Genotype likelihood ratio distributions and random match probability: Generalization, calculation and application

Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

July 10, 2017

Contact Information:

Dr. Mark W. Perlin
Chief Scientific Officer
Cybergenetics
160 North Craig Street, Suite 210
Pittsburgh, PA 15213 USA
(412) 683-3004
(412) 683-3005 FAX
perlin@cybgen.com

Abstract

Computers have revolutionized DNA evidence interpretation, replacing guesswork by sound statistical inference. Probabilistic reasoning resolves complex DNA mixtures, extracting contributor genotypes whose uncertainty is expressed through probability. Information theory can tell us much about these "probabilistic" genotypes, even before making a comparison to calculate a likelihood ratio (LR). A genotype's distribution of possible LR outcomes, under prior or posterior probability assumptions, shows the power and breadth of its match possibilities.

Genotype LR distributions can be rapidly computed by convolving independent locus distributions. The tail probability of the non-contributor distribution gives the chance that the evidence against a random person is as strong as it is against the suspect. This LR error is a generalized random match probability (RMP) for uncertain genotypes. A sexual assault case example applies these LR and RMP concepts to DNA mixture evidence and database search.

While the LR summarizes evidence, the RMP estimates error. Both statistical measures assist a trier of fact in understanding DNA evidence.

Table of Contents

Abstract 2

Introduction 5

Mathematical Background 9

Genotype Probability 9

Likelihood Ratio 10

Random Match Probability 11

Definite Evidence Genotype 12

Related Research 13

Non-contributor Distribution 14

Logarithmic Factor 14

Single Locus 15

Multiple Loci 17

Exclusionary Power 18

Contributor Distribution 20

Posterior Weighting 20

Inclusionary Power 21

Measuring Error 21

Average Factor Ratio 21

Random Match Probability 22

Non-contributor Distribution – Right Tail Probability 23

Contributor Distribution – Left Tail Probability 24

Case Example 24

Sexual Assault Information 25

Random Match Probability 26

DNA Database Identification 27

Verifying Distribution Accuracy 28

Conclusion 29

References 31

Appendix A: From Genotypes to Distributions 34

Discrete Probability 34

Measure and Integral 35

Pushforward Measure 36

Genotype Tail Probability 37

Appendix B: Population Genetics Adjustments 39

Co-ancestry Correction 39

Population Substitution 40

Tables 41

Legends 43

Figures 45

Introduction

In forensic deoxyribonucleic acid (DNA) analysis, biological evidence is collected from a crime scene. Most samples are mixtures of two or more people who contributed their DNA to the evidence (Butler 2000). The data generated from a mixture reflect the additive combination of these contributors, along with random effects from the laboratory procedures (Perlin and Szabady 2001).

A genotype is the genetic type of an individual's DNA, a pair of alleles at one or more genetic locations. Computers can mathematically separate mixture data into the genotypes of each contributor by explaining the quantitative data (Perlin and Sinelnikov 2009). Multiple explanations for an uncertain genotype have associated probabilities (Perlin, Legler et al. 2011). Genotype uncertainty also arises when analyzing small amounts of DNA, or in reconstructing someone's genotype from a kinship analysis of their relatives.

Before analyzing evidence data, there is a *prior* probability distribution for a genotype. This prior probability describes the relative prevalence of different genotype values in a human population. The prior probability is ascertained by reviewing hundreds of genotypes at every tested chromosome location (i.e., "locus"). The frequency counts of genotype values mirror their prevalence.

After analyzing informative evidence, the *posterior* genotype probability distribution accounts for the DNA data. This updating of prior genotype to posterior genotype by means of a data-explaining likelihood function is accomplished through Bayesian data analysis (Gelman, Carlin et al. 1995). More informative DNA data better concentrates genotype probability onto fewer possible values.

Two genotypes are compared, relative to a prior (population) genotype, to produce a match statistic that quantifies their association. A person's reference genotype has all probability placed on one genotype value. When an uncertain posterior evidence genotype is compared with a reference genotype, the match statistic sum reduces to a single term at the reference value. This term is the ratio of posterior to prior genotype probability, evaluated at the reference value.

Alan Turing called this posterior-to-prior ratio the (Bayes) "factor," and calculated it as a ratio of likelihoods (Good 1985). The factor numerically expresses how strongly the evidence supports a hypothesis. In DNA identification, the hypothesis is that a person contributed their DNA to the biological evidence. This hypothesis is mathematically assessed through the person's genotype and the evidence data.

Factors reside on a multiplicative scale. Taking logarithms converts factors to an additive information scale. When data are uninformative, the posterior probability does not move from the prior, so the ratio is one and the logarithm registers zero information. When posterior exceeds prior, the factor logarithm indicates positive identification information for a suspect. Larger numbers provide greater support. A posterior probability less than the prior gives a negative factor logarithm, reducing support for the identification hypothesis.

Turing used base ten logarithms, measuring information in "bans" – named for the British town Banbury that printed punch cards for his World War II cryptanalysis laboratory. Turing's statistician Jack Good called the logarithm (log) factor the "weight of evidence" (Good 1950).

DNA analysis tests multiple genome locations in a single reaction tube (Collins, Hennessy et al. 2004). The number of simultaneous tests has increased over two decades from four to twenty two. These locations are genetically independent of one another. This independence permits multiplication of locus factors, often producing astronomical DNA match statistics. On an additive scale, scientists combine information from aggregate DNA tests by adding together their independent log factors.

Every reference genotype has log factor information, evaluated as the logarithm of its posterior-to-prior probability ratio. In principle, Bayesian genotype probability is never zero and so the log factor is always well defined. (Computational practice may require some domain adjustment.) This genotype information is known before any match comparison is made.

The product set of aggregated multi-locus genotypes is large (around 10^{2L}, where L is the number of tested loci). Natural questions arise about log factor values distributed over the genotype set. What is the chance of a genotype having a log factor close to a particular information value? Or exceeding a specified value? Such genotype probabilities correspond to frequencies in a human population. Collapsing genotype probability and log factor values onto a one-dimensional distribution helps answer these questions (Appendix A).

The joint log factor distribution of aggregate DNA tests provides considerable information about an uncertain genotype. The *contributor* distribution weights the log factors by posterior genotype probability. The expected value of this distribution is the Kullback-Leibler (KL) statistic (Kullback and Leibler 1951) – a positive number that predicts the match statistic to the true contributor, and measures genotype divergence between the evidence and a random population. The spread of a bell-shaped contributor distribution, whether broad or narrow, indicates the possible range of match statistics when comparing the genotype with a true contributor.

The *non-contributor* distribution weights log factor values by prior genotype probability. This distribution shows the range of exclusionary match statistics. These (mainly negative) values are for a random person who has not contributed their DNA to the biological evidence.

Genotypes having distributions close to zero have little exclusionary power, because the negative log factors are small. More informative genotypes that can better exclude innocent people have more negative distributions, shifted away from zero information.

We can rapidly calculate the joint log factor distribution for contributors or non-contributors. The joint log factor is the sum of independent locus log factors. Each locus log factor distribution can be quickly assembled from prior and posterior genotype probabilities. It is well known that convolving the distributions of independent random variables yields the joint distribution of the sum of those variables (Feller 1968). Therefore, the joint distribution of the log factor sum can be rapidly computed by convolving the separate locus log factor distributions.

Match statistic error rates can be calculated for a particular genotype. This is done by numerically summing regions under the genotype's log factor distribution curve. The probability of genotypes having a factor value in an interval (e.g., around a reported match statistic) is the area under the curve within that interval. The probability of a random person having a match statistic at least as large as a suspect's value is the right tail area under a non-contributor distribution. By reducing aggregate genotype log factors to a simple distribution curve, probabilities of genotype sets correspond to areas of match statistics.

We begin with some "Mathematical Background," defining likelihood ratio (LR) and generalized random match probability (RMP) for uncertain genotypes. We show how to rapidly construct a "Non-contributor Distribution" for an uncertain genotype's log factor. We summarize similar results for the "Contributor Distribution." We discuss "Measuring Error" for genotype match using a log factor distribution. LR error size helps contextualize DNA match statistics for contributors and non-contributors. We present a sexual assault "Case Example" with DNA mixture evidence. The RMP error analysis illustrates how log factor distributions apply to forensic mixtures and investigative databases. The case also provides data for verifying distribution accuracy.

Mathematical Background

We use probability to represent genotype uncertainty, both before and after examining DNA data. We are interested in match information and its error. This section describes related mathematical ideas and research.

Genotype Probability

We often want to compare a questioned item *Q* with a known exemplar *K*, and measure a degree of association between them. Suppose each known exemplar corresponds to a certain identifying *type*. Let *X* be the set of all possible types. With DNA testing, *X* is the set of all genotypes. Based on data, we can determine the type of a questioned item up to probability.

In a human population, each genotype appears with some frequency. Let $p(x)$ be the probability $\Pr\{X=x\}$ of genotype *x* appearing in the population. This prior probability $p(x)$ is the chance that a questioned item has genotype *x*, before examining data. A *non-contributor* is someone who did *not* contribute DNA to biological evidence. Non-contributor genotypes follow the prior probability distribution $p(x)$.

Informative data changes our belief in an item's type (O'Hagan and Forster 2004). After examining feature data measured from a questioned item, $q(x)$ is the posterior probability

$\Pr\left\{X = x \middle| data\right\}$ that the item has genotype x. A *contributor* is someone who contributed DNA to biological evidence. Contributor genotypes follow the posterior probability distribution $q(x)$.

