.g. NRC's 1992 "ceiling principle").
  - Pro: It's easy and should give a conservative result.
  - Con: Impossible to justify this approach when you are capable of doing the analysis in another manner that gets your best estimate and a theoretically justified margin of error on your estimate. (So don't use this approach if you can avoid it).
- The "factor of ten" bounds (from NRC II).
  - Pro: Easy.
  - Con: As an ad-hoc approach with no theoretical justification, it gave decent results when just five or six markers were used. With more markers, a factor of ten isn't nearly big enough.

Where we are so far...

So far, we have talked about the two main sources of variability in our calculations.

- Sampling variability in the databases

  Solution: Attach confidence intervals or some measure of uncertainty to your estimates.

- Populations are composed of subpopulations which are difficult to identify or define.

  Solution: Always use formulae that include a conservative correction ($\theta = 0.03$) for this problem or report your results using a variety of values of θ.

An Example...

Suppose we have a mixed stain and we have the following two hypothesis:

H_p : The suspect and an unknown person contributed the evidence.
H_d : Two unknown people contributed the evidence.

The general approach is to calculate a likelihood ratio that compares the likelihood of the data (the DNA profile of the stain) under the two hypotheses.

Example, continued...

The likelihood ratio is:

$$\text{LR} = \frac{\Pr(\text{Evidence}|H_p)}{\Pr(\text{Evidence}|H_d)}$$

- This LR consists of two probabilities, each of which is the product of single-locus probabilities.
- Statistical theory works well for sums, not products, so we convert everything to the log scale. So now we work with the log-likelihood ratio (but we convert everything back to the original scale when we report it).

Example, continued...

Then our 95% confidence interval is:

$$\text{CI} = ln(\widehat{LR}) \pm 1.96\sqrt{\text{Var}[ln(\widehat{LR})]}$$

- There are two mysterious quantities:
  - $\widehat{LR}$ is our best estimate of the likelihood ratio. It comes from using the estimated allele frequencies from the database and the known probability equations that account for subpopulations (see Curran et al. 1999).
  - Var$[ln(\widehat{LR})]$ is a tricky thing. It used to be that the known formulae for this quantity all assumed that $\theta = 0$.

Luckily, just recently, Gary Beecham, a PhD student at North Carolina State University, worked out appropriate equations for Var[$ln(\widehat{LR})$].

- The DNAMIX-3 computer program has been updated to include these new formulae.

- The program is available at:

  `http://bioinformatics.ncsu.edu/brcwebsite/software_BRC.php`

- Calculations for both single-contributor and mixed stains are possible with this program.

Example, continued...

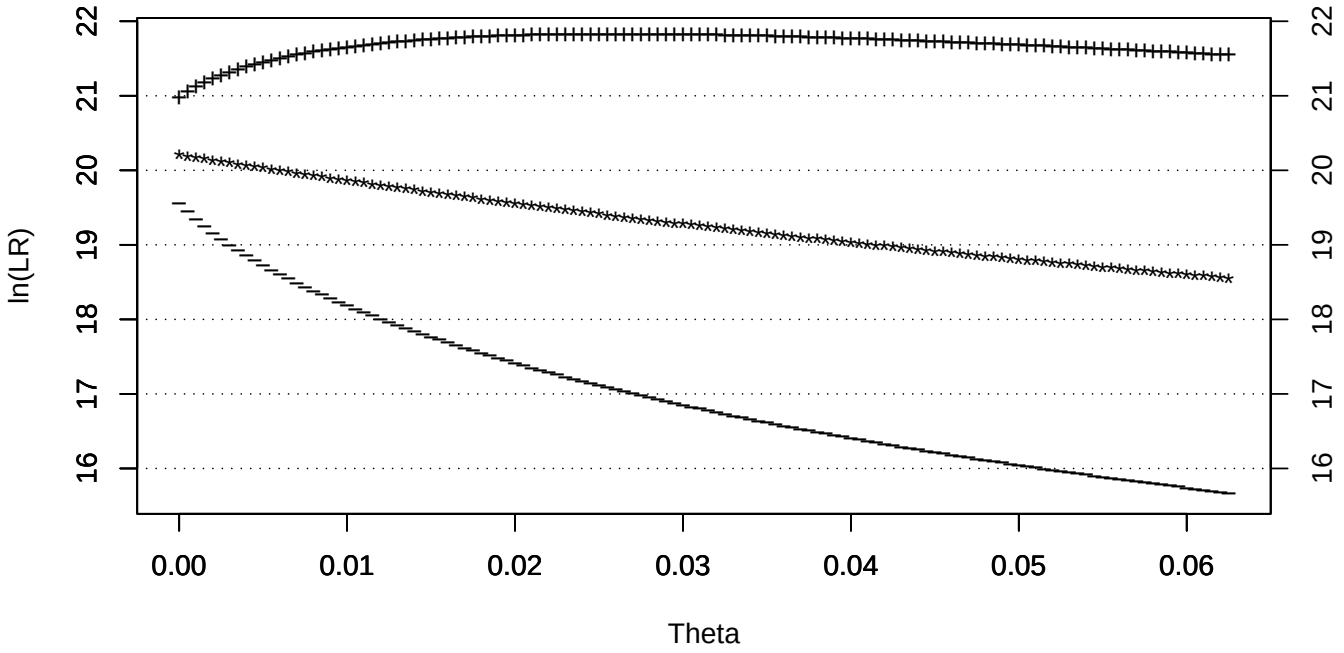
To look at the effects of θ on likelihood ratios and confidence intervals, we looked at the Caucasian database for the CODIS loci published by Budowle et al. (1999).

