Population Genetics & Statistics for Forensic Analysts

This course is provided free of charge and is part of a series designed to teach about DNA and forensic DNA use and analysis.

Find this course live, online at: http://dna.gov/training/populations

Updated: October 8, 2008

DNA
INITIATIVE

www.DNA.gov

About this Course

This PDF file has been created from the free, self-paced online course "Crime Scene and DNA Basics for Forensic Analysts." To learn more and take this and other courses online, go to http://www.dna.gov/training/online-training/. Most courses are free but you must first register at http://register.dna.gov.

If you already are registered for any course on DNA.gov, you may login directly at the course URL, e.g., http://letraining.dna.gov or you can reach the courses by using the URL http://www.dna.gov/training and selecting the "Login and view your courses" link.

Questions? If you have any questions about this file or any of the courses or content on DNA.gov, visit us online at http://www.dna.gov/more/contactus/.

Links in this File

Most courses from DNA.Gov contain animations, videos, downloadable documents and/or links to other useful Web sites. If you are using a printed, paper version of this course, you will not have access to those features. If you are viewing the course as a PDF file online, you may be able to use some of these features if you are connected to the Internet.

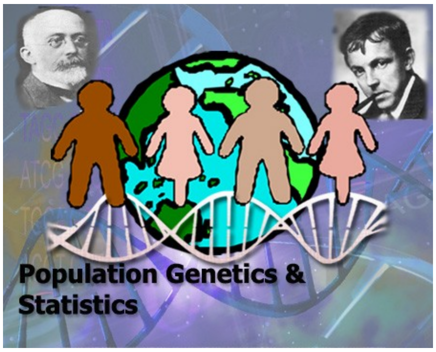
Animations, Audio and Video. Throughout this course, there may be links to animation, audio or video files. To listen to or view these files, you need to be connected to the Internet and have the requisite plug-in applications installed on your computer.

Links to other Web Sites. To listen to or view any animation, audio or video files, you need to be connected to the Internet and have the requisite plug-in applications installed on your computer.

Legal Policies and Disclaimers
See Legal Policies and Disclaimers for information on Links to Other Web Sites, Copyright Status and Citation and Disclaimer of Liability and Endorsement.

Population Genetics and Statistics for Forensic Analysts

Population Genetics and Statistics



This course provides information in the two lessons.

Population Theory. Learn the basic terms in population genetics, factors that can alter allele frequencies in a population, and calculations associated with the Hardy-Weinberg principle.

Statistics. Learn the importance and use of statistics and probability to the field of forensic science, how to use at least one of the accepted statistical approaches to evaluate DNA data, and how DNA population databases are constructed and used.

Population Theory

This module deals with calculating the frequency of **alleles** in relevant human populations. The knowledge of how human populations arise, interact, and develop is central to the understanding of genetics. Alleles are distributed according to the basic rules of Mendelian genetics and are sensitive to natural forces that can be quantitated when populations are studied. This section will introduce the fundamental terms, theories, models, and assumptions and forms the basis for the routine calculations performed on forensic samples. It is assumed that the student is familiar with basic genetics.

Read about Genetics in the Crime Scene and DNA Basics for Forensics Analysts PDF file.

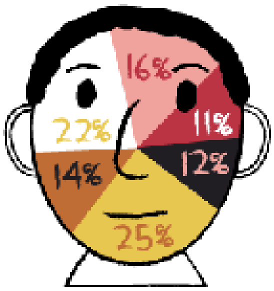
Objectives

Upon successful completion of this unit of instruction, the student shall be able to do the following:

- Define the basic terms in population genetics
- List factors that can alter allele frequencies in a population
- Perform calculations associated with the Hardy-Weinberg principle

Population Genetics

Population genetics is the study of the distribution of and change in allele frequencies under the influence of the four evolutionary forces: natural selection, genetic drift, mutation, and migration. Population structure is also taken into account.01



Descriptive statistics are used to summarize a collection of population data in a clear and concise manner. The collection of population data can be exhaustive and usually results in the accumulation of a considerable volume of information. The term "population" can be defined in many ways, but in the most common sense of the word, it is a collection of people or organisms of a particular species living in a geographic area.

In biology, a population generally denotes a group whose members breed primarily or solely among themselves. This is usually a result of a physical isolation, although biologically they could breed with any member of the species. A common misconception is that human populations fall into uniform groups with large biological differences, which lie along racial boundaries. Racial groups are not as easily defined. Depending on the classification, there can be between 3 and 200 races in the human population.02 Research has demonstrated that racial classifications are inadequate descriptors of the distribution of genetic variation in the human species.03 An alternative to classifying the human species by race is to divide them by ethnicity. Dividing populations along ethnic origins can present some of the same classification issues. The self-described ethnicity of an individual may be different from his or her familial origins. It is difficult to categorize population groups, and while neither race nor ethnicity is ideal, each is used to provide estimates of allele frequencies.

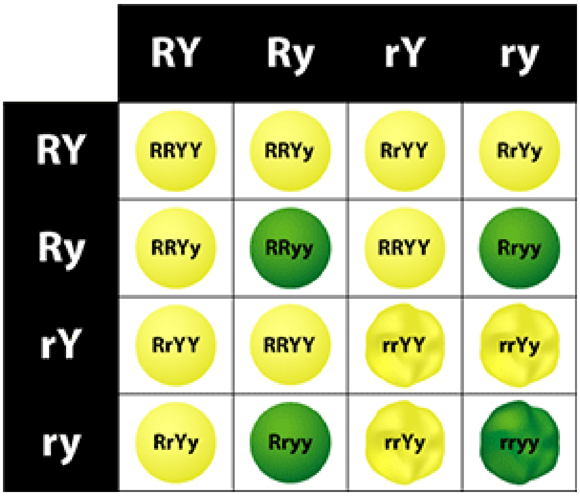
Heterozygosity

The variation in alleles is critical to the survival of a species and allows organisms to adapt to changing environments. Allele frequency, or the frequency at which alleles are found at any locus of interest, is used to estimate the frequency of a given genetic profile. Every diploid cell has two alleles, one inherited from each parent. If an individual has two different alleles at a specific locus, the individual is heterozygous at that locus; if the two alleles are the same, the individual is homozygous. Allele frequency is used to characterize the genetic diversity, or richness of the gene pool, in a population. Populations need variation. The measure of the amount of heterozygosity across loci can be used as a general indicator of the amount of genetic variability.

The number of possible genotypes from only a few loci is great, and can be calculated using the formula $k(k+1)/2$, where k is the number of alleles at a particular locus. The parameter k also represents the expected number of homozygous genotypes, and $k[(k-1)/2]$ represents the expected number of heterozygous genotypes. The observed heterozygosity can be compared to the expected heterozygosity, and the deviations between these values can indicate important population dynamics.

The online version of this course contains a multimedia [or downloadable] file. Visit this URL to view the file: http://beta.populations.dna.devis.com/m01/01/a

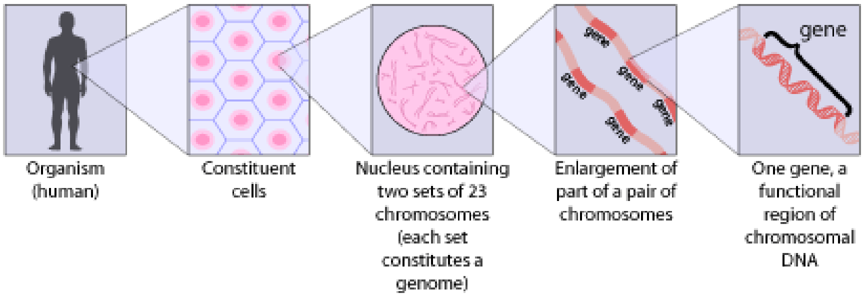
The exact genotypes can be determined using a Punnett square. A Punnett square is simply a grid that graphically represents expected genotypes. These grids are useful tools to visually determine projected genotypes and their frequencies.