Likelihood Ratio

The Bayes factor, or just *factor* $f(x)$, is the posterior to prior genotype probability ratio $q(x)/p(x)$. For any known genotype x_K, the factor $\alpha = f(x_K)$ expresses how much more a questioned item *Q* matches a known exemplar *K* than coincidence. The numerical association $f(x)$ is a *likelihood ratio* (LR) (Good 1950), which measures the probative force of evidence and factors out prior prejudice. Here is a routine proof showing that the factor is a LR.

For a suspect genotype *s*, the factor $f(s) = q(s)/p(s)$ is the posterior-to-prior probability ratio $\Pr\left\{X = s \middle| data\right\} / \Pr\left\{X = s\right\}$. By Bayes theorem, this ratio equals likelihood over total probability $\Pr\left\{data \middle| X = s\right\} / \Pr\left\{data\right\}$. Hypothesis H is that the suspect contributed his genotype *s* to the data, while the alternative $\overline{H}$ is that it was some unknown person. Therefore, the preceding expression is the $LR = \Pr\left\{data \middle| H\right\} / \Pr\left\{data \middle| \overline{H}\right\}$, a ratio of likelihoods.

The factor calculation can adjust prior population $p(x)$ and degree of relatedness (Appendix B). Match information is not tied to a particular choice of genotype prior.

The error set E_α is the subset of genotypes $\left\{x \in X \middle| f(x) \geq \alpha\right\}$ for which the factor $f(x)$ equals or exceeds factor α. When $\alpha = f(x_K)$ corresponds to a known exemplar K, E_α contains the genotypes x whose numerical LR association $f(x)$ with questioned DNA item *Q* is at least as great as with x_K.

Random Match Probability

We are often interested in the size of error set E_α, relative to the random population $p(x)$ distribution. We define *random match probability* (RMP) as $\Pr\{x \in E_\alpha\}$, the sum $\sum_{x \in E_\alpha} p(x)$ of prior probabilities $p(x)$, taken over all genotype values in error set E_α. A small RMP indicates a small chance of a false positive that inaccurately associates an exemplar genotype with a questioned evidence item.

For DNA mixtures, a generalized (or “restricted”) RMP accounts for quantitative data and parameters (Scientific Working Group on DNA Analysis Methods (SWGDAM) 2010). Our generalized RMP is defined for an uncertain genotype of one person, say, a contributor to a DNA mixture. All genotype values having an LR match statistic at least as large as a suspect’s LR level α are included in the RMP index set E_α. Each included genotype’s prior probability is added to the RMP sum.

Different RMP generalizations may employ a different set of allele pairs at a locus. Our RMP index set is optimal in several ways. Probabilistic computer inference *comprehensively* considers every genetically possible allele pair when forming the genotype index set. These allele pairs are for just *one contributor*, not mixture combinations. The index set contains the *matching allele pair*, since the suspect has minimal LR. The evidence against *other allele pairs* in the index set is as strong as it is against the suspect. The RMP sum is *minimal* for an index set containing the suspect’s genotype and all other genotypes matching the evidence as strongly.

Modern calculus integrates a function based on range values (Lebesgue 1902) instead of subdividing the domain (Riemann 1868); this improvement overcomes many technical issues. The distribution of a random variable has this function value perspective. Defining the RMP domain set E_α as the genotype pre-image of factor function *f* match statistic values shares this modern view (Wheeden and Zygmund 1977).

Definite Evidence Genotype

A *definite evidence genotype* concentrates all posterior probability onto one value. (This may occur, for example, with abundant DNA left by a single biological source.) For a matching suspect genotype *s*, the posterior probability $q(s)$ is one. The LR factor becomes $1/p(s)$, the reciprocal of the suspect genotype's prior probability.

For a definite genotype, classical RMP equals $p(s)$ (National Research Council 1996). This classical value agrees with our RMP sum $\sum_{x \in E_\alpha} p(x),\ \alpha = f(s)$, which collapses to the single term $p(s)$ (all other genotypes x have zero posterior probability $q(x)$, hence a factor $q(x)/p(x)$ of zero). Therefore $f(x) < f(s)$ for $x \neq s$, and the singleton set E_α contains only the suspect's genotype *s*.

Our generalized RMP cannot exceed $1/LR$. Markov's Inequality (Feller 1968) tells us that RMP $\sum_{x \in E_\alpha} p(x)$ is bounded above by $\frac{1}{\alpha}\sum_{x \in E_\alpha} f(x)p(x)$. But $f \cdot p = \frac{q}{p} \cdot p = q$, so the bound becomes $\frac{1}{\alpha}\sum_{x\in E_\alpha} q(x)$. The partial probability sum cannot exceed one, so we have $RMP \le 1/LR$, since $\alpha = LR$. Equality holds for a definite evidence genotype, i.e., $RMP = 1/LR$.

Related Research

Statisticians describe LR tail error (our RMP) as a probability of observing misleading statistical evidence (Royall 2000). Evidence and uncertainty "have different mathematical forms." Whereas an LR quantifies the strength of observed evidence, uncertainty in the LR is represented by probability (Sjerps, Alberink et al. 2016).

Forensic scientists have estimated LR error (our RMP) computationally (Gill, Curran et al. 2008, Slooten and Egeland 2015). Some approximated the LR distribution (Nothnagel, Schmidtke et al. 2010, Corradi and Ricciardi 2013). Monte Carlo simulation can count how frequently randomly generated genotypes exceed a reported match level (Slooten and Egeland 2014). Branch and bound algorithms help prune the search when genotype error set E_α is small (Dørum, Bleka et al. 2014), while divide and conquer methods can extend the search to larger sets (Kruijver 2015).

When genotyping systems consider all possible allele values independently of the data (Perlin, Legler et al. 2011) the search space may increase exponentially beyond the range of such combinatorial methods. Some scientists avoid exact error determination altogether, either by using a generic 1/LR upper bound (Taylor, Buckleton et al. 2015), or by electing to not report LR error (Kruijver, Meester et al. 2015, Taroni, Bozza et al. 2016).

The RMP can be viewed as a *p-value*, the probability that a non-contributor would attain an LR at least as large as the one observed for a suspect (Dørum, Bleka et al. 2014). Statistical hypothesis testing parallels decision-making in an adversarial justice system (Kenney 1988). There is an initial presumption of innocence, corresponding to the null hypothesis that a defendant is a non-contributor. The prosecutor's task is to prove this assumption false.

An LR summarizes the probative weight of identification evidence. The RMP measures the chance of false positive LR error, that an innocent non-contributor was incorrectly identified as a contributor. Jurors deliver a guilty verdict when they reject the null hypothesis (Saks and Neufeld 2011), finding a sufficiently small error "beyond reasonable doubt." The RMP may assist jurors in assessing such error.

Non-contributor Distribution

Genotypes have a match statistic probability distribution for people who did not contribute their DNA to evidence. This section constructs the logarithmic factor distribution at a single locus, and shows how convolution of independent loci efficiently calculates the multi-locus log factor distribution. We connect exclusionary power to standard concepts from information theory.

Logarithmic Factor

The logarithm of the Bayes factor is a standard additive measure of information (MacKay 2003). Additivity aids in understanding, visualizing, computing, combining, characterizing and communicating the match statistic. We examine the logarithmic distribution of match values for non-contributor genotypes that follow the prior probability distribution.

For each genotype $x \in X$, the match statistic is the Bayes factor $f(x) = q(x)/p(x)$. The logarithm of this function is the weight of evidence $\log\left[q(x)/p(x)\right]$, measured in ban units (Good 1950). We would like to see how these logarithmic values are distributed according to prior distribution $p(x)$ for non-contributors – random people in the population who have not contributed their DNA to the biological evidence.

We can map these $\log f(x)$ values from a multi-dimensional set of genotypes to a simpler one-dimensional real line. Building the probability mass function (pmf) amounts to depositing ordered pairs $\left(\log\left[q(x)/p(x)\right], p(x)\right)$ for every genotype $x \in X$ as points on a two-dimensional graph.

Single Locus

At a single genetic locus, genotype x is a pair of inherited alleles. Since the log factor $\log f(x)$ is the logarithm of a ratio $q(x)/p(x)$, we restrict attention to those genotypes *x* having prior denominator $p(x) > 0$ and posterior numerator $q(x) > 0$, giving a well-defined value. (With nonzero prior probability and likelihood, the posterior probability is also nonzero, and the log factor is defined everywhere.) Each well-defined genotype $x \in X$ deposits a *y*-axis ordinate amount $p(x)$ to the non-contributor distribution at *x*-axis abscissa location $\log f(x)$.

Adding together all the ordinate $p(x)$ probability amounts at abscissa location $y = \log f(x)$ gives the total probability mass at one point

$$u(y) = \sum_{\{x \in X \mid y = \log f(x)\}} p(x)$$

More compactly, we write $\log f^{-1}(y)$ for the set $(\log f)^{-1}(y)$ of genotypes $\{x \in X \mid y = \log f(x)\}$ having log factor value y. Then the non-contributor probability mass function is

$$u(y) = \sum_{x \in \log f^{-1}(y)} p(x)$$

The accumulation of probability mass for the $\log f$ distribution is shown in the Table 1 example. Each genotype possibility (x_1, x_2, x_3, x_4) is listed in the first column. The prior $p(x)$ and posterior $q(x)$ probabilities, before and after having seen data, respectively, are given in the next two columns. The Bayes factor $f(x)$ column contains the posterior-to-prior ratio $q(x)/p(x)$ of the preceding two columns. The last column is the logarithmic factor $\log f(x)$. The log factor is negative for exclusionary results where $f(x) < 1$, positive for inclusionary results with $f(x) > 1$, and zero when the factor $f(x)$ for genotype x is inconclusive.