In the first graph, we consider the two-contributor stain case where profile contains the three most common alleles at all thirteen loci.

The second plot shows a similar situation except that the profile contains the three least common alleles at each locus.

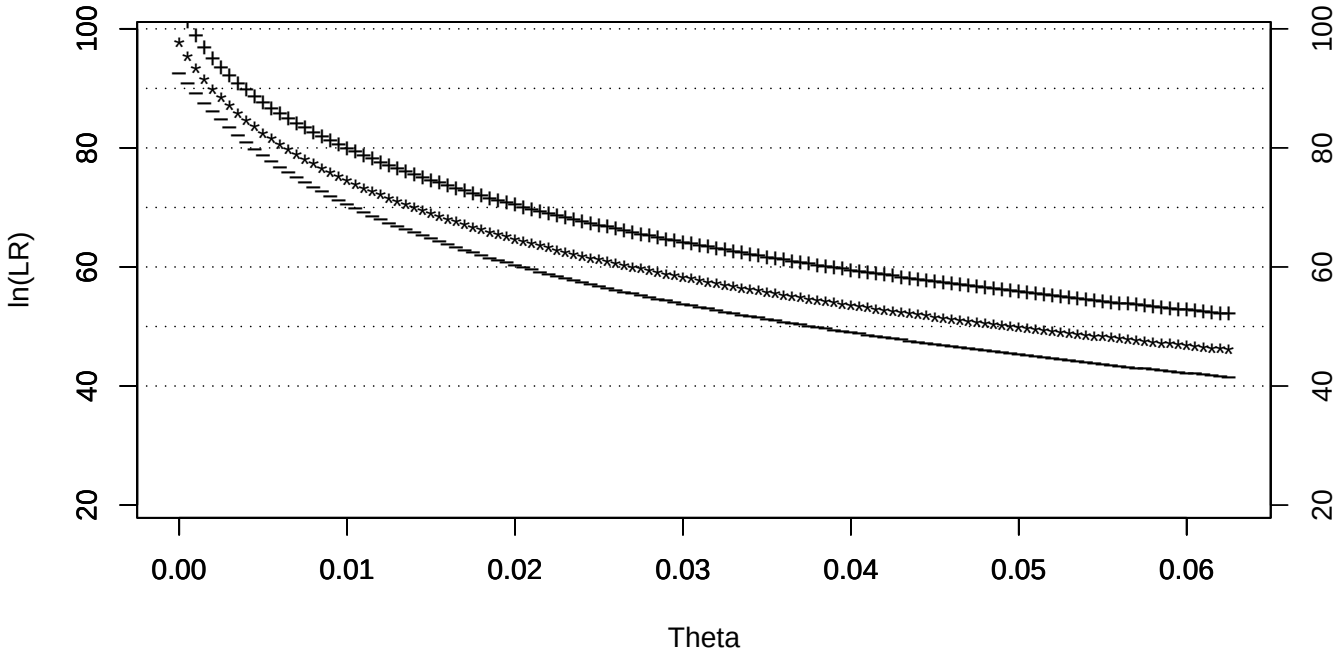
Graph of ln(LR) for a two-contributor mixture with common alleles

Figure 1: Ln(LR) and bounds for common evidence stain



Graph of ln(LR) for a two-contributor mixture with rare alleles

Figure 1: Ln(LR) and bounds for rare evidence stain



Conclusions

- It is important to use statistical techniques that recognize that allele frequencies are not known without error.
- It is critical that you take population structure into account.
- Methodology exists that accounts for both of these sources of uncertainty.

References

Balding DJ and RA Nichols (1997). Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 78: 583–589

Beecham GW and BS Weir (2004). Confidence interval of the likelihood ratio associated with mixed stain DNA evidence (submitted)

Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA and Keys KM (1999). Population data on the thirteen CODIS core short tandem repeat loci in Afican Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. Journal of Forensic Sciences 44: 1277–1286

Curran JM, Buckleton JS, Triggs CM and BS Weir (2002). Assessing uncertainty in DNA evidence caused by sampling effects. Science and Justice 42:29–37

Curran JM, Triggs CM, Buckleton J, and BS Weir (1999). Interpreting DNA Mixtures in Structured Populations. Journal of Forensic Sciences 44(5) 987–995

National Research Council (1992). DNA technology in forensic science. Washington, D.C.: National Academy Press.

National Research Council (1996). NRC II: The evaluation of forensic DNA evidence. Washington, D.C.: National Academy Press.