	RY	Ry	rY	ry
RY	RRYY	RRYy	RrYY	RrYy
Ry	RRYy	RRyy	RRYY	Rryy
rY	RrYY	RRYY	rrYY	rrYy
ry	RrYy	Rryy	rrYy	rryy

A gene pool is the unique set of alleles that could be found by analyzing the DNA from every member of a species or population. A large gene pool is often associated with considerable genetic diversity whereas a small gene pool is associated with poor genetic diversity and can lead to decreased fitness and an increased chance of extinction. Fitness is the probability of transmitting one's genes to the next generation in relation to the average probability for that population.04

Genetic diversity is greater *within* populations than *between* them. The study of genetics has revealed that approximately 85% of human genetic diversity comes from individuals of the same population. The remaining 15% is derived from diversity between continents (~10%) and to a lesser extent, diversity within continents (~5%).02 These data support the view that interracial variations do not represent major differences in the human genome, in contrast to what was previously thought.


gene
gene
gene
gene
gene
gene
Organism (human)
Constituent cells
Nucleus containing two sets of 23 chromosomes (each set constitutes a genome)
Enlargement of part of a pair of chromosomes
One gene, a functional region of chromosomal DNA

Hardy-Weinberg Principle

In 1908, two scientists, Godfrey H. Hardy, an English mathematician, and Wilhelm Weinberg, a German physician, independently worked out a mathematical relationship that related genotypes to allele frequencies.05

Godfrey H. Hardy Wilhelm Weinberg

Their mathematical concept, called the Hardy-Weinberg principle, is a crucial concept in population genetics. It predicts how gene frequencies will be inherited from generation to generation given a specific set of assumptions.06 The Hardy-Weinberg principle states that in a large randomly breeding population, allelic frequencies will remain the same from generation to generation assuming that there is no mutation, gene migration, selection, or genetic drift.04 This principle is important because it gives biologists a standard from which to measure changes in allele frequency in a population.

The Hardy-Weinberg principle can be illustrated mathematically with the equation:

$p^2+2pq+q^2 = 1$

Where 'p' and 'q' represent the frequencies of alleles. It is important to note that p added to q always equals one (100%).

To illustrate how the Hardy-Weinberg principle works, let us consider the MN blood group. Humans inherit either the M or the N antigen, which is determined by different alleles at the same gene locus. If we let the frequency of allele M=p and the frequency of the other allele N=q, then the next generation's genotypes will occur as follows:

- Frequency of MM genotype = p^2
- Frequency of MN genotype = 2pq
- Frequency of NN genotype = q^2

We can take a sample of the population and count the number of people with each genotype. For example, a sample of 5000 from Forensic Town, USA, has:

- 1460 individuals of type MM, that is 1460/5000 or 29.2%
- 2550 of type MN, that is 2550/5000 or 51%
- 990 of type NN, that is 990/5000 or 19.8%

If we apply the Hardy-Weinberg equation ($p^2 + 2pq + q^2 = 1$), we can calculate the allele frequencies as:

- Frequency of M = $p^2 + 0.5\ (2pq) = 0.292 + (0.5 \times 0.51) = 0.547$
- Frequency of N = $q = 1 - p = 1 - 0.547 = 0.453$

We can now calculate our expected genotype frequencies:

- MM = $p^2 = 0.5472 = 0.299$, or 1496 individuals in the sample
- MN = $2pq = 2 \times 0.547 \times 0.453 = 0.496$, or 2478 individuals
- NN = $q^2 = 0.4532 = 0.205$, or 1026 individuals

An example of how to determine whether a population is in the Hardy-Weinberg equilibrium:

Example

Suppose that scientists are observing a population of lab-bred flies, and discover a gene controlling eye color. The *R* allele produces regular-colored eye pigment, while the *r* allele produces red pigment. Individuals that are heterozygous (*Rr*) have pink eyes. In a population of 150 flies, 15 flies have red eyes, 90 have normal eye color, and 45 have pink eyes.

Is this population in Hardy-Weinberg equilibrium?

In order for a population to be considered to be in equilibrium, it must remain the same from generation to generation. Therefore, in order to determine if this population of fruit flies is in Hardy-Weinberg equilibrium, the genetic distribution of the current generation must be compared to a prediction of the genetic distribution of the next generation, as calculated using the Hardy-Weinberg equation.

Step 1: Determine the gene frequencies of the current generation.

Phenotype	Genotype	# of Individuals
Normal Eyes	*RR*	90
Red Eyes	*rr*	15
Pink Eyes	*Rr*	45

Given this information, calculating the allele frequencies is simply a matter of counting up all of the alleles.

- Remember, each parent carries *two* alleles, so the total # of alleles is twice the population.
- Also remember that *heterozygous* individuals carry one of *each* allele.

Taking these two factors into account,

f(R) = [(90 x 2) + (45)] / (150 x 2) = 225/300 = **0.75**

f(r) = [(15 x 2) + (45)] / (150 x 2) = 75/300 = **0.25**

Step 2: Determine the expected genotype frequencies for the next generation.

Plugging the frequencies of each allele into the Hardy-Weinberg equation, we find the expected numbers of each genotype in the population:

f(RR) = p2 = f(R) x f(R) = **0.5625**

f(rr) = q2 = f(r) x f(r) = **0.0625**

f(Rr) = 2pq = 2 x [f(R) x f(r)] = **0.375**

Multiplying each of these genotype frequencies with the total population number, we find that there should be:

- 84 normal-eyes flies (*AA*)
- 9 red-eyed flies (*aa*)
- 56 pink-eyed flies (*Aa*)

(Since partial individuals do not exist, the numbers are rounded off.)

Step 3: Compare the expected frequency with the original population numbers.

Comparing the expected numbers with the actual numbers of each phenotype, population geneticists can determine if populations are either in equilibrium (or very close to it) or are experiencing *disequilibrium* of some sort. In this example:

Phenotype	Genotype	Expected #	Observed #
Normal Eyes	*RR*	84	90
Red Eyes	*rr*	9	15
Pink Eyes	*Rr*	56	45

In this example, the population is not in equilibrium since the expected and observed values do not match. Once a population geneticist determines that a population is in disequilibrium, the reasons can be explored. Disequilibrium can be attributed to different possible mechanisms, depending on (1) the context of the population, and (2) the manner in which the population is skewed.

When a population meets all of the of the Hardy-Weinberg conditions, it is said to be in Hardy-Weinberg equilibrium (HWE). Human populations do not meet all of the conditions of HWE exactly, and their allele frequencies will change from one generation to the next and the population will evolve. How far a population deviates from HWE can be measured using the "goodness of fit" or chi-squared test (^{2}).

Mathematically the chi-squared test is represented:

^{2} = [(observed value – expected value)^2 / expected value]

Applying the above data to the chi-square test gives:

- ^{2} = [(990 - 1026)2 / 1026] + [(2550 – 2478)2 / 2478] + [(1460-1496)2 / 1496]
- ^{2} = [1296 / 1026] + [5184 / 2478] + [1296 / 1496]
- ^{2} = [1.263] + [2.092] + [0.866] = 4.221

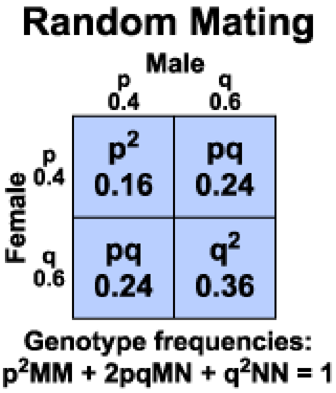
To determine what this chi-squared value means, we must next look at a chi-squared distribution table.

View the chi-squared distribution table.

Since we have two alleles, we therefore have 3 minus 1, or 2 degree of freedom. Degrees of freedom is a complex issue, but we could look at this in simple terms: if we have frequencies for three genotypes that are truly representative of the population then, no matter what we calculate for two of them, the frequency of the third must not be significantly different for what is required to fit the population.

Looking across the distribution table for 2 degrees of freedom, we find our chi-squared value of 4.221 is less than that required to satisfy the hypothesis that the differences in the O and E data did not arise by chance. Since the chi-squared value falls below the 0.05 (5%) significance cutoff, we can conclude that the Forensics Town population does not differ significantly from what we would expect for Hardy-Weinberg equilibrium of the MN blood group.

Random Mating



	Male p 0.4	Male q 0.6
Female p 0.4	p^2 0.16	pq 0.24
Female q 0.6	pq 0.24	q^2 0.36

Genotype frequencies:
$p^2MM + 2pqMN + q^2NN = 1$

One reason that Hardy-Weinberg does not always apply to humans is that random mating, a condition of HWE (Hardy-Weinberg Equilibrium), does not occur. Random mating implies that mating should be arbitrary with regard to the locus being considered. As with most mammals, humans tend to mate with individuals that are similar to themselves, especially with respect to evident or visible traits such as height, I.Q., and ethnicity. This constitutes non-random, positive assortative mating. In contrast, negative assortative mating is breeding between individuals with dissimilar genotypes, and is more rare.07 Positive assortative mating leads to an increase in homozygotes whereas negative assortative mating leads to an increase in heterozygotes. However, assortative mating is never complete. While humans often mate with individuals alike in characteristics such as physical attributes, they mate randomly with respect to other traits such as blood type and short tandem repeat (STR) genotype.