The Figure 1 histogram shows $(\log f(x), p(x))$ table row pairs binned at a deciban (i.e., $1/10$ of a ban) resolution. We begin the pair deposition process with genotype row 1 (Table 1). The logarithm of $1/2$ is around –0.3, defining a $\log f(x)$ bin "–0.3". After depositing genotype x_1, this bin initially contains the prior non-contributor probability $p(x_1) = 0.2$.

We continue with genotype row 2 (Table 1). Genotype x_2 has prior probability $p(x_2) = 0.3$, and a log factor $\log f(x_2)$ of around –0.3. Adding the second genotype's prior probability of 0.3 to bin "–0.3" gives a total log factor bin mass of 0.5 (Figure 1).

Genotype x_3 has a factor of 1, hence a zero log factor, placing a $p(x_3) = 0.25$ probability mass in bin "0". For genotype x_4, factor $q(x_4)/p(x_4) = 2$ has a log 2 factor of +0.3. We put its probability mass $p(x_4) = 0.25$ in bin "+0.3", completing the picture (Figure 1). The cumulative distribution function (CDF) shown in Figure 2 is a step function that monotonically increases from 0 to 1, incrementally adding a probability mass $p(x_i)$ at each abscissa point $\log f(x_i)$.

Multiple Loci

An experiment can test more than one feature. In DNA identification, multiple genetic loci are tested in a single reaction tube, generating data for a dozen or so loci simultaneously (Collins, Hennessy et al. 2004). Each tested locus l has its own locus genotype set X_l and prior probability function p_l. After testing, the single locus data can be analyzed to calculate the posterior probability q_l, factor f_l, and logarithmic factor $\log f_l$ functions. The previous section showed how to combine prior p_l and log factor $\log f_l$ to form a non-contributor locus pmf.

DNA testing uses short tandem repeat (STR) loci, where genotypes are pairs of alleles differentiated by sequence length (Weber and May 1989). The polymorphic loci used in forensic identification have many different alleles that help distinguish between people (Edwards, Civitello et al. 1991). The loci are chosen to be genetically independent of one another, either residing on different chromosomes or far apart on the same chromosome. This biological independence confers statistical *independence*, where events at one locus convey no information about events at another locus (Feller 1968). When testing multiple STR loci, independent results are multiplied together using the product rule.

The joint factor f over all L independent locus tests is the product $\prod_{l=1}^{L} f_l$ of the locus factors f_l. The logarithm of a product is the sum of the logarithms. Therefore,

$$\log f = \log\left(\prod_{l=1}^{L} f_l\right) = \sum_{l=1}^{L} \log f_l$$

Thus the joint match statistic $\log f$ is the sum of the logarithmic locus factors $\sum_{l=1}^{L} \log f_l$.

The joint probability density u of a sum of independent random quantities having pmfs u_1, u_2, …, u_L is the convolution of their pmfs (Feller 1968). That is,

$$u = u_1 * u_2 * \ldots * u_L$$

For discrete distributions, the convolution $u_1 * u_2$ is determined at value z as the sum

$$(u_1 * u_2)(z) = \sum_{y \in Y} u_1(y) \cdot u_2(z - y)$$

Convolving continuous distributions has a corresponding integral formulation.

The joint non-contributor distribution for joint factor f is readily computed by convolving the additive $\log f_l$ factor distribution functions of each locus l. Convolution is a fast built-in operation in many computer-programming languages, such as MATLAB (Natick, MA). Calculations can combine probability mass, cumulative distribution or integral transform functions to rapidly compute their convolution.

Exclusionary Power

The $\log f$ non-contributor distribution is an inherent property of an inferred "probabilistic" genotype. We can ascertain this distribution before making a match comparison to an exemplar genotype. Once posterior probability has been determined from the data, the $\log f$ distribution can be calculated immediately. The non-contributor distribution describes the power of the genotype to statistically exclude non-contributors.

With informative data, posterior $q(x)$ is different from prior $p(x)$, i.e., $q \neq p$. The average match statistic must then be exclusionary, as shown here. The average non-contributor $\log f$ is the expected value

$$E_p[\log f] = \sum_{x \in X} p(x) \log \frac{q(x)}{p(x)}$$

over multi-locus genotype tuples $(x_1, x_2, ..., x_L)$ in *X*. Since the logarithm of a reciprocal is the negative of the logarithm, this expected value equals

$$= -\sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

The sum is the *relative entropy* of p and q, which equals the expected value under prior probability *p* of the logarithmic ratio of probability functions *p* and *q*

$$= -E_p\left[\log \frac{p}{q}\right]$$

This expression is the negative value of the Kullback–Leibler *divergence* KL_{pq} between p and q (Kullback and Leibler 1951), or

$$= -\mathrm{D}[p \,||\, q]$$

Since $\mathrm{D}[p \,||\, q] > 0$ when $p \neq q$ (applying Jensen's inequality to the concave logarithm function), we have that $E_p[\log f] < 0$. So the expected non-contributor match value is exclusionary. A larger KL_{pq} indicates greater genotype exclusionary power.

Contributor Distribution

Genotype match statistics also have a probability distribution for the person who contributed their DNA to evidence. We apply ideas from the previous section to construct contributor distributions and discuss their inclusionary power.

Posterior Weighting

We can similarly examine the logarithmic distribution of match values for contributor genotypes, now weighted by posterior probability $q(x)$. These logarithmic values are associated with posterior genotypes of people who are more likely to have contributed their DNA to the biological evidence.

We build this distribution by layering log factor and posterior probability, the pairs $\left(\log\left[q(x)/p(x)\right], q(x)\right)$, for all genotypes $x \in X$. For each abscissa location $y = \log f(x)$, the ordinate contributor probability mass is

$$v(y) = \sum_{x \in \log f^{-1}(y)} q(x)$$

The joint contributor distribution of the additive log factor is then obtained by convolving these independent locus log factor distributions.

Inclusionary Power

The average contributor $\log f$ match statistic is derived from a genotype as the expected value relative to posterior probability q as

$$E_q[\log f] = \sum_{x \in X} q(x) \log \frac{q(x)}{p(x)}$$

This relative entropy is the KL divergence between probability distributions q and p

$$= D[q \,||\, p]$$

The positive KL number is the expected inclusionary LR information for posterior genotype pmf q, relative to population prior p. Large KL values correspond to more informative genotypes that have greater inclusionary power.

Measuring Error

There are close relationships between information measures and sums that pertain to the extreme set E_α. We explore those connections here to calculate RMP. We apply these ideas to assess error in inclusionary and exclusionary statistical match.

Average Factor Ratio

On a genotype set *E*, the average value of function f with respect to probability p is

$$\underset{E}{\operatorname{avg}}_p f = \frac{\underset{E}{\operatorname{sum}} f \cdot p}{\underset{E}{\operatorname{sum}} p}$$

When f is a Bayes factor $\frac{q}{p}$, $f \cdot p$ is $\frac{q}{p} \cdot p$. Cancelling the prior function p reduces $f \cdot p$ to the posterior distribution q, and so the average factor is a ratio of probability sums

$$\underset{E}{\operatorname{avg}}_p f = \frac{\underset{E}{\operatorname{sum}} q}{\underset{E}{\operatorname{sum}} p}$$

This function average over a set helps describe error values in an easy way.

Random Match Probability

Exchanging the left hand side with the right hand side denominator, we have

$$\underset{E}{\operatorname{sum}} p = \frac{\underset{E}{\operatorname{sum}} q}{\underset{E}{\operatorname{avg}}_p f}$$

For LR error analysis, we can examine the extreme subset $E_\alpha = \left\{ x \in X \middle| f(x) \geq \alpha \right\}$ of genotype domain elements having function values at least as great as factor α.

In forensic DNA, this subset E_α is the set of genotypes x whose match statistic $f(x)$ is greater than or equal to the match statistic $\alpha = f(x_K)$ for a known person of interest *K*. The RMP measures the "random man" size $\underset{E_\alpha}{\operatorname{sum}} p$ of this extreme genotype set E_α.

Writing the above equality with extreme set E_α, the RMP is the ratio

$$\underset{E_\alpha}{\operatorname{sum}} p = \frac{\underset{E_\alpha}{\operatorname{sum}} q}{\underset{E_\alpha}{\operatorname{avg}}_p f}$$

Algebraically separating numerator and denominator into multiplicative factors gives

$$\operatorname*{sum}_{E_\alpha} p = \frac{1}{\operatorname*{avg}_{p}\limits_{E_\alpha} f} \cdot \operatorname*{sum}_{E_\alpha} q$$

Multiplying through by 1, written as α/α , we then have

$$\operatorname*{sum}_{E_\alpha} p = \frac{1}{\alpha} \cdot \frac{\alpha}{\operatorname*{avg}_{p}\limits_{E_\alpha} f} \cdot \operatorname*{sum}_{E_\alpha} q$$

The left hand side RMP is less than or equal to factor reciprocal $1/\alpha$. This inequality is immediate because the right hand side is $1/\alpha$ multiplied by factors bounded by one: a minimum cannot exceed an average, and partial probability cannot exceed total probability.

In a specific match comparison, small shrinkage factors can make RMP much smaller than the generic $1/LR$ bound. The Case Example below illustrates such error estimate improvement.

Non-contributor Distribution – Right Tail Probability

RMP is the *prior* size of genotype error set E_α . We calculate this false positive *specificity* statistic as the right tail probability of non-contributor pmf $u(y)$ when $y \geq \log\alpha$. To determine match error for contributor genotypes we find an area under the distribution curve.