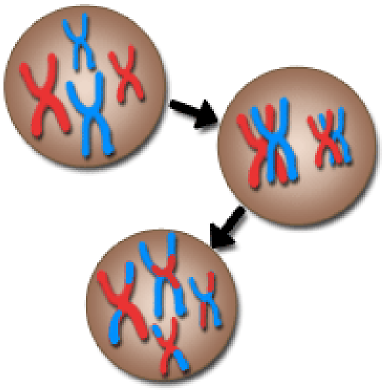
The online version of this course contains a multimedia [or downloadable] file on random mating by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m01/02/a/. You must have a user name and password to view the online course.

Mutation

The ultimate source of evolution is mutation, a permanent, heritable change in the nucleotide sequence of a chromosome, usually in a single gene. Mutations occur at random and can vary in their effect. They may be neutral with no phenotypic expression, or cause variations to an individual's phenotype, which may range from small-scale to large-scale. Although mutations can affect an individual's survival, evolution is driven forward only if this mutation can be passed on to the next generation, thereby affecting that generation's survival rates as well.

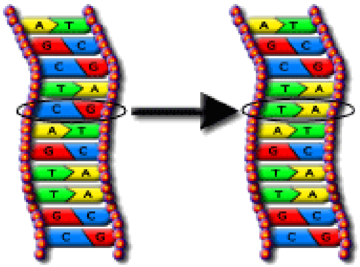
Recombination is another source of variation in a population. It is a process whereby two homologous chromosomes exchange some of their genetic material producing two chromosomes that are genetically unique from the original, or parental, chromosomes. Recombination enlarges the amount of genetic diversity in the population by increasing the number of alleles at any given genetic locus.08

Both mutations and recombination can alter the allelic frequencies from generation to generation and, at least in theory in small populations, can affect HWE. While mutation and recombination add to the variation within a population, their effects are limited.



Recombination

While migration and recombination add to the variation within a population, their effects are limited.



Mutation

Migration and Gene Flow

Allele frequencies will change if migration occurs into or away from the population. The effect of migration on HWE (Hardy-Weinberg Equilibrium) is dependent on the difference in allele frequencies between the donor and recipient populations.09

Gene flow is another way to introduce genetic variability to a population. Similar to migration, it occurs when members of one gene pool mate with members of another gene pool, which can lead to an alteration of the allele frequencies.04

Genetic Drift and Natural Selection

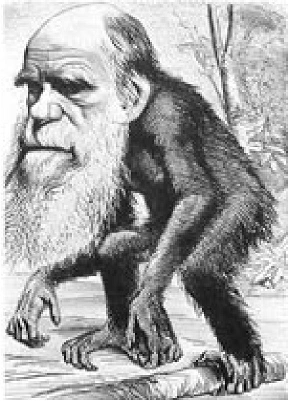
Genetic Drift

Allele frequencies in small populations do not generally reflect those of larger populations since too small of a set of individuals cannot represent all of the alleles for the entire population. Genetic drift occurs when the population size is limited and therefore by chance, certain alleles increase or decrease in frequency. This can result in a shift away from Hardy-Weinberg equilibrium (HWE). Unlike natural selection, genetic drift is random and rarely produces adaptations to the environment.10

The online version of this course contains a multimedia [or downloadable] file on genetic drift presented by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m01/02/d/. You must have a user name and password to view the online course.

Natural Selection

Although population genetics by itself is important, one of the objectives of this field is to assess how changes in allele frequencies affect the evolution of a population. Evolution in its modern form was first explored by Charles Darwin in 1859. In his book *On the Origin of Species*, Darwin outlined what he called "descent with modification" and what we now refer to as evolution. He speculated that all species evolved from a common ancestor. Over time, faced with new environments and habitats, populations of species acquired modifications, which allowed them to better adapt to their environment.11



Darwin termed these changes within populations, natural selection, and he proposed the idea of "survival of the fittest." Individual variations which proved beneficial would be preserved within a population, whereas variations that were lethal to the organism would be destroyed. Under natural selection, some individuals in a population have modifications that allow them to more successfully survive and reproduce, making their adaptations more common as a whole due to their increased reproductive success. Over a long period of time, this change in the characteristics of a population can lead to the production of a new species.11

View an animation on Natural Selection.

Darwin 's theory of evolution can be summarized in three main principles:

1. **Principle of variation:** Among individuals within any population, there is variation in morphology, physiology, and behavior.
2. **Principle of heredity:** Offspring resemble their parents more than they resemble unrelated individuals.
3. **Principle of selection:** Some forms are more successful at surviving and reproducing than other forms in a given environment.12

It is important to remember that evolution occurs at the population level, not at the individual level.

To see an example of mutation and natural selection at work, consider the case of the peppered moth. Prior to the Industrial Revolution in England, the peppered moth was found almost entirely in its light-colored form. Its color provided camouflage against the lichen-covered trees, preventing the moths being seen by predators. The pollution from the industrial revolution caused much of the lichen on the trees to die. As a result, the light moths became more visible to birds, whereas the dark colored moths (which arose from a mutation) were able to blend in better with the trees and avoid being eaten. The result was that the population of dark moths, which was about 1% in 1848, increased to about 90% by 1959!13

Inbreeding

Inbreeding is the mating of related individuals and can alter gene frequency in a population. When related individuals mate, the child can inherit identical copies of the gene through both parental lineages, resulting in an increase in homozygosity at any given locus.04

The degree of a relationship between two persons can be measured using the inbreeding coefficient, F. With inbreeding, the expected heterozygosity is reduced by a fraction, F, and that of homozygotes are increased

(NRC, National Research Council).

Inbreeding Coefficient	
F	
Parent/child	$^{1}/_{4}$
Siblings	$^{1}/_{4}$
First cousins	$^{1}/_{16}$

Population genetics studies show some substructure within racial groups. Mating tends to occur between persons who are likely to share some common ancestry. Allele frequencies have not yet been homogenized because people tend to mate within these subgroups (NRC). If there is subdivision within a population, it will lead to decreased heterozygosity within that subgroup. Population substructure may exist, and can be adjusted for with the use of a correction factor, theta (). (NRC). This will be covered in more detail in course: Population Genetics & Statistics.

Read about Population Subgroups in *NRC*.

Read about Subpopulation Theory in *NRC*.

Inbreeding can allow recessive alleles to become homozygous; therefore, unusual recessive diseases are much more common among the children. Inbred populations deviate from expected Hardy-Weinberg frequencies and there is a greater chance that certain alleles can become fixed within the population. When an allele becomes fixed in a population, all individuals are homozygous at this locus because no other alleles exist.

The online version of this course contains a multimedia [or downloadable] file on inbreeding by Greggory LeBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m01/02/e/. You must have a user name and password to view the online course.

Copyrighted material used with permission of the author, Simon C Kao, MD, The University of Iowa, and www.vh.org.

The Amish population of Lancaster County, Pennsylvania, is an example of an inbreeding group of individuals. As a result, the Amish suffer from a variety of genetic disorders including Ellis - van Creveld (EVC) syndrome, a disease caused by inheritance of two mutated copies of the EVC gene. Symptoms of the disease include short-limbed dwarfism with polydactyly (additional fingers or toes), bone malformations in the wrist, heart defects, and prenatal eruption of the teeth.14

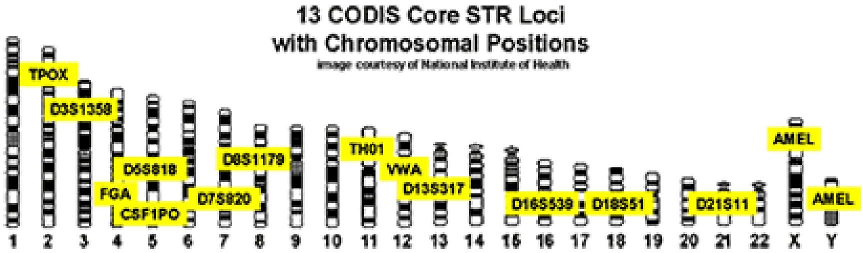
Linkage Equilibrium

Linkage is the tendency of genes or other DNA sequences at specific loci to be inherited together as a consequence of their physical proximity on a single chromosome.15 It is possible to predict linkage by estimating the distance between genes or segments of DNA using recombination frequency. In this way, one can construct genetic maps called linkage maps.

View an animation about linkage equilibrium.

Genetic recombination occurs when two homologous chromosomes exchange parts of their DNA. This often happens in gametes (egg and sperm) so that new combinations of alleles can be passed on to the next generation. Recombination occurs at random. If there is a large distance between two segments of DNA, there is a good chance that recombination will occur between them. However, if the two sequences are close together, recombination will rarely occur between them (the two sequences will tend to stay together rather than being split apart by recombination). In the latter case, the recombination frequency is low and the two DNA sequences are considered to be linked. High recombination frequencies are found for DNA sequences far enough apart that they are not in linkage with each other and are considered to be in linkage equilibrium.

The recombination fraction is defined as the proportion of recombinants that define genetic distance. Two loci that show 1% recombination are defined as being 1 centimorgan (cM) (in honor of Thomas Hunt Morgan, 1933 Nobel Laureate in Physiology or Medicine) apart on a genetic (or linkage) map. In general, genetic distance can be related to physical distance, with 1cM being approximately equal to 1Mb or 1 million base pairs of DNA.07


13 CODIS Core STR Loci with Chromosomal Positions
image courtesy of National Institute of Health

The STR loci used to type biological fluids in forensic laboratories are not physically linked, apart from D5S818 and CSF1PO. However, none of the loci, including the latter pair, are genetically linked (i.e. they are all in linkage equilibrium with one another). Therefore, it is possible to multiply the genotypic frequencies over several loci to come up with a combined genotypic frequency for the entire profile.03

Statistics

Introduction

Statistics are important to the forensic scientist because they provide a means to quantitate uncertainty, to summarize large amounts of data into understandable terms, and to make decisions based on these summaries in the presence of uncertainty and variation. Statistics provide a means for forensic scientists to convey the significance of DNA testing results. When comparing the DNA profiles obtained from forensic evidence to that from a known individual, statistics can be used to assess the significance of an association.

Objectives

Upon successful completion of this unit of instruction, the student shall be able to do the following:

- Describe the importance and use of statistics and probability in the field of forensic science
- Use at least one of the accepted statistical approaches to evaluate DNA data
- Describe how DNA population databases are constructed and used

The online version of this course contains a multimedia [or downloadable] file on statistical probabilities by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/00. You must have a user name and password to view the online course.

Population Databases

There are numerous approaches to statistical interpretation of forensic DNA typing results. The approaches outlined in this module are advocated by the 1996 National Research Council Report (NRC II) and the associated formulas are provided in the Federal Bureau of Investigation's Combined DNA Index System (CODIS) software program. In addition to these methods of interpretation, there are several books and publications available that propose other approaches and provide more detail regarding the handling of forensic DNA evidence.01-07

Read NRC II.

Population Databases

Population databases allow for estimations of how rare or common a DNA profile may be in a particular population. As mentioned in the last module, defining a population can be difficult. When compiling a database, it is important to have a population size that is sufficiently large to capture the most common alleles, and one that will yield enough samples of the common alleles to permit the calculation of reliable frequency estimates. Ideally, a database should contain several hundred samples; the National Research Council suggested using between 120-150 individuals with the expectation that this will yield between 240-300 alleles.[07] However, 200 alleles has become the *de facto* minimum size for a database.[05] In populations with no existing data, it may be possible to use data corresponding to other similar populations, depending on the context of case.

The online version of this course contains a multimedia [or downloadable] file on population databases by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/01. You must have a user name and password to view the online course.

The CODIS software, PopStats, has the following population databases:

- African American
- Asian
- Caucasian
- Hispanic
- Native American

The allele frequencies in the populations were gathered from the CODIS 13 core short tandem repeat (STR) loci.[08,09]

Confidence Intervals

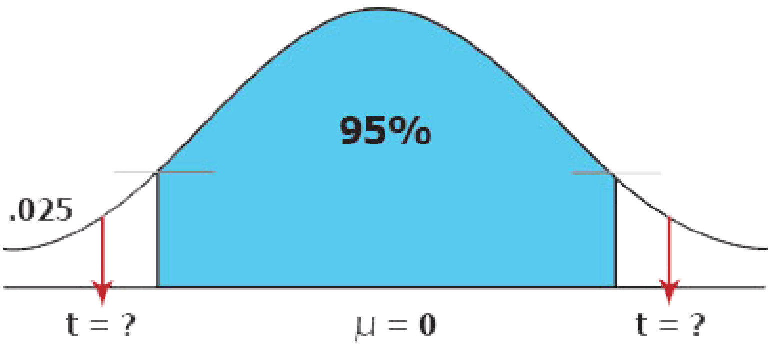
In a statistical context, populations are defined by parameters such as mean and variance. These parameters are estimated from the data obtained from population samples. Here we are interested in the frequency of an allele in the population, estimated from its proportion in the sample set. As the sample size increases, the estimate will become a more dependable measure of the true frequency of the allele in the population.[10] Classical statistics deals with the relationship between sample size and the reliability of a parameter estimate by calculating a confidence interval, which furnishes a range of plausible values for the unknown parameter.

The online version of this course contains a multimedia [or downloadable] file on confidence intervals by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/02/. You must have a user name and password to view the online course.

Confidence Interval

$z\sqrt{[p(1-p)]/n}$

Where z is found from statistical tables and depends on the desired range (e.g. 95%); and p is the proportion observed in the sample of size n



95% Confidence Interval

The online version of this course contains a multimedia [or downloadable] file on estimation by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/02. You must have a user name and password to view the online course.

Theta Correction

In the previous module, inbreeding and population substructure were introduced. Population substructure can be dealt with in the same manner as inbreeding. The 1996 National Research Council Report (NRC II) proposed using formulas 4.4a and 4.4b to deal with population substructure. Formulas analogous to those noted below (with theta) are used for the inbreeding coefficient, F. These formulas do not require that the subpopulations be distinct or mate at random.07 The value of theta () can be positive or negative, but must be less than or equal to 1.

Read about the inbreeding coefficient, F eslewhere in this PDF file.

- Homozygote: $p_i^2 + p_i (1-p_i)_{ii}$ (formula 4.4a from the NRC II)

- Heterozygote: $2p_i p_j (1- _{ij})$, $i \neq j$ (formula 4.4b from the NRC II)

Read about formula 4.4a and 4.4b from NRC II.

The parameter can be defined as the probability that two alleles in different people, in the same subpopulation, are identical by descent.05 Strictly the two alleles can originate from the same individual or, more likely, two different individuals. The value of can be determined for a given population. Generally, = 0.01 can be used for large populations, and = 0.03 for small, isolated populations.

Frequentist Approach

Statistical Approaches

There are generally three statistical approaches used in forensic DNA interpretation:05

- Full Bayesian
- Frequentist logical (or likelihood ratio)

- Logical (or likelihood ratio)

The scientific community's knowledge of population genetics provides the framework for each of these methods. Each approach requires the use of allele frequencies, which are calculated from various databases. This module will focus on the frequentist and logical approaches, the most widely used statistical approaches in the United States.

Frequentist Approach

The frequentist approach uses probabilities or genotype frequencies to address statistical questions. Coincidence probability, or random match probability (RMP), and the probability of exclusion are each frequentist approaches that are applied by the forensic science community.

Probability theory was first introduced in seventeenth century France when two mathematicians, Blaise Pascal and Pierre de Fermat, debated over problems from games of chance.[10] The idea was that a person who understands decision making in the face of uncertainty has an advantage over someone that does not.[01]

Assuming that all outcomes are equally probable, then the probability of event A is the number of ways event A can happen divided by the total number of possible outcomes.[11]

Probability

For example, what is the probability of drawing an ace from a shuffled deck of 52 playing cards?

S=4 aces in the deck N=52 cards in the deck S/N = 4/52 = 1/13

The frequentist approach requires a basic understanding of some factors about probability:

$0 \leq$ Probability of an event A ≤ 1

- The probability can be any single value between 0 and 1 inclusive.[02] If an event is impossible, then the probability value is 0; if the event is certain to occur it has a probability of 1. All other events that could potentially occur have probabilities that lie somewhere between 0 and 1. It is common practice to multiply the probability of an event by 100 to obtain a percentage of probability.

Probability of an event A or B =
Probability (A) + Probability (B)

- If two events, A and B, are mutually exclusive, then the probability that one or the other of them (A or B) is true is equal to the sum of the probabilities of A and B.[02]

Probability of an event A and B =
Probability (A) x Probability (B)

- The probability that two independent events, A and B, will both take place is calculated by multiplying the probability of the events A and B.[01] This is the product rule.