Suppose a contributor's genotype x_K has an inclusionary LR of α . A small genotype set size $\Pr\{E_\alpha\}$ shows that few non-contributors have a match statistic larger than α . The small (prior) right tail probability value supports the contributor having contributed DNA to the evidence. The match statistic α is far away from the bulk of non-contributor match scores. The small RMP indicates the observed match statistic is specific for a genotype match.

Contributor Distribution – Left Tail Probability

The *posterior* genotype match probability of contributor pmf $v(y)$ when $y \geq \log \alpha$ is

$$\sum_{y \geq \log \alpha} v(y) = \sum_{\{x \mid f(x) \geq \alpha\}} q(x)$$

This genotype set size relates to the statistical *sensitivity* of the match statistic. It measures the size $\Pr\{E_\alpha | \text{data}\}$ of set E_α after examining data.

Suppose a non-contributor's genotype x_K has an exclusionary LR α. A large genotype set probability $\Pr\{E_\alpha | \text{data}\}$ would show that most true genotypes have a match statistic larger than α. Equivalently, a small genotype set probability $1 - \Pr\{E_\alpha | \text{data}\}$ shows few true genotypes having a match statistic as small as α. This gives low probability for a true contributor genotype. So a small (posterior) left distribution tail area $1 - \Pr\{E_\alpha | \text{data}\}$ supports the hypothesis that someone did not contribute DNA to the evidence.

Case Example

We apply the RMP concept of LR error to a sexual assault case in Britain. The example illustrates how error determination helps with DNA database investigation and evidence. The data are used to verify convolution accuracy.

Sexual Assault Information

On New Year's night, 1 January 2014, a woman was sexually assaulted when walking home through a park at 3 a.m. in Southampton, England. The police collected vaginal swabs from the victim, and submitted them to a forensic laboratory for DNA testing with an SGMplus® STR kit (Applied Biosystems, Foster City, CA). Searching the DNA evidence against England's national DNA database (NDNAD) identified 13 candidate suspects based on allele similarity. Other non-biological factors, such as geographical location, singled out homeless Stuart Ashley Burton as the likely perpetrator.

Cybergenetics (Pittsburgh, PA) applied the TrueAllele® Casework software to the SGMplus data, separating out the genotypes of two contributors. The major 85% contributor matched the victim with a statistic of a trillion. Comparing the minor 15% genotype distribution $q(x)$ with Burton's known genotype x_k, relative to a Caucasian population $p(x)$, gave an LR (Bayes factor) $f(x_K)$ of 67,890, with $\log f(x_K)$ = 4.8318 ban.

TrueAllele can bin locus $\log f_l(x)$ values for genotypes x, weighted by prior probabilities $p_l(x)$, to form non-contributor densities $u_l(y)$ along a $y = \log f$ scale. Convolving these u_l locus densities produces a joint non-contributor distribution $u(y)$, as shown in Figure 3. This u distribution has an average exclusionary power

$$\mathrm{E}_p\left[\log \frac{q(x)}{p(x)}\right]$$

of $-KL_{pq}$ = –3.4397 ban, with a standard deviation of 1.6253 ban.

Random Match Probability

Burton's $f\left(x_K\right)$ match statistic of 67.9 thousand has $\log f\left(x_K\right)$ of 4.8318 ban, which gives a right tail RMP of 0.9197x10^{-6}, or 1/1,087,000. Therefore, the chance that a non-contributor (someone who did not contribute their DNA to the vaginal swab evidence) has a match statistic of 67.9 thousand or more, is one in 1.087 million. This precise RMP is 16 times smaller than the generic $1/\alpha$ (reciprocal LR) error estimate of one in 67.9 thousand. The two shrinkage factors locate where RMP improves on the 1/LR error estimate.

The first shrinkage factor says the minimum $\alpha = f\left(x_K\right)$, or $\min\limits_{E_\alpha} f$, cannot exceed the average factor $\operatorname*{avg}_{p}\limits_{E_\alpha} f$. Substituting into the expression

$$\min_{E_\alpha} f \le \operatorname*{avg}_{p}\limits_{E_\alpha} f = \frac{\sum\limits_{x \in E_\alpha} f(x)p(x)}{\sum\limits_{x \in E_\alpha} p(x)}$$

where $\sum f \cdot p = \sum q$, the match values yield

$$67{,}890 \le 236{,}500 = \frac{0.2175}{0.9197 \times 10^{-6}}$$

A larger genotype error set E_α can increase RMP, leading to a greater divergence between $\min f$ and $\operatorname{avg} f$. Here the $\operatorname{avg} f / \min f$ improvement is 3.4841.

The second shrinkage factor says the contributor distribution tail probability on E_α cannot exceed one. The right contributor tail (Figure 4, white region) has probability mass

$$\sum_{x \in E_\alpha} q(x) = 0.2175$$

The reciprocal $1/\operatorname{sum} q$ equals 4.5969. A smaller right tail gives a higher factor gain.

Combining the two shrinkage factors as 3.4841 times 4.5969 equals 16.016. There is a 16-fold improvement from the generic error 1/LR = 1/67,890 upper bound to the exact RMP = 1/1,087,000. Not one in a million people would have a genotype that reached the reported match statistic by chance. The small RMP is persuasive that the match statistic does not falsely include an innocent person (Koehler 2001).

DNA Database Identification

When a DNA database search of evidence returns multiple people $k = 1,...,K$, they can be differentiated by their match statistic. For each retrieved known genotype x_k, determining the posterior-to-prior probability ratio $q(x_k)/p(x_k)$ gives a Bayes factor of $f(x_k)$ that can be used to compare the retrieved genotypes (Table 2, column 2).

In the Southampton rape case, the genotypes show mainly positive $\log f(x_k)$ values (Table 2, column 3 & Figure 5). This is because they were all retrieved from a database search through allelic similarity to the same evidence genotype $q(x)$. However, relative to the evidence, Burton's genotype has a $\log f$ value of 4.8318. This $\log(LR)$ is over 4 ban greater than the 0.4731 ban match statistic average of the other twelve, and over 3 ban away from the largest neighboring value of 1.7455 ban (Table 2, column 3).

The RMP can provide further information useful for differentiating between similar genotypes found from a database search. The RMPs of the 12 less likely suspects range from 1 in 11, to 1 in 808 (Table 2, last column). However, Burton's LR error is 1 in 1.087 million, which is highly specific. Unlike the other suspects, his RMP shows it is extremely unlikely that he is a non-contributor whose genotype produced the 67,890 match statistic by chance.

Based on the DNA match statistics, and other evidence, Burton pleaded guilty to the New Year's Day sexual assault. He was sentenced to twelve years in prison.

Verifying Distribution Accuracy

To verify RMP accuracy, a cumulative distribution for the evidence genotype was independently calculated by Monte Carlo simulation. Ten thousand non-contributor genotypes were randomly drawn from a Great Britain Caucasian (GBC) population. The TrueAllele VUIer software compared the evidence genotype with these randomly simulated exemplars, relative to a GBC population, to calculate match statistics and their base ten logarithms. A co-ancestry theta of 1% was used.

Empirical CDFs for the convolution-based $\log f$ values (blue) and the Monte Carlo simulated values (red) are shown in Figure 6. The two CDF curves are quite similar. A two-sample Kolmogorov-Smirnov (K-S) test (Massey 1951) accepted the null hypothesis that the data curves are from the same continuous distribution (p = 0.2475). The K-S statistic was 0.0102, with a critical value of 0.0136.

The two distributions are the same, statistically. But whereas convolving probability functions gives exact values throughout the entire log factor range, Monte Carlo approximation has limited sampling in the sparse probability tail regions. Since error determination focuses on the tail regions, exact convolution is better suited for determining accurate RMP probability than is Monte Carlo simulation.

Conclusion

Forensic scientists make comparisons to quantify associations between objects. However, they do not directly compare these objects. Nor do they directly compare the data derived from features of these objects. Rather, they compare the underlying types that produce feature data. With DNA evidence, we compare genotypes inferred from observable phenotypic data.

The strength of genotype association is quantified in the DNA match statistic. The statistic is the probability of a match between the evidence and a suspect, relative to coincidence. This LR value measures the strength of DNA evidence against the suspect.

We also care about LR error. RMP is a standard DNA statistic for conveying potential error. Stated in LR evidentiary terms, generalized RMP is the probability that the evidence against a random person is as strong as it is against the suspect. A low probability inspires confidence in match specificity, while a high probability of spurious match raises doubt.

Twenty years ago, the RMP statistic summarized simple DNA evidence having a definite genotype. For a genotype having a population prior probability of p, the LR is $1/p$ and the RMP is p. The strength of simple DNA evidence is simply expressed through the single number p.

More complex DNA data introduces uncertain genotypes. Mathematical mixture separation or kinship analysis can represent this uncertainty using probability. A contributor’s genotype at a genetic locus is a probability distribution over allele pairs. At the suspect’s genotype, the posterior evidence genotype probability is q. The LR generalizes to q/p, the ratio of posterior to prior genotype probabilities. The reduction from 1 (definite) to q (uncertain) in numerator posterior probability expresses reduced DNA evidence strength.

RMP quantifies LR error. Genotype uncertainty introduces multiple possibilities into an RMP sum. The LR factor function *f* defines these genotype possibilities. The genotype candidates form a set E_α of genotypes *x* having an LR value $f(x)$ that is at least as large as the suspect's LR value α .

Instead of one prior probability *p* for a definite genotype, with genotype uncertainty the RMP becomes a sum $\sum_{x \in E_\alpha} p(x)$ of prior probabilities over many genotype values. Reciprocal equality between LR and RMP becomes the inequalities $LR \le 1/RMP$ and $RMP \le 1/LR$. The RMP sum becomes a tail probability of the genotype's non-contributor LR distribution.