The term odds is often used incorrectly; odds are not the same thing as probability. The odds on an event occurring is the ratio of two competing probabilities- the probability that the event will occur and the probability that it will not occur. If the probability of the event is p, the probability that it will not occur is 1 – p. The odds are therefore p/(1-p).[01,03]

Coincidence Approach

A good description of the coincidence approach, also referred to as the random match probability, is given in *Forensic DNA Evidence Interpretation*,05 and reads, "The coincidence approach proceeds to offer evidence against a proposition by showing that the evidence is unlikely if this proposition is true. Hence it supports the alternative proposition. The less likely the evidence under the proposition the more support given to the alternative." The alternative is that the match occurs by chance.

The coincidence approach assesses whether or not the match between the DNA profile obtained from the evidence and the DNA profile obtained from the suspect occurs by chance (coincidence). Allele frquencies are calculated using data that assumes Hardy-Weinberg equilibrium and no linkage disequilibrium.

Read about allele frequencies eslewhere in this PDF file.

Homozygote frequencies are determined by p^2 and heterozygote frequencies by 2pq (Hardy-Weinberg formulas). The STR loci used by the forensic science community are not linked, and, therefore, the genotypes from each locus can be multiplied (i.e. the product rule) to obtain the frequency of the combined individual genotypes.

👓

Read about formulas 4.1a and 4.1b from NRC II (1996 National Research Council Report).

The online version of this course contains a multimedia [or downloadable] file on determining the genotype frequency by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/04/a/. You must have a user name and password to view the online course.

As discussed previously, a theta correction can be applied to the formulas. For homozygous loci, the NRC II report recommends using Equation 4.4a with a conservative value of . The 2p rule was originally recommended by the NRC for VNTRs (Variable Number Tandem Repeats) based on the ambiguity of allele calls; however, since this ambiguity does not exist with STR loci it is not necessary.

The NRC II panel assumed that theta is positive for all pairs of alleles. It noted that for heterozygotes, the Hardy-Weinberg calculation is generally an overestimate. The assumption is that Hardy-Weinberg proportions always give overestimates of heterozygotes when > 0.

The panel surmised that for homozygotes (using equation 4.4a) with small allele frequencies, a small value of can introduce a large change in the genotype frequency.

Laboratories are not compelled to choose one formula over another, but analysts should be familiar with the relevant discussions on this topic.

Example Calculation

Homozygotes: $p^2 + p(1-p)$ (Homozygote example w/ theta)

An evidence sample has a genotype at D3S1358 of 16, 16, and a of 0.03 is used.

The frequency of the 16 allele in Caucasians = 0.2315.

Genotype frequency = $(0.2315)^2 + [(0.2315)(1-0.2315)(0.03)]$

Genotype frequency = $0.0535 + [(0.2315)(0.7685)(0.03)]$

Genotype frequency = $0.0535 + 0.0053$

Genotype frequency = 0.0588

Heterozygote: $2p_ap_b(1- {}_{ab})$, $a \neq b$

(Heterozygote example w/ theta)

An evidence sample has a genotype at D3S1358 of 15, 17.

The frequency of the 15 allele = 0.2904 and the frequency of the 17 allele = 0.2000 in African Americans and a of 0.03 is used.

Genotype frequency = $[2(0.2904)(0.2000)]\ [(1-0.03)\ ]$

Genotype frequency = $(0.11616)(0.97)$

Genotype frequency = 0.1127

Heterozygote: $2p_ap_b$

An evidence sample has a genotype at D3S1358 of 15, 17.

The frequency of the 15 allele = 0.2904 and the frequency of the 17 allele = 0.2000 in African Americans.

Genotype frequency = $[2(0.2904)(0.2000)]$

Genotype frequency = 0.1161

The combined frequency across multiple loci (non-linked) can be calculated by using the product rule, multiplying the respective frequencies together.

Example:

Evidence sample locus with locus frequencies

Locus	Locus Frequency
D3S1358	0.07534
vWA	0.02725
FGA	0.04453

- Total frequency = (frequency of D3S1358)(frequency of vWA)(frequency of FGA)
- Total frequency = (0.07534)(0.02725)(0.04453)
- Total frequency = 0.0000914
- Probability = 1/0.0000914 = 1 in 10,941 unrelated people
- Probability of observing the given genotype is reported as 1 in 10,941 unrelated people.

Probability of Exclusion

The probability of exclusion provides an estimate of the proportion of the population that has a genotype composed of at least one allele not observed in the mixed profile. 12,13 This approach, for practical purposes, is considered conservative and analysts do not need to make assumptions about the mixture or number of contributors. It has, however, been criticized because it does not make use of all of the available genetic data.12

The steps for applying the probability of exclusions (PE) are to sum all of the homozygotes and to sum all the heterozygotes that are represented within the mixed profile.

For a given locus, assuming Hardy-Weinberg equilibrium,

$PE_{locus} = 1\text{-}p^2$

Or if Hardy-Weinberg equilibrium is not assumed,

$PE_{locus} = 1\text{-}p^2 - p(1\text{-}p)$

The probability of exclusion (PE) is calculated for each homozygous locus by repeating the above steps, performing the equivalent calculation for heterozygotes, and summing the calculated values.

Total Probability of Exclusion without Theta

$PE = 1 - (p_1 + p_2 + ... + p_n)^2$

$PEmix = p_{locus(1)} + p_{locus(2)} + ... + p_{locus(n)}$

Total Probability of Exclusion with Theta

$PE = 1 - (p_1 + p_2 + ... + p_n)^2 - [p_1(1\text{-}p_1) + p_2(1\text{-}p_2) + ... + p_n(1\text{-}p_n)]$

$PEmix = p_{locus(1)} + p_{locus(2)} + ... + p_{locus(n)}$

Alleles with Low Frequencies

The NRC II (1996 National Research Council Report) report recommends using a five event minimum allele frequency for rare alleles. This frequency is calculated using 5/2n where n is the number of alleles in the database.

The online version of this course contains a multimedia [or downloadable] file on alleles with low frequencies by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/04/c/. You must have a user name and password to view the online course.

Logical Approach

Bayes Theorem

Many statisticians have employed what is known as Bayesian probability, after the British clergyman Thomas Bayes, which is based on probability as a measure of one's degree of belief.01 This type of probability is conditional in that the outcome is based on knowing information about other circumstances and is derived from Bayes Theorem, published in 1764.14

Conditional Probability

Conditional probability, by definition, is the probability P of an event A given that an event B has occurred.15 Mathematically, conditional probabilities can be represented as: P(A | B) where "|" means "given that."

Example:

Take the example of a die with six sides. If one was to throw the die, the probability of it landing on any one side would be 1/6. This probability, however, assumes that the die is not weighted or rigged in any way, and that all of the sides contain a different number. If this were not true, then the probability would be conditional and dependent on these other factors.

Likelihood Ratio

Conditional probabilities can assist when estimating the probability that evidence came from an identified source. The probability estimate is based on calculation of a likelihood ratio (LR).03 The likelihood ratio is the ratio of two probabilities of the same event under different hypotheses. Thus for events A and B, the probability of A given that B is true (hypothesis #1), divided by the probability of event A given that B is false (hypothesis #2) gives a likelihood ratio. The likelihood ratio is a ratio of probabilities, and can take a value between zero and infinity.02 The higher the ratio, the more likely it is that the first hypothesis is true.

In forensic biology, likelihood ratios are usually constructed with the numerator being the probability of the evidence if the identified person is the source of the evidence, and the denominator being probability of the evidence if an unidentified person is the source of the evidence.

The results can be interpreted as follows:

- LR < 1 — the genetic evidence has more support from the denominator hypothesis
- LR = 1 — the genetic evidence has equal support from both numerator and denominator hypotheses
- LR > 1 — the genetic evidence has more support from the numerator hypothesis16

These likelihood ratios can be translated into verbal equivalents that depict, in a relative way, the strength of the particular likelihood ratio in consideration.16,03 These verbal equivalents, however, are only a guide.17

Table of Verbal Equivalents

Limited evidence to support	LR <1-10
Moderate evidence to support	LR 10-100
Moderately strong evidence to support	LR 100-1000
Strong evidence to support	LR 1000-10000
Very strong evidence to support	LR >10000

The following equation can be used to determine the probability of the evidence given that a presumed individual is the contributor rather than a random individual in the population.

$$LR = P(E/H_1) / P(E/H_0)$$

$P(E/H_1)$ is the probability of the evidence given a presumed individual is the contributor.

$P(E/H_0)$ is the probability of the evidence given the presumed individual is not the contributor of the evidence.

In the case of a single source sample, the hypothesis for the numerator (the suspect is the source of the DNA) is a given, and thus reduces to 1. This reduces to:

LR = 1/ $P(E/H_0)$ which is simply 1/P, where P is the genotype frequency.

The use of the likelihood ratio for single source samples is simply another way of stating the probability of the genotype and, while stated differently, is the same as the random match probability approach.

Likelihood Ratio - Mixtures

Although likelihood ratios can be used for determining the significance of single source crime stains, they are more commonly used in mixture interpretation. The following example show the likelihood without a theta correction. A theta correction can be applied to the likelihood ratio calculation. Refer to NRCII formulas 4.10a and 4.10b.

Example of Two Person Mixture

Source	D3S1358	vWA	FGA	D8S1179	D21S11
Evidence	15	16,17	19,23	12,16	30,31.2,32.2
Victim	15	16,17	19	12,16	30,32.2
Suspect	15	16	19,23	16	31.2

Two explanations are possible for the above mixture:

- H_1 (also stated as ⊂) &#8212; Contributors were the victim and the suspect
- H_0 (also stated as ⊄) &#8212; Contributors were the victim and an unknown individual

The evidence is certain under H_1. Under H_1 the probability of the evidence depends on the chance of obtaining the evidence alleles (and no other alleles) from an unknown individual.

The table below shows all possible genotypes for an unknown individual, given the genotypes of the evidence and the victim.

Possible Genotypes

Locus	Evidence	Victim	Unknown Individual
D3S1358	15,15	15,15	15,15
vWA	16,17	16,17	16,16 or 16,17 or 17,17
FGA	19,23	19,19	19,23 or 23,23
D8S1179	12,16	12,16	12,12 or 12,16 or 16,16
D21S11	30,31.2,32.2	30,32.2	31.2,31.2; 30,31.2; or 31.2,32.2

The table below shows the equations used to determine the P (E/H_1) and P (E/H_0) assuming Hardy-Weinberg Equilibrium, where p=allele frequency.

Equations

Locus	$P(E/H_1)$	$P(E/H_0)$
D3S1358	1	P^2_{15}
vWA	1	$P^2_{16} + P^2_{17} + 2p_{16}\ p_{17}$

FGA	1	$p^2_{23} + 2p_{19}\,p_{23}$
D8S1179	1	$p^2_{12} + p^2_{16} + 2p_{12}\,p_{16}$
D21S11	1	$p^2_{31.2} + 2p_{30}\,p_{31.2} + 2p_{31.2}\,p_{32.2}$
TOTAL (Product)	1	Product of above

For the above example, the following frequencies were used to determine $P(E/H_1)$.

Frequencies

Locus	Allele	Frequency	Allele	Frequency	Allele	Frequency
D3S1358	15	0.2463				
vWA	16	0.2015	17	0.2627		
FGA	19	0.0561	23	0.1581		
D8S1179	12	0.1454	16	0.0138		
D21S11	30	0.2321	31.2	0.0994	32.2	0.1122

The following table shows the calculations for $P(E/H_1)$ given the above allele frequencies.

Calculations

Locus	$P(E/H_1)$	$P(E/H_0)$
D3S1358	1	$(0.2463)^2 = 0.0607$
vWA	1	$(0.2015)^2 + (0.2627)^2 + 2(0.2015)(0.2627) = 0.2154$
FGA	1	$(0.1581)^2 + 2(0.0561)(0.1581) = 0.0427$
D8S1179	1	$(0.1454)^2 + (0.0138)^2 + 2(0.1454)(0.0138) = 0.0253$
D21S11	1	$(0.0994)^2 + 2(0.2321)(0.0994) + 2(0.0994)(0.1122) = 0.0783$
TOTAL	1	0.0000011

(Product)

To determine the likelihood ratio, the above numbers are inserted into the previous formula as follows:

$LR = P(E/H_1) / P(E/H_0)$

$LR = 1/0.0000011$

LR= 909,091

The results are 909,091 times more likely if the victim and the suspect are the contributors of the mixture rather than the victim and a random individual in the population.18

NOTE: For mixtures with more than one unknown, review *Interpreting DNA Evidence*: *Statistical Genetics for Forensic Scientists*, Evett, I.W. and Weir, B.S., Sinauer Associates, Inc., 1998.

The use of any formula for mixture interpretation should only be applied to cases in which the analyst can reasonably assume "that all contributors to the mixed profile are unrelated to each other, and that allelic dropout has no practical impact.01

Likelihood Ratio

Calculations may also be performed to determine the probability of seeing a certain profile given that it has been seen once already, which is a factor to be considered in regard to the integrity of database information.

This type of conditional probability uses the equations 4.10a and 4.10b from NRC II (1996 National Research Council Report) report.

Read about equation 4.10a in NRC.

Read about equation 4.10b in NRC

- For homozygotes use:
  $P(A_i\ A_i\ |A_i\ A_i) = [2\&theta;+(1-\&theta;)p_i][3\&theta;+(1-\&theta;)p_i]/[(1+\&theta;)(1+2\&theta;)]$
- For heterozygotes use:
  $P(A_i\ A_j\ |A_i\ A_j) = 2[\&theta;+(1-\&theta;)p_i][\&theta;+(1-\&theta;)p_j]/[(1+\&theta;)(1+2\&theta;)]$

Note:

where &theta; is the average of the parameters of &theta; over all genotypes

As with forensic samples, a five-event minimum allele frequency should be used for rare alleles.

Parentage and Relatedness

Forensic science laboratories may be involved in paternity and relatedness cases. In these cases, DNA evidence is generally interpreted using likelihood ratios, comparing probabilities of the evidence under alternative propositions.01, 03

DNA parentage and relatedness testing can be conducted in a forensic science laboratory to

- Resolve questioned paternity

- Assist investigation of alleged rape or incest
- Assist in the identification of a missing person or unidentified remains
- Determine other familial relationships, such as maternity or sibling relatedness
- Discriminate between identical and fraternal twins.19

The rules of parentage testing are as follows:

1. The child cannot have a genetic marker that is absent in both parents.
2. The child must inherit a pair of genetic markers from each parent.
3. The child cannot have a pair of identical genetic markers, unless both parents have the marker.
4. The child must have the genetic marker, if that marker is present as an identical pair (homozygous) independently in both parents.19

The above are conditional on no mutation having occurred. Analysts need to be aware that inconsistencies at a locus could be due to a mutation. Mutations are rare events, and in general, two inconsistencies are sufficient to exclude in paternity cases. This module will not address statistics involving mutations. Note that the American Association of Blood Banks provides formulas that account for the inclusion of mutations into statistical paternity calculations.

To determine if the alleged father is the true biological father, the DNA profiles of the child, mother, and alleged father are compared. A child inherits two different alleles at each genetic locus&#8212;one from the mother and one from the father. If a child has an allele that the mother does not have, this obligate allele has to come from the biological father.20 The results are either an exclusion &#8212;the alleged father is not the biological father&#8212;or an inclusion.17

If the alleged father has the same allele as the obligate allele, a Paternity or System Index (PI or SI) can be calculated. This is the relative probability that the alleged father and not an unrelated, randomly selected male of the same ethnic background transmitted the obligate allele to the child.

This is a likelihood ratio and is presented in the formula X | Y, where X is the chance that the alleged father could transmit the obligate allele and Y is the chance that an unrelated man of the same race could have the allele.21 X is assigned the value of 1 if the alleged father is homozygous for the allele of interest and 0.5 if the alleged father is heterozygous.

The probability of an unrelated, randomly selected man possessing the obligate allele is determined by using a database that lists the frequency distribution of individual alleles. If there is more than one obligate allele, the individual paternity indexes can be multiplied and the total across all loci is called the Combined Paternity Index (CPI). This is a measure of the strength of the genetic evidence and is an odds ratio, not a probability.20

An interpretation of the CPI is as follows:

- CPI can range from 0 to infinity.
- If CPI is between 0-1, the genetic evidence is more consistent with non-paternity than paternity.
- If CPI>1, the genetic evidence is more consistent with paternity than non-paternity.16

It is normal practice to establish a threshold value for CPI, above which it is accepted that the tested man is the true biological father. This threshold is 1000 in Europe, but can be as low as 100 in the USA.19

Probability of Paternity

Sometimes a likelihood ratio is converted into a probability. This probability is known as the probability of paternity (POP).20 This formula tests the hypothesis that the alleged father is indeed the biological father of the child. For example, a POP of 99% reflects a 99% probability that the hypothesis is correct and a 1% probability that it is not. The CPI is used in the Bayes formula along with another variable called a prior probability (PP), which represents the social evidence.20 Testing labs typically use a value of 0.5 for the PP on the basis that this is a neutral, unbiased value.16

Mathematically, POP = (100) (CPI) (PP) / [(CPI) (PP) + (1 &#8211; PP)]19

POP is not widely used in the United States. A more common approach, similar to the frequentist probability of exclusion, is the Random Man Not Excluded (RMNE) statistic. This is the proportion of the population that could contribute all of the obligate alleles and therefore could not be excluded, or would be falsely included.19 A single locus RMNE is calculated by $1-(1-p)^2$. Combining the RMNE statistics over all loci gives the combined RMNE (CRMNE), which is equivalent to the CPI. The value of the CRMNE is typically small (less than one), and is analogous to 1-CRMNE or exclusionary power (EP). EP represents the probability of excluding a falsely accused man.19

Although the previous terms and statistics are specific to parentage testing, similar methods are used to estimate the relatedness in other situations, such as identification of human remains and missing persons.