At a locus, we can assemble the (contributor or non-contributor) LR log factor distribution from prior and posterior genotype probabilities. Convolving independent locus distributions rapidly calculates the joint factor LR distribution. RMP is then computed as a right tail probability (beyond the suspect's LR value) of the non-contributor distribution.

These LR and RMP methods are not limited to DNA identification. In every forensic modality, observable features produce data, from which underlying types can be ascertained. Combining a forensic type's prior and posterior probabilities can construct its LR log factor distribution. The LR distribution tells us about the type's strength of evidence, and facilitates RMP and other error calculations. This statistical analysis applies to all types of forensic features, whether discrete or continuous (Appendix A).

In the courtroom, forensic experts can deluge a trier of fact with DNA details. The LR summarizes all the evidence against a defendant in a single number. RMP describes the probability of LR error. Multiplying RMP by a relevant population size gives the number of innocent people for whom the DNA evidence is as strong as it is against the defendant. LR and RMP statistics can both help a trier of fact to understand the evidence.

References

Butler, J. M. (2000). Forensic DNA Typing: Biology and Technology Behind STR Markers. New York, Academic Press.

Collins, P. J., L. K. Hennessy, C. S. Leibelt, R. K. Roby, D. J. Reeder and P. A. Foxall (2004). "Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR Identifiler PCR Amplification Kit." J Forensic Sci **49**(6): 1265-1277.

Corradi, F. and F. Ricciardi (2013). "Evaluation of kinship identification systems based on short tandem repeat DNA profiles." Journal of the Royal Statistical Society: Series C (Applied Statistics) **62**(5): 649–668.

Dørum, G., Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø and T. Egeland (2014). "Exact computation of the distribution of likelihood ratios with forensic applications." Forensic Science International: Genetics **9**: 93-101.

Edwards, A., A. Civitello, H. Hammond and C. Caskey (1991). "DNA typing and genetic mapping with trimeric and tetrameric tandem repeats." Am. J. Hum. Genet. **49**: 746-756.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications. New York, John Wiley & Sons.

Gelman, A., J. B. Carlin, H. S. Stern and D. Rubin (1995). Bayesian Data Analysis. Boca Raton, FL, Chapman & Hall/CRC.

Gill, P., J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker and J. Lambert (2008). "Interpretation of complex DNA profiles using empirical models and a method to measure their robustness." Forensic Science International: Genetics **2**(2): 91-103.

Good, I. J. (1950). Probability and the Weighing of Evidence. London, Griffin.

Good, I. J. (1985). Weight of evidence: a brief survey. Bayesian Statistics. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith. Amsterdam, Elsevier Science Publishers B.V. (North-Holland). **2:** 249-270.

Kenney, J. (1988). "Hypothesis testing: guilty or innocent." Quality Progress **55**(1): 55-57.

Koehler, J. J. (2001). "When are people persuaded by DNA match statistics?" Law and Human Behavior **25**(5): 493-513.

Kruijver, M. (2015). "Efficient computations with the likelihood ratio distribution." Forensic Science International: Genetics **14**: 116-124.

Kruijver, M., R. Meester and K. Slooten (2015). "p-Values should not be used for evaluating the strength of DNA evidence." Forensic Science International: Genetics **16**: 226-231.

Kullback, S. and R. A. Leibler (1951). "On information and sufficiency." Ann Math Stat **22**(1): 79-86.

Lebesgue, H. (1902). "Integral, length, area." Annals of Mathematics **7**(3): 231-359.

MacKay, D. J. (2003). Information Theory, Inference and Learning Algorithms. Cambridge, UK, Cambridge University Press.

Massey, F. T. (1951). "The Kolmogorov-Smirnov test for goodness of fit." Journal of the American Statistical Association **46**(253): 68–78.

National Research Council (1996). Evaluation of Forensic DNA Evidence: Update on Evaluating DNA Evidence. Washington, DC, National Academies Press.

Nothnagel, M., J. Schmidtke and M. Krawczak (2010). "Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci " Int J Legal Med **124**: 205-215

O'Hagan, A. and J. Forster (2004). Bayesian Inference. New York, John Wiley & Sons.

Ott, J. (1991). Analysis of Human Genetic Linkage. Baltimore, Maryland, The Johns Hopkins University Press.

Perlin, M. W., M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose and B. W. Duceman (2011). "Validating TrueAllele® DNA mixture interpretation." J Forensic Sci **56**(6): 1430-1447.

Perlin, M. W. and A. Sinelnikov (2009). "An information gap in DNA evidence interpretation." PLoS ONE **4**(12): e8327.

Perlin, M. W. and B. Szabady (2001). "Linear mixture analysis: a mathematical approach to resolving mixed DNA samples." J Forensic Sci **46**(6): 1372-1377.

Riemann, B. (1868). "On the representability of a function by a trigonometric series." Proceedings of the Royal Philosophical Society at Göttingen **13**: 87-132.

Royall, R. (2000). "On the probability of observing misleading evidence." Journal of the American Statistical Association **95**(451): 760-768.

Saks, M. J. and S. Neufeld (2011). "Convergent evolution in law and science: the structure of decision-making under uncertainty." Law, Probability and Risk **10**(2): 133-148.

Scientific Working Group on DNA Analysis Methods (SWGDAM) (2010). Interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories. http://www.forensicdna.com/assets/swgdam_2010.pdf, FBI Laboratory.

Sjerps, M. J., I. Alberink, A. Bolck, R. D. Stoel, P. Vergeer and J. H. van Zanten (2016). "Uncertainty and LR: to integrate or not to integrate, that's the question." Law Probab Risk **15**(1): 23-29.

Slooten, K.-J. and T. Egeland (2014). "Exclusion probabilities and likelihood ratios with applications to kinship problems." Int J Legal Med **128**(3): 415-425.

Slooten, K.-J. and T. Egeland (2015). "Exclusion probabilities and likelihood ratios with applications to mixtures." Int J Legal Med **130**(1): 39-57.

Taroni, F., S. Bozza, A. Biedermann and C. Aitken (2016). "Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio." Law Probability and Risk **15** (2): 1-18.

Taylor, D., J. Buckleton and I. Evett (2015). "Testing likelihood ratios produced from complex DNA profiles." Forensic Science International: Genetics **16**: 165-171.

Weber, J. and P. May (1989). "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction." Am. J. Hum. Genet. **44**: 388-396.

Wheeden, R. L. and A. Zygmund (1977). Measure and Integral: An Introduction to Real Analysis. New York, Marcel Dekker.

Appendix A: From Genotypes to Distributions

Reducing genotype probability and log factor values to a one-dimensional distribution simplifies and accelerates error determination. We show that discrete genotype match error is a tail probability. This correspondence is a special case of the "pushforward measure." For more general forensic comparisons, we show how to construct this measure, and apply it to contributor and non-contributor log factor distributions.

Discrete Probability

With discrete genotypes, the specificity error for false positive matches is RMP – the size of genotype error set E_α. This set size also equals the right tail probability of the non-contributor pmf $u(y)$ when $y \geq \log\alpha$. We show here that the (easily calculated) tail probability $\sum_{y\geq\log\alpha} u(y)$ equals the size of the genotype error set $\Pr\{x \in E_\alpha\}$.

First express the $u(y)$ tail probability as

$$\sum_{y\geq\log\alpha} u(y) = \sum_{y\geq\log\alpha}\left[\sum_{x\in\log f^{-1}(y)} p(x)\right]$$

by expanding pmf u as a sum of prior probabilities $p(x)$ over non-contributor genotypes x sharing log factor y. Writing out the inner summation for the set of genotypes x, we have

$$= \sum_{y\geq\log\alpha} \sum_{\{x\in X \mid \log f(x) = y\}} p(x)$$

This reduces to the combined sum

$$= \sum_{\{x \in X \mid \log f(x) \geq \log \alpha\}} p(x)$$

Exponentiating the logarithms on both sides of the summation's set condition inequality, the expression equals

$$= \sum_{\{x \in X \mid f(x) \geq \alpha\}} p(x)$$

Since E_α is the genotype error set $\{x \in X | f(x) \geq \alpha\}$, we obtain the prior probability sum

$$= \sum_{x \in E_\alpha} p(x)$$

which equals the RMP set size

$$= \Pr\{x \in E_\alpha\}$$

Measure and Integral

The Lebesgue theory of measurable sets, functions and integrals generalizes continuous functions and Riemann integration to handle pathological situations (Wheeden and Zygmund 1977). Lebesgue measure and integration work with finite and infinite sets, over discrete and continuous domains, and eliminate technical issues involving sets of measure zero and infinite discontinuities.

For a set *X*, a σ-*algebra* Σ of subsets of *X* contains *X*, and is closed under set complementation and countable set unions. A *measure* μ is a nonnegative function on measureable subsets *E* in Σ for which $\mu\left(\bigcup E_K\right) = \sum \mu(E_K)$, whenever $\{E_K\}$ is a countable family of disjoint sets in Σ.

A *measure space* is a triple (X,Σ,μ). A real-valued function $f(x)$ defined for *x* in a measureable set *E* in Σ is a *measureable function* when $\{x \in E | f(x) \geq \alpha\}$ is a measureable set for all finite real numbers $\alpha \in \mathbb{R}$. The *Lebesgue integral* $\int_E f\, d\mu$ of a measureable function *f* over a measureable set *E* with respect to measure μ is more robust than its Riemann counterpart, and enjoys many useful convergence properties.