The online version of this course contains a multimedia [or downloadable] file on paternity indexes by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/06/a/. You must have a user name and password to view the online course.

Paternity Calculations

In paternity calculations the likelihood ratio is the same as the paternity index. The three tables below show the calculations of the likelihood ratio for different possible combinations of genotypes assuming that the mother, alleged father and father are unrelated.

In the tables

- GC is the child's genotype
- GM is the mother's genotype
- GAF is the alleged father's genotype
- H_p is the mother and alleged father are the true parents
- H_d is the mother and an unrelated male are the true parents
- LR numerator is P(GC|GM, GAF, H_p)
- LR denominator is P(GC|GM, GAF, H_d).01

Likelihood Ratio and Paternity Calculation

GC	**GM**	**GAF**	**Num**	**Denominator**	**LR**
A_i A_i	A_i A_i	A_i A_i	1	p_i	1/ p_i
		A_i A_j, j&ne; i	1/2	p_i	$1/2p_i$

		A_j A_k, k&ne;i, j	0	p_i	0
A_i A_i	A_i A_j, i &ne; j	A_i A_i	1/2	p_i /2	1/ p_i
		A_i A_j, j&ne;i	1/4	p_i /2	1/2 p_i
		A_j A_k, k&ne;i, j	0	p_i /2	0
A_i A_j, i&ne;j	A_i A_i	A_j A_j	1	p_j	1/p_j
		A_j A_k, k&ne;j	1/2	p_i	1/2 p_j
		A_k A_l, k, l&ne;j	0	p_i	0
A_i A_j, i&ne;j	A_i A_j, i&ne;j	A_i A_i	1/2	$(p_i + p_j)/2$	$1/(p_i + p_j)$
		A_i A_j	1/2	$(p_i + p_j)/2$	$1/(p_i + p_j)$
		A_j A_k, k&ne;j	1/4	$(p_i + p_j)/2$	$1/(2(p_i + p_j))$
		A_k A_l, k, l&ne;i, j	0	$(p_i + p_j)/2$	0
A_i A_j, i&ne;j	A_i A_k, k&ne;i, j	A_j A_j	1/2	p_j /2	1/ p_j
		A_j $A_{l,l}$ &ne;j	1/4	p_j /2	1/2 p_j
		A_k A_l, k, l&ne;j	0	p_j /2	0

Example:

Locus	**Mother**	**Child**	**Alleged Father**
D3S1358	16	16	15, 16
vWA	17, 18	17, 18	18, 20
FGA	25, 27	23, 25	21, 23
D18S51	14, 18	18	18, 20

The following frequencies were used:

Locus	**Allele**	**Frequency**	**Allele**	**Frequency**
D3S1358	16	0.2315		
vWA	17	0.2627	18	0.2219
FGA	23	0.1581		
D18S51	18	0.0918		

- D3S1358 &#8212; LR = 1/2p_{16} = 1/(2 x 0.2315) = 2.1598
- vWA &#8212; LR = 1/[2(p_{17} + p_{18})] = 1/[2(0.2627 + 0.2219)] = 1.0317
- FGA &#8212; LR = 1/2p_{23} = 1/(2 x 0.1581) = 3.1625
- D18S51 &#8212; LR = 1/2p_{18} = 1/(2 x 0.0918) = 5.4466

Therefore, LR = 2.1598 x 1.0317 x 3.1625 x 5.4466 = 38.3815

In other words, it is 38 times more likely that the child's DNA profile would be observed if the alleged father is the true father rather than an unrelated individual selected at random from the Caucasian population.

The online version of this course contains a multimedia [or downloadable] file on paternity calculations by Greggory LaBerge. Visit this URL to view the file: http://beta.populations.dna.devis.com/m02/06/b/. You must have a user name and password to view the online course.

Non-autosomal DNA Analysis

The use of mitochondrial DNA (mtDNA) and Y-chromosome markers has become more prevalent in forensic science. Both methods characterize non-autosomal DNA, which does not follow the basic rules of Mendelian genetics. The statistical assessment can still be applied when comparing the DNA profiles obtained from forensic evidence to that from a known individual. A common approach is simply to state the number of occurrences of a mtDNA sequence or Y-STR type present in the database. This approach is commonly referred to as the counting method. This is the approach adopted by the FBI for mtDNA analysis.[22] Likelihood ratios provide another approach to statistical assessment of non-autosomal DNA results.[05]

Source Attribution

Source attribution of evidence does not require that the profile be unique, only that there is reasonable scientific certainty regarding the source of the evidence.[16] The term unique can have several meanings, but typically it means a circumstance that is only one of its kind. Some laboratories, based on the statistical frequencies of a profile, report a particular individual as the source of an evidentiary sample. The Federal Bureau of Investigation has outlined an approach to assess source attribution of an evidentiary profile, published in Forensic Science Communications.

Read the article, "Source Attribution of a Forensic DNA Profile" in *Forensic Science Communications*.

Author: Christina M. Mailloux

Christina M. Mailloux received her B.S. degree from Tufts University in biology in 2002. She subsequently went to work for the National Institutes of Health under a Postbaccalaureate Intramural Research Training Award (IRTA), where she engaged in research on mammalian genetics. Christina is currently pursuing her PhD in Human Medical Genetics at the University of Colorado Health Sciences

Center, where she studies vitilgo and other associated disorders.

Author: Greggory LaBerge

Greggory LaBerge, M.Sc. is the Commander and Director of the Denver Police Department Crime Laboratory. He is also affiliated with the University of Colorado Health Sciences Center - Human Medical Genetics Program as a Ph.D. candidate conducting statistical genetics research on autoimmune disease.

Author: Christina M. Mailloux

Christina M. Mailloux received her B.S. degree from Tufts University in biology in 2002. She subsequently went to work for the National Institutes of Health under a Postbaccalaureate Intramural Research Training Award (IRTA), where she engaged in research on mammalian genetics. Christina is currently pursuing her PhD in Human Medical Genetics at the University of Colorado Health Sciences Center, where she studies vitilgo and other associated disorders.

Author: Greggory LaBerge

Greggory LaBerge, M.Sc. is the Commander and Director of the Denver Police Department Crime Laboratory. he is also affiliated with the University of Colorado Health Sciences Center - Human Medical Genetics Program as a Ph.D. candidate conducting statistical genetics research on autoimmune disease.

Example

Determining Hardy-Weinberg Equilibrium

Suppose that scientists are observing a population of lab-bred flies, and discover a gene controlling eye color. The *R* allele produces regular-colored eye pigment, while the *r* allele produces red pigment. Individuals that are heterozygous (*Rr*) have pink eyes. In a population of 150 flies, 15 flies have red eyes, 90 have normal eye color, and 45 have pink eyes.

Is this population in Hardy-Weinberg equilibrium?

In order for a population to be considered to be in equilibrium, it must remain the same from generation to generation. Therefore, in order to determine if this population of fruit flies is in Hardy-Weinberg equilibrium, the genetic distribution of the current generation must be compared to a prediction of the genetic distribution of the next generation, as calculated using the Hardy-Weinberg equation.

Step 1: Determine the gene frequencies of the current generation.

Phenotype	Genotype	# of Individuals
Normal Eyes	*RR*	90
Red Eyes	*rr*	15
Pink Eyes	*Rr*	45

Given this information, calculating the allele frequencies is simply a matter of counting up all of the alleles.

- Remember, each parent carries *two* alleles, so the total # of alleles is twice the population.
- Also remember that *heterozygous* individuals carry one of *each* allele.

Taking these two factors into account,

f(R) = [(90 x 2) + (45)] / (150 x 2) = 225/300 = **0.75**

f(r) = [(15 x 2) + (45)] / (150 x 2) = 75/300 = **0.25**

Step 2: Determine the expected genotype frequencies for the next generation.

Plugging the frequencies of each allele into the Hardy-Weinberg equation, we find the expected numbers of each genotype in the population:

f(RR) = p2 = f(R) x f(R) = **0.5625**

f(rr) = q2 = f(r) x f(r) = **0.0625**

f(Rr) = 2pq = 2 x [f(R) x f(r)] = **0.375**

Multiplying each of these genotype frequencies with the total population number, we find that there should be:

- 84 normal-eyes flies (*AA*)
- 9 red-eyed flies (*aa*)
- 56 pink-eyed flies (*Aa*)

(Since partial individuals do not exist, the numbers are rounded off.)

Step 3: Compare the expected frequency with the original population numbers.

Comparing the expected numbers with the actual numbers of each phenotype, population geneticists can determine if populations are either in equilibrium (or very close to it) or are experiencing *disequilibrium* of some sort. In this example:

Phenotype	**Genotype**	**Expected #**	**Observed #**
Normal Eyes	*RR*	84	90
Red Eyes	*rr*	9	15
Pink Eyes	*Rr*	56	45

In this example, the population is not in equilibrium since the expected and observed values do not match. Once a population geneticist determines that a population is in disequilibrium, the reasons can be explored. Disequilibrium can be attributed to different possible mechanisms, depending on (1) the context of the population, and (2) the manner in which the population is skewed.

Works Cited & Online Links

1. Bindon, J. R. 2003. Population genetics. Lecture notes, Univ. of Alabama. http://www.as.ua.edu/ant/bindon/ant270/lectures/POPGEN.pdf

2. Barbujani, G., A. Magagni, E. Minch, and L. L. Cavalli-Sforza. 1997. An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* 94 (9):4516&#8211;19.
3. Tishkoff, S. A., and K. K. Kidd. 2004. Implications of biogeography of human populations for "race" and medicine. *Nat Genet Suppl. no.* 36 (11): S21&#8211;S27.
4. Thompson, Margaret W., Roderick R. McInnes, and Huntington F. Willard. 1991. *Thompson & Thompson genetics in medicine.* 5th ed. Philadelphia: W. B. Saunders.
5. Dorak, M. Tevfik. 2005. *Basic population genetics.* http://dorakmt.tripod.com/genetics/popgen.html
6. McClean, Phillip. 1997. *Population and evolutionary genetics.* http://www.cc.ndsu.nodak.edu/instruct/mcclean/plsc431/popgen/ popgen1.htm
7. Griffiths, Anthony J. F., et al. 1998. *An introduction to genetic analysis.* 6thed. New York: W. H. Freeman.
8. Lodish, Harvey, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James E. Darnell. 2000. *Molecular cell biology.* New York: W. H. Freeman.
9. Buckleton, John S., Christopher M. Triggs, Simon J. Walsh, eds. 2004. *Forensic DNA evidence interpretation.* Boca Raton, FL: CRC Press.
10. National Biological Information Infrastructure. 2005. Introduction to population genetics. *Genetic biodiversity.*
11. Corstjens, H., H. A. Billiet, J. Frank, and K. C. Luyben. 1996. Variation of the pH of the background electrolyte due to electrode reactions in capillary electrophoresis: Theoretical approach and in situ measurement. *Electrophoresis* 17 (1):137&#8211;43.
12. Darwin, Charles. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London: John Murray. Also available online at http://pages.britishlibrary.net/ charles.darwin/texts/origin1859/origin_fm.html.
13. Biology-Online.Org. 2000. Natural selection in action. *Genetics and evolution* http://www.biology-online.org/2/11_natural_selection.htm
14. Strachan, T., and A. P. Read. 1999. *Human molecular genetics.* 2nd ed. New York: John Wiley & Sons.
15. Jobling, M. A., and P. Gill. 2004. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics* 5 (10):739&#8211;51.

Online Links

- Population Subgroups in NRC
  http://newton.nap.edu/books/0309053951/html/99.html
- Subpopulation Theory in NRC
  http://newton.nap.edu/books/0309053951/html/102.html

1. Evett, Ian W., and Bruce S. Weir. 1998. *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.
2. Aitken, C. G. G. 1995. *Statistics and the evaluation of evidence for forensic scientists*. Chichester, UK: John Wiley & Sons.
3. Aitken, C. G. G., and D. A. Stoney. 1991. *The use of statistics in forensic science*. New York: Ellis Horwood.
4. Butler, John M. 2001. *Forensic DNA typing: Biology and technology behind STR markers*. San Diego, CA: Academic Press.
5. Buckleton, John, Christopher M. Triggs, and Simon J. Walsh, eds. 2005. *Forensic DNA evidence interpretation*. Boca Raton, FL: CRC Press.
6. Butler, John M. 2005. *Forensic DNA typing: Biology, technology, and genetics of STR markers.* 2nd ed. Burlington, MA: Elsevier Academic Press.
7. National Research Council Committee on DNA Forensic Science: An Update. 1996. *The evaluation of forensic DNA evidence*. Washington, DC: National Academy Press.
8. Lim, S. E., W. F. Tan-Siew, C. K. Syn, H .C. Ang, S. T. Chow, and B. Budowle. 2005. Genetic data for the 13 CODIS STR loci in Singapore Indians. *Forensic Sci Int* 148 (1) : 65&#8211;7.
9. Budowle, B., B. Shea, S. Niezgoda, and R. Chakraborty. 2001.CODIS STR loci data from 41 sample populations. *J Forensic Sci* 46 (3): 453&#8211;89.
10. Grinstead, Charles M., and J. Laurie Snell. 1997. *Introduction to probability*. 2nd rev. ed. Providence, RI: American Mathematical Society. http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book-5-17-03.pdf.
11. Glosser, G. 2005. Lesson on introduction to probability. In *Mrs. Glosser's Math Goodies*. http://www.mathgoodies.com/ lessons/vol6/intro_probability.html, (2005).
12. DNA Advisory Board. 2000. Statistical and population genetics issues affecting the evaluation of the frequency of occurrence of DNA profiles calculated from pertinent population databases. *Forensic Science Communications* 2 (3). http://www.fbi.gov/hq/lab/fsc/backissu/july2000/dnastat.htm.
13. Devlin, B. 1993. Forensic inference from genetic markers. *Stat Methods Med Res* 2 (3): 241&#8211;62.
14. Bayes, T. 1763. An essay toward solving a problem in the doctrine of chances. In *Philosophical Transactions of the Royal Society of London 53: 370&#8211;418.* http://www.stat.ucla.edu/history/essay.pdf.
15. Devore, Jay L. 2000. *Probability and statistics for engineering and the sciences*. 5th ed. Pacific Grove, CA: Duxbury Press.
16. LaBerge, G. 2004. *Analysis of DNA forensic evidence*. PowerPoint presentation.
17. Evett, I. W., G. Jackson, J. A. Lambert, and S. McCrossan. 2000. The impact of the principles of evidence interpretation on the structure and content of statements. *Sci Justice* 40 (4): 233&#8211;9.
18. City of Phoenix Police Department Laboratory Services Bureau Forensic Biology Protocol(s).
19. MDS Diagnostic Services. 2005. *DNA parentage testing*. Patients/TestInfo/DNA_Parentage.asp.
20. Primorac, D., and M. S. Schanfield. 2000. Application of forensic DNA testing in the legal system. *Croatian Medical Journal* 41 (1): 32&#8211;46.
21. Ostrowski, R. 2003. Paternity indices. Paper presented at Statistics of DNA Profiling forum, Ohio. http://bioforensics.com/conference/Paternity/.

22. Isenberg, A. R., and J. M. Moore. 1999. Mitochondrial DNA analysis at the FBI Laboratory. *Forensic Science Communications* 1 (2). http://www.fbi.gov/hq/lab/fsc/backissu/july1999/dnalist.htm.

- American Association of Blood Banks (AABB)
  http://www.AABB.org
- Formula 4.1a and 4.1b in NRC II
  http://fermat.nap.edu/books/0309053951/html/92.html
- Formula 4.10a in NRC
  http://www.nap.edu/books/0309053951/html/114.html
- Formula 4.10b in NRC
  http://www.nap.edu/books/0309053951/html/115.html
- Formula 4.4a and 4.4b in NRC II
  http://www.nap.edu/books/0309053951/html/102.html
- NRC II
  http://www.nap.edu/books/0309053951/html
- "Source Attribution of a Forensic DNA Profile"
  http://www.fbi.gov/hq/lab/fsc/backissu/july2000/source.htm