Pushforward Measure

Let (X,Σ,μ) be a measure space, and *f* a measureable function from *X* to $\mathbb{R}^1$. Let $(\mathbb{R},\mathscr{B})$ be the space of real numbers with the Borel σ-algebra $\mathscr{B}$. Then the *pushforward measure* $f_*(\mu)$ is a nonnegative measure defined as

$$f_*(\mu)(B) = \mu\left(f^{-1}(B)\right)$$

for every Borel set *B* in $\mathscr{B}$. The resulting measure space $(\mathbb{R},\mathscr{B},f_*(\mu))$ is the one induced on $\mathbb{R}$ by measure space (X,Σ,μ) and function *f*.

In particular, for any $\alpha \in \mathbb{R}$, consider the half infinite interval

$$[\alpha,\infty) = \{y \in \mathbb{R} | y \geq \alpha\}$$

Taking the inverse image of this Borel set under function *f*, we have that

$$E_\alpha = f^{-1}([\alpha,\infty)) = \{x \in X | f(x) \geq \alpha\}$$

is the subset of *X* having $f(x) \geq \alpha$. The measure of this subset E_α

$$\mu(E_\alpha) = \mu\left(f^{-1}([\alpha,\infty))\right)$$

describes the size of E_α. Given a measure space (X,Σ,μ), the pushforward measure $f_*(\mu)$ specifies a *distribution function* on $\mathbb{R}$.

When μ is a probability measure, so that $\mu(X)=1$,

$$\begin{aligned}\mu(E_\alpha) &= \mu\left(f^{-1}([\alpha,\infty))\right)\\ &= f_*(\mu)([\alpha,\infty))\\ &= \int_\alpha^\infty df_*(\mu)\end{aligned}$$

Thus subset probability $\mu(E_\alpha)$ is pushed forward from (X,Σ,μ), and equals the tail probability $\int_\alpha^\infty df_*(\mu)$ on the half infinite interval $[\alpha,\infty)$ in the measure space $\left(\mathbb{R},\mathscr{B},f_*(\mu)\right)$. On any probability space (X,Σ,μ), the pushforward measure for f induces a probability distribution function on the real line. Product measures correspond to tuples of elements drawn from independent spaces.

Genotype Tail Probability

The tail probability $\sum_{y\geq\log\alpha} u(y)$ equals the genotype error set probability $\Pr\{x\in E_\alpha\}$. This equality follows immediately from measure theory when the tail probability is defined through a pushforward measure.

In general, pushing *f* forward from *X* to $\mathbb{R}$ reduces the problem of finding a measure $\mu(E_\alpha)$ in a multi-dimensional probability space (X,Σ,μ) to that of calculating an integral in the one-dimensional probability space $\left(\mathbb{R}^1,\mathscr{B},f_*(\mu)\right)$. This integral is the right tail probability $\int_\alpha^\infty df_*(\mu)$ of *f* starting from point α, which is the same as one minus the cumulative distribution

$\int_{-\infty}^{\alpha} df_*(\mu)$ of f ending at α. The pushforward dimension reduction translates subset probability in (X,Σ,μ) into a simpler integral over $\mathbb{R}^1$.

When the measure μ is the prior distribution $p(x)$, the pushforward measure $f_*(p)$ describes the non-contributor factor distribution. When μ is the posterior $q(x)$, pushforward measure $f_*(q)$ gives the contributor factor distribution. Using $\log f$ in place of f pushes forward onto $\mathbb{R}$ the corresponding log factor distribution.

Appendix B: Population Genetics Adjustments

There are well-known adjustments that can be made for population genetics. The co-ancestry correction for relatedness lowers the match statistic. Any population's allele frequencies can be substituted into a match statistic, replacing the original genotype prior.

Co-ancestry Correction

All people share a common ancestry, more so in closely related populations. Therefore, human genotypes are not entirely independent of each other. The usual Hardy-Weinberg equilibrium population probability for a genotype *ij*

$$\Pr\{X = ij\} = \begin{cases} p_i^2,\ i = j \\ 2p_ip_j,\ otherwise \end{cases}$$

assumes independent mating, and therefore requires some adjustment.

A simple and effective correction is to introduce a co-ancestry coefficient θ that measures the degree of inbreeding within a population. Then the prior genotype probabilities become (Ott 1991)

$$\Pr\{X = ij \mid \theta\} = \begin{cases} p_i^2 + \theta p_i(1-p_i),\ i = j \\ 2(1-\theta)p_ip_j,\ otherwise \end{cases}$$

accounting for an increase in homozygote ($i = j$) genotypes, with a commensurate decrease in heterozygotes ($i \neq j$).

Population Substitution

Bayesian genotype inference updates population prior $p(x)$ to a posterior $q(x)$. This update is mediated though a likelihood function

$$l(x) = \Pr\{data \mid X = x, ...\}$$

based on observed DNA data, where

$$q(x) \propto l(x)\,p(x)$$

So for a different population $p_o(x) \neq p(x)$ having different allele frequencies, the posterior $q_o(x) \neq q(x)$ is different as well. We can transform one genotype posterior $q_o(x)$ based on prior $p_o(x)$ to a new $q(x)$ based on $p(x)$. This is easily done through the likelihood function l using Bayes theorem by writing posterior function q as

$$q(x) = \frac{l(x)\,p(x)}{\sum_{y \in X} l(y)\,p(y)}$$

for any prior function p. In a vectorized computer language, q can be calculated over the entire domain $x \in X$ in one step.

In the MATLAB programming language, for example, likelihood l and prior p column vectors are combined as `l.*p` over `l'*p` to produce the posterior genotype probability vector q. In practice, one can first exhaustively compute a genotype $q_o(x)$ using any prior $p_o(x)$, and later swap in a new population $p(x)$ using Bayes theorem. Changing populations does not require extensive genotype re-computation.

Tables

Table 1. Forming $\log f$ factors from prior and posterior genotype probability.

Type	**Prior**	**Posterior**	**Factor**	**log factor**
x	*p(x)*	*q(x)*	*f(x)*	*log f(x)*
1	0.20	0.10	0.5	-0.301
2	0.30	0.15	0.5	-0.301
3	0.25	0.25	1.0	0.000
4	0.25	0.50	2.0	0.301

Table 2. The LR match statistics and RMP error probabilities in the Southampton case. Each row represents a different retrieved DNA database genotype, with "SB" the accused. The last column's "one in" value is the reciprocal of the RMP given in the preceding column.

Item	LR	log(LR)	RMP	one in:
1	1/(17.7)	-1.2485	0.09155110	11
2	1/(2.72)	-0.4339	0.03595410	28
3	1.21	0.0824	0.01818210	55
4	1.54	0.1878	0.01569030	64
5	2.01	0.3025	0.01330630	75
6	3.35	0.5248	0.00958381	104
7	3.35	0.5248	0.00958381	104
8	5.21	0.7166	0.00713871	140
9	5.90	0.7709	0.00655932	152
10	17.8	1.2513	0.00297871	336
11	17.9	1.2535	0.00296855	337
12	55.6	1.7455	0.00123809	808
SB	67,890	4.8318	0.00000092	1,087,000

Legends

Figure 1. Histogram of binned $\left(\log f(x), p(x)\right)$ pairs constructs a probability mass function at a locus. The *x*-axis is the factor expressed in logarithmic ban units, while the *y*-axis is probability mass.

Figure 2. Cumulative probability represents a log factor distribution at a locus. The *x*-axis is the factor expressed in logarithmic ban units, while the *y*-axis is cumulative probability.

Figure 3. The Southampton case joint non-contributor distribution (blue line) for a genotype separated from DNA mixture evidence data. The *x*-axis is the factor expressed in logarithmic ban units, while the *y*-axis is probability density. Computed by the TrueAllele computer and displayed in its user interface.

Figure 4. The Southampton case joint contributor distribution (blue line). The green arrow indicates the log factor value of the suspect's genotype. The left (red region) and right (white region) tail probabilities are shaded. The *x*-axis is the factor expressed in logarithmic ban units, while the *y*-axis is probability density. (Rendered by TrueAllele.)

Figure 5. The Southampton case joint non-contributor distribution (blue line) for a genotype separated from mixture data. The green arrows indicate the log factor values of retrieved DNA database genotypes. The *x*-axis is the factor expressed in logarithmic ban units, while the *y*-axis is probability density. (Rendered by TrueAllele.)

Figure 6. CDFs for convolution-based $\log f$ values (blue line) and Monte Carlo simulated values (red line). The *x*-axis is the factor expressed in logarithmic ban units, while the *y*-axis is cumulative probability. (Rendered by MATLAB.)

Figures

Figure 1

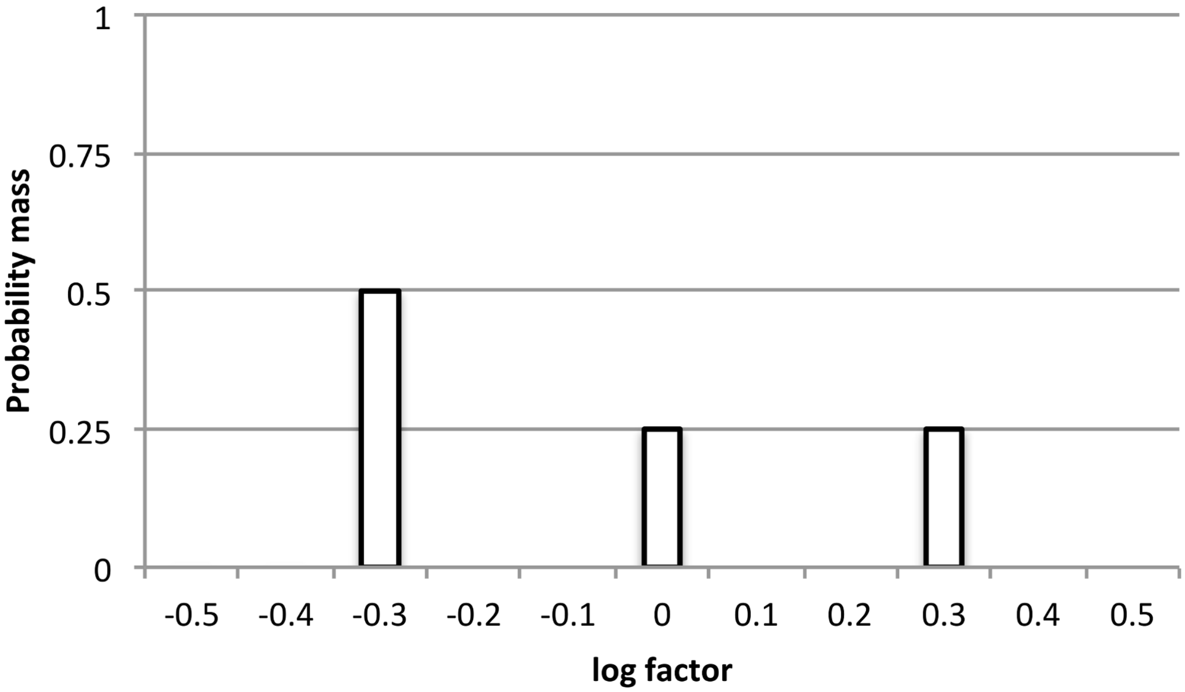


Figure 2

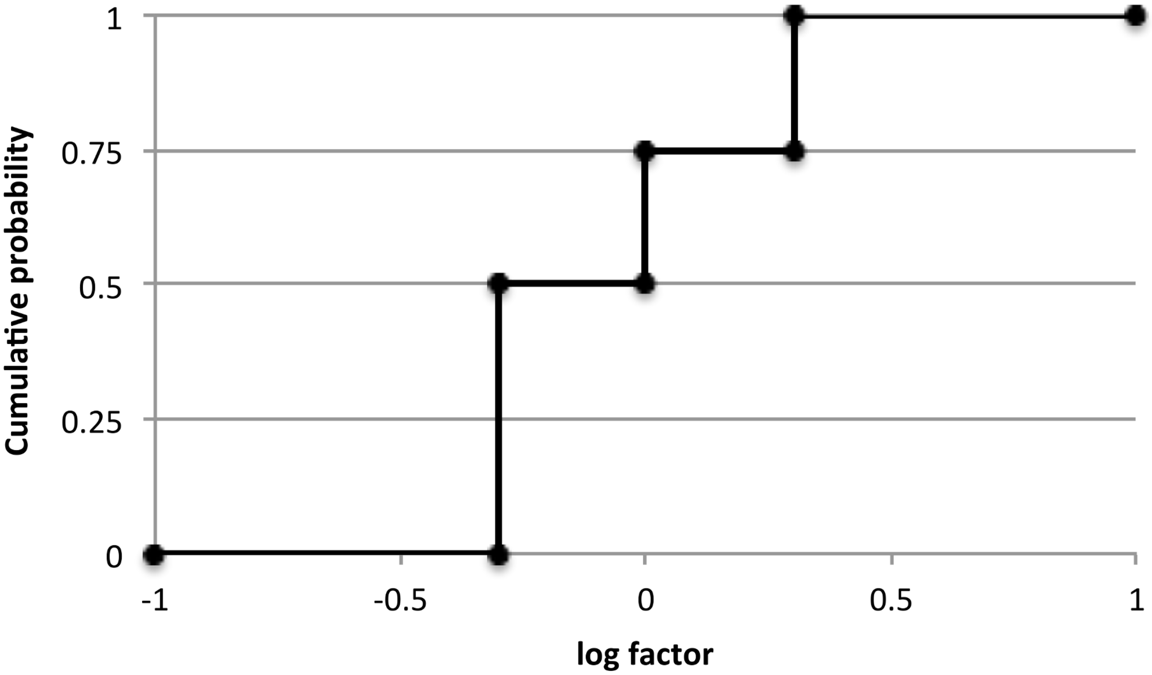


Figure 3

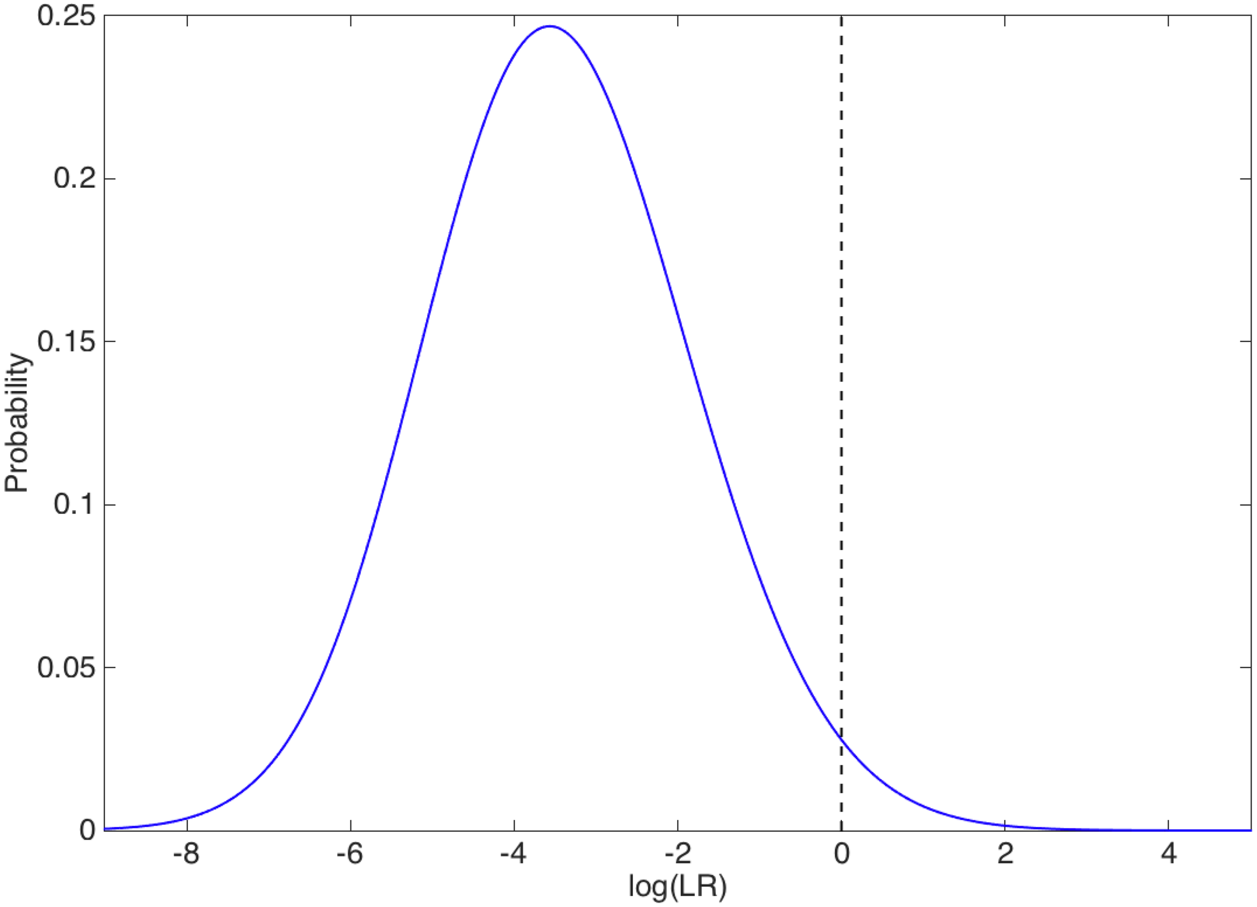


Figure 4

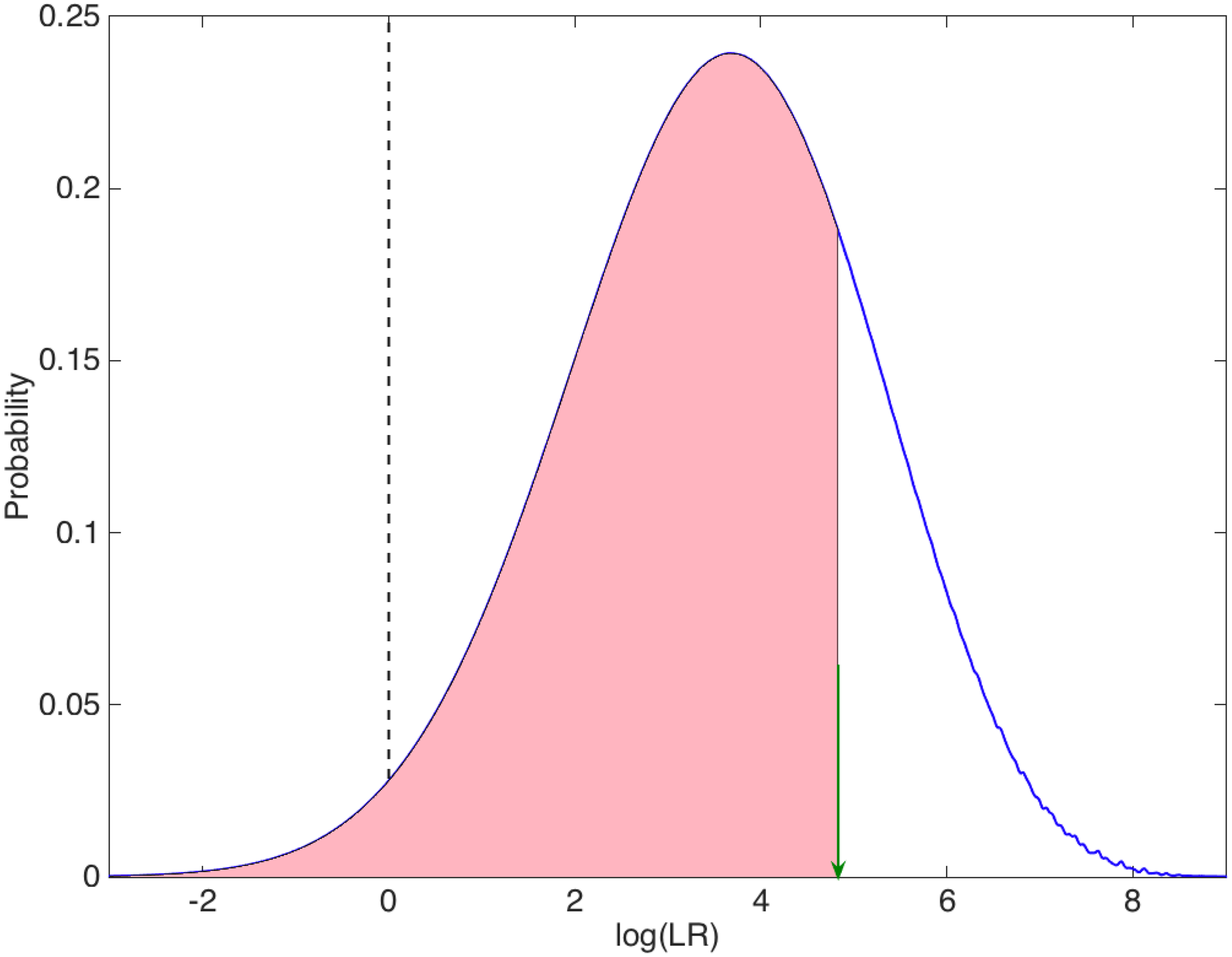


Figure 5

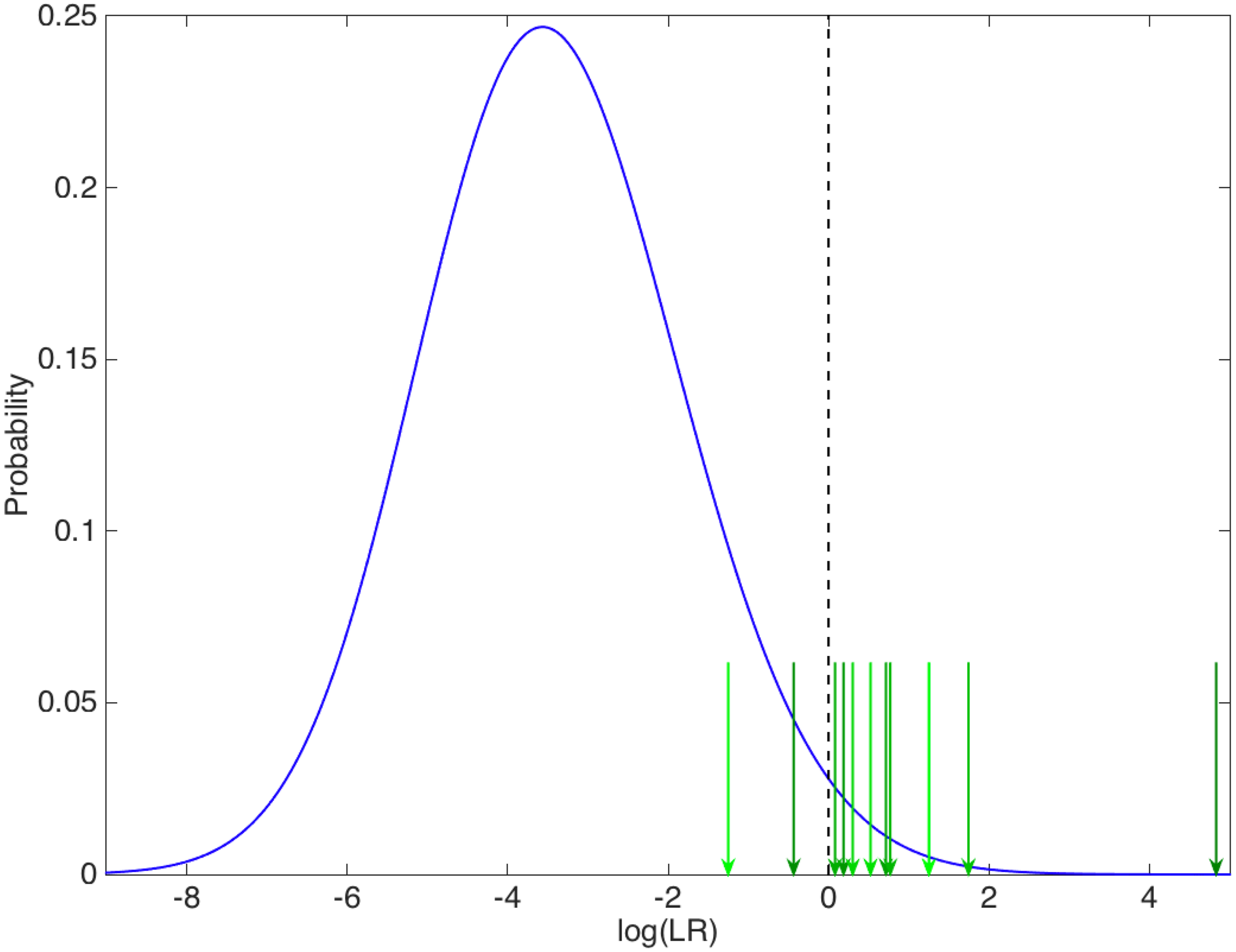
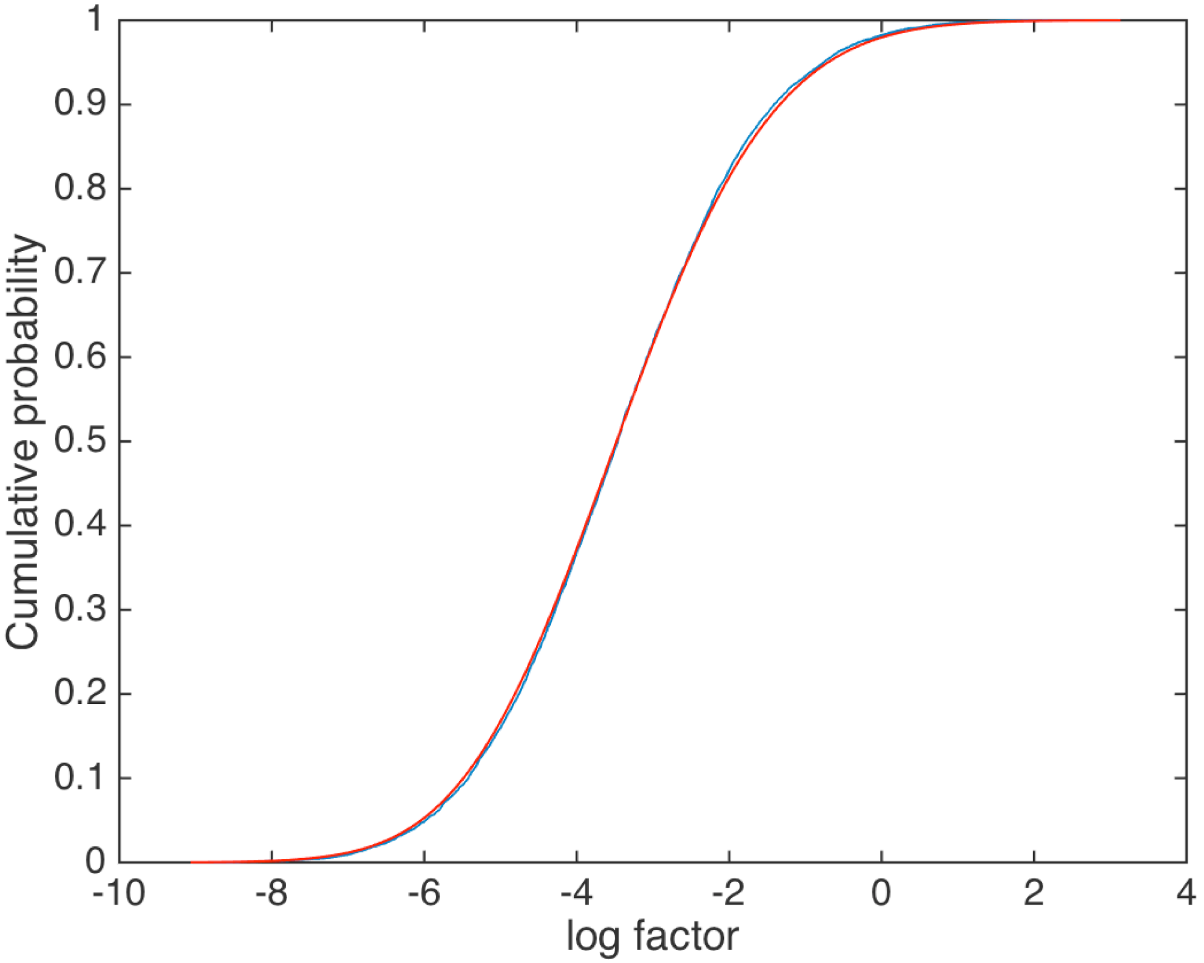


Figure 6


Cumulative probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10
-8
-6
-4
-2
0
2
4
log factor